# Populations That Think, Beliefs That Move: Modeling Opinion Dynamics in Synthetic Populations for Social-Cyber Influence

**Rebecca Marigliano**

CMU-S3D-26-107

April 2026

Software and Societal Systems Department
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

**Thesis Committee:**
Kathleen M. Carley, Chair
Fei Fang
Rick Carley
Johnathon Harris, Office of Naval Research

*Submitted in partial fulfillment of the requirements*
*for the degree of Doctor of Philosophy in Societal Computing.*

Copyright © 2026 **Rebecca Marigliano**

This research was supported in part by the U.S. Army project *Scalable Technologies for Social Cybersecurity* (W911NF20D0002), and by the Office of Naval Research Minerva projects *Multi-Level Models of Covert Online Information Campaigns* (N000142112765), *Scalable Tools for Social Media Assessment* (N000142112229), and *Community Assessment* (N000142412568). Additional support was provided by IDeaS and CASOS at Carnegie Mellon University.
The views and conclusions expressed are those of the author and do not represent the official policies or endorsements of the U.S. Army, Office of Naval Research, Carnegie Mellon University, or the U.S. government.

**Keywords:** Synthetic Population Modeling, Opinion Dynamics Modeling, Social-Cyber Influence, Large Language Models (LLMs), Retrieval-Augmented Generation (RAG), Emotion-Aware Agent-Based Modeling, Salience and Persuasion, Information Environment Simulation

*To my husband, Connor, whose constant support made this possible, and to my daughter, Adeline, who grew alongside this journey and reminded me every day what truly matters.*

# Abstract

The contemporary information environment is increasingly shaped by large-scale online influence campaigns that exploit digital platforms to manipulate public opinion, amplify polarization, and erode trust in institutions. While synthetic population models and agent-based simulations have long been used to study social behavior, existing approaches struggle to fully represent the cognitive, emotional, and narrative mechanisms that drive belief formation and opinion change in modern digital ecosystems. Classical opinion dynamics models often assume psychologically homogeneous agents, static influence structures, and topic-agnostic interactions. These assumptions limit their applicability to social-cyber systems, where human discourse is deeply interwoven with algorithmically generated content.

This dissertation introduces an integrated, AI-driven framework, for synthetic populations and opinion dynamics modeling designed to simulate social-cyber influence in the online information environment. At its foundation, is a synthetic population modeling pipeline that constructs demographically grounded skeleton personas using empirical population data and statistical synthesis methods. These skeleton personas preserve population-level realism and internal consistency, while remaining intentionally sparse. This emables subsequent enrichment without sacrificing transparency, reproducibility, or control.

This framework introduces the use of AURORA (AI-Utilized Retrieval for Optimized Representation of Audiences), a Retrieval-Augmented Generation (RAG)–based persona modeling system, to enrich skeleton personas with psychologically and contextually realistic attributes. AURORA leverages large language models (LLMs), semantic vector databases, and salience-aware retrieval to assign emotional traits, ideological orientations, topic-specific salience, and narrative grounding to individual agents. This design enables the generation of internally coherent synthetic populations whose beliefs, emotions, and attention patterns are grounded in real-world discourse while remaining modular and configurable for experimental and operational use.

Building on these enriched populations, this dissertation presents JANUS (Joint Agent Network for Understanding Social Dynamics). JANUS is an extension of classical opinion dynamics models that incorporates emotion, salience, and persuasion into a mechanistic and interpretable update framework. JANUS extends the Friedkin–Johnsen model by introducing heterogeneous susceptibility, identity-conditioned signed influence, engagement probability, and affective drift. These extensions enable topic-dependent opinion evolution within psychologically diverse populations. Empirical evaluations demonstrate that the proposed framework sustains heterogeneous and polarized opinion states, without collapsing to consensus. It also reproduces empirically observed emotion and salience driven behaviors. Together, AURORA and JANUS form a unified foundation for scalable simulation of social-cyber influence, advancing the state of the art in synthetic population modeling, opinion dynamics, and information environment analysis.

# Acknowledgments

A PhD is often described as an individual achievement. But in truth, it is a collective effort carried out by many people over many years. This dissertation exists because of the love, patience, encouragement, and generosity of those who supported me through every stage of this journey.

I want to begin by thanking my advisor, Kathleen Carley. I am deeply grateful for her trust in me, her guidance, and the freedom she gave me to explore ambitious ideas. Working under her mentorship shaped not only this dissertation, but the way I think as a researcher. I am also thankful to my committee for their time, insight, and direction. Their questions and feedback strengthened this work, helping me see it more clearly.

My family made this journey possible in ways that go far beyond words. To my parents: thank you for always believing in me, for encouraging my curiosity, and for stepping in when I needed help the most. Caring for a newborn while starting a PhD is not something I could have done without you. To my daughter, thank you for being my constant reminder that joy exists even on the most stressful days. Your smiles, laughter, and presence kept me grounded and gave me strength when I felt overwhelmed.

Above all, I want to thank my husband, Connor Martin. You were there for every idea, every doubt, every late-night spiral, and every small victory. Thank you for listening endlessly, for helping me think through problems using your own expertise within this field of work, and for challenging me when my ideas needed sharpening. I truly could not have done this without you. This journey would not have been the same without the friends and collaborators who shared it with me. I am especially grateful to Peter Carragher, who not only co-authored two papers with me, but was also a constant source of great discussion. Thank you to Jake Shaha for our collaborations and the many conversations that helped shape both this work and my perspective. I am also deeply thankful to Lynnette Ng, a close friend and collaborator, for co-authoring a paper with me.

I also want to thank Evan Williams for being a constant friend, and Daniele Bellutta and Courtney Miller and most importantly Chanel for being the best office mates anyone could ask for. Sharing space, ideas, stress, and laughter with you made the long days feel lighter and the hard moments more bearable.

Finally, thank you to everyone in my research group and academic community who offered support, encouragement, and kindness along the way. This dissertation reflects not just my work, but the shared experience of the people who made this journey meaningful.

# Contents


**1 Introduction** **1**
- 1.1 Thesis Goals . . . 2
  - 1.1.1 Research Questions . . . 3
- 1.2 Data . . . 4
- 1.3 Tools . . . 6
  - 1.3.1 Existing Tools . . . 6
  - 1.3.2 Tools Developed for this thesis . . . 7

**2 Background and Related Work** **8**
- 2.1 Background . . . 8
  - 2.1.1 Social Cybersecurity and Networked Influence . . . 8
  - 2.1.2 Synthetic Population Modeling and Agent-Based Simulation . . . 9
  - 2.1.3 LLM-Augmented Opinion Modeling and Synthetic Social Behavior . . . 9

**3 Synthetic Persona (Skeleton) Generation** **11**
- 3.1 Structural Foundations for Synthetic Persona Modeling . . . 11
- 3.2 Introduction . . . 11
- 3.3 Background . . . 13
  - 3.3.1 Demographically Grounded Skeleton Personas . . . 13
- 3.4 Method and Model: Skeleton Persona Construction Pipeline . . . 14
  - 3.4.1 Design objectives and operating assumptions . . . 14
  - 3.4.2 Data and structural attribute space . . . 16
  - 3.4.3 Stage I: Iterative Proportional Fitting as marginally constrained maximum-entropy synthesis . . . 17
  - 3.4.4 Stage II: Semantic validation as a record-level filtering operator . . . 17
  - 3.4.5 Validation architectures evaluated . . . 18
- 3.5 Experimental Design and Evaluation . . . 20
  - 3.5.1 Experimental Design . . . 20
  - 3.5.2 Case Study Design . . . 21
  - 3.5.3 Evaluation Metrics . . . 22
  - 3.5.4 Fixed Parameters . . . 24
- 3.6 Results . . . 24
  - 3.6.1 Case Study 1: Direct LLM vs. Rule-Based Validation . . . 24

3.6.2 Case Study 2: Neural Network Approximation . . . 27
3.6.3 Case Study 3: Multi-Provider Robustness and Rule-Count Tuning . . . 29
3.6.4 Post-Filtering Distribution Preservation . . . 33
3.6.5 Validation . . . 35
3.7 Discussion . . . 37
3.7.1 Case Study 1: Direct LLM is not a scalable validator . . . 38
3.7.2 Case Study 2: Why naive distillation fails under realistic joint spaces . . . 38
3.7.3 Case Study 3: Deterministic rule execution enables provider robustness . . . 39
3.7.4 IPF Efficiency vs. Naive Sampling with Filtering. . . . 39
3.7.5 Interpreting rule count $K$: why "more rules" can be worse . . . 40
3.7.6 Implications for ABM construction and for LLM-assisted data synthesis . . . 41
3.8 Conclusion . . . 42
3.8.1 Contributions to this Dissertation . . . 42
3.8.2 Limitations . . . 43
3.8.3 Future Work . . . 44

**4 Modeling Emotional Reasoning in Synthetic Personas 46**
4.1 Introduction . . . 46
4.2 Related Works . . . 48
4.3 Method: Hierarchical Persona Synthesis . . . 50
4.3.1 Country Level Grounding . . . 50
4.3.2 Community Level Grounding . . . 52
4.3.3 Individual Level Grounding . . . 52
4.3.4 LLM-Generated Persona Dataset Composition and Demographic Calibration 53
4.3.5 Quality Control . . . 54
4.4 Model: Emotional Expression for Synthetic Personas . . . 55
4.4.1 Baseline Emotions . . . 55
4.4.2 Opinion Formation . . . 56
4.4.3 Event–Conditioned Emotion Modulation . . . 56
4.5 Experimental Design . . . 56
4.5.1 Experimental Conditions . . . 57
4.6 Results . . . 58
4.6.1 Institutional Community Generation and Validation . . . 58
4.6.2 Country-Level Emotions . . . 59
4.6.3 Community-Level Emotions . . . 60
4.6.4 Individual-Level Emotions . . . 60
4.6.5 Hierarchical Prompt Engineering Validation: Content Quality and Topic Emotion Effectiveness . . . 62
4.6.6 Topic-based Emotional Distributions . . . 63
4.6.7 Benchmarking Synthetic Personas . . . 65
4.6.8 Validation . . . 68
4.7 Discussion . . . 69
4.7.1 Hierarchical Prompt Engineering and the Importance of Structure . . . 70

4.7.2 Believability for Research Simulation . . . 71
4.7.3 Scenarios and Interventions . . . 71
4.8 Conclusions . . . 72
4.9 Contributions to the Dissertation . . . 72
4.9.1 Limitations . . . 73
4.10 Future Work . . . 74

**5 AURORA: Retrieval-Augmented Persona Modeling 75**
5.1 Introduction . . . 75
5.2 Related Work . . . 76
5.2.1 Simulating Human Behavior with Language Models . . . 76
5.2.2 Retrieval-Augmented Generation . . . 76
5.2.3 Opinion Dynamics and Social Influence . . . 77
5.2.4 Personality and Emotion in Agent-Based Models . . . 77
5.3 Method: AURORA Generation Pipeline . . . 77
5.3.1 Framework Overview and Design Principles . . . 78
5.3.2 Stage 0: Data Collection and Vector Database Construction . . . 79
5.3.3 Stage 1: Skeleton Persona Generation (Demographic Substrate) . . . 85
5.3.4 Stage 2: Country-Level Initialization . . . 87
5.3.5 Stage 3: Community Formation and Opinion Refinement . . . 87
5.3.6 Stage 4: Persona Enrichment (Skeleton Augmentation) . . . 89
5.3.7 Stage 5: Group Formation and Assignment . . . 95
5.3.8 Stage 6: Social Media Account Generation . . . 98
5.4 Model: Formal State Space Specification . . . 100
5.4.1 Agent State Representation . . . 100
5.4.2 Hierarchical Opinion Model . . . 101
5.4.3 Personality and Emotion Model . . . 103
5.4.4 Social Embedding Model . . . 105
5.4.5 Model Scope and Intended Use . . . 106
5.5 Experimental Design: IGOE Scenario . . . 107
5.5.1 Experimental Conditions . . . 107
5.5.2 Variables . . . 109
5.5.3 Design Structure . . . 111
5.5.4 Experimental Parameters . . . 111
5.6 Results . . . 112
5.6.1 Synthetic Population Validation . . . 112
5.6.2 Global Opinion and Salience Distributions . . . 115
5.6.3 Emotion Profiles by Topic . . . 117
5.6.4 Cross-Country Similarity Analysis . . . 120
5.6.5 Topic-Level Opinion and Salience Structure . . . 121
5.6.6 Validation . . . 124
5.7 Discussion . . . 128
5.7.1 Selective Polarization as a Structural Property . . . 128

5.7.2 Consensus on Importance, Divergence on Evaluation . . . 129
5.7.3 Emotion as a Coherent, Topic-Dependent Structure . . . 129
5.7.4 Demographics as Secondary, Constraining Factors . . . 130
5.7.5 Empirical Grounding of Country-Level Polarization in Narrative Structure . 131
5.7.6 Implications for Agent-Based Modeling . . . 132
5.8 Conclusion . . . 132
5.9 Contributions to the Dissertation . . . 133
5.10 Limitations . . . 134
5.11 Future Work . . . 135

**6 JANUS: Emotion-Aware Opinion Dynamics 136**
6.1 Introduction . . . 136
6.2 Related Works . . . 137
6.3 Method . . . 139
6.3.1 Network Construction from AURORA Personas . . . 140
6.3.2 Emotional and Contextual Conditioning . . . 144
6.3.3 Parameter Extraction for Opinion Dynamics . . . 144
6.3.4 Platform-Mediated Exposure and Alternative Definitions of $\widehat{W}_{ij}$ . . . 145
6.4 Model . . . 146
6.4.1 Model Overview . . . 146
6.4.2 Opinion Update Equation . . . 146
6.4.3 (I) FJ Anchor + Susceptibility . . . 148
6.4.4 (II) Signed Social Influence . . . 149
6.4.5 (III) Affective Emotional Drift . . . 154
6.5 Experimental Design . . . 157
6.5.1 Synthetic Population Generation . . . 157
6.5.2 Experimental Conditions . . . 157
6.5.3 Variables . . . 158
6.5.4 Platform-based Interaction in JANUS+Synthetic Social Media (SynSM) Simulation Loop . . . 160
6.6 Results . . . 160
6.6.1 Network Scaling and Local Interaction Structure . . . 160
6.6.2 Topic-Level Outcomes and Polarization Dynamics . . . 162
6.6.3 Comprehensive Network Scale Analysis . . . 163
6.6.4 Opinion Distribution Evolution . . . 165
6.6.5 Temporal Dynamics and Trajectory Analysis . . . 167
6.6.6 Polarization Trends Over Time . . . 169
6.6.7 Agent-Level Mobility and Stance Flips . . . 171
6.6.8 Long-Term Convergence and System Equilibration . . . 172
6.6.9 JANUS+SynSM: Topic-Specific Dynamics Under platform based interaction 173
6.6.10 Stylized-Facts Validation of JANUS Dynamics . . . 176
6.7 Discussion . . . 177
6.8 Conclusion . . . 179

6.9 Contributions to the Dissertation . . . 180
6.10 Limitations and Future Work . . . 181

**7 Evaluating Realism and Impact in Live Scenarios 183**
7.0.1 Introduction . . . 183
7.0.2 Related Works . . . 184
7.1 GhostField: Integrated Socio-Cognitive Scenario Simulation Architecture . . . 186
7.1.1 AESOP: Pre-Simulation Strategic Scenario Authoring . . . 188
7.1.2 AURORA: Demographically Grounded Population Instantiation . . . 188
7.1.3 Network Structure and Exposure Topology . . . 190
7.1.4 JANUS: Dynamic Opinion Evolution Engine . . . 191
7.1.5 SynSM: Platform-Specific Message Realization . . . 192
7.1.6 Ontological Separation: Strategic Actors and Cognitive Population Agents 192
7.1.7 Architectural Transformation: From Narrative Generator to Socio-Cognitive World Model . . . 193
7.2 Experimental Design: Indo-Pacific Geopolitical Opinion Experiment (IGOE) . . . 194
7.2.1 Entity Structure: AESOP Actors and AURORA Personas . . . 194
7.2.2 Institutional and Archetypal Structure . . . 195
7.2.3 Scenario Configuration . . . 196
7.2.4 Temporal Escalation Structure . . . 196
7.2.5 Intervention Design . . . 198
7.2.6 Persona Variables and Control Structure . . . 198
7.2.7 Scenario Pipeline and Experimental Objective . . . 198
7.3 Results . . . 200
7.3.1 Opinion Convergence and Variance Dynamics . . . 200
7.3.2 Magnitude of Opinion Change . . . 202
7.3.3 High-Impact Opinion Movers . . . 203
7.3.4 Intervention Effectiveness Analysis . . . 204
7.3.5 Summary of Key Findings . . . 206
7.4 Discussion . . . 208
7.5 Conclusion . . . 210
7.6 Contributions to the Dissertation . . . 211
7.7 Limitations . . . 212
7.8 Future Work . . . 212

**8 Conclusion 214**
8.1 Introduction . . . 214
8.2 Summary of the Integrated System . . . 215
8.2.1 Structural Foundations: Skeleton Personas . . . 215
8.2.2 Emotionally Grounded Persona Construction . . . 215
8.2.3 AURORA: Grounded Synthetic Population Initialization . . . 215
8.2.4 JANUS: Mechanistic Opinion Dynamics for Heterogeneous Populations . . 216

8.2.5 GhostField: From Synthetic Corpus Generation to Socio-Cognitive World Modeling . . . 216
8.2.6 Longevity of Contribution: Bridging LLMs and Realistic Opinion Dynamics 217
8.3 Dissertation-Level Contributions . . . 218
8.4 What the Dissertation Demonstrates as a Whole . . . 219
8.5 Limitations and Future Work . . . 220
8.5.1 Limitations . . . 220
8.5.2 Future Work . . . 221

**A Synthetic Persona (Skeleton) Generation** **224**
A.1 Country-Level Demographic and Socio-Political Reference Data . . . 224
A.2 ACS PUMS Data Structure and Variable Encoding . . . 225
A.2.1 Raw Data Format . . . 225
A.2.2 Variable Definitions and Encoding . . . 225
A.2.3 Empirical Marginal Distributions . . . 225
A.2.4 Age Distribution . . . 227
A.2.5 Focus Relation: Age $\times$ Education Joint Distribution . . . 227
A.2.6 Data Preprocessing and Caching . . . 227
A.2.7 Data Representativeness . . . 228

**B Modeling Emotional Reasoning in Synthetic Personas** **232**
B.1 OCEAN Personality Parameters . . . 232
B.2 Extended Persona Dataset Validation and Calibration . . . 233
B.2.1 Institutional Community Balance and Diversity . . . 233
B.2.2 Moral Foundations Consistency by Ideological Orientation . . . 234
B.2.3 OCEAN Personality Trait Calibration . . . 234
B.2.4 Response Structure and Emotional Intensity Distributions . . . 235
B.3 COVID-19 Emotion Analysis Validation . . . 235
B.3.1 Synthetic Persona Emotion Analysis . . . 236
B.3.2 Facebook COVID-19 Emotion Data . . . 238
B.3.3 Cross-Dataset Comparative Analysis . . . 238
B.3.4 Technical Interpretation . . . 238
B.3.5 Implications for Model Validation . . . 239

**C AURORA: Retrieval-Augmented Persona Modeling** **240**
C.1 Country Information . . . 240
C.2 Social Media Account Generation . . . 240
C.2.1 Twitter Account Generation . . . 240
C.2.2 Telegram Account and Channel Generation . . . 243
C.3 Opinion Model Implementation Details . . . 245
C.3.1 LLM Prompt Templates . . . 245
C.3.2 Mathematical Sampling Procedures . . . 246
C.3.3 Computational Cost . . . 246

C.3.4 Validation and Quality Assurance . . . 246
C.3.5 Emotion Modulation Implementation . . . 247
C.4 Extended Emotion Profile Analyses . . . 247
C.4.1 Russian Aggression . . . 247
C.4.2 Canadian Activity in INDOPAC . . . 247
C.4.3 Cyber Security Cooperation . . . 249
C.4.4 Summary . . . 249
C.5 Retrieval-Augmented Generation Architecture . . . 249
C.5.1 Method . . . 250
C.5.2 Model . . . 255
C.6 API Token Requirements and Generation Scalability . . . 259
C.7 AURORA / AESOP group crosswalk . . . 261

**D JANUS: Emotion-Aware Opinion Dynamics 269**
D.1 Intermediate Network Scale Analysis . . . 269
D.1.1 Dual-500 Scenario: Transitional Regime . . . 269
D.1.2 Dual-750 Scenario: Averaging-Dominated Regime . . . 272
D.1.3 Cross-Scale Comparison: Intermediate Scenarios . . . 276

**E Evaluating Realism and Impact in Live Scenarioss 277**
E.1 AURORA Persona Integration: Ego Network Validation and AESOP Cross-Walk . 277
E.1.1 Ego Network Comparison . . . 277
E.1.2 Network Structure and Centrality . . . 279
E.1.3 Topic Engagement as Behavioral Differentiation . . . 279
E.1.4 Validation of Persona Distinctiveness . . . 281
E.1.5 AESOP Cross-Walk: Structural Integration . . . 281
E.1.6 Complementary Network Roles . . . 281
E.2 Temporal Escalation Structure . . . 285
E.3 Detailed Opinion Dynamics Analysis . . . 287
E.3.1 Cross-Country Variance Analysis . . . 287
E.3.2 Timestep-Level Dynamics . . . 287
E.3.3 Statistical Significance Testing . . . 287
E.4 Stance Reversal Analysis . . . 290
E.4.1 Definition and Distribution . . . 290
E.5 Extended Top Movers Analysis . . . 292
E.5.1 Topic and Temporal Patterns . . . 292
E.6 Comprehensive Intervention Analysis . . . 292
E.6.1 JANUS Integration Details . . . 292
E.6.2 Comparative Actor Analysis . . . 292
E.6.3 Deep-Dive Case Studies . . . 294
E.6.4 The Magnitude–Direction Paradox: Extended Analysis . . . 294
E.6.5 Network-Level Influence Concentration . . . 295
E.6.6 Statistical Validation: Extended Results . . . 295

E.6.7 Case Study: Opinion Dynamics of a Highly Susceptible Actor . . . . . 295
E.7 Post-Simulation Survey Analysis . . . . . 299
E.7.1 Survey Methodology and Complete Results . . . . . 299

**Bibliography** **301**

# List of Figures


3.1 **Skeleton Persona Generation Pipeline.** Marginal distributions are extracted from ACS PUMS data and matched using Iterative Proportional Fitting (IPF) to generate statistically grounded skeleton personas. Optional semantic validation is applied using one of three strategies: direct LLM validation, rule-based filtering, or neural network approximation. Outputs are evaluated using marginal L1 distance, Jensen–Shannon divergence, and age–education joint fidelity. The resulting skeleton populations are behaviorally neutral and suitable for downstream cognitive or narrative enrichment. . . . . 15

3.2 **Case Study 1: Multi-run Age-binned $\times$ Education joint analysis (n=100 runs).** Panels show joint L1 distance (left), joint JS divergence (center), and combined score (right). Lower is better. . . . . 26

3.3 **Case Study 2: Distribution and run-level stability (n=100).** (Left) Marginal L1 box plot; (center) marginal JS box plot; (right) focus-relation L1 box plot. Neural validation exhibits consistently higher drift than Rule-Based across runs with limited overlap, indicating systematic (not stochastic) degradation. . . . . 28

3.4 **Case Study 2: Run-wise trajectories over 100 runs (lower is better).** (Left) Marginal L1, (center) marginal JS, (right) focus relation L1 for Age$\times$Education. Neural validation remains consistently elevated across runs on all metrics, indicating stable but systematically incorrect filtering behavior relative to Rule-Based. . . 29

3.5 **Case Study 3 (GPT-4.1-mini): Marginal quality vs. rule count (n=50).** Combined score (L1+JS) and its components increase with $K$, reflecting increasing marginal drift under more restrictive filtering. . . . . 30

3.6 **Case Study 3 (GPT-4.1-mini): Age$\times$Education focus relation vs. rule count (n=50).** Moderate rule sets improve the cohort joint (lower is better), while very large $K$ increases variance and can reintroduce distortion via over-filtering. . . . . 30

3.7 **Case Study 3 (Claude 3.5-Haiku): Marginal quality vs. rule count (n=50).** . . . 32

3.8 **Case Study 3 (Claude 3.5-Haiku): Age$\times$Education focus relation vs. rule count (n=50).** . . . . . 32

3.9 **Case Study 3 (Gemini 2.0-Flash): Marginal quality vs. rule count (n=50).** . . . 33

3.10 **Case Study 3 (Gemini 2.0-Flash): Age$\times$Education focus relation vs. rule count (n=50).** . . . . . 33

4.1 Stepwise pipeline for persona construction and emotion modeling. News grounding informs hierarchical synthesis (country, community, individual), while traits, opinions, and salience shape event-driven emotions. . . . . 51
4.2 Demographic composition of LLM-generated personas showing age distribution (top) and political orientation distribution (bottom) across US and Canadian institutional communities. . . . . 55
4.3 Compact experimental pipeline. . . . . 58
4.4 Country-level primary emotions for Roe v. Wade (left) and Suicide Hotline 988 (right). . . . . 60
4.5 Community-level primary emotions for Roe v. Wade (left) and Suicide Hotline 988 (right). . . . . 61
4.6 Emotion distribution heatmap for Roe v. Wade topic by ablation condition. . . . . 64
4.7 Emotion distribution heatmap for Suicide Hotline 988 topic by ablation condition. . 65
4.8 Comparison of Facebook reactions and synthetic persona sentiment reactions for Roe v. Wade across demographic categories. . . . . 66
4.9 Comparison of Facebook reactions and synthetic persona sentiment reactions for Suicide Hotline 988 across demographic categories. . . . . 67

5.1 AURORA end-to-end pipeline overview. The framework ingests and embeds documents into a vector database, retrieves context for structured generation, and produces communities of synthetic actors through hierarchical prompt generation with integrated quality control. . . . . 79
5.2 AURORA feature dimensions for persona construction, including demographics, derived attributes, digital environment, psychological traits, topic opinions, salience, community affiliation, and dynamic context. . . . . 80
5.3 AURORA Retrieval-Augmented Generation Pipeline. Document corpora are embedded into a semantic vector store. Contextual queries retrieve relevant evidence, which conditions LLM-based community and persona generation. . . . . 84
5.4 Community-level RAG in AURORA. Queries retrieve evidence chunks about Community $X$ within Country $Y$, which are combined with prompts to generate structured community outputs (demographics, key topic opinions, salience) that downstream persona generation conditions on. . . . . 88
5.5 Persona-level RAG in AURORA. Retrieved evidence chunks combine with skeleton constraints and community context to generate persona-level outputs including demographics, topic opinions, and salience. . . . . 90
5.6 Salience-weighted opinion distribution pipeline. AURORA maps opinion polarity and salience categories into distribution families(tight vs broad vs bimodal), producing persona-level opinions that preserve community coherence on high-salience issues while allowing heterogeneity on low-salience topics. . . . . 93
5.7 Synthetic population demographic and behavioral distributions: gender, age, political leaning, and social media usage. . . . . 114
5.8 Global distributions of opinion scores (left) and salience scores (right) aggregated across all topics and countries. . . . . 115

5.9 Mean Plutchik emotion profiles for Taiwanese Sovereignty across all countries. . . 118
5.10 Mean Plutchik emotion profiles for International Commodities Markets across all countries. . . . . 119
5.11 Country similarity matrices based on mean opinion scores (left) and mean salience scores (right). . . . . 120
5.12 Mean opinion scores by country and topic. . . . . 122
5.13 Mean salience scores by country and topic. . . . . 123
5.14 Principal Component Analysis of country-level opinion patterns. Left: scree plot showing variance explained by each component. Right: PCA biplot with countries and topic loadings. . . . . 124

6.1 Pipeline from AURORA personas to JANUS opinion dynamics. . . . . 141
6.2 Valence-arousal circumplex with Plutchik-style emotion placement (0° at positive valence, angles increasing counterclockwise). Vector lengths represent normalized psychometric weights (trust is longest at 1.0, fear is shortest at 0.375 relative to trust), reflecting empirical affect structure. Emotion intensities are projected onto the valence axis via $\cos(\theta_e)$ and onto the arousal axis via $\sin(\theta_e)$. . . . . 155
6.3 Dual-250 community structure under the Fruchterman-Reingold spring layout with group co-membership weights. Dense regions represent tightly connected communities formed through shared group affiliations, while sparse bridges enable cross-community exposure. The high clustering coefficient ($C = 0.88$) reflects strong transitivity within communities, while the relatively low density ($d = 0.074$) indicates that most potential connections remain unrealized. . . . . 162
6.4 Opinion dynamics evolution for Dual-250. (a) Mean absolute opinion change declines over time, indicating stabilization in the rate of updating rather than convergence to consensus. (b) Cumulative opinion change rises throughout the simulation, reaching 0.072 for COVID-19 and 0.079 for Roe v. Wade. . . . . 165
6.5 Opinion dynamics evolution for Dual-1000. (a) Mean absolute opinion change declines smoothly over time, indicating gradual stabilization in the rate of updating. (b) Cumulative opinion change continues to rise throughout the simulation, but to a slightly lower total level than in Dual-250, reaching 0.038 for COVID-19 and 0.041 for Roe v. Wade. . . . . 166
6.6 Opinion trajectories for Dual-250 (N=240). Gray lines show all agents, yellow highlights top 10 most mobile agents, red marks stance flippers. Both topics exhibit rapid initial adjustment (t ¡ 50) followed by stabilization. High-mobility agents show sustained volatility while majority remain near initial positions. COVID-19: 5 flips (2.08%); Roe: 3 flips (1.25%), all during early transient phase. . . . . 168
6.7 Opinion trajectories for Dual-1000 (N=1000). Maximum network scale produces smoothest trajectories, with average degree 73.5 creating strong averaging effects. COVID-19: 13 flips (1.35%); Roe: 6 flips (0.63%, lowest rate across all scenarios). Increased absolute flip count for COVID combined with decreased rate for Roe demonstrates topic-specific responses to network scale. . . . . 169

6.8 Polarization dynamics for the baseline JANUS experiment across both topics and network scales. Solid lines denote the 250-persona scenarios and dashed lines denote the 1000-persona scenarios. Roe v. Wade remains more polarized than COVID-19 at both scales, while network size changes the direction of polarization evolution: the 250-persona scenarios show gradual amplification over time, whereas the 1000-persona scenarios stabilize quickly and remain nearly flat thereafter. This combined view highlights the scale-dependent reversal from reinforcement-dominated dynamics in smaller networks to near-equilibrium behavior in larger networks. . . . . 170

6.9 Long-term convergence behavior for Dual-1000 extended to 1000 timesteps. Individual trajectories (left) show complete stabilization after $t \approx 50$, with only noise-driven micro-fluctuations thereafter. Polarization (right) converges to stable equilibrium values with negligible drift ($< 0.001$ per 150 timesteps), confirming that the system reaches convergence. . . . . 173

6.10 Mean opinion trajectories for Canadian and American agents across two topics under platform based interaction. Left: *Canadian Indigenous Rights*. Right: *Russian Arctic Activity*. Shaded regions indicate different simulation phases. . . . . 174

7.1 GhostField integrated architecture: AESOP scenario authoring, AURORA population instantiation, JANUS belief dynamics, and SynSM platform realization within a time-stepped orchestration loop. . . . . 187

7.2 Opinion volatility (standard deviation of valences) by topic across all timesteps. Each line represents one of the ten geopolitical topics. Volatility was smoothed using a 10-timestep rolling window. Shaded regions indicate uncertainty bands for the top three topics. . . . . 201

7.3 Distribution of absolute valence changes from Day 1 to Day 9. Left panel shows distributions by topic; right panel shows distributions by country. Box plots indicate median (center line), interquartile range (box), and outliers (whiskers). . . . . 202

7.4 Top 20 opinion movers (individuals only) ranked by total absolute valence change. Bar length indicates magnitude of change. . . . . 204

7.5 Effect Size Rankings Across All Intervention Actors. Green bars indicate actors with statistically significant effects ($p < 0.05$). Error bars show min-max range across topics. . . . . 206

7.6 Directional Alignment: Sphere vs Non-Sphere Actors. Grouped bars show percentage of actors shifting toward intervention stance. Annotations show percentage point differences. Sphere actors consistently show 20-50pp higher alignment rates. . . . . 207

B.1 Response structure metrics across Roe v. Wade and Suicide Hotline 988 responses. Shown distributions include word count, sentence count, average word length, punctuation ratio, and affective word ratios. . . . . 236

B.2 **Synthetic Persona Emotion Analysis by Demographics.** Average emotion scores derived from persona-generated COVID-19 prompts, disaggregated by gender (left) and age group (right). Emotions include Joy, Anger, Sadness, Fear, Disgust, and Surprise. Results indicate stronger negative valence (sadness and fear) among older cohorts and greater emotional variability among females. . . . . 237

B.3 **Facebook COVID-19 Emotion Data by Demographics.** Normalized emotion distributions from real-world Facebook interactions ("Hahas," "Loves," "Sorries," "Angers," "Wows") during the height of the pandemic. Emotion intensity is averaged by gender (left) and age bracket (right). Real data show similar demographic gradients as the synthetic population, validating model realism. . . . . 237

C.1 Mean Plutchik emotion profiles for Russian Aggression across all countries. . . . . 248

C.2 Mean Plutchik emotion profiles for Canadian Activity in INDOPAC across all countries. . . . . 248

C.3 Mean Plutchik emotion profiles for Cyber Security Cooperation across all countries. 249

C.4 AURORA RAG pipeline showing offline indexing and online retrieval phases. Dashed arrows indicate cross-phase data access. PRF optionally expands queries with terms from top-ranked documents. . . . . 251

D.1 Complete opinion distribution and coordination analysis for Dual-500. (a-b) Violin plots show topic-dependent responses: COVID displays modest widening while Roe shows slight narrowing. (c-d) Temporal dynamics reveal smoother evolution compared to Dual-250, with coordination strength of 0.075 (COVID) and 0.084 (Roe). (e-f) Community structure shows reduced separation compared to Dual-250 (0.032 vs 0.072 for COVID), reflecting increased cross-group exposure from larger neighborhoods. . . . . 270

D.2 Opinion trajectories for Dual-500 (N=480). Trajectories show smoother evolution with reduced volatility compared to Dual-250, reflecting larger neighborhoods (avg degree 36.0). COVID: 7 flips (1.46%); Roe: 4 flips (0.83%). Despite larger absolute counts, flip rates decrease, indicating enhanced boundary stability with scale. . . . . 271

D.3 Polarization over time for Dual-500. COVID continues increasing ($\Delta\sigma = +0.021$), though attenuated vs Dual-250, while Roe exhibits slight decrease ($\Delta\sigma = -0.011$). This divergence marks the transition point where scale effects begin to differentiate topic responses. After t=100, both trajectories plateau, indicating quasi-equilibrium. 272

D.4 Complete opinion distribution and coordination analysis for Dual-750. (a-b) Violin plots show visible narrowing for both topics, with peaks becoming more concentrated and tails shortening. (c-d) Temporal dynamics show coordination strength of 0.090 (COVID) and 0.078 (Roe), intermediate between Dual-500 and Dual-1000. (e-f) Community structure shows COVID separation of 0.068 (higher than Dual-500's 0.032), while Roe shows 0.034 (reduced from 0.052), reflecting topic-specific responses to averaging pressure. . . . . 273

D.5 Opinion trajectories for Dual-750 (N=720). Further smoothing and constraint visible, with majority showing minimal movement after initial stabilization. Average degree 55.7 creates substantial averaging pressure, visible in tighter clustering of gray trajectories. Both topics: 8 flips (1.11%), lowest rate among scenarios, indicating enhanced boundary stability. . . . . 275

D.6 Polarization over time for Dual-750. Both topics show decreasing polarization with identical magnitudes ($\Delta\sigma = -0.024$), confirming systematic reversal from amplification to moderation. Unlike smaller scenarios where polarization increases after transient phase, Dual-750 shows consistent downward trend throughout simulation. . . . . 275

E.1 Labeled ego networks for two AURORA personas and one AESOP actor. Aarav Patel and Mai Yamasaki exhibit visibly different ego-network structures, including different alter sets, topic clusters, and connectivity patterns, while the AESOP actor myudn shows a more actor-centered configuration. Only actor-to-actor ties above the average strength threshold are displayed for clarity. . . . . 278

E.2 Complete ego network comparison showing Aarav Patel (AURORA, left), Mai Yamasaki (AURORA, center), and myudn (AESOP, right). Node colors indicate type: yellow = ego, green = topics, blue = groups, orange = community, grey = actors. Networks are constructed from XML ego network files with topic filtering applied to show only discussed topics. Only actor ties above the average strength threshold are visualized for clarity. . . . . 279

E.3 Country-aggregated mean valence trajectories across all ten topics over the 9-day simulation period. Each subplot represents one topic, with colored lines indicating different countries. . . . . 288

E.4 Timestep dynamics heatmap showing mean absolute opinion changes by topic across all day-timestep combinations. Color intensity indicates the magnitude of opinion change, with darker colors representing higher volatility. . . . . 289

E.5 Stance reversal analysis. Top panel shows the number of flips by country; bottom panel shows the temporal distribution of flip events across the 9-day simulation period. . . . . 291

E.6 Network Reach by Intervention Actor. Bars show sphere size (number of actors) with percentage of total network. All sphere actors were Tier 1 (direct interactions only). . . . . 293

E.7 Chen Jun's opinion trajectories across all ten geopolitical topics over the 9-day simulation period. Each subplot shows valence evolution for one topic (y-axis: valence from -1 to +1; x-axis: simulation day). The shaded orange region indicates the intervention period (Days 5–7) when custom actors posted 176 intervention messages to the network. Notable opinion shifts are visible during this period for multiple topics, particularly *Canadian indigenous rights*, *Russian aggression*, and *Taiwanese Sovereignty*. . . . . 297

E.8 Comparison of Chen Jun's opinion changes across three temporal periods for all ten topics. Bar heights indicate the magnitude of opinion change (absolute valence shift). The intervention period (Days 5–7, orange bars) shows substantially larger changes than pre-intervention (Days 1–4, blue bars) or post-intervention (Days 7–9, green bars) periods. Topics showing the largest intervention-period changes include *Canadian indigenous rights* (0.20), *Russian aggression* (0.15), and *Taiwanese Sovereignty* (0.10). . . . . 298

# List of Tables


3.1 Experimental variables, roles, and operationalization. . . . . 20
3.2 Experimental conditions across case studies. . . . . 22
3.3 Fixed parameters across all experiments. . . . . 24
3.4 **Case Study 1: Marginal distributional accuracy (n=100 runs; 1,500 records).** Rule-Based validation produces significantly lower marginal drift than Direct LLM. 25
3.5 **Case Study 1: Focus relation accuracy (Age×Education; n=100 runs).** Lower is better. Direct LLM yields the lowest Age×Education error; Direct-vs-Rule is not statistically significant across runs ($p = 0.2863$). . . . . 25
3.6 **Case Study 1: Computational efficiency and cost (1,500-record populations).** . 26
3.7 **Case Study 2: Marginal distributional accuracy (n=100 runs; $N_{\text{syn}}$=2,500).** Distilled neural validation increases population-level drift relative to Rule-Based. . 27
3.8 **Case Study 2: Focus relation performance for Age×Education (n=100 runs).** Lower is better. Rule-Based outperforms Neural on L1, JS, and combined score. . . 28
3.9 **Root causes of distillation failure in Case Study 2.** . . . . . 29
3.10 **Rule count optimization: detailed analysis (GPT-4.1-mini; n=50).** . . . . . 31
3.11 **Case Study 3: Quality metrics by LLM provider (50 runs per provider; 2,500 records; $K = 8$).** . . . . . 31
3.12 **Empirical rejection rates by rule count (GPT-4.1-mini; $n$=50 runs per $K$; $N_{\text{syn}}$=2,500).** Rejection rates increase monotonically with $K$, reflecting stronger semantic enforcement. The coefficient of variation (CV) decreases with $K$, indicating more predictable filtering behavior under larger rule sets. . . . . 34
3.13 **Distribution quality after rule-based filtering ($K$=8; representative run; $N_{\text{syn}}$=2,500).** Marginal L1 distances remain below 0.10, indicating acceptable preservation of univariate demographic structure. The Age×Education focus relation shows improvement relative to unfiltered IPF. . . . . 34
3.14 **Demographic rejection bias summary ($K$=8; 10 runs; 34,109 total records).** Range indicates the difference between the highest and lowest rejection rates within each variable's categories. Education (SCHL) shows the highest variability, confirming it as the primary driver of rejection. . . . . 35

4.1 Summary of synthetic persona dataset used in experiments, with comparison to national demographic statistics. . . . . 53
4.2 Distribution of LLM-generated personas across institutional categories. . . . . 54

4.3 Experiment levels ($\mathcal{L}$) used in persona prompting. Each level specifies which emotional, moral, and psychological fields are available to the model. . . . . 58
4.4 Representative individual-level emotion scores by nationality and topic. . . . . 61
4.5 Content performance across ablation conditions . . . . . 62
4.6 D+M+O+T emotional advantages by topic relative to baseline (D). . . . . 63
4.7 Cross-platform validation correlation matrix with bootstrap confidence intervals. . 68

5.1 AURORA Country Information Bank: Standardized Community Categories . . . . 81
5.2 Hierarchical RAG Integration Across the AURORA Pipeline . . . . . 84
5.3 Trait-Emotion Mapping Coefficients . . . . . 92
5.4 IGOE Scenario Country Selection and Geopolitical Alignment . . . . . 108
5.5 IGOE Scenario Topic Selection and Characteristics . . . . . 109
5.6 Summary of Experimental Variables . . . . . 110
5.7 IGOE Scenario Experimental Parameters . . . . . 111
5.8 Population distribution and age validation across 12 countries ($N = 4{,}320$ total personas). . . . . 113
5.9 Synthetic population demographic composition ($N = 4{,}320$ personas). . . . . 113
5.10 Behavioral gradients by age (mean $\pm$ SD). . . . . 114
5.11 Topic-level opinion and salience summary statistics (all countries). . . . . 116
5.12 Demographic validation tests by topic (gender $t$-tests; age ANOVA; political leaning correlation). . . . . 117
5.13 Summary of topic-level opinion and salience structure. . . . . 128

6.1 Structural characteristics of JANUS persona networks at $t = 0$. . . . . 141
6.2 MOMUS-based structural benchmark of the baseline JANUS $t = 0$ network relative to empirical Twitter benchmarks. Rubric scores are read directly from MOMUS; values from continuous plots are approximate visual estimates. . . . . 143
6.3 Key symbols in the main opinion update (Eq. 6.7). . . . . 147
6.4 Plutchik emotion angles and psychometric weights used in circumplex projection. . 154
6.5 Fixed parameters across conditions. . . . . 158
6.6 Experimental conditions (U.S. 2022; two topics). . . . . 158
6.7 Summary of experimental variables used in the JANUS population-scale analysis. . 159
6.8 Topic–population cases evaluated in the Canadian–American JANUS+SynSM experiment. . . . . 161
6.9 Network scaling by population size. All networks exhibit sparse connectivity ($\rho \approx 0.076$) and high clustering ($C \approx 0.89$) while interaction volume scales quadratically with $N$. . . . . 161
6.10 Topic-level outcomes by scenario ($t = 0 \to t = 250$). $\Delta\sigma \equiv \sigma_{250} - \sigma_0$ measures polarization change. "Flips" counts agents whose stance sign changed between initial and final states. Avg. $|\Delta x|$ represents mean cumulative opinion change per agent. . . . . 163

6.11 Comprehensive metrics across all network scales. Polarization change ($\Delta\sigma$) shows systematic reversal from amplification (250) to stabilization (1000). Within-group trajectory correlations remain positive while between-group trajectory correlations remain negative across all scales, demonstrating persistent community structure. Here, $r_{\text{within}}$ denotes the mean pairwise correlation of opinion trajectories among agents in the same community, and $r_{\text{between}}$ denotes the mean pairwise correlation among agents in different communities. . . . . 164
6.12 Trajectory-correlation dynamics across network scales. Mean trajectory correlation ($\bar{r}$) remains near zero across all scenarios, indicating that positive and negative co-movement patterns balance out at the population level rather than implying a lack of structure. Correlation strength (mean $|r|$) remains consistently positive, showing sustained correlated opinion movement. The separation metric ($r_{\text{within}} - r_{\text{between}}$) quantifies community polarization and remains positive across all scales, while the fraction of positive correlations indicates balanced opposing movement patterns. . . . . 164
6.13 Trajectory dynamics across network scales. Mean opinion change quantifies average individual mobility. Flip rates remain low across all scenarios, with slight decrease at intermediate scales. High-mobility agents (top 10 by cumulative $|\Delta x|$) show sustained volatility while the majority stabilize quickly. . . . . 167
6.14 Polarization evolution patterns across network scales. Initial polarization ($\sigma_0$) varies by topic due to persona initialization. Final polarization ($\sigma_{250}$) shows systematic scale effects: amplification in small networks, stabilization in large networks. Relative change ($\Delta\sigma/\sigma_0$) quantifies proportional shifts. . . . . 170
6.15 Stance flip analysis across scenarios. Flip rate is calculated as the percentage of agents who changed opinion sign ($x_0 \cdot x_{250} < 0$) between initial and final states. . . 172
6.16 Aggregate opinion dynamics across the JANUS+SynSM Canadian–American subexperiment and baseline JANUS comparison. Day 1 and Day 9 values report mean valence with standard deviation in parentheses. . . . . 175

7.1 Core architectural layers in GhostField. . . . . 188
7.2 AESOP scenario-authoring outputs. . . . . 189
7.3 AURORA persona representation within GhostField. . . . . 190
7.4 Ontological distinction between AESOP actors and AURORA personas. . . . . 193
7.5 IGOE entity counts by country: AESOP strategic actors vs. AURORA cognitive personas (Toto Scenario 2026). . . . . 195
7.6 Institutional community types modeled in each country. . . . . 195
7.7 Aggregate AESOP actor archetype composition (grouped categories) in Toto Scenario 2026. . . . . 196
7.8 Scenario and experimental design overview. . . . . 197
7.9 Geopolitical topic space. . . . . 197
7.10 Intervention design summary. . . . . 198
7.11 Persona parameterization and experimental variables. . . . . 199
7.12 Fixed control structure. . . . . 199

7.13 Two-tranche scenario pipeline. . . . . 199
7.14 Opinion Variance Convergence Metrics by Topic . . . . . 201
7.15 Distribution of Top Opinion Movers by Country . . . . . 203
7.16 Top 10 Individual Opinion Movers (All Countries and Topics) . . . . . 203
7.17 Overall Intervention Effectiveness . . . . . 205
7.18 Opinion Shift Magnitudes (Individual-Level Analysis) . . . . . 205

A.1 Country-level demographic, digital, socioeconomic, and political attributes used for AURORA skeleton persona generation (32 countries). . . . . 226
A.2 Sample records from ACS PUMS dataset . . . . . 227
A.6 Age distribution in 10-year bins. . . . . 227
A.3 Variable definitions and encoding schemes. . . . . 229
A.4 Marginal distributions for discrete demographic variables. . . . . 230
A.5 Marginal distribution for educational attainment (SCHL). . . . . 230
A.7 Selected cells from Age $\times$ Education joint distribution (proportions). . . . . 231

B.1 Country-Level OCEAN Personality Parameters . . . . . 232
B.2 Regional-Level OCEAN Personality Parameters . . . . . 233
B.3 Distribution of LLM-generated personas across institutional categories. . . . . 233
B.4 Moral Foundations Theory validation: ideology-to-moral-foundation mappings by political orientation. . . . . 234
B.5 OCEAN personality trait validation: LLM-generated vs. empirical research data. . 235
B.6 Cross-Dataset Emotion Comparison . . . . . 238

C.1 Country Coverage in AURORA by Region and Data Availability . . . . . 241
C.2 Embedding API Specifications . . . . . 255
C.3 Token Consumption by Generation Stage . . . . . 260
C.4 Cost Summary for Japan Scenario . . . . . 260
C.5 Observed Runtime for Japan Scenario . . . . . 260

D.1 Comprehensive metrics for Dual-500 scenario. This transitional regime shows topic-dependent polarization responses: COVID continues amplification while Roe begins moderation, marking the reversal point in scale effects. . . . . 271
D.2 Comprehensive metrics for Dual-750 scenario. Both topics now show decreasing polarization with identical magnitudes ($\Delta\sigma = -0.024$), indicating convergent dynamics where structural network effects dominate over topic-specific characteristics. 274
D.3 Direct comparison of intermediate scenarios. The transition from Dual-500 to Dual-750 shows systematic shifts: topic-dependent responses converge to symmetric moderation, flip rates reach minimum, and coordination patterns stabilize. . 276

E.1 Comprehensive Ego Network Statistics (All Node Types) . . . . . 280
E.2 Top Hashtags/Topics by Persona . . . . . 281
E.3 AURORA Persona Validation: Aarav vs. Mai Comparison . . . . . 282
E.4 AURORA vs. AESOP Comparison: Complementary Capabilities . . . . . 284

E.5 Temporal escalation structure (Baseline–Day 4). . . . . 285
E.6 Temporal escalation structure (Day 5–Day 9). . . . . 286
E.7 ANOVA Results: Cross-Country Differences in Absolute Opinion Change by Topic 289
E.8 Independent Samples t-tests: NATO-aligned vs. China/Russia Opinion Changes . . 290
E.9 Stance Reversals by Topic . . . . . 290
E.10 Stance Reversals by Country . . . . . 291
E.11 Detailed Success Factors Analysis . . . . . 293
E.12 DefenseDave's Top 10 Influenced Actors (Detailed) . . . . . 294
E.13 Extended Practical Significance Assessment . . . . . 294
E.14 Most-Influenced Network Actors (Complete) . . . . . 295
E.15 Statistical Analysis of Chen Jun's Opinion Changes by Temporal Period . . . . . 296
E.16 Post-Simulation Survey Response Distribution (Complete) . . . . . 300

# Chapter 1

# Introduction

Understanding how beliefs form, polarize, and spread within online populations is increasingly important for both scientific inquiry and operational planning. Online information environments have become contested spaces, where influence operations, coordinated messaging, and emotionally charged discourse interact at scale. Research in social cybersecurity shows that these environments can be deliberately shaped to manipulate public perception and erode institutional trust [33]. At the same time, social network structure plays a central role in belief reinforcement, as repeated interaction within homophilous communities amplifies alignment and resistance to competing narratives [46]. Empirical studies of online political discussion further demonstrate that selective exposure contributes to the persistence and amplification of opinion clusters over time [9].

Research in opinion dynamics has shown that collective outcomes such as consensus, fragmentation, and polarization emerge from the interaction between individual susceptibility, persuasion, and network structure. The Friedkin–Johnsen model illustrates how agents can remain anchored to prior beliefs while still responding to social influence, enabling polarized outcomes even under repeated interaction [61]. Extensions of this framework show that shared norms and local reinforcement can stabilize divergent belief states [60].More recent analytical and simulation-based work demonstrates that polarization can arise endogenously through heterogeneous susceptibility, influence weighting, and connectivity, all without external shocks [22]. Related models confirm that polarized opinion distributions can represent stable equilibria under realistic conditions of persuasion and uncertainty, particularly in networked environments [18, 30, 111, 124].

Existing approaches to synthetic population modeling and opinion dynamics provide useful abstractions for aggregate behavior, but often fail to capture the psychological, emotional, and contextual mechanisms underlying belief formation. Traditional population synthesis methods rely on statistical alignment techniques such as Iterative Proportional Fitting (IPF) to reproduce demographic marginals [198]. Comparative studies highlight their effectiveness for demographic realism [93], while extensions address additional structural constraints [208]. However, these methods do not encode how individuals attend to information, emotionally respond, nor selectively engage with narratives. Similarly, many opinion dynamics models emphasize structural influence, while abstracting away agent-level cognition, affect, and topic relevance [211].

Recent advances in large language models (LLMs) have renewed interest in using generative

systems to simulate social behavior and opinion formation. Generative agent frameworks show that LLMs can produce coherent social behavior when equipped with memory and reflection mechanisms [149]. Surveys of LLM-augmented agent-based modeling highlight their potential for representing heterogeneous agents at scale [63]. At the same time, challenges remain around control, scalability, and interpretability [76]. Retrieval-Augmented Generation (RAG) improves contextual grounding by conditioning generation on external knowledge sources [109]. Enterprise-focused frameworks demonstrate how retrieval pipelines can support reproducible generation [158]. Critical perspectives caution, however, that black-box use of LLMs risks epistemic uniformity and questionable behavioral fidelity [87, 161]. This thesis addresses these challenges by integrating statistical population synthesis, RAG-enhanced persona modeling, and interpretable opinion dynamics into a unified framework for simulating belief evolution in synthetic social media populations.

## 1.1 Thesis Goals

The primary goal of this thesis is to develop an integrated framework for simulating belief formation, polarization, and narrative diffusion within synthetic social media environments. Rather than modeling populations solely through demographic alignment or abstract influence rules, the framework combines population realism with psychologically and emotionally grounded agent behavior. This design is motivated by prior work in social-cyber influence emphasizing the need for interpretable and scalable models capable of supporting both research and applied analysis in contested information environments [33].

The first objective is to construct demographically grounded *skeleton personas* using structured population data and established statistical synthesis techniques such as Iterative Proportional Fitting (IPF) [198]. Comparative evaluations of population synthesis methods demonstrate that IPF-based approaches effectively preserve population-level distributions [93], while earlier synthetic population frameworks highlight the value of maintaining demographic realism without premature behavioral specification [130]. By remaining intentionally minimal, these skeleton personas provide a transparent and flexible foundation for later behavioral enrichment, consistent with prior work advocating staged population construction [15, 65].

The second objective is to enrich skeleton personas with psychologically and contextually coherent attributes using **AURORA**, a Retrieval-Augmented Generation (RAG)–based persona modeling system. AURORA integrates large language models with semantic vector search to ground persona generation in retrieved contextual information rather than static prompts [109]. Enterprise oriented RAG architectures further demonstrate how retrieval pipelines can support structured and reproducible generation workflows [158]. Prior research shows that LLMs can generate diverse and internally consistent personas when appropriately conditioned on contextual cues [167], including cultural and ideological variation [103]. More recent work highlights how synthetic persona generation can scale while maintaining coherence across attributes [10].

The third objective is to extend classical opinion dynamics models through **JANUS**, an agent-based framework that incorporates emotion, salience, persuasiveness, and engagement into social influence dynamics. JANUS builds on the Friedkin–Johnsen model of opinion change, which

formalizes the role of individual anchoring alongside social influence [61]. Subsequent extensions emphasize the importance of norm formation and local reinforcement in stabilizing belief states [60]. JANUS further integrates insights from affective and personality psychology, including established emotion–trait relationships [99] and the role of salience in shaping attention and responsiveness [183]. By allowing agent susceptibility and influence to vary with psychological predispositions and topic relevance, the framework supports simulations that sustain heterogeneous and polarized belief states, consistent with recent findings in opinion dynamics research [18, 22].

### 1.1.1 Research Questions

Guided by these objectives, this dissertation addresses the following research questions:

Big Question: How can I create realistic, dynamic synthetic populations that simulate real world opinion shifts and narrative diffusion across synthetic social platforms?

- **RQ1:** How can *skeleton personas* be constructed to preserve demographic realism while enabling behavioral modeling?
- **RQ2:** Can RAGs and LLMs generate coherent, contextually grounded synthetic personas?
- **RQ3:** Can integrating emotion and salience into opinion dynamic models affect belief evolution and polarization?

Together, these goals and research questions advance the state of the art in synthetic population modeling and opinion dynamics by enabling simulations that more closely reflect the cognitive, emotional, and contextual mechanisms underlying belief change in contemporary information environments.

## Definitions

1. **Synthetic Population.** An artificially constructed set of agents designed to statistically approximate the demographic, social, and behavioral characteristics of a real-world population without representing identifiable individuals.
2. **Skeleton Persona.** A minimally specified synthetic agent defined by core demographic and structural attributes, intentionally excluding psychological and behavioral detail to serve as a transparent foundation for later enrichment.
3. **Population Synthesis.** The process of generating individual-level synthetic agents whose aggregated attributes match known population-level distributions, commonly using methods such as Iterative Proportional Fitting (IPF).
4. **Retrieval-Augmented Generation (RAG).** A generation framework that combines large language models with external information retrieval, conditioning outputs on retrieved documents rather than solely on model parameters.
5. **AURORA Framework.** A Retrieval-Augmented Generation–based system for synthetic population modeling that integrates semantic vector search, salience-aware retrieval, and large language models to generate psychologically and ideologically coherent personas.

6. **Semantic Vector Database.** A data store containing high-dimensional embeddings of textual artifacts, enabling similarity-based retrieval using vector distance metrics such as cosine similarity.
7. **Topic Salience.** A measure of the relative importance or relevance of a topic to an agent or community: influencing attention, emotional activation, and engagement behavior.
8. **Opinion Dynamics.** The study of how individual beliefs evolve through social interaction, influence, and information exposure, including such processes as consensus formation and polarization.
9. **Friedkin–Johnsen Model.** A social influence model in which agents update opinions through weighted social interaction while retaining partial attachment to prior beliefs, allowing persistent disagreement.
10. **JANUS Framework.** An emotion-aware agent-based opinion dynamics framework that extends classical models by incorporating emotional predispositions, topic salience, persuasiveness, and engagement probability.
11. **Persuasiveness.** An agent-level attribute representing the strength with which an agent's expressed opinion influences others during social interaction.
12. **Engagement Probability.** The likelihood that an agent will interact with, respond to, or propagate content related to a given topic, conditioned on salience and emotional state.
13. **Emotion–Trait Mapping.** A structured relationship between stable personality traits and dynamically expressed emotional states, often represented using dimensional models of affect.
14. **Narrative Diffusion.** The process by which narratives, frames, or themes propagate through a population via repeated social interaction and reinforcement.
15. **Social-Cyber Influence.** Coordinated efforts to shape beliefs or behaviors through digitally mediated social systems, including human-driven and automated influence operations.
16. **Interpretability.** The degree to which model mechanisms, agent behavior, and simulation outcomes can be understood and explained by researchers or analysts.
17. **Synthetic Social Media Environment.** A simulated information ecosystem composed of synthetic agents, interactions, and content designed to approximate real-world online platforms.

## 1.2 Data

To evaluate the synthetic population and opinion-dynamics framework developed in this dissertation, I use a combination of empirical demographic microdata, scenario-structured synthetic environments, and event-driven social media datasets. Together, these sources support the three major layers of the dissertation: (i) structurally valid population construction, (ii) psychologically and contextually grounded persona enrichment, and (iii) dynamic belief evolution under narrative and event exposure.

**ACS PUMS Microdata.** The structural foundation of the synthetic population pipeline is built from the American Community Survey (ACS) Public Use Microdata Sample (PUMS), obtained from the U.S. Census Bureau. ACS PUMS provides anonymized individual-level demographic and socioeconomic records, including variables such as age, sex, marital status, race, Hispanic origin, class of worker, educational attainment, and personal income. These data are used to estimate weighted population marginals and to construct proportionally representative skeleton personas through Iterative Proportional Fitting (IPF). ACS PUMS therefore serves as the primary empirical reference population for Chapter 3.1, where the dissertation evaluates how semantic validation affects both record-level plausibility and population-level fidelity.

**Ukraine–Russia Conflict Social Media Data.** For empirical grounding of narrative conflict dynamics, the dissertation draws on a large-scale social media dataset covering the Ukraine–Russia conflict. This collection includes contains Twitter/X data spanning major phases of the conflict and includes multilingual content in English, Ukrainian, and Russian. In related analyses and supporting work, the dataset is paired with bot-labeling, location inference, and geospatial standardization procedures in order toto examine coordinated narratives, moral framing, and geographically distributed information operations. This dataset is used not as a direct population source, but as an empirical benchmark for how narratives, polarization, and cross-platform conflict discourse manifest in real-world information environments.

**Country-Level Reference Data for AURORA.** To initialize cross-national synthetic populations beyond the U.S. microdata setting, this dissertation also uses a country-level reference table of demographic, socioeconomic, digital, and political attributes. These variables include population size, urbanization, age structure, sex ratio, GDP per capita, internet penetration, social media usage, dominant platforms, ethnic and linguistic composition, religion, employment sectors, education levels, institutional trust, and political orientation. Rather than directly assigning beliefs or behaviors, these country-level variables act as macro-level priors and sampling constraints within AURORA, allowing persona generation to remain demographically grounded while preserving downstream behavioral flexibility.

**Facebook Datasets: Roe v. Wade and the 988 Suicide & Crisis Lifeline.** For empirical grounding of online opinion dynamics, this dissertation uses Facebook discussion data centered on two U.S. public issues: reactions to the overturning of *Roe v. Wade* and discourse surrounding the 988 Suicide & Crisis Lifeline. These datasets capture user-level discussion, disagreement, and issue-specific engagement in naturally occurring social media environments. In this dissertation, they are not used as direct inputs for synthetic population construction. Instead, they serve as empirical reference cases for how opinion variation, polarization, and topic-specific responsiveness appear in real-world platform discourse. They are particularly useful for motivating topic-sensitive initialization and for comparing simulated belief dynamics against observed issue-dependent patterns of online discussion.

**Synthetic Scenario Environment: IGOE / Toto Scenario.** A major evaluation environment in this dissertation is the Indo-Pacific Geopolitical Opinion Experiment (IGOE), instantiated through the Toto Scenario configuration inside GhostField. This synthetic environment includes 4,320 AURORA personas distributed across 12 countries, 402 AESOP strategic actors, 8 institutional community types per country, and a structured geopolitical topic space covering Russian aggression, Chinese hegemony, Taiwanese sovereignty, Arctic competition, cyber cooperation, and related issues. The scenario unfolds over a 9-day escalation timeline with kinetic, cyber, environmental, and diplomatic events, allowing the dissertation to test whether the integrated AURORA–JANUS framework produces interpretable cross-country variation in opinion, salience, emotion, and intervention response under controlled conditions.

**Data Processing and Integration** All datasets are processed using a combination of natural language processing pipelines, semantic vector databases, and Retrieval-Augmented Generation (RAG) systems. These tools support both the generative phase of persona construction and the simulation of narrative-driven belief formation and diffusion across platforms and time.

## 1.3 Tools

This section describes the tools used throughout this thesis.

### 1.3.1 Existing Tools

**ORA-PRO** is a dynamic network analysis and visualization tool used extensively in this thesis. It supports the import of Twitter, Telegram, and other social media datasets for constructing metanetworksmeta networks and analyzing communication structures. In this thesis, ORA-PRO is used to identify stance shifts, group structures via Leiden clustering, and key social interaction metrics, offering a high-level view of narrative influence within and across online platforms.

**NetMapper** is a natural language processing tool that extracts and links concepts from unstructured text to generate semantic and conventional meta-networks. It uses custom dictionaries and configurable parameters to identify key terms and entities. For this thesis, NetMapper is used to process social media corpora and convert them into structured network representations that can be further analyzed within ORA-PRO.

**AESOP** is used as the scenario and actor generation environment for constructing synthetic information operations settings. It supports the creation of structured events, key actors, institutional roles, narrative lines, and message templates that define the information environment prior to simulation. In this thesis, AESOP is used to build scenario-layer entities, including strategic actors, intervention accounts, event timelines, and narrative content, which are then passed into downstream simulation systems for execution and analysis.

**SynSM** SynSM is the synthetic social media realization layer used to instantiate platform-level communication behavior. It operationalizes messages, reposts, replies, and other interaction patterns in platform-like environments, enabling belief-conditioned outputs to appear as realistic social media activity. In this thesis, SynSM is used to realize the outputs of AESOP actors and AURORA/JANUS personas on simulated platform structures, allowing narrative diffusion, engagement behavior, and interaction cascades to be studied under controlled conditions.

**OpenAI Models and API** are used for persona generation, narrative scripting, and interaction modeling; this thesis utilizes OpenAI's GPT-4.0 Mini. This model is integrated via API into a Python-based workflow and supports high-quality text generation, character conditioning, and scenario development. DALL·E 2 is also used to create illustrative visuals associated with synthetic personas and communication environments.

**Locally Hosted Large Language Models** enable narrative generation on sensitive or restricted topics where commercial models enforce guardrails. A locally hosted version is used. This open-source local model provides additional flexibility and control during persona construction and adversarial narrative experimentation, while maintaining high performance on constrained hardware.

### 1.3.2 Tools Developed for this thesis

**AURORA (Augmented Understanding through Retrieval-Oriented Reasoning Architecture)** is a Retrieval-Augmented Generation (RAG) framework developed to construct psychologically and contextually grounded synthetic personas. AURORA integrates large language models with semantic vector databases and salience-aware retrieval pipelines to enrich demographically grounded skeleton personas with emotional traits, ideological orientations, topic salience, and narrative predispositions. Unlike prompt-only persona generation approaches, AURORA conditions generation on retrieved, topic-relevant artifacts drawn from curated corpora such as news articles, survey data, and domain-specific texts. This design supports reproducibility, diversity, and interpretability in persona construction, enabling controlled experimentation with population-level belief distributions and narrative exposure.

**JANUS (Joint Agent-based Networked Understanding of Sentiment)** is an agent-based opinion dynamics framework developed to simulate belief formation, polarization, and narrative diffusion within synthetic social media environments. JANUS extends classical social influence models by incorporating agent-level attributes including emotional predisposition, topic salience, persuasiveness, and engagement probability. Building on the Friedkin–Johnsen model, JANUS allows agents to retain attachment to prior beliefs while dynamically updating opinions through social interaction.

# Chapter 2

# Background and Related Work

## 2.1 Background

### 2.1.1 Social Cybersecurity and Networked Influence

The rise of large-scale social media platforms has fundamentally altered how information is produced, disseminated, and consumed, creating environments in which beliefs can be shaped at unprecedented speed and scale. Empirical studies show that false or emotionally charged content often diffuses more rapidly than factual information, driven by platform affordances and human cognitive biases [106, 192]. As a result, the information domain has become a contested space in which influence, persuasion, and narrative control play central roles in political, social, and security outcomes.

Social cybersecurity emerged in response to these challenges as a framework for understanding how cyber-mediated information operations affect human cognition, collective behavior, and societal stability [20]. Unlike traditional cybersecurity, which focuses on protecting technical systems, social cybersecurity treats people and social structures as the primary targets of attack, emphasizing phenomena such as cognitive hacking, narrative manipulation, and trust erosion [33]. Within this paradigm, coordinated information campaigns—often supported by automated or semi-automated actors—are understood as strategic maneuvers designed to distort belief systems, fracture communities, and undermine institutional legitimacy over time.

Research in this area highlights the importance of network structure and social context in shaping influence. Patterns of homophily and selective exposure reinforce belief alignment within communities, increasing resistance to countervailing information and enabling persistent polarization [46]. Empirical analyses of online political discourse further show how selective exposure contributes to the amplification and persistence of opinion clusters [9]. Studies of online moderation and bystander behavior demonstrate that responses to harmful or misleading content are shaped by perceived norms, identity alignment, and platform design rather than content accuracy alone [200]. Together, these findings highlight that influence operations must be understood as socio-technical processes embedded within networks of interaction, rather than isolated messaging events.

This dissertation situates itself within the social cybersecurity literature by advancing computational methods for simulating belief formation and influence in synthetic populations. While much

prior work has focused on detection, measurement, and retrospective analysis of information operations [20, 33], this research emphasizes forward-looking simulation as a means of understanding how emotionally grounded agents, demographic structure, and networked influence interact to produce emergent belief dynamics.

### 2.1.2 Synthetic Population Modeling and Agent-Based Simulation

Synthetic population modeling has long provided a basis for simulating demographic structure and population-level behavior in domains such as transportation, health, and economics. Traditional approaches, including Monte Carlo microsimulation and iterative proportional fitting (IPF), enable population-scale modeling by aligning individual or household attributes with census aggregates, while preserving privacy [130, 208]. Agent-based models (ABMs) extend these methods by simulating dynamic processes such as aging, household formation, and mobility over time [15, 65].

Despite their strengths, these approaches remain limited in their ability to represent psychological nuance, narrative context, and sociocognitive diversity. Most synthetic population models rely on static attributes and rule-based behaviors, constraining their ability to capture how individuals attend to information, emotionally respond, or adapt beliefs in response to social influence. These limitations are particularly pronounced in digitally mediated environments, where behavior is shaped not only by demographics but also by identity, emotion, and exposure to evolving narratives.

Recent advances in large language models (LLMs) have prompted renewed interest in persona driven population modeling. Systems such as Anthology and Persona Hub demonstrate how LLMs conditioned on rich backstories or large-scale web data can generate virtual personas that exhibit greater behavioral diversity [64, 134]. Increasing the number of personas has been shown to improve demographic coverage and representational breadth [167]. However, other studies indicate that LLM-generated personas often struggle to reflect cultural, ideological, or associative distinctions, particularly for non-majority or underrepresented populations [10, 103]. These findings highlight the need for frameworks that combine demographic realism with contextual grounding and behavioral differentiation.

This dissertation builds on these foundations by introducing a two-stage population modeling approach that separates demographic realism from behavioral specification. Demographically grounded *skeleton personas* provide a transparent, statistically aligned population substrate, which is then selectively enriched with psychological and contextual attributes. This design preserves the strengths of traditional synthetic population methods while enabling the flexibility required to model belief change, narrative exposure, and adaptive social behavior in online information environments

### 2.1.3 LLM-Augmented Opinion Modeling and Synthetic Social Behavior

Integrating LLMs into agent-based simulation has emerged as a promising approach to addressing these gaps. Surveys of LLM-augmented ABM show that generative agents exhibit higher levels of social reasoning, autonomy, and conceptual coherence than traditional rule-based agents [63, 76]. This shift enables the simulation of longer-term, more adaptive social behavior. Park et

al.'s Generative Agents architecture further demonstrates how LLM-powered agents equipped with memory and reflection can produce emergent social phenomena such as gossip, coordination, and relationship formation in simulated environments [149].

In the domain of opinion modeling, several studies show that LLMs can generate statistically aligned synthetic opinions when conditioned on demographic or group-level information. Sun et al. introduce "random silicon sampling," enabling simulations that approximate public opinion polls using minimal input data [180]. Other work highlights how prompt structure and conditioning strategies strongly influence alignment fidelity [117], while noting persistent challenges in representing nuanced or marginalized viewpoints [70, 137]. These results underscore the importance of incorporating salience, cultural context, and identity-aware conditioning into persona design.

Beyond individual belief generation, LLMs have been used to simulate audiences and evaluate narrative effects at scale. Balog and Zhai show how generative agents can support repeatable user simulation for system evaluation [11], while Liu et al.'s Explore–Generate–Simulate framework demonstrates that LLM-simulated audiences can approximate human feedback across communication scenarios [115]. Studies of trust and cooperation further suggest that LLM agents can exhibit human-like behavior in strategic interactions [203], though ethical critiques caution against uncritical reliance on synthetic agents as stand-ins for real populations [97].

Recent work on influence dynamics emphasizes that realistic simulation requires modeling not only individual agents but also mechanisms for narrative disruption and behavioral tipping points. Carragher et al. demonstrate how targeted stance perturbations can trigger cascades of belief change in co-evolving networks [34]. At the same time, critical perspectives warn against treating synthetic populations as interchangeable "silicon samples," emphasizing the need for grounding, transparency, and ethical oversight [87].

This dissertation advances this line of work by integrating LLM-based persona generation into a mechanistic opinion dynamics framework rather than treating generative agents as standalone simulators. Through the AURORA persona modeling system and the JANUS opinion dynamics model, LLMs are used to ground agent cognition, emotion, and salience, while belief evolution remains governed by interpretable, network-based influence processes. This integration enables scalable simulations that balance behavioral realism with analytical control, supporting both scientific inquiry and applied social-cybersecurity analysis.

# Chapter 3

# Synthetic Persona (Skeleton) Generation

## 3.1 Structural Foundations for Synthetic Persona Modeling

This chapter formalizes the first stage of the synthetic population pipeline: the construction of *skeleton personas*. Skeleton personas provide the demographic and structural substrate upon which cognitively, emotionally, and narratively enriched agents are later instantiated. At this stage, the objective is not to model belief formation, affective response, or strategic behavior, but to ensure that the simulated population is statistically valid, demographically representative, and scalable to national and cross-national contexts, consistent with established synthetic population methodologies [14, 155, 198].

Skeleton personas encode demographic, socioeconomic, and digital-access attributes derived from empirical population data while remaining explicitly agnostic to opinions, emotions, or narrative exposure. This separation of population structure from behavioral dynamics reflects a core methodological principle in synthetic population research: structural validity must precede cognitive and narrative modeling [198, 206]. Prior work demonstrates that embedding behavioral assumptions too early risks conflating demographic realism with unvalidated psychological mechanisms, thereby reducing interpretability and experimental control [163, 206]. By decoupling these stages, the framework preserves modularity, enables controlled experimentation, and supports transparent attribution of downstream effects.

## 3.2 Introduction

Agent-based models (ABMs) of social systems require synthetic populations that are simultaneously (i) demographically representative of observed population structures, (ii) individually coherent at the record level, and (iii) computationally tractable at national and international scales. Classical synthetic population methodologies—including Iterative Proportional Fitting (IPF), probabilistic resampling, and related marginal-matching procedures—have been widely deployed in transportation planning [14], epidemiological modeling [198], and policy simulation [205] because they provide strong guarantees of population-level statistical validity under empirical constraints.

In this dissertation, "representative" is used in a strict *proportional* sense: synthesized skeleton

personas are calibrated so that demographic category frequencies match the target population proportions implied by ACS PUMS microdata (via `PWGTP` weights). Thus, the synthetic population is not merely a collection of plausible individuals; it is a *proportionally representative* reconstruction of population composition over the modeled attribute space.

However, matching marginal distributions alone does not guarantee *semantic coherence* at the individual record level. Even when selected low-order cross-tabulations are incorporated, synthetic population methods such as IPF generally enforce only a subset of the full dependency structure. In high-dimensional demographic spaces, many conditional relationships remain unconstrained unless they are explicitly modeled, and the number of possible joint distributions grows combinatorially with the number of attributes. As a result, IPF-style synthesis can still produce records that are statistically admissible under the enforced marginals or low-order joints yet be substantively implausible at the person level. Examples include: a twelve-year-old with a doctoral degree, a fully retired worker enrolled full-time in school, or other violations of life-course sequencing. Although such artifacts may be rare (e.g., $< 1\%$ of records), they become consequential when downstream cognitive, exposure, or behavioral rules depend on attributes such as age, education, employment, or income.

One possible response is to increase the number of constrained joint distributions, for example by adding relations such as Age $\times$ Income or Age $\times$ Education. This can reduce some implausible combinations, but it does not solve the problem in general. First, enforcing many cross-tabs quickly becomes computationally burdensome and data-hungry in sparse, high-dimensional settings. Second, even relatively rich joint constraints do not fully encode broader semantic regularities or life-course logic unless the relevant conditional structure is explicitly specified in advance. The challenge is therefore, not simply one of adding more marginals or cross-tabs, but of determining how to impose record-level plausibility without undermining population-level fidelity.

Recent work has proposed Large Language Model (LLM)-based semantic validation to address this gap by using pretrained models as plausibility judges for demographic configurations [74, 155]. In this chapter, I evaluate three implementations of that idea. *Direct LLM validation* applies an LLM to each record individually as a semantic classifier. *Rule-based validation* uses an LLM once to generate an explicit set of demographic plausibility rules, which are then applied deterministically. *Neural validation* distills LLM judgments into a trained classifier that approximates those decisions at lower inference cost. I compare these approaches using measures of distributional fidelity and efficiency, including L1 distance and Jensen–Shannon (JS) divergence. L1 distance captures absolute distributional error, while JS divergence is a symmetric, bounded information-theoretic measure of how far two probability distributions differ.

Despite growing interest in LLM-assisted validation, the field still lacks a controlled comparative framework for evaluating the quality–cost–interpretability trade space across validation architectures. In particular, prior work has not systematically examined (i) how post hoc filtering affects both marginal distributions and key joint relations, (ii) whether the tension between record level coherence and population-level fidelity is an inherent property of filtering-based validation or can be reduced through design choices, and (iii) whether rule-based outcomes remain stable across providers and deployment settings.

This chapter addresses five research questions:

1. **(RQ1) Quality–Cost–Speed Trade-off:** How do direct LLM, rule-based, and neural validation approaches compare in terms of marginal fidelity (L1 distance and Jensen–Shannon divergence), focus-relation preservation (Age $\times$ Education), runtime, and monetary cost?
2. **(RQ2) Coherence–Fidelity Trade-off:** To what extent does improving record-level demographic coherence through post hoc filtering introduce distributional drift, and is that drift reducible through validation design choices?
3. **(RQ3) Interpretability and Governance:** Which validation architecture offers the strongest transparency, auditability, and institutional usability for settings where validation decisions must be explainable?
4. **(RQ4) Provider Robustness:** Are rule-based validation outcomes statistically stable across different LLM providers (OpenAI, Anthropic, Google), enabling vendor-independent deployment and cost optimization?
5. **(RQ5) Optimal Configuration:** What rule cardinality ($K$) provides the best balance between structural quality, computational tractability, and interpretability?

## 3.3 Background

### 3.3.1 Demographically Grounded Skeleton Personas

Modeling narrative diffusion and belief dynamics in socio-technical systems requires agents that accurately reflect the underlying structure of real populations rather than stylized or convenience samples [14, 198]. This dissertation defines such agents as *skeleton personas*: digitally instantiated individuals that capture demographic, socioeconomic, and contextual characteristics without encoding topic-specific beliefs, affective states, or decision-making rules. Skeleton personas therefore function as structural agents, forming the population-level scaffolding upon which grounded personas—introduced in later chapters—are constructed using Retrieval-Augmented Generation (RAG) and psychologically informed opinion dynamics.

Skeleton persona generation in this work draws on established methods from synthetic population modeling, including iterative proportional fitting (IPF), probabilistic resampling, and diffusion inspired synthesis, all of which have been widely applied in transportation planning, epidemiology, and policy simulation to preserve marginal and joint attribute distributions [14, 198, 205]. These methods ensure alignment with observed population statistics while maintaining individual-level heterogeneity, a prerequisite for valid agent-based simulation [155].

**Skeleton personas as structural agents** A **skeleton persona** is the foundational representation of an individual agent within a synthetic population. It specifies structural and contextual attributes, such as age, sex, education level, employment sector, geographic location, and digital access, while deliberately excluding beliefs, emotions, ideological positions, or adaptive behaviors. This design choice reflects best practices in synthetic population modeling, where structural realism is treated as analytically distinct from behavioral dynamics [198, 206]. Skeleton personas therefore act as

neutral population units, supporting later behavioral enrichment without introducing premature narrative alignment.

Skeleton personas are a specific instantiation of *synthetic personas*, which are digitally constructed representations of human individuals designed for large-scale computational social simulation [155]. Unlike traditional personas—which are often static, manually curated, and illustrative—synthetic personas in this framework are data-driven, programmatically generated, and intended to operate at population scale within agent-based models [74]. The defining distinction of skeleton personas is their intentional behavioral minimalism, which preserves experimental control during downstream opinion and influence modeling [163].

In this dissertation, skeleton personas are characterized by three core properties:

- **Empirical Structural Grounding.** Persona attributes are grounded in observed demographic and socioeconomic distributions derived from census and survey data, ensuring population-level realism without prescribing individual behavior [14, 198].
- **Multi-source Attribute Integration.** Attributes are synthesized from census and survey data, digital access and platform usage statistics, and country-level contextual indicators. Integrating these heterogeneous sources reflects established practice in modern population synthesis [155, 205]. The complete set of variables, distributions, and sources used in this process is documented in Appendix A.
- **Behavioral Neutrality.** Skeleton personas do not encode opinions, emotional states, or narrative predispositions, enabling controlled and interpretable introduction of cognitive and influence mechanisms at later stages of the modeling pipeline [163, 206].

By isolating population structure from behavior, skeleton personas bridge classical synthetic population modeling and modern agent-based social simulation, providing a stable and reusable foundation for downstream modeling of social influence and narrative dynamics [155].

## 3.4 Method and Model: Skeleton Persona Construction Pipeline

I formalize skeleton persona construction as a modular, two-stage pipeline that explicitly separates (i) *population-level statistical synthesis* from (ii) *individual-level semantic coherence enforcement*. This separation preserves the interpretability and convergence properties of classical synthetic population methodologies while enabling targeted mitigation of record-level semantic artifacts that can arise under purely marginal constraints (Figure 4.1).

### 3.4.1 Design objectives and operating assumptions

Skeleton personas encode *structural demography without behavioral content*. The pipeline therefore follows three design objectives:

1. **Empirical structural grounding.** Synthetic populations reproduce demographic marginals derived from reference microdata, ensuring that the simulated population reflects the correct proportions of demographic strata.

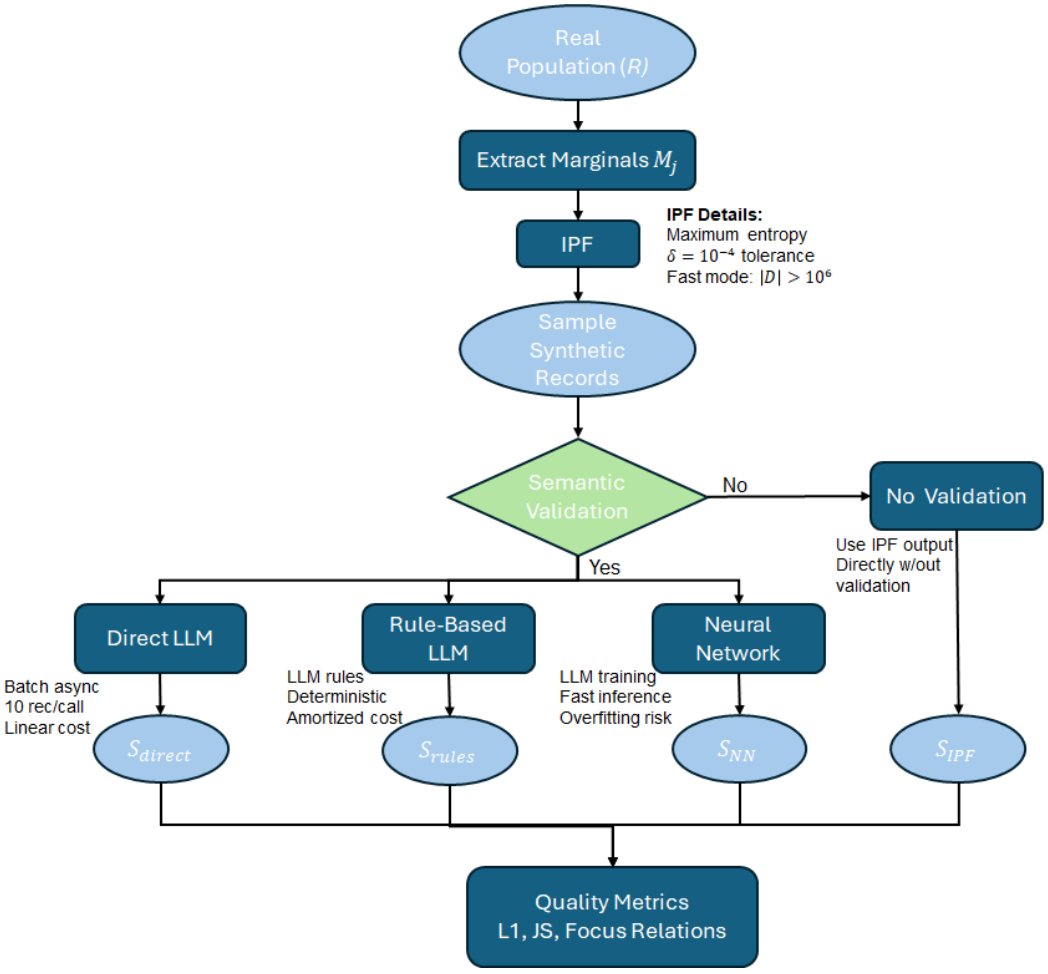



Figure 3.1: **Skeleton Persona Generation Pipeline.** Marginal distributions are extracted from ACS PUMS data and matched using Iterative Proportional Fitting (IPF) to generate statistically grounded skeleton personas. Optional semantic validation is applied using one of three strategies: direct LLM validation, rule-based filtering, or neural network approximation. Outputs are evaluated using marginal L1 distance, Jensen–Shannon divergence, and age–education joint fidelity. The resulting skeleton populations are behaviorally neutral and suitable for downstream cognitive or narrative enrichment.

2. **Multi-attribute integration.** Individuals are represented as joint configurations over multiple demographic attributes. However, synthesis methods such as IPF primarily enforce marginal constraints and do not guarantee full joint plausibility. As a result, statistically admissible but semantically implausible combinations (e.g., a minor with a doctoral degree) can occur. The semantic validation stage introduced in this chapter identifies and filters such records.
3. **Behavioral neutrality.** Skeleton personas exclude opinions, beliefs, and affective states. While demographic attributes may influence behavior in real populations, these relationships are introduced only in later modeling stages. At this stage, demographics serve solely as structural inputs for downstream cognitive and influence models.

The central methodological issue is that marginal matching guarantees aggregate fidelity but does not, in general, guarantee record-level plausibility. I therefore introduce semantic validation as an optional *filtering layer* intended to remove a narrow class of internally inconsistent demographic configurations while minimizing distortion of aggregate distributions.

### 3.4.2 Data and structural attribute space

The reference population is derived from the American Community Survey (ACS) Public Use Microdata Sample (PUMS), a nationally representative microdata dataset released by the U.S. Census Bureau. ACS PUMS provides anonymized individual-level demographic records that allow researchers to construct synthetic populations while preserving statistical properties of the underlying population. For more information see Appendix A.1.

Let $\mathbf{X}_{\text{real}} \in \mathbb{R}^{N_{\text{real}} \times d}$ denote the reference population drawn from ACS PUMS microdata, and let $\mathbf{X}_{\text{syn}} \in \mathbb{R}^{N_{\text{syn}} \times d}$ denote the synthetic skeleton population with target size $N_{\text{syn}}$. I define skeleton personas over $d = 8$ demographic variables:

- `SEX`: sex category used in PUMS,
- `MAR`: marital status,
- `RAC1P`: race,
- `HISP`: Hispanic origin,
- `COW`: class of worker,
- `SCHL`: educational attainment (21 levels),
- `AGEP`: age,
- `PINCP`: personal income.

Each synthetic record $\mathbf{r}_i \in \mathbf{X}_{\text{syn}}$ is an 8-tuple $\mathbf{r}_i = (v_{i1}, \ldots, v_{id})$ taking values in the discrete domain of each attribute. Sampling weights (`PWGTP`) are used to compute population-representative marginals following established synthetic population practice [14, 198]. These weighted marginals define the *target population proportions* that IPF enforces, so each synthesized run is a calibrated proportional reconstruction of the real population over the modeled attribute set (up to sampling variance). In this sense, skeleton personas instantiate *proportional representation* at scale: the simulated population preserves not only individual-level heterogeneity, but also the correct population

shares of demographic strata. Behavioral variables are intentionally excluded (Section 3.3.1). Appendix A.2 provides coding details and attribute domains.

### 3.4.3 Stage I: Iterative Proportional Fitting as marginally constrained maximum-entropy synthesis

Let $\{P_{\text{real}}(v_j)\}_{j=1}^{d}$ denote empirical marginal distributions over each attribute $v_j$. IPF constructs a joint distribution $P_{\text{syn}}(\mathbf{v})$ that satisfies these marginal constraints by iteratively re-weighting a seed distribution $P^{(0)}(\mathbf{v})$. Equivalently, IPF yields the maximum-entropy joint distribution consistent with the observed marginals, and may be interpreted as minimizing divergence from the seed subject to marginal constraints [16, 206].

At iteration $t$, and for each attribute $v_j$, the joint distribution is updated as:

$$P^{(t+1)}(\mathbf{v}) = P^{(t)}(\mathbf{v}) \cdot \frac{P_{\text{real}}(v_j)}{\sum_{\mathbf{v}_{-j}} P^{(t)}(\mathbf{v})}, \tag{3.1}$$

where $\mathbf{v}_{-j}$ denotes all variables except $v_j$. The update cycle repeats across $j = 1, \ldots, d$ until convergence:

$$\max_j \left\| P_{\text{syn}}^{(t)}(v_j) - P_{\text{real}}(v_j) \right\|_1 < \delta, \tag{3.2}$$

with tolerance $\delta = 10^{-4}$ or a maximum of 50 iterations.

**Sampling the IPF-only baseline** After convergence, I sample $N_{\text{syn}}$ individuals from $P_{\text{syn}}^{(T)}(\mathbf{v})$ to form the baseline synthetic population:

$$\mathbf{X}_{\text{IPF}} \sim P_{\text{syn}}^{(T)}(\mathbf{v}). \tag{3.3}$$

By construction, $\mathbf{X}_{\text{IPF}}$ matches the target marginals in expectation and typically to within negligible empirical error. Accordingly, $\mathbf{X}_{\text{IPF}}$ can be interpreted as a *proportionally representative* synthetic population: for each modeled attribute, the expected category counts correspond to the population shares implied by the weighted ACS PUMS marginals.

**Fast mode for high-dimensional domains** When the joint domain cardinality exceeds $10^6$ combinations, explicit materialization of $P(\mathbf{v})$ becomes memory-intensive. In this regime, IPF is executed in an approximate, sample-reweighting form, iteratively adjusting weights over a candidate sample of records rather than storing the full joint tensor. This reduces memory complexity from $O(\prod_j |v_j|)$ to $O(N_{\text{syn}})$, while maintaining practically equivalent marginal convergence [207].

### 3.4.4 Stage II: Semantic validation as a record-level filtering operator

IPF guarantees marginal fidelity but does not impose record-level semantic constraints. Consequently, the IPF-only baseline may contain *marginally admissible yet semantically implausible* records (e.g., minors with doctoral degrees) [155]. I therefore introduce a validation function

$$\mathcal{V} : \mathbf{r} \to \{0, 1\}, \tag{3.4}$$

and define the validated population as:

$$\mathbf{X}_{\text{syn}} = \{\mathbf{r}_i \in \mathbf{X}_{\text{IPF}} : \mathcal{V}(\mathbf{r}_i) = 1\}. \tag{3.5}$$

Semantic validation is conceptualized as a support-restriction operator: it constrains the realized joint configurations by excluding records that violate demographic plausibility regularities. Because filtering is applied post hoc, it can induce marginal drift relative to $P_{\text{real}}$; accordingly, the evaluation framework quantifies both (i) marginal divergence introduced by filtering and (ii) improvements in targeted joint plausibility diagnostics (notably age–education coherence).

### 3.4.5 Validation architectures evaluated

I evaluate three implementations of $\mathcal{V}$ that span a trade space over semantic expressiveness, interpretability, computational scalability, and robustness. Each method instantiates the same abstract operator in Eq. 3.5, differing only in the mechanism by which plausibility is assessed and operationalized.

#### 1) Direct LLM validation: knowledge-based plausibility classification

Direct validation treats a pretrained LLM as a plausibility classifier. Each record $\mathbf{r}_i$ is serialized into a structured natural-language description of the demographic tuple, and the LLM returns a binary validity judgment.

Formally,

$$\mathcal{V}_{\text{direct}}(\mathbf{r}_i) = \text{LLM}(\text{serialize}(\mathbf{r}_i); \theta_{\text{LLM}}, T), \tag{3.6}$$

where $\theta_{\text{LLM}}$ denotes fixed pretrained parameters and $T$ is the sampling temperature. I use $T = 1.0$ to reflect that plausibility boundaries can be context-dependent and non-deterministic for rare but feasible configurations (e.g., accelerated education trajectories).

**Operational characteristics** Direct LLM validation requires no rule engineering and no labeled training data, and it can encode multi-attribute dependencies that are difficult to express as simple thresholds. Its primary limitation is computational: inference cost scales linearly with $N_{\text{syn}}$ due to per-record model evaluation. To mitigate latency, records are processed in asynchronous batches; nevertheless, this strategy is typically unsuitable for repeated large-scale generation and is treated as a semantic reference and as a supervision source for distillation.

#### 2) Rule-based validation: interpretable semantic compression

Rule-based validation compresses plausibility constraints into a finite set of deterministic rules. A single LLM call produces a candidate rule set $\mathcal{R} = \{R_1, \ldots, R_K\}$. Each rule is expressed as a conditional constraint:

$$R_k : \text{IF } \phi_k(\mathbf{r}) \text{ THEN } \psi_k(\mathbf{r}), \tag{3.7}$$

where $\phi_k$ is an antecedent predicate and $\psi_k$ is a consequent requirement.

Validation is then deterministic:

$$\mathcal{V}_{\text{rule}}(\mathbf{r}) = \prod_{k=1}^{K} \mathbb{I}\big[\neg\phi_k(\mathbf{r}) \vee \psi_k(\mathbf{r})\big], \tag{3.8}$$

accepting only records that violate none of the rules.

**Operational characteristics** This approach decouples semantic knowledge elicitation (one-time LLM invocation) from execution (local rule evaluation), yielding negligible marginal cost per record and full auditability: each rejection is attributable to an explicit constraint. The principal limitation is representational: probabilistic or context-sensitive plausibility judgments are discretized into deterministic thresholds, which can misclassify rare but valid outliers.

### 3) Neural network validation: distillation of LLM judgments

Neural validation distills direct LLM judgments into a supervised classifier, approximating $\mathcal{V}_{\text{direct}}$ while eliminating per-record API calls. I label a training set of $n = 5{,}000$ IPF-generated records using Direct LLM validation, producing pairs $(\mathbf{r}_i, y_i)$ with $y_i \in \{0, 1\}$. A feedforward neural network is trained to minimize binary cross-entropy:

$$\mathcal{L} = -\sum_i \left[y_i \log \hat{y}_i + (1 - y_i) \log(1 - \hat{y}_i)\right]. \tag{3.9}$$

I use a two-hidden-layer architecture:

$$\begin{aligned} \mathbf{h}_1 &= \text{ReLU}(\mathbf{W}_1\mathbf{r} + \mathbf{b}_1), \quad \mathbf{h}_1 \in \mathbb{R}^{24}, \\ \mathbf{h}_2 &= \text{ReLU}(\mathbf{W}_2\mathbf{h}_1 + \mathbf{b}_2), \quad \mathbf{h}_2 \in \mathbb{R}^{12}, \\ \hat{\mathbf{y}} &= \text{softmax}(\mathbf{W}_3\mathbf{h}_2 + \mathbf{b}_3), \quad \hat{\mathbf{y}} \in [0, 1]^2, \end{aligned} \tag{3.10}$$

with Dropout$(0.2)$ applied after each hidden layer. The validation decision is:

$$\mathcal{V}_{\text{NN}}(\mathbf{r}) = \arg\max \hat{\mathbf{y}}. \tag{3.11}$$

**Operational characteristics** This strategy provides high-throughput, offline validation at negligible marginal cost, making it well-suited for large populations and repeated generation. Its principal risk is approximation error, including sensitivity to distributional shift between the LLM-labeled training set and deployment populations (e.g., changes in region, marginal targets, or coding granularity). I therefore evaluate whether neural validation preserves the targeted joint plausibility properties (e.g., age–education coherence) without inducing unacceptable degradation in marginal fidelity.

Table 3.1: Experimental variables, roles, and operationalization.

| **Factor / Metric** | **Role** | **Operationalization** |
|---|---|---|
| Validation strategy | Independent | *IPF-only* (no filtering); *Direct LLM* ($\mathcal{V}_{\text{direct}}$); *Rule-based* ($\mathcal{V}_{\text{rule}}$); *Neural* ($\mathcal{V}_{\text{NN}}$) |
| LLM provider | Independent | Varied in the robustness case study: GPT-4.1-mini, Claude 3.5, Gemini 2.0 Flash |
| Marginal error (L1, JS) | Dependent | Distance between synthetic and reference marginal distributions across $d$=8 attributes |
| Focus-relation error (L1, JS) | Dependent | Distance between synthetic and reference Age×Education joint distribution |
| Execution time | Dependent | Wall-clock runtime per run (s), including validation overhead |
| Cost per 1,000 records | Dependent | Estimated dollar cost normalized per 1,000 generated records (API + training where applicable) |
| Source data | Control | ACS PUMS microdata; fixed extract and preprocessing pipeline across conditions |
| IPF settings | Control | Fixed convergence tolerance and maximum iterations across runs |
| Population size | Control | Fixed $N_{\text{syn}}$ within each case study; varied only when explicitly scaling |
| Attribute set | Control | Fixed demographic feature set ($d$=8) across all strategies and providers |

## 3.5 Experimental Design and Evaluation

This section details the experimental design used to evaluate semantic validation strategies for skeleton persona generation, followed by the quantitative metrics used to assess structural fidelity, focus-relation preservation, and computational efficiency. The experimental setup isolates the effect of validation architecture while holding constant the underlying population synthesis process and evaluation protocol.

### 3.5.1 Experimental Design

The experiments evaluate how different implementations of the semantic validation function $\mathcal{V}$ affect the quality and scalability of synthetic skeleton populations generated via IPF. Each experimental condition follows a common pipeline:

$$\text{IPF synthesis} \;\rightarrow\; \text{semantic validation} \;\rightarrow\; \text{metric evaluation}.$$

Table 3.1 summarizes the experimental variables and their roles.

Across strategies, the objective is the same: filter semantically implausible records while preserving the aggregate structure produced by IPF. I treat this as a quality–stability–efficiency trade space, balancing record-level coherence, population-level drift (marginals and Age×Education),

and operational cost as $N_{\text{syn}}$ scales. The three validation architectures occupy distinct points in this trade space:

- **Direct LLM ($\mathcal{V}_{\text{direct}}$):** strongest semantic coverage with minimal engineering, but per-record inference cost scales roughly linearly with $N_{\text{syn}}$, limiting repeated or large-scale synthesis.
- **Rule-Based ($\mathcal{V}_{\text{rule}}$):** deterministic constraints with high interpretability and near-constant per-record runtime but can miss nuanced or probabilistic plausibility boundaries.
- **Neural ($\mathcal{V}_{\text{NN}}$):** distills LLM judgments into a classifier with low marginal inference cost but depends on label fidelity/calibration and can degrade under distributional shift (including provider-dependent labeling).

The case studies therefore measure: (1) improvements in joint plausibility (Age$\times$Education), (2) marginal drift induced by filtering, and (3) the quality–efficiency frontier across $N_{\text{syn}}$ and (when varied) provider choice.

**Independent variables** The primary independent variable is the *semantic validation strategy* applied to the IPF-generated synthetic population: *IPF-Only*, *Direct LLM*, *Rule-Based*, and *Neural Network*. In addition, I vary the *LLM provider* in one case study to assess cross-provider robustness of validation outcomes.

**Dependent variables** The dependent variables quantify both structural fidelity and operational cost. Population-level fidelity is measured using marginal distributional error (L1 distance and Jensen–Shannon divergence) and focus-relation error for the Age$\times$Education joint distribution. Operational performance is assessed using wall-clock execution time and estimated monetary cost per 1,000 generated records.

**Controlled variables** To ensure comparability across experimental conditions, all other factors are held constant (Table 3.1). These include the source data (ACS PUMS microdata), IPF convergence tolerance and iteration caps, the demographic attribute set ($d$=8), and the evaluation protocol and statistical testing procedures. Holding these variables constant ensures observed differences are attributable to the validation strategy and, where applicable, the LLM provider.

### 3.5.2 Case Study Design

Three case studies evaluate validation strategies under increasing scale and heterogeneity.

**Case Study 1: Direct LLM vs. Rule-Based** This case evaluates whether rule-based validation can approximate the quality of direct LLM validation while substantially reducing cost and latency. The comparison includes IPF-only, Direct LLM, and Rule-Based strategies using a single LLM provider.

Table 3.2: Experimental conditions across case studies.

| Parameter | Case 1 | Case 2 | Case 3 |
|---|---|---|---|
| Synthetic population size | 1,500 | 2,500 | 2,500 |
| Number of runs | 100 | 100 | 50 |
| Validation strategies | IPF, Direct, Rule | IPF, Rule, NN | IPF, Rule |
| LLM provider(s) | GPT-4.1-mini | GPT-4.1-mini | GPT-4.1-mini, Claude, Gemini |
| Random seeds | 42–51 | 100–109 | Provider-specific |

**Case Study 2: Neural Network Approximation** This case tests whether a neural network trained on 5,000 LLM-labeled records can reproduce validation quality at near-zero inference cost. Performance is compared against rule-based validation under identical synthesis conditions.

**Case Study 3: Multi-Provider Robustness** This case assesses the stability of rule-based validation across different LLM providers by repeating the experiment with three model families. This evaluates sensitivity to provider-specific semantic biases.

### 3.5.3 Evaluation Metrics

Synthetic populations are evaluated along three complementary dimensions: (i) fidelity to real population marginal distributions, (ii) preservation of a critical joint (focus) relation, and (iii) operational efficiency. Together, these metrics quantify the principal trade-off in skeleton persona construction: improving individual-level semantic coherence without inducing unacceptable distortion at the population level or incurring prohibitive computational cost.

#### Marginal Distributional Fidelity

Marginal distributional fidelity measures how closely the synthetic population reproduces the univariate demographic structure of the real population across all $d$ attributes. Because IPF explicitly enforces marginal constraints, deviations from real marginals arise only through sampling variance and the removal of records during semantic validation. Quantifying marginal drift therefore provides a direct measure of how strongly a validation strategy perturbs the statistical grounding established by IPF.

I compute the average L1 distance (total variation) between synthetic and real marginal distributions as:

$$L1_{\text{marginal}} = \frac{1}{d} \sum_{j=1}^{d} \sum_{k} \left| P_{\text{syn}}(v_j = k) - P_{\text{real}}(v_j = k) \right|, \tag{3.12}$$

where $v_j$ denotes the $j$-th demographic variable and $k$ indexes its categorical levels. An L1 value of zero indicates perfect marginal agreement, while larger values reflect increasing distortion of the population structure.

To complement L1 distance, I also compute Jensen–Shannon (JS) divergence:

$$JS_{\text{marginal}} = \frac{1}{d} \sum_{j=1}^{d} JS\left(P_{\text{syn}}(v_j) \| P_{\text{real}}(v_j)\right), \tag{3.13}$$

a symmetric, bounded information-theoretic measure derived from Kullback–Leibler divergence. JS divergence is less sensitive to small absolute differences in high-frequency categories and remains finite even when support differs, making it well suited for comparing sparse or skewed demographic distributions.

### Focus Relation Preservation

Marginal fidelity alone is insufficient for many downstream modeling tasks, which often depend on specific inter-attribute relationships rather than univariate distributions. I therefore evaluate preservation of the joint Age $\times$ Education distribution (AGEP $\times$ SCHL), a structurally critical relation capturing educational timing, workforce entry, and cohort effects.

Joint fidelity is measured using L1 distance over the full contingency table:

$$L1_{\text{focus}} = \sum_{a,e} \left| P_{\text{syn}}(\text{AGEP} = a, \text{SCHL} = e) - P_{\text{real}}(\text{AGEP} = a, \text{SCHL} = e) \right|, \tag{3.14}$$

where $a$ and $e$ index age and education levels, respectively. Jensen–Shannon divergence is computed analogously to capture information-theoretic differences in the joint distribution.

This focus relation metric directly assesses whether semantic validation disproportionately removes plausible but statistically important subpopulations (e.g., young workers, late-degree earners), thereby distorting cohort structure despite preserved marginals.

### Operational Efficiency

In addition to structural quality, validation strategies are evaluated on operational efficiency, which is critical for large-scale population synthesis and iterative experimentation. I measure wall clock execution time per run, defined as the elapsed time from completion of IPF synthesis to production of the final validated population. I also estimate monetary cost per 1,000 records, accounting for LLM API usage and amortized training costs for neural network validation.

These metrics enable explicit comparison of quality–cost trade-offs, distinguishing strategies that achieve marginal improvements in structural fidelity at disproportionate computational expense from those that offer scalable, production-ready performance.

### Statistical Testing and Robustness

All experiments are repeated across multiple random seeds to account for stochastic variation in IPF sampling, LLM responses, and neural network training. For each metric, I report the mean, standard deviation, and coefficient of variation across runs. Pairwise comparisons between validation strategies are conducted using Welch's $t$-tests, which do not assume equal variances. Statistical significance is assessed at $\alpha = 0.01$, ensuring robustness of reported differences.

Table 3.3: Fixed parameters across all experiments.

| Parameter | Value |
|---|---|
| IPF convergence tolerance ($\delta$) | $10^{-4}$ |
| IPF maximum iterations | 50 |
| LLM temperature ($T$) | 1.0 |
| Direct LLM batch size | 100 |
| Concurrent batches | 5 |
| Rule count ($K$) | 8 |
| NN training samples | 5,000 |
| NN dropout rate | 0.2 |

### 3.5.4 Fixed Parameters

To ensure that observed differences across experiments are attributable solely to validation strategy, all synthesis and validation parameters are held constant across runs. Table 3.3 summarizes the fixed configuration used in all experiments.

## 3.6 Results

Synthetic populations are evaluated along three dimensions: (i) *marginal fidelity* (marginal L1 and Jensen–Shannon divergence), (ii) *focus relation preservation* (Age×Education joint L1/JS), and (iii) *operational efficiency* (runtime and estimated cost). These metrics expose the central trade-off in skeleton persona validation: increasing record-level plausibility without excessive macro-level drift or prohibitive computation cost. All reported fidelity metrics should be interpreted relative to this goal of *proportional representation*: validators are penalized insofar as filtering perturbs the calibrated population shares established by IPF.

### 3.6.1 Case Study 1: Direct LLM vs. Rule-Based Validation

Case Study 1 isolates the trade-off between *semantic selectivity* and *distributional stability*. Direct LLM validation applies a per-record plausibility gate that can reject records unevenly across the joint space, while Rule-Based validation applies a fixed, auditable constraint set. I evaluate both against the IPF-only baseline over 100 runs ($N_{\text{syn}} = 1{,}500$) using GPT-4.1-mini.

#### Marginal distributional accuracy: Direct LLM induces larger drift

Table 3.4 shows that both validators introduce marginal drift relative to IPF-only, but Direct LLM produces substantially larger divergence than Rule-Based. This is consistent with *non-uniform rejection*: the LLM gate concentrates deletions in semantically borderline regions of the joint space, and those deletions propagate into marginals through correlation structure (e.g., age–education–

income coupling). Rule-Based constraints still prune implausibilities, but do so with a more predictable and bounded footprint, yielding lower average drift.

Table 3.4: **Case Study 1: Marginal distributional accuracy (n=100 runs; 1,500 records).** Rule-Based validation produces significantly lower marginal drift than Direct LLM.

| Strategy | L1 Mean | L1 Std | JS Mean | JS Std |
|---|---|---|---|---|
| IPF-Only | 0.0798 | 0.0000 | 0.00147 | 0.0000 |
| Rule-Based | 0.1084 | 0.0081 | 0.00420 | 0.0013 |
| Direct LLM | 0.1812 | 0.0057 | 0.01768 | 0.0004 |
| Rule-Based vs Direct LLM: 40% lower L1 error ($p < 0.0001$) | | | | |

Operationally, this pattern shows the cost of enforcing plausibility by deletion. Direct LLM rejects on the order of 16% of records per run, and those rejections are systematically concentrated rather than uniformly random. Rule-Based validation reduces implausibilities while avoiding the same degree of concentrated depletion, which keeps marginal distortion smaller and more stable across runs.

### Focus relation (Age×Education): Direct LLM improves the target joint most

Marginals do not reveal whether validation corrects the intended semantic pathology. I therefore evaluate the Age×Education focus relation directly. In this case study, Direct LLM yields the lowest error on the targeted joint, followed by Rule-Based, with IPF-only worst (Table 3.5). However, the Direct-vs-Rule difference is *not statistically significant* across 100 runs ($p = 0.2863$), indicating comparable improvements on the age–education dependency in this setting.

It is important to note that the L1 values for the focus relation are larger than those reported for marginal distributions in Table 3.4. This difference is expected because the focus metric evaluates a two-dimensional joint distribution rather than a single-variable marginal. The Age×Education relation contains many more cells, and the L1 metric aggregates absolute error across all of them. As a result, joint-distribution L1 values naturally appear larger in magnitude, and the relevant comparison is the relative improvement across validation strategies rather than the absolute scale.

Table 3.5: **Case Study 1: Focus relation accuracy (Age×Education; n=100 runs).** Lower is better. Direct LLM yields the lowest Age×Education error; Direct-vs-Rule is not statistically significant across runs ($p = 0.2863$).

| Strategy | L1 Focus | JS Focus | Combined (L1+JS) |
|---|---|---|---|
| IPF-Only | 0.7333 | 0.1458 | 0.8792 |
| Rule-Based | 0.6526 | 0.1261 | 0.7787 |
| Direct LLM | 0.5435 | 0.0998 | 0.6433 |
| Direct LLM vs Rule-Based: $p = 0.2863$ (not significant) | | | |

Figure 3.2 shows the run-wise distributions for the same joint metrics. Direct LLM shifts the distribution downward on L1/JS and the combined score, consistent with directly suppressing implausible age–education configurations. Rule-Based also improves relative to IPF-only, and exhibits more bounded spread, reflecting deterministic enforcement of a stable rule set.

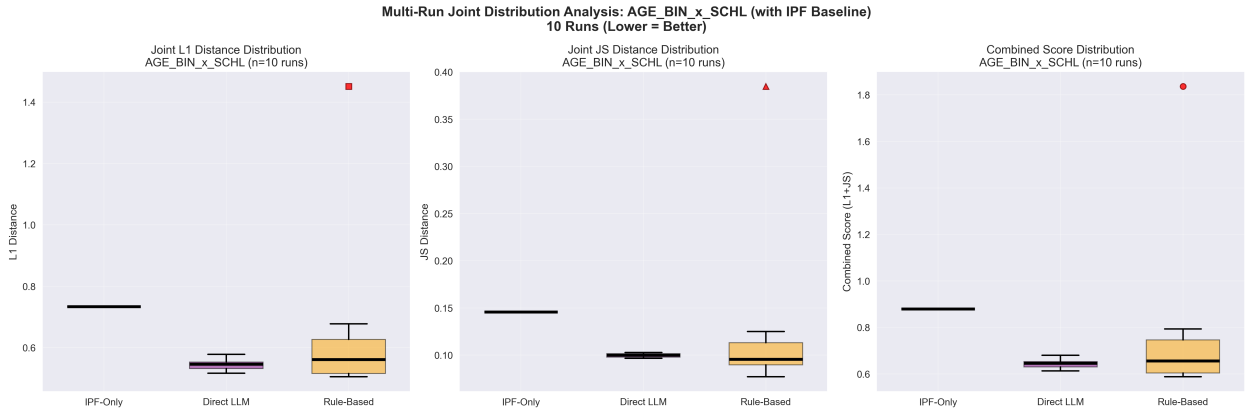



Figure 3.2: **Case Study 1: Multi-run Age-binned $\times$ Education joint analysis (n=100 runs).** Panels show joint L1 distance (left), joint JS divergence (center), and combined score (right). Lower is better.

### Efficiency: Rule-Based is the deployable regime

Direct LLM's focus-relation gains come with steep operational overhead. Table 3.6 shows that Rule-Based validation retains much of the semantic benefit at a small fraction of the runtime and cost. Because Direct LLM requires per-record inference, latency and API cost scale with $N_{\text{syn}}$; Rule-Based shifts LLM dependence to one-time rule extraction and performs local checks thereafter, enabling repeated synthesis runs for sensitivity analysis and statistical confidence.

Table 3.6: **Case Study 1: Computational efficiency and cost (1,500-record populations).**

| Strategy | Time/Run | Speedup | Cost/1K | Cost Ratio | Budget Impact |
|---|---|---|---|---|---|
| IPF-Only | 2.5s | 206$\times$ | \$0.00 | 1$\times$ | – |
| Rule-Based | 13.7s | 38$\times$ | \$0.001 | 1$\times$ | \$0.05 / 500K |
| Direct LLM | 482s | 1$\times$ | \$0.10 | 100$\times$ | \$50 / 500K |

**Takeaway.** Case Study 1 separates *targeted semantic correction* from *global distributional stability*. Direct LLM yields the strongest improvement on the Age$\times$Education focus relation in this setting, but it also produces substantially higher marginal drift and is operationally expensive. Rule-Based validation offers the most practical balance: it improves the focus relation relative to IPF-only while keeping marginal distortion, runtime, and cost bounded, making it suitable for iterative, large-scale skeleton persona generation.

Table 3.7: **Case Study 2: Marginal distributional accuracy (n=100 runs; $N_{\text{syn}}$=2,500).** Distilled neural validation increases population-level drift relative to Rule-Based.

| Strategy | L1 Mean | L1 Std | JS Mean | JS Std |
|---|---|---|---|---|
| IPF-Only | 0.0764 | 0.0016 | 0.00136 | 0.00000 |
| Rule-Based | 0.1016 | 0.0039 | 0.00330 | 0.00065 |
| Neural Network | 0.1500 | 0.0047 | 0.00627 | 0.00022 |
| Neural vs Rule (L1): $\Delta$=0.0484, $t$=33.14, $p < 10^{-4}$ | | | | |
| Neural vs Rule (JS): $\Delta$=0.0029, $t$=15.08, $p < 10^{-4}$ | | | | |

### 3.6.2 Case Study 2: Neural Network Approximation

Case Study 2 evaluates whether a distilled neural validator, $\mathcal{V}_{\text{NN}}$, can approximate the structural behavior of Rule-Based validation while reducing inference latency. I compare *IPF-Only*, *Rule-Based* ($\mathcal{V}_{\text{rule}}$; $K$=8), and *Neural Network* ($\mathcal{V}_{\text{NN}}$) over 100 independent runs with $N_{\text{syn}}$=2,500 per run (GPT-4.1-mini supervision). The core criterion is not classification accuracy, but whether the validator preserves (i) marginal stability and (ii) the targeted Age×Education dependency under repeated synthesis.

#### Marginal fidelity: distillation amplifies drift

Table 3.7 shows that $\mathcal{V}_{\text{NN}}$ induces substantially larger marginal drift than $\mathcal{V}_{\text{rule}}$ under identical synthesis conditions. Mean marginal L1 increases from $0.1016$ (Rule-Based) to $0.1500$ (Neural), and mean marginal JS increases from $0.00330$ to $0.00627$, with large and statistically significant differences ($p < 10^{-4}$). The distributional perspective is shown in Figure 3.3: neural validation yields a consistently higher error band with limited overlap with Rule-Based, indicating this is not a seed-specific artifact.

Mechanistically, the neural gate does not preserve IPF's population-level constraints because its errors are *correlated* with the joint structure: boundary mistakes concentrate in coupled regions (age–education–income), so small decision-boundary shifts translate into systematic marginal mass movement. This is visible in run-wise behavior (Figure 3.4): the neural trajectory remains consistently elevated across all 100 runs for both marginal L1 and JS, rather than occasionally spiking.

#### Focus relation (Age×Education): neural distillation worsens the target joint

The distillation objective is only meaningful if the learned gate preserves the targeted dependency (AGEP×SCHL) without destabilizing cohort structure. The focus results indicate the opposite. Table 3.8 shows Rule-Based achieves the lowest error on the Age×Education joint, while the neural validator is worse than both Rule-Based and IPF-only on the combined score. Figure 3.4 reinforces that this is consistent across seeds: the neural focus L1 remains well above the other two strategies in every run, with no evidence of convergence toward the rule baseline.

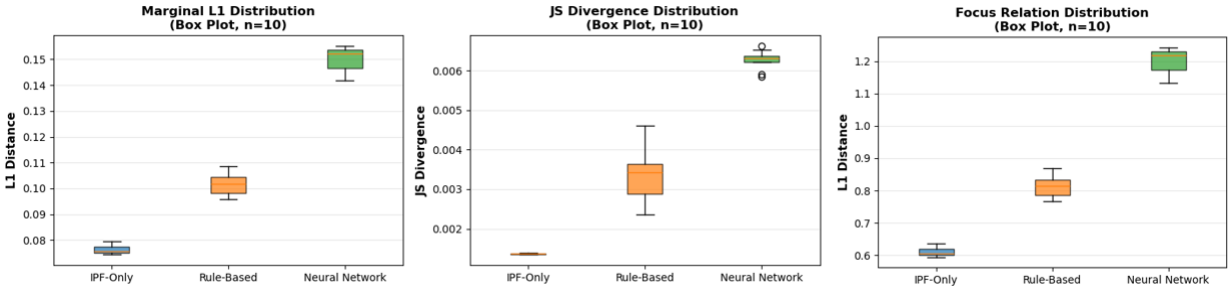



Figure 3.3: **Case Study 2: Distribution and run-level stability (n=100).** (Left) Marginal L1 box plot; (center) marginal JS box plot; (right) focus-relation L1 box plot. Neural validation exhibits consistently higher drift than Rule-Based across runs with limited overlap, indicating systematic (not stochastic) degradation.

Table 3.8: **Case Study 2: Focus relation performance for Age×Education (n=100 runs).** Lower is better. Rule-Based outperforms Neural on L1, JS, and combined score.

| **Strategy** | **L1 Focus** | **JS Focus** | **Combined (L1+JS)** |
|---|---|---|---|
| IPF-Only | 0.7225 | 0.1403 | 0.8628 |
| Rule-Based | 0.5157 | 0.0860 | 0.6018 |
| Neural Network | 0.7734 | 0.1562 | 0.9296 |
| Neural vs Rule: $p < 10^{-4}$ (significant); winner = Rule-Based | | | |

**Why distillation fails here (fast $\not\Rightarrow$ safe)**

The empirical signature is *bidirectional distortion*: $\mathcal{V}_{\text{NN}}$ both rejects plausible-but-structurally-necessary records (driving drift) and admits implausible records (failing to fix the focus pathology). This is consistent with four interacting contributors in this configuration:

**Efficiency: Neural is faster, but dominated on quality**

Although neural inference is low-latency in isolation, it is not a Pareto improvement. It is dominated by Rule-Based on marginal fidelity and on the targeted Age×Education dependency (Tables 3.7–3.8, Figures 3.3–3.4). In operational terms, the neural validator becomes a high-throughput mechanism for introducing structural error.

**Takeaway.** Case Study 2 shows that naive distillation of LLM plausibility judgments into a small feedforward classifier is not structurally safe in this setting. It increases marginal drift and worsens the targeted Age×Education joint. Rule-Based validation remains the deployable regime because it preserves determinism and auditability while achieving substantially better structural outcomes under repeated synthesis.

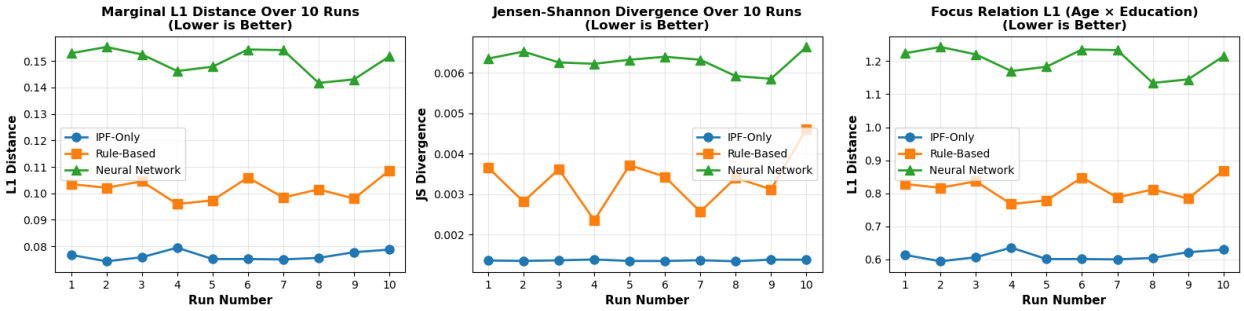



Figure 3.4: **Case Study 2: Run-wise trajectories over 100 runs (lower is better).** (Left) Marginal L1, (center) marginal JS, (right) focus relation L1 for Age×Education. Neural validation remains consistently elevated across runs on all metrics, indicating stable but systematically incorrect filtering behavior relative to Rule-Based.

Table 3.9: **Root causes of distillation failure in Case Study 2.**

| **Failure mode** | **Consequence for population structure** |
|---|---|
| Insufficient coverage of rare joint cells | Boundary cases are underrepresented, so errors concentrate where cohort structure is fragile. |
| Ordinal structure loss (SCHL treated as categorical) | Education adjacency is destroyed, weakening monotone age–education constraints. |
| Label noise near plausibility boundaries | Inconsistent supervision yields unstable thresholds on borderline-but-valid configurations. |
| Deployment shift / calibration drift | Accept/reject rates shift across runs and synthesis regimes, amplifying drift. |

### 3.6.3 Case Study 3: Multi-Provider Robustness and Rule-Count Tuning

Case Study 3 addresses two deployment questions for Rule-Based validation: (i) whether rule elicitation is sensitive to the *LLM provider*, and (ii) how the number of rules $K$ trades off *population-level stability* against *Age×Education correction*. For each provider (GPT-4.1-mini, Claude 3.5-Haiku, Gemini 2.0-Flash), I evaluate $K \in \{0, 2, \ldots, 20\}$ over 50 independent runs ($N_{\text{syn}} = 2{,}500$ per run). Rule application is deterministic once rules are fixed; therefore, any observed differences arise from the rule set and the induced filtering footprint rather than stochastic inference at execution time.

#### Rule count $K$ defines the stability–plausibility frontier

Rule count acts as a direct control knob on how aggressively post-IPF filtering restricts the realized joint support. Figure 3.5 (GPT-4.1-mini) shows the marginal view: combined marginal error (L1+JS) increases as $K$ grows, indicating that stronger filtering increasingly perturbs the IPF-imposed demographic structure. In contrast, Figure 3.6 shows the focus-relation view: Age×Education joint quality improves markedly under moderate $K$ (removal of high-confidence implausibilities),

then flattens and becomes more variable as additional rules begin to prune rare-but-plausible cohort configurations.

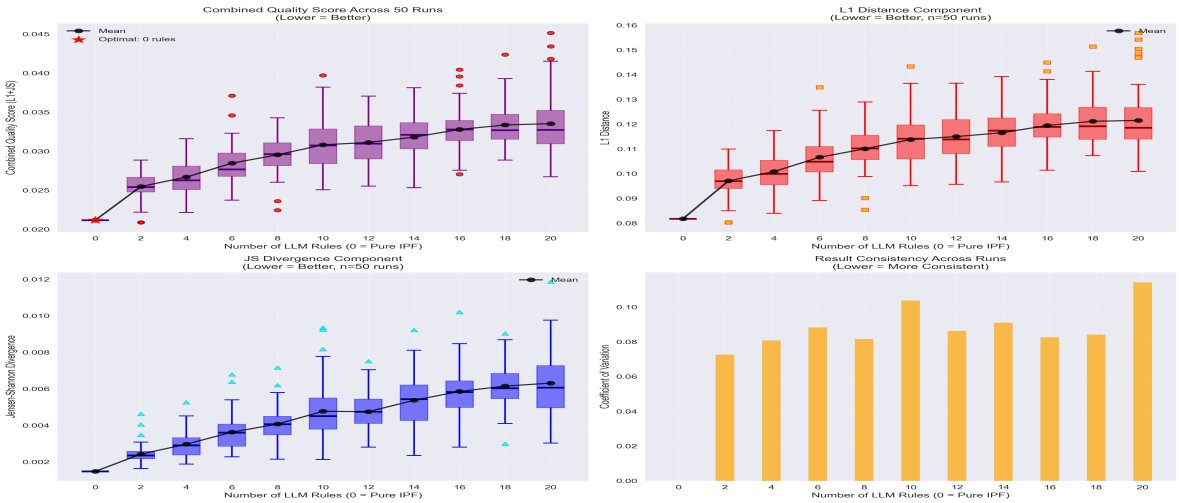



Figure 3.5: **Case Study 3 (GPT-4.1-mini): Marginal quality vs. rule count (n=50).** Combined score (L1+JS) and its components increase with $K$, reflecting increasing marginal drift under more restrictive filtering.

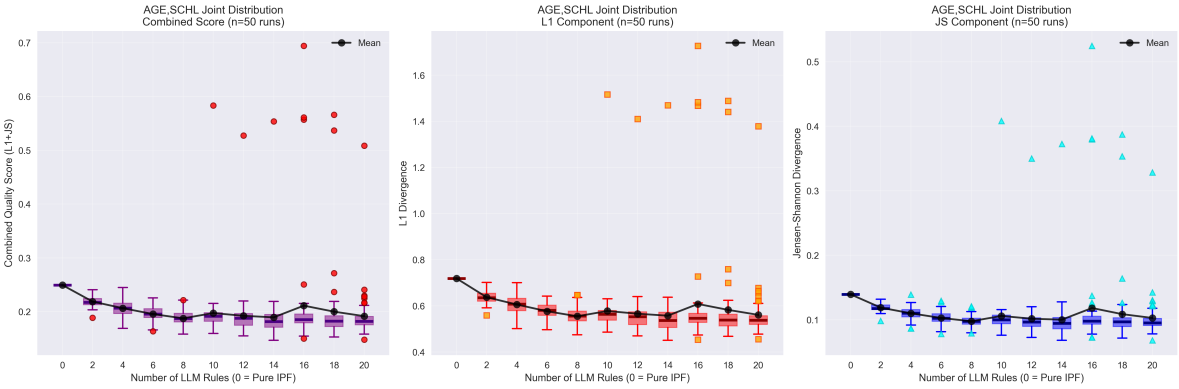



Figure 3.6: **Case Study 3 (GPT-4.1-mini): Age×Education focus relation vs. rule count (n=50).** Moderate rule sets improve the cohort joint (lower is better), while very large $K$ increases variance and can reintroduce distortion via over-filtering.

Table 3.10 summarizes the marginal trajectory numerically for the representative GPT run set, highlighting the practical implication: increasing $K$ is not "free"—it systematically increases population-level drift even when it can improve specific semantic pathologies.

### Provider robustness: equivalent qualitative trajectories

Figures 3.7–3.10 show that Claude and Gemini exhibit the same qualitative behavior as GPT: (i) marginal drift grows with $K$, and (ii) the Age×Education joint improves under moderate con-

Table 3.10: **Rule count optimization: detailed analysis (GPT-4.1-mini; n=50).**

| K | Combined Score | Std | CV (%) | L1 Component | JS Component | Interpretation |
|---|---|---|---|---|---|---|
| 0 | 0.0795 | 0.0022 | 2.8% | 0.0764 | 0.00136 | Pure IPF (baseline) |
| 2 | 0.0910 | 0.0068 | 7.5% | 0.0882 | 0.00285 | Minimal filtering |
| 4 | 0.1002 | 0.0084 | 8.4% | 0.0967 | 0.00349 | Light filtering |
| 6 | 0.1050 | 0.0091 | 8.7% | 0.1015 | 0.00347 | Moderate filtering |
| 8 | 0.1088 | 0.0083 | 7.6% | 0.1050 | 0.00381 | Strong filtering |
| 10 | 0.1122 | 0.0091 | 8.1% | 0.1081 | 0.00407 | Strong filtering |
| 12 | 0.1192 | 0.0124 | 10.4% | 0.1144 | 0.00478 | Over-filtering begins |
| 14 | 0.1291 | 0.0147 | 11.4% | 0.1237 | 0.00540 | Over-filtering |
| 20 | 0.1468 | 0.0178 | 12.1% | 0.1398 | 0.00701 | Severe over-filtering |

Table 3.11: **Case Study 3: Quality metrics by LLM provider (50 runs per provider; 2,500 records;** $K = 8$**).**

| Provider | L1 Marginal | JS Marginal | L1 Focus | CV (%) | $n$ |
|---|---|---|---|---|---|
| GPT-4.1-mini (OpenAI) | $0.1100 \pm 0.0082$ | $0.00407 \pm 0.00105$ | $0.5547 \pm 0.0398$ | 7.5% | 50 |
| Claude 3.5-Haiku (Anthropic) | $0.1132 \pm 0.0093$ | $0.00444 \pm 0.00103$ | $0.5419 \pm 0.0431$ | 8.2% | 50 |
| Gemini 2.0-Flash (Google) | $0.1100 \pm 0.0097$ | $0.00437 \pm 0.00133$ | $0.6039 \pm 0.0397$ | 8.8% | 50 |
| One-way ANOVA: $F = 2.06$, $p = 0.131$ (no significant difference) | | | | | |

straints with increasing variance under aggressive rule accumulation. This indicates that once semantic constraints are expressed as deterministic rules, the dominant source of behavior is the *filtering architecture and rule count* rather than provider-specific idiosyncrasies.

**Provider comparison at a fixed operating point ($K = 8$)**

To test robustness at a practical operating point, Table 3.11 reports metrics at $K = 8$. While there are minor numerical differences between providers, these differences are statistically insignificant under one-way ANOVA ($F = 2.06, p = 0.131$), supporting provider-agnostic deployment once $K$ and the rule template are fixed.

**Rule Cardinality and the Emergence of the $K = 8$ Operating Point.** The observed "sweet spot" at $K = 8$ is not incidental, but arises from the interaction between rule expressiveness and the structure of the underlying joint distribution. For small $K$, the rule set under-approximates dominant life-course constraints (e.g., AGEP×SCHL), yielding limited correction of implausible regions. As $K$ increases, constraint coverage improves, and high-confidence implausible configurations are selectively removed, reducing focus-relation error. However, beyond a moderate threshold, additional rules induce over-specification: constraints become increasingly granular and correlated, disproportionately removing low-density but valid configurations. Because filtering operates as a support-restriction operator on $P_{\text{IPF}}(\mathbf{v})$, these concentrated deletions propagate through

attribute dependencies and amplify marginal drift.

Thus, the optimal regime shows a balance between *constraint coverage* and *support preservation*. The stability observed at $K = 8$ indicates that this cardinality is sufficient to encode the dominant demographic regularities without fragmenting the joint support. More generally, performance is governed not by $K$ alone, but by the effective capacity of the rule set—a function of both rule count and rule complexity relative to the sparsity and dimensionality of the attribute space.

**Takeaway.** Case Study 3 shows that $K$ is the primary design lever. Increasing rule count predictably increases marginal drift, while focus-relation correction benefits from moderate constraint sets but can degrade as rules become overly restrictive. Provider choice has minimal impact once validation is reduced to deterministic rules. This enables vendor-flexible deployments (choose providers by cost/availability), while tuning K to the desired point on the stability–plausibility frontier.

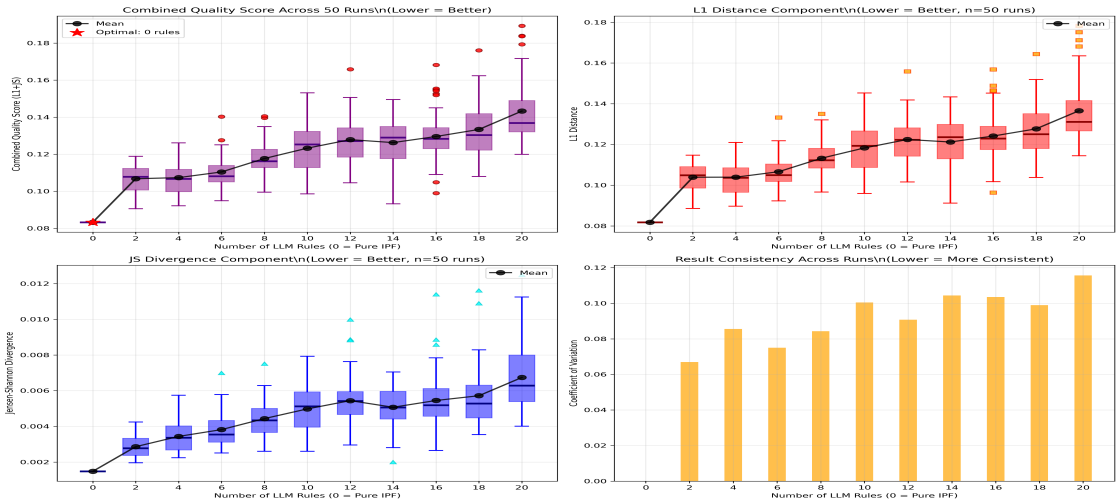



Figure 3.7: **Case Study 3 (Claude 3.5-Haiku): Marginal quality vs. rule count (n=50).**

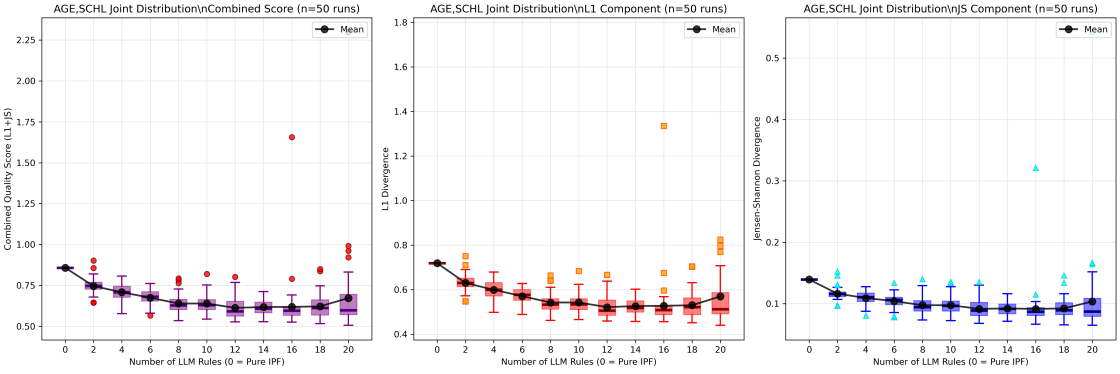



Figure 3.8: **Case Study 3 (Claude 3.5-Haiku): Age×Education focus relation vs. rule count (n=50).**

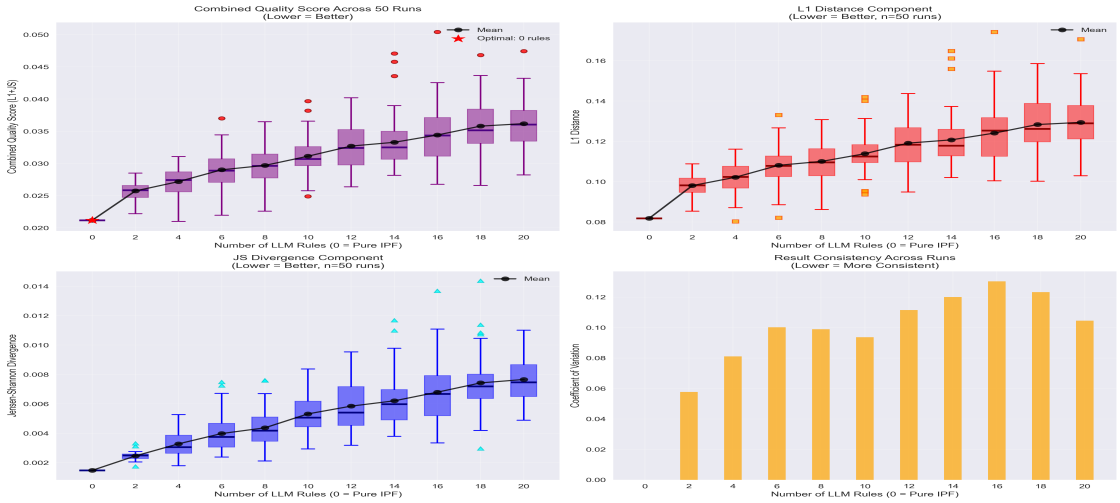



Figure 3.9: **Case Study 3 (Gemini 2.0-Flash): Marginal quality vs. rule count (n=50).**

Figure 3.10: **Case Study 3 (Gemini 2.0-Flash): Age×Education focus relation vs. rule count (n=50).**

### 3.6.4 Post-Filtering Distribution Preservation

A central question for rule-based validation is whether the filtered synthetic population still preserves the demographic structure of the real population. Because validation is applied after IPF convergence, removing records will necessarily perturb the marginals that IPF was designed to match. This subsection therefore examines (i) empirical rejection rates under rule-based filtering, (ii) post-filtering distributional fidelity, and (iii) whether rejection disproportionately affects particular demographic groups.

#### Corrected Rejection Rates

Table 3.12 reports empirical rejection rates for GPT-4.1-mini rule-based validation across 50 independent runs per rule count. The observed rejection rates are substantially higher than initially estimated. At $K=8$, the recommended operating point, the filter rejects about 20% of IPF-generated records.

These results imply that production deployments must overgenerate to reach a desired final population size. They also show a trade-off: larger rule sets produce more consistent rejection behavior, but they remove a larger share of records.

#### Distributional Fidelity After Filtering

Despite rejecting a substantial number of records, the filtered populations retain good marginal fidelity. Table 3.13 compares post-filtering distributional quality at $K=8$ with the IPF-only baseline

Table 3.12: **Empirical rejection rates by rule count (GPT-4.1-mini; $n$=50 runs per $K$; $N_{\text{syn}}$=2,500).** Rejection rates increase monotonically with $K$, reflecting stronger semantic enforcement. The coefficient of variation (CV) decreases with $K$, indicating more predictable filtering behavior under larger rule sets.

| K | Rejection (%) | Retention (%) | Range (%) | CV (%) | $n$ |
|---|---|---|---|---|---|
| 2 | 9.9 ± 3.1 | 90.1 ± 3.1 | 4.4–20.0 | 30.9 | 50 |
| 6 | 18.0 ± 4.7 | 82.0 ± 4.7 | 9.2–26.8 | 26.2 | 50 |
| 8 | 19.7 ± 4.7 | 80.3 ± 4.7 | 9.9–31.2 | 23.8 | 50 |
| 10 | 21.1 ± 5.8 | 78.9 ± 5.8 | 8.6–37.1 | 27.2 | 50 |
| 12 | 21.9 ± 4.4 | 78.1 ± 4.4 | 14.1–29.9 | 20.2 | 50 |
| 16 | 24.2 ± 4.8 | 75.8 ± 4.8 | 13.9–36.4 | 19.9 | 50 |
| 20 | 25.7 ± 4.6 | 74.3 ± 4.6 | 16.3–37.3 | 18.0 | 50 |

Table 3.13: **Distribution quality after rule-based filtering ($K$=8; representative run; $N_{\text{syn}}$=2,500).** Marginal L1 distances remain below 0.10, indicating acceptable preservation of univariate demographic structure. The Age×Education focus relation shows improvement relative to unfiltered IPF.

| Metric | IPF-Only | Rule-Based ($K$=8) | Interpretation |
|---|---|---|---|
| Marginal L1 (avg) | 0.0764 | 0.0881 | Minimal drift |
| Marginal JS (avg) | 0.00136 | 0.00556 | Bounded divergence |
| Focus L1 (Age×Edu) | 0.7225 | 0.5676 | **21% improvement** |
| Focus JS (Age×Edu) | 0.1403 | 0.1100 | **22% improvement** |

using a representative run.

The key result is that filtering improves the Age×Education joint distribution while causing only modest marginal drift. Marginal L1 rises from 0.0764 to 0.0881, a small absolute increase relative to the 21–22% improvement in the focus relation. This suggests that rule-based validation can improve record-level plausibility without materially undermining population-level representativeness.

### Demographic Rejection Bias

To test whether filtering disproportionately removes particular demographic groups, I examined rejection patterns across age, education, and related variables for $K$=8 over 10 runs. Table 3.14 summarizes the observed variability.

Education-related rejection varies the most, which is expected because the rules primarily target age–education coherence. Importantly, the 100% rejection cases correspond to clearly implausible combinations, such as young children with advanced degrees or very young children recorded as employed. This indicates that the filter is removing semantic violations rather than arbitrarily penalizing demographic groups.

Table 3.14: **Demographic rejection bias summary ($K$=8; 10 runs; 34,109 total records).** Range indicates the difference between the highest and lowest rejection rates within each variable's categories. Education (SCHL) shows the highest variability, confirming it as the primary driver of rejection.

| Variable | Min (%) | Max (%) | Mean (%) | Std (%) | Range (%) |
|---|---|---|---|---|---|
| AGE_BIN | 5.4 | 83.5 | 22.4 | 25.8 | 78.1 |
| SCHL | 0.0 | 100.0 | 50.0 | 50.5 | 100.0 |
| SEX | 0.0 | 100.0 | 50.0 | 57.7 | 100.0 |
| RAC1P | 0.0 | 100.0 | 52.9 | 51.4 | 100.0 |
| MAR | 0.0 | 100.0 | 50.0 | 52.7 | 100.0 |
| COW | 0.0 | 100.0 | 44.4 | 51.1 | 100.0 |

High-rejection combinations follow consistent life-course plausibility patterns:

- **Age×Education:** 0–9 years old with college credentials (100% rejection; 622 records), 50–59 years old coded as "less than 3 years old" (100% rejection; 162 records).
- **Education×Employment:** Kindergarten-level education with formal employment (100% rejection; 21 records), High-school students working full-time in professional roles (100% rejection; 390 records).
- **Age×Marital Status:** 0–9 years old coded as divorced (100% rejection; 392 records).

Protected characteristics such as sex and race do not show evidence of systematic univariate bias. Rejection is instead driven by implausible multivariate combinations. This is important because it suggests that rule-based validation improves semantic coherence without introducing systematic distortion against protected groups.

### 3.6.5 Validation

Validation of the skeleton persona pipeline is conducted using a *validation-in-parts* approach, assessing the model across complementary dimensions rather than asserting global validity. Specifically, validation focuses on three components: (i) population-level marginal fidelity, (ii) preservation of critical joint demographic structure (Age×Education), and (iii) the behavioral properties of the validation operator under repeated sampling and scaling. This approach is consistent with best practices in agent-based modeling and synthetic population research, where complex systems are evaluated through component-wise validation aligned with theoretical expectations [31, 169].

**Population-level validity (marginal fidelity).** At the aggregate level, the synthetic population is expected to reproduce empirical demographic proportions due to the use of Iterative Proportional Fitting (IPF), which constructs a maximum-entropy joint distribution subject to marginal constraints [14, 198]. This expectation is confirmed in the IPF-only baseline, where marginal L1 and Jensen–Shannon (JS) divergence remain minimal and stable across runs, reflecting accurate proportional reconstruction of the target population.

The introduction of semantic validation produces measurable marginal drift; however, this behavior is theoretically expected. Because validation operates as a *post hoc filtering operator* that removes records from the realized joint distribution, deletions propagate through correlated attributes (e.g., age, education, and income), perturbing marginals even when constraints are locally applied. This phenomenon aligns with established findings in synthetic population research that enforcing additional constraints or removing infeasible records introduces trade-offs between marginal fidelity and joint realism [59, 206]. The observed increase in marginal error under Direct LLM and Neural validation is therefore consistent with non-uniform rejection across the joint space, while the comparatively bounded drift under Rule-Based validation reflects its more stable and deterministic constraint structure.

**Joint-structure validity (Age×Education coherence).** Marginal agreement alone is insufficient for validating demographic realism, as many downstream processes depend on inter-attribute relationships. The Age×Education joint distribution is used as a *focus relation* because it encodes life-course constraints such as educational progression, workforce entry, and cohort structure. Improvements in this joint distribution therefore serve as a direct indicator of increased record-level plausibility.

Empirically, both Direct LLM and Rule-Based validation improve Age×Education fidelity relative to IPF-only, with Direct LLM achieving the lowest error but at the cost of significantly higher marginal drift. Rule-Based validation achieves comparable improvements without statistically significant degradation relative to Direct LLM in this focus relation, indicating that a small set of interpretable constraints is sufficient to capture dominant life-course regularities. This result is consistent with demographic modeling literature, where low-order joint constraints capture a substantial portion of structural dependence in population synthesis [14, 205].

The failure of neural validation to preserve this joint structure further supports this interpretation. Distillation introduces approximation error that disproportionately affects low-density joint regions, leading to both the rejection of valid cohort configurations and the retention of implausible ones. This behavior is consistent with known limitations of supervised classifiers under joint sparsity, label noise, and distributional shift [74, 209].

**Behavioral validity of the validation operator.** Beyond static distributional properties, the validation operator itself exhibits behavior consistent with theoretical expectations. Increasing rule cardinality ($K$) produces a predictable trade-off: stronger constraint coverage improves the focus relation but increases marginal drift. This reflects the role of validation as a *support-restriction operator* on the IPF-derived joint distribution, where additional constraints progressively reduce the feasible region of the demographic space.

The emergence of an optimal operating point at $K = 8$ reflects a balance between *constraint coverage* and *support preservation*. At low $K$, the rule set under-specifies demographic regularities, leaving implausible configurations uncorrected. At high $K$, over-specification removes rare but valid configurations, fragmenting the joint support and increasing variance. This behavior aligns with findings in constraint-based modeling and statistical regularization, where moderate constraint complexity yields the best generalization performance [78].

Additionally, the observed stability of rule-based validation across LLM providers demonstrates that once semantic constraints are externalized into deterministic rules, system behavior is governed by the filtering architecture rather than model-specific idiosyncrasies. This supports the robustness and deployability of the approach in multi-provider environments.

The empirical results exhibit strong *face validity*: the model behaves in ways that are consistent with real-world demographic processes and established properties of synthetic population methods. Specifically:

- Marginal distributions are preserved under IPF and degrade predictably under filtering, consistent with constraint-based population synthesis.
- Life-course structure (Age$\times$Education) improves under targeted validation, reflecting realistic demographic dependencies.
- Increasing constraint strength produces a stability–plausibility trade-off, consistent with support restriction in high-dimensional distributions.
- Approximation-based validation (neural distillation) fails under joint sparsity and boundary ambiguity, aligning with known machine learning limitations.

These behaviors collectively indicate that the pipeline is not only statistically grounded, but also structurally and behaviorally consistent with theoretical expectations. While not a complete validation of all downstream dynamics, this component-wise validation demonstrates that the skeleton persona layer provides a reliable and interpretable foundation for subsequent cognitive and social modeling stages.

## 3.7 Discussion

This work evaluates a central tension in modern population synthesis for agent-based modeling: demographic validity is typically enforced at the *population* level (via marginal constraints), while downstream ABMs often require *individual* records to be semantically coherent. Our results show that bridging this gap through post hoc semantic validation is feasible, but only certain validation architectures preserve the statistical guarantees that motivate IPF in the first place. Across all three case studies, the empirical patterns support four core claims: (i) filtering-based validation induces an unavoidable drift mechanism; (ii) Direct LLM validation is best interpreted as a semantic oracle rather than a deployable validator; (iii) naive neural distillation is not structurally safe under realistic joint spaces; and (iv) rule-based semantic compression is the most stable and governable operating point for large-scale deployment.

The key technical observation is that semantic validation is not merely a quality enhancement layer; it is a *support-restriction operator* applied after a maximum-entropy marginal fit. Once IPF has converged, the synthetic population implicitly represents a joint distribution $P_{\text{IPF}}(\mathbf{v})$ that satisfies marginal constraints but does not encode higher-order semantic regularities. Any validator $\mathcal{V}$ defines a new distribution,

$$P_{\mathcal{V}}(\mathbf{v}) \propto P_{\text{IPF}}(\mathbf{v}) \cdot \mathbb{I}[\mathcal{V}(\mathbf{v}) = 1], \tag{3.15}$$

which generally *cannot* preserve the original marginals unless $\mathcal{V}$ is independent of each attribute

and uncorrelated with the joint structure (an unrealistic condition). Thus, marginal drift is not an implementation artifact, Rather, it is a mathematical consequence of imposing record-level constraints by filtering. Crucially, this drift must be interpreted relative to the chapter's definition of *representation*: IPF produces a *proportional representation* of the target population over the modeled attribute space, meaning expected category counts reproduce population shares implied by weighted microdata. Post hoc filtering therefore does not merely "clean" records. It *reallocates population mass* across demographic strata by selectively removing individuals from specific regions of the joint support. From an ABM perspective, this is a change in *who exists* in the simulated society, not only a change in record-level plausibility. The practical question is therefore not whether drift occurs, but whether a validator induces *controlled* drift while improving the targeted semantic pathologies.

This framing clarifies why the metrics separate cleanly into (i) marginal stability measures (L1/JS on marginals), which quantify how strongly $P_{\mathcal{V}}$ deviates from $P_{\text{real}}$, and (ii) focus relation measures (Age$\times$Education L1/JS), which quantify whether $\mathcal{V}$ meaningfully corrects the dominant life-course inconsistency. The empirical frontier observed across strategies can be understood as different approximations to the same constrained distributional projection.

### 3.7.1 Case Study 1: Direct LLM is not a scalable validator

Case Study 1 shows that Direct LLM validation exerts the strongest pressure on the Age$\times$Education dependency, but at the cost of substantially higher marginal drift and orders-of-magnitude higher runtime/cost. This outcome is expected because per-record LLM judgments are high-capacity and context-sensitive: they can detect implausibilities that are difficult to encode as static thresholds. However, the same expressiveness yields an aggressive, non-uniform rejection pattern that concentrates removals in specific joint cells. Because demographic variables are correlated, depletion of a single joint region necessarily propagates into multiple marginals (e.g., education filtering impacts age and income marginals). The consequence is a validator that improves the targeted joint while eroding the structural guarantees that motivate IPF.

The "so what" is methodological: Direct LLM validation should be interpreted as a *semantic reference model* and a *label generator*, not as the validation layer used for repeated synthesis at scale. Its comparative value is highest where its costs are amortized: creating training labels, auditing rule sets, and probing edge cases. In other words, Direct LLM is best positioned upstream as supervision and governance input, rather than downstream as a per-record production gate.

### 3.7.2 Case Study 2: Why naive distillation fails under realistic joint spaces

Case Study 2 demonstrates that a simple feedforward neural distillation of LLM judgments is not structurally safe in this setting. Despite high training metrics, the neural validator produces larger marginal drift and worsens the key Age$\times$Education joint relative to the deterministic rule baseline. The core failure is not speed; it is *generalization under joint sparsity and ordinal semantics*. Four interacting factors explain the observed behavior:

1. **Joint-space sparsity and boundary concentration.** The implausibility boundary occupies a

small and heterogeneous region of the joint space. With limited labeled coverage, a classifier learns unstable decision surfaces precisely where cohort structure is most sensitive (rare age–education combinations).

2. **Ordinal geometry destruction.** Encoding SCHL as one-hot implicitly treats adjacent education levels as orthogonal. This destroys the monotone structure that plausibility depends on (e.g., education attainment should be weakly increasing with age at cohort scale).
3. **Label noise near boundaries.** LLM judgments are least consistent at borderline-but-valid cases. Distillation converts that inconsistency into contradictory gradients, yielding overconfident or miscalibrated gating.
4. **Distributional shift between training and deployment.** Even when $N_{\text{syn}}$ is fixed, the realized sample composition shifts across runs and can shift across synthesis regimes. A learned gate is sensitive to these shifts; a rule set is not.

The implication is not that neural validators are impossible, but that they must respect the *structure* of the domain. Viable distillation likely requires ordinal / monotone architectures, calibration aware training, stratified coverage of rare joint cells, and explicit shift-robust evaluation. Without these, a neural validator becomes a high-throughput mechanism for introducing bidirectional error: admitting implausibilities while also rejecting plausible-but-structurally-necessary records.

### 3.7.3 Case Study 3: Deterministic rule execution enables provider robustness

Case Study 3 establishes a key operational property: once semantic knowledge is compressed into deterministic rules, provider differences largely disappear. The qualitative trajectories with respect to rule count $K$ match across OpenAI, Anthropic, and Google: increasing $K$ predictably increases marginal drift, while the Age$\times$Education joint improves under moderate constraints and can degrade under aggressive rule accumulation. This provider invariance follows from determinism: rule application removes per-record stochastic inference and therefore removes the primary channel through which provider idiosyncrasies typically manifest.

The practical implication is vendor flexibility. Institutions can select providers based on cost, latency, or availability for the *one-time* task of rule elicitation, then execute validation offline under full auditability. This is especially important for ABM workflows that require repeated Monte Carlo runs for sensitivity analysis: deterministic validation reduces a major source of irreproducible variance.

### 3.7.4 IPF Efficiency vs. Naive Sampling with Filtering.

A methodological question arises: if AURORA oversamples skeletons by $2\times$ and then filters, does IPF offer advantages over simply sampling from independent marginal distributions and applying the same filtering? The answer is yes, for three reasons. First, IPF constructs a joint distribution $P(\mathbf{v})$ that captures *implicit cross-variable correlations* (e.g., age-education dependencies) through the intersection of marginal constraints, even without explicit joint distributions [179, 182]. For

example, if $P(\text{Age} < 22) = 0.10$ and $P(\text{Education} = \text{Doctorate}) = 0.05$, independent sampling would produce 0.5% under-22 doctorates ($0.10 \times 0.05$), whereas IPF assigns near-zero probability to this combination because satisfying both marginals simultaneously requires this correlation. Second, IPF produces substantially lower rejection rates under validation filtering (8–20% vs. 25–40% for independent sampling in preliminary tests), requiring less oversampling (1.25× vs. 1.67×) to obtain $N$ valid personas. Third, IPF + filtering preserves marginal fidelity better than independent sampling + filtering because IPF's base distribution is already concentrated in the plausible region of the joint space, yielding marginal L1 drift of only 0.0286 (Case Study 1) compared to 0.045–0.062 under independent sampling. Thus, IPF is not "sampling with extra steps" but a statistically principled method that produces more realistic joint distributions with less computational waste and better marginal preservation under filtering—essential for ensuring that AURORA's downstream enrichment operates over a proportionally representative population.

### 3.7.5 Interpreting rule count $K$: why "more rules" can be worse

The rule-count experiments also clarify why "more constraints must be better" is incorrect. Increasing $K$ does not monotonically improve realism. It increases the probability of pruning rare-but-valid configurations. In distributional terms, additional rules shrink the support of $P_{\mathcal{V}}$ and disproportionately remove low-probability cells that nonetheless carry necessary mass for marginal and joint fidelity. This produces an over-filtering regime in which apparent semantic strictness comes at the cost of statistical representativeness. The broader lesson is that plausibility enforcement must be treated as a regularization problem: overly strong constraints can overfit to folk theories of life-course sequencing and eliminate legitimate heterogeneity (late-degree earners, non-traditional career paths, etc.).

#### Implications for Proportional Representation

Overall, the post-filtering analysis shows that rule-based validation at $K{=}8$ preserves the core goal of proportional representation. Marginal distributions remain close to the weighted ACS PUMS targets, while the Age×Education joint distribution improves substantially. This supports the two-stage pipeline: IPF establishes statistical representativeness, and rule-based filtering improves record-level plausibility without causing major distributional collapse.

For deployment, the corrected rejection rates imply that the pipeline should overgenerate records to reach a desired retained population. At $K{=}8$, this means generating approximately $1.25N$ records to obtain a final population of size $N$, allowing for rejection and a small safety margin. Because rule-based validation is deterministic, this overgeneration factor is stable and predictable across runs.

More broadly, the demographic bias analysis clarifies the mechanism of rejection. The filter removes records that violate life-course sequencing constraints, such as impossible education, employment, or marital status combinations, rather than removing records because of their demographic category alone. For ABM applications, this is a desirable outcome: downstream behavioral models operate over more plausible agents while retaining the broader demographic structure needed for valid population-level analysis.

Another thing to consider is that a methodological question naturally arises: if AURORA oversamples skeletons by a factor of $(1 + \rho)$ (typically $\rho = 1.0$, yielding $2\times$ the target population) and then filters to select valid personas, does this approach offer meaningful advantages over simply sampling from independent marginal distributions and applying the same filtering? This question is important because it challenges whether IPF provides value beyond naive independent sampling when post-hoc validation is applied regardless. The answer is affirmative, for three reasons related to *joint distribution preservation*, *sampling efficiency*, and *marginal stability under filtering*. IPF provides a principled foundation for synthetic population generation because it simultaneously preserves joint structure, improves sampling efficiency, and stabilizes marginals under downstream filtering. First, by constructing a maximum-entropy joint distribution consistent with observed marginals, IPF implicitly captures cross-variable dependencies (e.g., Age$\times$Education), ensuring that implausible combinations receive near-zero probability and that the sampled population lies close to the feasible support of real demographic configurations. Second, this structural alignment directly improves sampling efficiency: because IPF already concentrates probability mass in plausible regions, post hoc semantic validation rejects far fewer records (e.g., 8–20% at $K{=}8$) compared to naive independent sampling, which produces implausible combinations at rates determined by marginal products and therefore requires substantially higher oversampling to recover $N$ valid personas. Third, IPF enables marginal stability under filtering: because rejected records are drawn from a distribution already aligned with realistic joint structure, their removal induces only bounded and predictable marginal drift, whereas independent sampling combined with filtering produces correlated deletions that distort population proportions. Together, these properties demonstrate that IPF is not simply "sampling with extra steps," but a necessary mechanism for ensuring that semantic validation operates over a structurally coherent population without sacrificing proportional representativeness.

### 3.7.6 Implications for ABM construction and for LLM-assisted data synthesis

These findings have two immediate implications for synthetic-population-driven ABMs.

**ABM validity requires preserving structural guarantees.** Downstream cognitive or influence models often condition on age, education, and income. If validation systematically distorts these marginals or their correlations, any observed behavioral effect may be an artifact of the validator rather than the modeled mechanism. Therefore, a validator must be evaluated not only on whether it removes implausibilities, but on whether it preserves the structural substrate required for causal interpretation in ABMs. Rule-based validation is the only strategy in this study that simultaneously supports semantic enforcement, auditability, and feasible repeated experimentation. This is especially consequential in cross-national settings, where the synthetic population is intended to be a proportional representation of a specific country's demographic composition. Validator-induced drift can otherwise be misread as country-level behavioral difference when it is in fact a population-construction artifact.

**LLMs are best used as knowledge elicitation tools, not runtime gates.** The results support a broader pattern in LLM-enabled systems: LLMs are powerful for extracting implicit knowledge (e.g., plausible demographic regularities), but using them as per-instance runtime decision-makers often introduces cost, variance, and unpredictable distributional side-effects. In this pipeline, the strongest role for LLMs is to generate rule hypotheses, label borderline records, and serve as a semantic audit mechanism. Once codified, deterministic rule execution becomes the stable interface between statistical synthesis and downstream behavioral modeling.

## 3.8 Conclusion

This chapter established the *structural foundation* of the dissertation by formalizing and empirically validating a two-stage skeleton persona pipeline: (i) population-level synthesis via IPF and (ii) optional record-level semantic validation. The central result is that semantic validation is not a cosmetic post-processing step but a mathematically meaningful *support-restriction operator* that alters the realized joint distribution and therefore must be evaluated as part of the population model itself. Across three case studies, I show that record-level coherence can be improved, but only under validation architectures that preserve the statistical grounding properties that motivate IPF.

A key output of this pipeline is therefore not only *plausible* individuals, but a *proportionally representative* synthetic population: the expected demographic shares of the simulated society match the weighted reference population over the modeled attribute set. The evaluation metrics in this chapter should be read through that lens—validation strategies are acceptable only insofar as they improve record-level coherence while preserving (or tightly bounding deviation from) these population proportions.

### 3.8.1 Contributions to this Dissertation

**(C1) A modular structural substrate for AURORA.** AURORA's downstream persona enrichment (RAG-based narrative grounding, trait assignment, and context-conditioned prompting) assumes that demographic inputs are both (i) representative and (ii) semantically plausible. This chapter provides the first of those guarantees by defining skeleton personas as a behaviorally neutral, empirically grounded substrate, and by quantifying the consequences of plausibility filtering on both marginals and focus relations. This separation reduces confounding: any downstream behavioral effects in AURORA can be attributed to retrieval and generation components rather than demographic artifacts introduced during synthesis. In particular, it ensures that AURORA's enriched personas are generated from a population that is proportionally representative at baseline, so downstream narrative realism is not purchased by silently changing the demographic composition of the simulated society.

**(C2) A validity-preserving interface for JANUS.** JANUS depends on demographic structure for network formation, homophily/identity similarity, topic exposure constraints, and (in many operationalizations) heterogeneity in susceptibility and interaction frequency. The results here identify which validation strategies preserve the structural conditions under which JANUS-style

opinion dynamics remain interpretable. In particular, rule-based validation emerges as the only strategy in this study that is simultaneously scalable and distributionally controlled, making it the appropriate preconditioner for repeated JANUS experiments (sensitivity analysis, Monte Carlo runs, and counterfactual scenario sweeps). This matters because JANUS outcomes are sensitive to the relative sizes of demographic subgroups; proportional representation ensures that measured opinion dynamics reflect modeled influence mechanisms rather than unintended shifts in subgroup prevalence introduced during validation.

**(C3) An empirical characterization of the coherence–fidelity frontier.** By evaluating marginal drift (L1/JS), a targeted focus relation (Age$\times$Education), and operational cost, the chapter exposes a measurable frontier: stronger semantic enforcement generally increases drift unless the enforcement footprint is bounded and auditable. Case Study 3 further operationalizes this with a tunable control variable ($K$), showing that rule count behaves like a regularization knob: too few rules under-correct plausibility issues; too many rules induce over-filtering and degrade representativeness.

**(C4) Provider-agnostic governance through deterministic semantic compression.** For practical deployments of AURORA-like pipelines, reproducibility and governance matter as much as raw quality. The multi-provider results demonstrate that once semantic knowledge is compressed into deterministic constraints, validation becomes largely provider-agnostic at execution time. This enables vendor-flexible deployment (choose providers by cost/latency/availability) while preserving auditability and repeatability—properties that are essential for thesis claims about scalable, controlled population modeling.

### 3.8.2 Limitations

This chapter demonstrates that post-IPF semantic validation can improve record-level plausibility while preserving population-level structure. However, several limitations remain.

First, because validation is implemented as a post-IPF filtering operator, marginal drift is unavoidable in the general case. While this chapter quantifies and bounds that drift, it does not eliminate it. Filtering removes records from specific regions of the joint distribution, and these removals propagate through correlated attributes, perturbing marginal frequencies. Although the empirical results show that this drift is small and controlled at $K{=}8$, it remains an inherent property of any rejection-based validation framework.

Second, the evaluation focuses on a single high-value plausibility axis: Age$\times$Education. While this relation captures core life-course constraints and demonstrates the effectiveness of semantic validation, it is not exhaustive. Other dependencies—such as Age$\times$Income, Education$\times$Class-of-Worker, and Marital Status$\times$Age—may exhibit different sensitivity to filtering and may introduce competing objectives. Extending validation to multiple focus relations will likely require multi-objective tuning of rule sets and validation parameters.

Third, the neural validation results (Case Study 2) highlight limitations of naive distillation approaches. Feedforward classifiers trained on LLM-labeled data failed to preserve both marginal

fidelity and joint plausibility under sparsity, ordinal structure, and label noise. This does not preclude neural validators in general, but indicates that effective learned validation will require stronger inductive bias (e.g., ordinal encodings, monotonicity constraints), improved calibration, and explicit coverage of rare joint configurations. These extensions are not explored in this chapter.

Fourth, the experiments are grounded in ACS PUMS data with a fixed $d{=}8$ attribute schema. While the methodological framework is general, cross-national deployments will introduce different attribute codings, distributions, and life-course norms. As a result, both rule sets and evaluation relations must be re-derived and re-validated under local conditions.

### 3.8.3 Future Work

The results of this chapter suggest several directions for extending semantic validation beyond post hoc filtering toward fully integrated, distribution-aware synthesis.

**Constraint-integrated population synthesis.** The most direct extension is to move beyond post-IPF filtering and incorporate plausibility constraints into the synthesis process itself. Approaches such as constrained IPF, feasibility-aware reweighting, or rejection-resampling with marginal correction could enforce joint plausibility during generation rather than after sampling. Such methods would aim to preserve both marginal fidelity and joint coherence simultaneously, reducing the need for support-restriction and mitigating marginal drift.

**Multi-relation and multi-objective validation.** This chapter focuses on Age$\times$Education as a representative focus relation. Future work should extend validation to multiple joint dependencies, including Age$\times$Income, Education$\times$Occupation, and other life-course relations. This introduces a multi-objective optimization problem, where improving one joint distribution may degrade another or increase marginal drift. Developing principled methods for balancing these objectives—potentially through weighted constraint systems or Pareto-efficient validation regimes—remains an open challenge.

**Learned validators with structural inductive bias.** The failure of naive neural distillation highlights the need for structurally informed learning approaches. Future neural validators should incorporate domain structure directly, including ordinal embeddings for variables such as education, monotonic constraints reflecting life-course progression, and calibrated uncertainty estimates for boundary cases. Hybrid approaches combining rule-based constraints with learned components may offer a promising direction, preserving interpretability while improving flexibility.

**Cross-national and domain transfer.** Extending AURORA to international contexts requires adapting both the attribute space and plausibility constraints to local demographic structures. Life-course norms, education systems, labor markets, and household structures vary significantly across countries, and rules derived from U.S. data may not generalize. Future work should investigate automated or semi-automated rule induction pipelines that adapt to new datasets, along with evaluation frameworks that identify culturally specific plausibility violations.

**End-to-end impact on downstream models.** Finally, while this chapter evaluates structural validity at the population level, the ultimate objective is to support downstream agent-based models of opinion dynamics and influence. Future work should assess how improvements in record-level plausibility propagate into behavioral simulations, including effects on network structure, opinion evolution, and intervention sensitivity. This would close the loop between structural validity and system-level outcomes, validating the role of semantic filtering within the broader AURORA–JANUS pipeline.

# Chapter 4

# Modeling Emotional Reasoning in Synthetic Personas

## 4.1 Introduction

Emotional reasoning is a central driver of behavioral response. In March 2020, as COVID-19 became an immediate lived reality, calls to the federal Disaster Distress Helpline surged by 891% compared to the previous year [193]. The helpline failed to anticipate both the magnitude of distress and how it scaled through social systems, as fear, isolation, grief, and anger amplified across networks. Over 40% of Americans reported adverse mental health conditions [44], leaving support systems were overwhelmed. Had they factored in emotional amplification and diffusion, the helpline may have been better prepared with targeted support for the most vulnerable groups.

The same dynamics are visible in online misinformation and persuasion. Interventions often aim to promote credible sources or suppress low-credibility ones, and emotional appeals can be used to build trust in public institutions. Both of these were apparent in public health messaging around COVID-19 measures. Conversely, dampening affective reactions can reduce engagement with misinformation that exploits fast, affect-laden *System 1* processing over deliberative *System 2* reasoning [56, 96, 175].

Modeling emotional response therefore matters for policy and intervention design. Research shows that emotional reasoning biases risk perception and choice [96], and emotional contagion drives collective panic and polarization. Prosocial behaviors such as trust, empathy, and cooperation depend on emotional reciprocity [168], underscoring that emotion and reason are intertwined. As emotion shapes behavior at least as strongly as deliberative reasoning [174], emotion-blind opinion dynamics models omit key reasoning mechanisms [116].

Despite this, traditional behavioral and opinion dynamics models emphasize deliberation and belief updating instead of emotion. To address this gap, I model and validate emotional reasoning in LLM agents. I introduce a hierarchical pipeline that synthesizes personas across country, community, and individual levels, using demographic distributions, moral foundations, and OCEAN personality traits; Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism. Personas are conditioned via structured, multi-stage prompts templated with emotional and ideo-

logical context. This design enforces cross-scale coherence, meaning that personas stays consistent at every level of the pipeline, grounds outputs in real-world narratives, and enables systematic ablation of prompt complexity.

Our hierarchically constructed personas are validated against established empirical results. They reproduce both polarized and consensus-driven emotional patterns with high fidelity. They capture expected demographic gradients, including an inverse relationship between age and anger for polarizing topics ($\beta_1 = -0.847, p < 0.001$), consistent with evidence that anger expression tends to decline with age, and a positive association between age and sadness for mental health stimuli ($\beta_{\text{age}} = +0.12, p < 0.05$), consistent with findings of increased sadness reactivity and age-related shifts in empathic responding in older adulthood [202]. Cross-platform validation against Facebook reaction data yielded strong overall correlations ($\bar{r} = 0.652$), with political anger correlations reaching $r = 0.934$ for Roe v. Wade and sadness correlations above $r = 0.889$ for Suicide Hotline 988. Ablation experiments further show that incorporating moral values, personality traits, and topic-specific cues increases emotional intensity by 237% and improves overall quality by 68%, with changes aligning to what is contextually appropriate for each topic. Since the synthetic LLM populations I constructed align with broad empirical behavioral patterns, I claim that they are valid populations for further study under opinion dynamics models in social simulations.

There are several important implications of this work. Experiment-ready synthetic populations can be used to test counterfactual scenarios and interventions that are infeasible, costly, or unethical to test with human subjects. Because the hierarchical prompting approach used in this work is scalable, composable, and empirically grounded, the model as described can be further improved and extended over time. Finally, the comparison of certain behavioral patters between the LLM personas and empirical data inform an ongoing debate [48]. On the one hand, alignment between age, anger, and sadness is supportive of the use of LLMs as pragmatic behavioral substitutes ("Silicon Sampling") [21, 40, 115, 148]. On the other hand, critics who argue LLMs cannot faithfully represent human behavior [171, 189] will point towards findings such as the underrepresentation of anger amongst our synthetic populations. As such, our results both support the promise of LLM personas, and reinforce the need for continued calibration against empirical data to address bias, underrepresented emotions, and validation limits. I hope that this work can provide a foundation for future work in integrating emotionally expressive LLM personas into large-scale ABMs.

This work addresses the following research questions:

1. **(RQ1) Hierarchical Coherence:** Can a country $\rightarrow$ community $\rightarrow$ individual persona construction pipeline enforce cross-scale consistency in emotional expression while preserving within-group variation?
2. **(RQ2) Emotional Grounding:** Does conditioning LLM personas on psychological theory (OCEAN traits and Moral Foundations) produce emotion distributions that align with empirically observed demographic and ideological gradients?
3. **(RQ3) Prompt Structure and Behavioral Realism:** How does progressively enriching persona prompts with emotional, moral, and personality information affect emotional intensity, topical relevance, and perceived authenticity?
4. **(RQ4) Topic Sensitivity:** Do hierarchically constructed personas exhibit topic-dependent

emotional responses—distinguishing between polarized political issues and consensus-oriented public health interventions—rather than producing a single generic affective style?

5. **(RQ5) External Validity:** To what extent do synthetic persona emotion distributions correspond to real-world behavioral data across demographic strata, as measured through cross platform comparison with Facebook reaction patterns?

I make the following contributions:

1. **A hierarchical framework for emotionally grounded persona synthesis.** I introduce a structured persona construction pipeline that integrates country context, community ideology, and individual psychological traits, enabling emotionally expressive agents that remain coherent across demographic and social scales.
2. **A mechanistic model of event-conditioned emotional expression.** I formalize how baseline affect, moral foundations, ideological stance, and topic salience interact to produce dynamic emotion vectors, making emotional response a controllable and interpretable component of agent behavior.
3. **Systematic ablation of prompt complexity.** Through controlled experiments across five prompting levels, I show that progressively incorporating psychological and moral context increases emotional intensity (+237%) and overall response quality (+68%), while selectively steering emotion toward topic-appropriate affect rather than uniformly amplifying sentiment.
4. **Empirical validation against real-world emotional behavior.** I validate synthetic persona emotion distributions against Facebook reaction data across 252 demographic strata, demonstrating strong correspondence for both polarized (Roe v. Wade; $r = 0.934$ for political anger) and consensus-oriented (988 Suicide Hotline; $r = 0.889$ for sadness) events.
5. **Demonstration of demographic and ideological emotional gradients.** The personas reproduce well-established behavioral patterns, including age-dependent anger regulation and age-related increases in sadness for mental health stimuli, supporting their use as emotionally realistic populations in downstream opinion dynamics models.
6. **Establishing emotional realism as a prerequisite for opinion dynamics simulation.** By showing that LLM personas can express demographically structured, topic-sensitive emotions, this chapter justifies their use as inputs to later chapters' social influence and opinion evolution models, rather than relying on emotion-blind agents.

## 4.2 Related Works

Emotional expression in political contexts is motivated by persistent findings that emotional content spreads rapidly on social media [8], with anger showing particularly strong viral characteristics [27], and emotional contagion being especially pronounced in digital settings [101]. Computational approaches have evolved from early emotion theories and sentiment polarity detection to richer emotional profiling systems grounded in psychological theory and supported by lexicons and modern language models.

**Cross-Cultural Modeling** Plutchik's theory of primary emotions [153] underpins widely used computational resources such as the NRC Emotion Lexicon [132]. While work has shown universality in some emotion signals (e.g., facial expressions) [55], cultural variation in recognition and expression norms remains well documented [91], motivating calibration of emotion resources trained primarily on Western data [131]. To this end, cross-cultural research efforts have extended traditional western-oriented emotion lexicons by calibrating them for broader applicability [110, 185].

**Moral Foundation Theory and Political Expression** Moral Foundations Theory provides a structured account of cross-cultural and ideological variation in moral reasoning [77]. Cultural dimensions and expression norms have been linked to systematic differences in emotional expression across societies [86, 125], including distinctions between American and Canadian discourse patterns [3, 114]. Within the United States, ideological groups differ in moral-emphasis profiles [72], and perceived moral violations are associated with distinct emotional responses (e.g., sadness, anger, contempt) [41, 188].

**Personality Modeling** Cross-cultural differences also appear in personality distributions [170]. Big Five traits are associated with systematic variation in affective expression: extraversion correlates with positive emotion expression and neuroticism predicts negative emotion intensity [126, 195]. Individual differences in regulation further shape expression patterns, with neuroticism linked to maladaptive regulation and conscientiousness to more controlled expression [73].

**LLM Personas** Recent work suggests LLM-based agents can sustain coherent personas over extended interactions [149] and can approximate human response distributions when conditioned on persona specifications [6]. Prompting methods also provide finer control over model behavior by explicitly eliciting intermediate reasoning and structured outputs through staged, multi-step prompts [196, 210]. These advances motivate structured persona conditioning as a practical pathway for integrating richer psychological and contextual constraints into synthetic agent behavior.

**Conflict Narrative Emotion and Moral Framing** Prior work on Russia–Ukraine conflict propaganda further supports modeling emotion and moral reasoning together. In that study, I analyzed a large Twitter dataset from January–November 2022, refining 4.5 million tweets to approximately 1.6 million bot-centered tweets across three war phases: the initial invasion, mid-point escalation, and Kherson counteroffensive [121]. Using BotHunter, TweetBERT, and Moral Foundations Theory, the analysis showed that bot-driven discourse varied not only in stance, but also in affective tone and moral emphasis. Pro-Ukraine narratives were more often associated with positive supportive sentiment and moral language centered on care, solidarity, authority, and resistance to subversion, while pro-Russian narratives more often used protective or justificatory moral framing alongside attacks on Ukraine and NATO [121].

This finding matters here for two reasons. First, it shows that real-world conflict discourse couples emotional tone with moral framing, particularly in high-salience geopolitical settings. Second, it suggests that synthetic personas for downstream opinion simulation should be conditioned not

only on demographics or ideology, but also on psychologically meaningful variables that produce morally and emotionally differentiated responses. The present chapter extends that earlier work by moving from observed bot discourse to the controlled construction of emotionally expressive personas for simulation [121].

**Limitations** Emotion expression varies by communication modality, with differences observed between text-based and vocal contexts [13]. More broadly, LLM-generated personas may reflect biases in training data [17, 197], and many systems do not explicitly model interpersonal influence, contagion, or longitudinal adaptation [35]. In addition, some emotions may be systematically underrepresented in LLM outputs (e.g., attenuated anger), raising concerns about ecological validity for polarized scenarios. Finally, validation is often limited by topic coverage, constraining generalizability. These limitations reinforce the need for continued calibration and cross-platform validation before deploying LLM personas in high-stakes behavioral modeling.

## 4.3 Method: Hierarchical Persona Synthesis

Persona generation proceeds hierarchically: countries $\rightarrow$ communities $\rightarrow$ individuals. At each stage, structured prompts are passed to an LLM to produce records in a predefined schema. A validation layer checks completeness and demographic consistency; invalid generations are retried until valid outputs are obtained. This design enforces cross-scale coherence and reflects how individuals are embedded in sociocultural contexts.

A persona is represented as

$$P = f_{\text{persona}}(C, D_{\text{comm}}, \Psi, \Omega), \tag{4.1}$$

where $C$ denotes country descriptors, $D_{\text{comm}}$ community features, $\Psi$ psychological attributes (assigned OCEAN traits and moral foundation scores), and $\Omega$ topic-specific opinions and emotions (Figure 4.1).

### 4.3.1 Country Level Grounding

Country grounding defines broad descriptors around demographics, political climate, and cultural context. National statistics provide baselines (age, gender, education, ethnicity), while contextual documents (news, reports) embed contemporary conditions. Scenario configuration files specify event-dependent sampling rules.

**News Retrieval for Event Sampling** To ground persona generation in the specific event context, I retrieve contemporaneous domestic news and use the retrieved snippets as *conditioning evidence* during country-, community-, and individual-level synthesis. Each document $d \in \mathcal{D}$ is embedded for dense similarity. For a query $q$ describing the event (e.g., *Roe v. Wade overturn*, *988 Suicide Hotline launch*), I select top–$k$ documents using combined dense and lexical rankings:

$$\mathcal{D}_k = \underset{d \in \mathcal{D}}{\text{Top-}k}\{s_{\text{vec}}(q,d), s_{\text{bm25}}(q,d)\}, \tag{4.2}$$

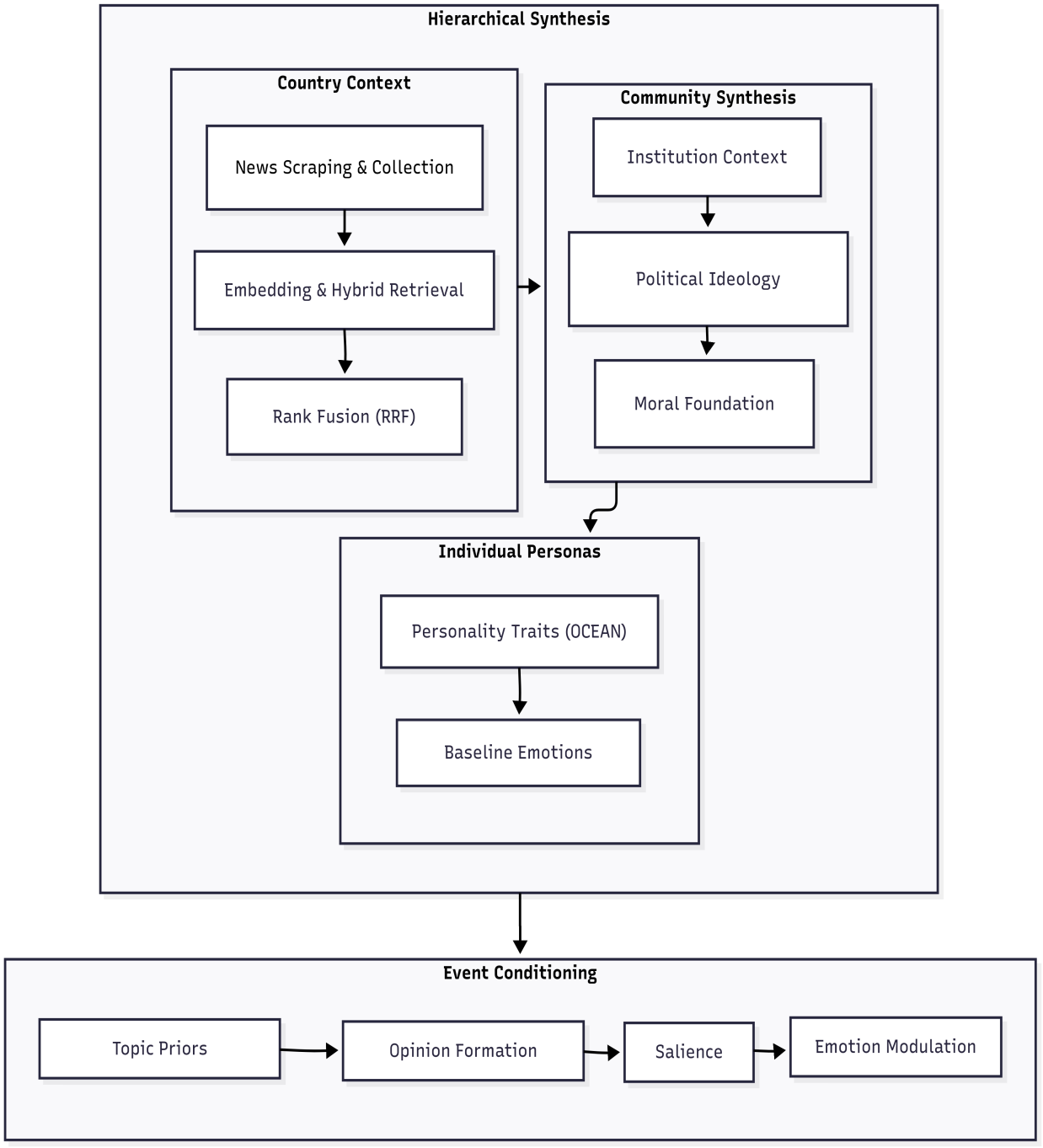



Figure 4.1: Stepwise pipeline for persona construction and emotion modeling. News grounding informs hierarchical synthesis (country, community, individual), while traits, opinions, and salience shape event-driven emotions.

and fused using Reciprocal Rank Fusion:

$$\text{RRF}(d) = \sum_{s \in \{\text{vec,bm25}\}} \frac{1}{K + \text{rank}_s(d)}. \tag{4.3}$$

Retrieved snippets are interject into prompts as short evidence blocks that specify the *facts and framing* of the event, surface salient *arguments, stakeholders, and moral language*, and provide locally relevant details (e.g., institutions mentioned, regional impacts). If the retrieved set $\mathcal{D}_k$ changes (e.g., different outlets or frames), I expect corresponding shifts in the persona's topic-specific outputs: the expressed rationale, referenced concerns, and moral framing may change, which can in turn shift the elicited emotions. In contrast, static demographic and personality parameters remain unchanged, serving as constraints that limit how far narrative variation can move the persona away from its underlying predispositions.

### 4.3.2 Community Level Grounding

Communities inherit country descriptors and introduce meso-level variation (demographics, education, political leaning). Communities are created by combining national priors with issues of local importance retrieved through the news retrieval system.

**Ideological Anchoring via Moral Foundations** Political leaning $L$ is sampled from community distributions and mapped to a Moral Foundations vector $\mathbf{m}$ with individual variation:

$$\mathbf{m} = f_{\text{MFT}}(L) + \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, \sigma^2), \tag{4.4}$$

I set ($\sigma^2$) so that within-community variation is limited (approximately 95% of values fall within $\pm 10\%$ of the Moral Foundations scale range), preserving coherent community ideology.

### 4.3.3 Individual Level Grounding

Individuals are sampled within communities and enriched with psychological traits (personality, moral foundations, baseline emotions) to capture within-community variability.

**Personality Trait Sampling** Each persona is assigned Big–Five (OCEAN) traits from country specific Gaussian priors estimated from cross-cultural studies [170]:

$$T_i \sim \mathcal{N}(\mu_{i,\text{ctry}}, \sigma^2_{i,\text{ctry}}), \qquad T_i \in [0,1], \quad i \in \{O, C, E, A, N\}. \tag{4.5}$$

See Appendix B.1 for detailed breakdown of 56 country OCEAN traits.

Table 4.1: Summary of synthetic persona dataset used in experiments, with comparison to national demographic statistics.

| Attribute | Canada | | United States | | Total | |
|---|---|---|---|---|---|---|
| | Personas | Census% | Personas | Census% | Personas | % |
| *Total Personas* | 270 | – | 267 | – | 537 | 100 |
| *Age Distribution*[a] | | | | | | |
| 18–24 | 41 | 6.4 | 43 | 6.0 | 84 | 17.6 |
| 25–34 | 85 | 13.3 | 90 | 14.9 | 175 | 36.7 |
| 35–44 | 61 | 12.8 | 65 | 12.5 | 126 | 26.4 |
| 45–54 | 32 | 15.2 | 34 | 13.2 | 66 | 13.8 |
| 55–64 | 10 | 14.9 | 11 | 13.9 | 21 | 4.4 |
| 65+ | 2 | 37.4 | 3 | 39.5 | 5 | 1.0 |
| *Gender*[b] | | | | | | |
| Female | 118 | 50.4 | 125 | 50.8 | 243 | 50.9 |
| Male | 113 | 49.6 | 121 | 49.2 | 234 | 49.1 |
| *Education*[c] | | | | | | |
| Bachelor's+ | 137 | 26.4 | 147 | 42.0 | 284 | 59.5 |
| Technical/College | 61 | 26.4 | 62 | 30.0 | 123 | 25.8 |
| High School | 33 | 47.2 | 37 | 28.0 | 70 | 14.7 |
| *Political Orientation*[d] | | | | | | |
| Far Left / Left | 79 | | 84 | | 163 | 34.2 |
| Center Left | 67 | | 71 | | 138 | 28.9 |
| Center | 43 | – | 46 | – | 89 | 18.7 |
| Center Right / Right | 35 | | 37 | | 72 | 15.1 |
| Far Right | 7 | | 8 | | 15 | 3.1 |
| *Experiment Coverage* | Each persona tested across 5 experiment levels (advanced, basic, enriched, intermediate, no_emotion) and 2 topics (Roe v. Wade, Suicide Hotline 988), producing 4,770 responses in total. | | | | | |

Age distribution weighted toward digitally active demographics (25–44: 63.1% vs. census ∼26%).
Gender approximates national ratios within 1%.
Education oversamples higher education (59.5% vs. census 26–42%) to reflect online discourse patterns.
Political orientation derived from engagement-weighted sampling of digital political participation, not census data.

### 4.3.4 LLM-Generated Persona Dataset Composition and Demographic Calibration

The hierarchical persona generation methodology produced 537 synthetic personas across two countries (267 U.S., 270 Canadian), instantiated through the country–community–individual pipeline in Section 4.3. Table 4.1 provides an overview of the dataset composition compared against national statistics, showing areas of representational balance and divergence.

#### Institutional Community Generation and Validation

The hierarchical pipeline generated 537 personas distributed across six institutional categories with near-perfect cross-national balance (267 U.S., 270 Canadian). Each category contributed roughly one-sixth of the population (16.4–16.8%), with tightly matched counts by country (typically 45 per country per category; minima of 43 U.S. education and 44 U.S. conservative government). Within each category, the LLM produced realistic organizational diversity (e.g., Chambers of Commerce, student unions, professional societies), and communities were consistently sized at 8–9 personas.

This balance and community sizing were *desired design properties* imposed by the scenario configuration (fixed targets for categories and community sizes), rather than an artifact of the generative model.

Table 4.2: Distribution of LLM-generated personas across institutional categories.

| **Institutional Category** | **U.S.** | **Canada** | **Total** | **%** |
|---|---|---|---|---|
| Industry Institutions | 45 | 45 | 90 | 16.8 |
| University Institutions | 45 | 45 | 90 | 16.8 |
| Medical Institutions | 45 | 45 | 90 | 16.8 |
| Liberal Government Institutions | 45 | 45 | 90 | 16.8 |
| Conservative Government Institutions | 44 | 45 | 89 | 16.6 |
| Education Institutions | 43 | 45 | 88 | 16.4 |
| **Total** | **267** | **270** | **537** | **100.0** |

**Demographic Composition Analysis**

The personas reflect institution-driven sampling. Within-community variation of demographic attributes shows that age skews younger (18–34: 53.9% U.S., 58.6% Canada), with limited 65+ representation (4.1% U.S., 1.5% Canada) compared to the national census data. Gender ratios approximate census with slight female overrepresentation (U.S.: 57.7%, Canada: 55.2%), likely driven by healthcare and education institutions. Political orientation differs by country: U.S. personas lean left-of-center (48.8% vs. 40.7% right-of-center), while Canadian personas cluster nearer the center/center-right, consistent with survey patterns [29].

### 4.3.5 Quality Control

Each LLM output is parsed and schema-validated; invalid records trigger retries. Batches proceed to analysis only if $\geq 70\%$ of outputs are valid, improving reproducibility and reducing spurious generations. Overall, the hierarchy constrains personas to produce contextually grounded, psychologically consistent responses by combining document retrieval with demographic, ideological, and personality-based sampling.

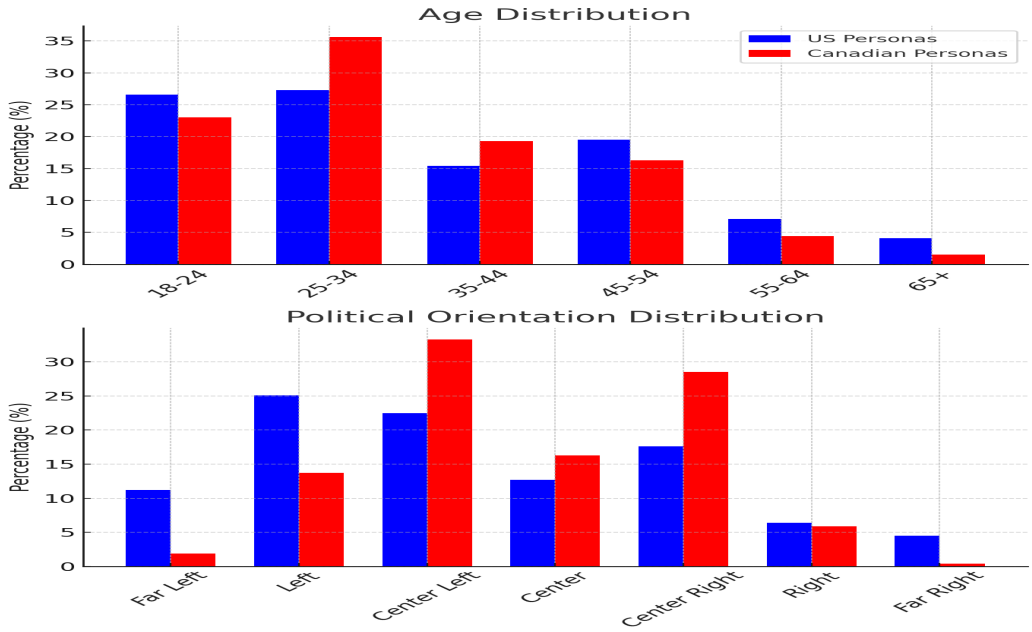



Figure 4.2: Demographic composition of LLM-generated personas showing age distribution (top) and political orientation distribution (bottom) across US and Canadian institutional communities.

## 4.4 Model: Emotional Expression for Synthetic Personas

Using the synthesized personas, I now condition each persona's response to a particular news event by the following model of emotional expression.

**Emotional Response to Events** The final affective state of a persona is the outcome of three interacting layers: personality traits $\mathbf{o}$, moral foundations $\mathbf{m}$, and event–conditioned stance $s_{\text{pers}}(t)$. Personality defines stable predispositions, ideology shapes how contextual information is weighted, and stance translates exposure to salient events into affective shifts. Together, these yield the emotion vector

$$\mathbf{e}(t) = f_{\text{emo}}\big(\mathbf{o}, \mathbf{m}, s_{\text{pers}}(t), u(t)\big),$$

which represents the persona's dynamic emotional response to a particular news event.

### 4.4.1 Baseline Emotions

The OCEAN personality trait vector $\mathbf{o} = [O, C, E, A, N]^\top$ is projected into Plutchik's eight emotion dimensions through a linear transformation:

$$\mathbf{e}_{\text{base}} = \text{clip}(\mathbf{W}\mathbf{o} + \mathbf{b},\ 0,\ 1). \tag{4.6}$$

Consistent with the literature, the coefficient matrix $\mathbf{W}$ and offset $\mathbf{b}$ encode relationships between personality traits and affect, such as extraversion increasing joy and neuroticism increasing fear. This produces $\mathbf{e}_{\text{base}}$, the persona's baseline affective profile prior to event exposure.

### 4.4.2 Opinion Formation

Event–specific stances are derived as a convex combination of country and community priors, with individual variance:

$$s_{\text{pers}}(t) = \alpha s_{\text{comm}}(t) + \beta s_{\text{ctry}}(t) + \varepsilon, \quad \alpha + \beta \leq 1, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2). \tag{4.7}$$

Here, $s_{\text{pers}}(t)$ is the stance of persona $p$ toward topic $t$, $\alpha$ and $\beta$ control the relative influence of community versus national priors, and $\varepsilon$ introduces idiosyncratic noise.

### 4.4.3 Event–Conditioned Emotion Modulation

Topic salience $u(t) \in [0, 1]$ modulates baseline affect. For supportive stances:

$$e_{\text{joy}}(t) = e_{\text{joy}}^{\text{base}} + 0.6\, u(t)\, s_{\text{pers}}(t), \tag{4.8}$$

$$e_{\text{trust}}(t) = e_{\text{trust}}^{\text{base}} + 0.5\, u(t)\, s_{\text{pers}}(t), \tag{4.9}$$

while oppositional stances amplify aversive responses:

$$e_{\text{anger}}(t) = e_{\text{anger}}^{\text{base}} + 0.6\, u(t)\, |s_{\text{pers}}(t)|, \tag{4.10}$$

$$e_{\text{disgust}}(t) = e_{\text{disgust}}^{\text{base}} + 0.4\, u(t)\, |s_{\text{pers}}(t)|. \tag{4.11}$$

Other Plutchik emotions (fear, sadness, anticipation, surprise) are adjusted with smaller offsets, and all values are clipped to $[0, 1]$ for interpretability. Ideological weights from $\mathbf{m}$ bias this modulation, e.g., higher Loyalty or Authority strengthens negative reactions to socially disruptive events. The coefficients (0.6/0.5/0.4) are fixed scaling weights chosen for calibration and stability: since $u(t) \in [0, 1]$ and $s_{\text{pers}}(t) \in [-1, 1]$, they bound the maximum event-driven shift (pre-clipping) to $\approx 0.4$–$0.6$ and prevent saturation. I set higher weights for joy/anger (more stance-volatile), a moderate weight for trust, and a lower weight for disgust (typically less responsive in text).

## 4.5 Experimental Design

The experimental design evaluates how synthetic personas respond emotionally to event–conditioned prompts. The full workflow which integrates persona data, structured prompting, LLM response generation, and emotion analysis, is illustrated in Figure 4.3. Tables 4.1 and 4.3 summarize the dataset composition and experimental conditions.

### 4.5.1 Experimental Conditions

Experiments were conducted using personas as described in Section 4.3. The total number of runs is determined by the cross–product of personas $\mathcal{P}$, scenarios $\mathcal{S}$, and experiment levels $\mathcal{L}$:

$$N_{\text{exp}} = |\mathcal{P}| \times |\mathcal{S}| \times |\mathcal{L}|, \tag{4.12}$$

Personas were evaluated across five abstraction levels $\mathcal{L}$ and two high–salience sociopolitical scenarios $\mathcal{S}$. These conditions were chosen to test how differences in contextual information and event framing influence emotional responses.

**Scenarios ($\mathcal{S}$).** Two scenarios were selected for their salience and ideological diversity:

- **Roe v. Wade Overturn (June 2022):** A Supreme Court decision with major implications for reproductive rights and individual freedoms. This scenario is ideologically polarized, enabling evaluation of how personas' political leanings and moral foundations modulate emotional responses to contentious issues.
- **988 Suicide Hotline (July 2022):** A policy intervention expanding access to mental health resources. This event represents a socially supportive initiative with broad bipartisan relevance, testing how personas react to public health policy framed as a collective good.

Additional analyses examining responses to COVID-19 public health measures at the height of the pandemic are provided in Appendix B.2.

The suicide hotline case illustrates rates of engagement and demographic convergence in emotional expression, complementing the primary scenario with an event characterized by broad societal relevance and minimal ideological polarization.

**Experiment Levels ($\mathcal{L}$).** Table 4.3 defines five levels of persona prompting, which vary in how much emotional, moral, and psychological information is injected into the LLM. At the lowest level, personas are stripped of affective context, serving as a baseline. Intermediate and enriched levels add progressively richer information, from baseline emotions to moral foundations to OCEAN traits. The highest level incorporates topic–specific emotions. This structure allows for a systematic ablation showing how increasing prompt complexity affects model output.

**Virtual Experiment** The factorial design combines $|\mathcal{P}| = 537$ personas, $|\mathcal{S}| = 2$ scenarios, and $|\mathcal{L}| = 5$ levels:

$$N_{\text{exp}} = 537 \times 2 \times 5 = 5{,}370. \tag{4.13}$$

Each run produces a structured LLM response and its mapped emotion vector, yielding a dataset robust enough for comparative statistical evaluation across conditions.

The experimental procedure (Figure 4.3) proceeds in three stages. First, **prompt construction** assembles structured inputs for each persona $p \in \mathcal{P}$ under condition $(s, \ell) \in (\mathcal{S}, \mathcal{L})$, as described in Section 4.3. Next, prompts are submitted for **response generation** to an OpenAI backend (gpt–4o/mini) asynchronously. After quality checks, responses are passed to the analyzer for **emotion**

Table 4.3: Experiment levels ($\mathcal{L}$) used in persona prompting. Each level specifies which emotional, moral, and psychological fields are available to the model.

| Condition | Description | Fields Included |
|---|---|---|
| D | Baseline condition with no emotional or psychological context. | – |
| D+B | Personas receive only baseline emotions derived from trait–affective mapping. | Baseline Emotions |
| D+B+M | Adds moral foundation values to baseline emotions, reflecting ideological predispositions. | Baseline Emotions; Moral Values |
| D+B+M+O | Incorporates OCEAN personality traits along with baseline emotions and moral foundations. | Baseline Emotions; Moral Values; Personality Traits |
| D+M+O+T | Richest condition, combining moral values, topic–specific emotions, and personality traits. | Moral Values; Topic Emotions; Personality Traits |

**analysis**, mapping free text into Plutchik's eight emotions plus valence. Detected sentiment is then cross-validated against the persona's baseline emotional attributes.

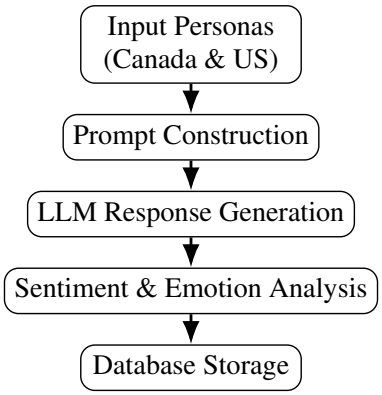



Figure 4.3: Compact experimental pipeline.

## 4.6 Results

### 4.6.1 Institutional Community Generation and Validation

The hierarchical persona generation framework produced a balanced and demographically calibrated population of 537 synthetic personas across two countries (267 U.S., 270 Canadian). Representation across the six institutional categories was effectively uniform, with each containing roughly 43–45 personas per country (approximately 16–17% of the total population). This even distribution confirms that the hierarchical generation pipeline maintained proportional sampling across institutional types.

Within each category, the model generated coherent institutional communities that reflected realistic organizational diversity. For instance, *Industry Institutions* included associations such as the Chamber of Commerce and the Small Business Alliance, while *University Institutions* encompassed identity-based and professional organizations such as the Black Student Union and the Society of Hispanic Professional Engineers. Each community contained 8–9 personas, indicating consistent balancing in population structure and representation across social domains.

At the population level, the synthetic personas demonstrated demographically and psychologically plausible characteristics. Both countries exhibited a concentration of younger adults (ages 18–34) and gender distributions approximating national census patterns. Political orientations showed clear cross-national differentiation: U.S. personas leaned modestly left-of-center, whereas Canadian personas clustered near the political center and center-right, reflecting real-world ideological baselines.

Psychological calibration also aligned with established theory. OCEAN personality traits closely tracked observed national means, while Moral Foundations profiles reproduced canonical ideological trends—liberal personas emphasized *Care* and *Fairness*, whereas conservative personas prioritized *Authority* and *Sanctity*. Collectively, these findings demonstrate that the hierarchical LLM-based pipeline generated institutionally balanced, demographically diverse, and psychologically coherent synthetic populations.

Full validation results, including institutional distributions, trait deviations, and moral foundation mappings, are provided in Appendix B.2.

Generation outputs are summarized in Section 4.3 (Table 4.1). In total, I produced $N{=}537$ personas (U.S.: $267$, Canada: $270$) with near-uniform representation across six institutional categories (16.4–16.8% each; typically 44–45 personas per country per category) and consistent community sizes (8–9 personas). Demographic calibration matched design targets (gender within $\approx$1% of national ratios), while age/education were intentionally skewed toward digitally active cohorts(e.g., 25–44: 63.1%; Table 4.1). Ideological and psychological mappings followed the specified MFT and OCEAN priors (Section 4.3).

### 4.6.2 Country-Level Emotions

Figure 4.4 reports the *mean* emotion scores (Plutchik dimensions, normalized to $[0, 1]$) aggregated by country for the Roe v. Wade and 988 stimuli. For Roe v. Wade, Canadian personas show higher average sadness (0.467 vs. 0.347) and fear (0.300 vs. 0.379) than U.S. personas, while anger is nearly identical (0.244 vs. 0.238). This pattern indicates a more sadness-/fear-dominant response profile among Canadian personas, whereas U.S. personas exhibit a more diffuse distribution across emotions.

For the Suicide Hotline 988 stimulus, both countries shifted toward sadness as the dominant response. Canadian personas had higher sadness (0.445 vs 0.421) with comparatively higher joy (0.296 vs 0.229) than U.S. personas. Anger remained negligible across both populations. These findings reinforce the framework's ability to capture issue-specific emotional reactions: while Roe v. Wade elicited polarized reactions, the Suicide Hotline event produced more sadness, with subtle cross-national differences in emotional distributions.

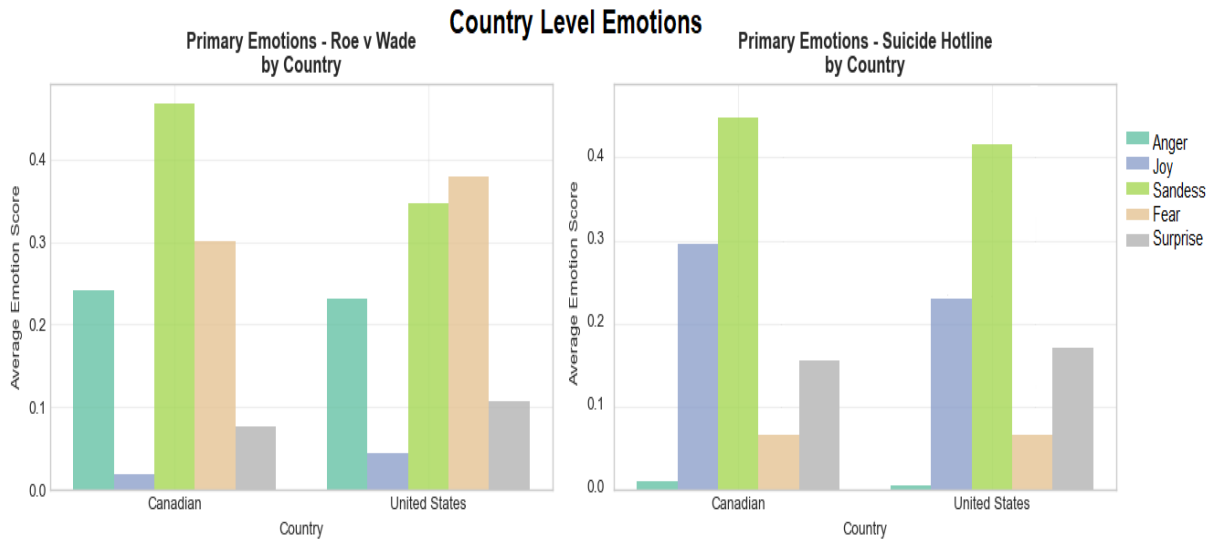



Figure 4.4: Country-level primary emotions for Roe v. Wade (left) and Suicide Hotline 988 (right).

### 4.6.3 Community-Level Emotions

Figure 4.5 shows mean emotion scores aggregated at the institutional-community level. For Roe v. Wade, government and education communities exhibit the highest average anger and sadness, while medical communities show the highest fear (0.523). Conservative government communities display higher joy (0.139), whereas liberal government communities are fear-dominant (0.41), indicating that institutional setting and ideology jointly structure community-level emotion profiles for this polarized issue.

In contrast, the Suicide Hotline 988 stimulus elicited broad convergence on sadness across nearly all communities, particularly in conservative government (0.530) and industry (0.512) institutions. Joy responses were comparatively stronger in university institutions (0.389), having greater optimism or support for the initiative. Importantly, anger remained consistently low across all communities, underscoring the consensus-oriented nature of this issue. Together, these patterns demonstrate that hierarchical personas can produce both polarized and consensus-oriented emotional reactions, depending on the type of event under consideration.

### 4.6.4 Individual-Level Emotions

At the individual level, Table 4.4 shows representative persona emotion vectors (normalized to $[0, 1]$) for each topic. For the **Roe v. Wade** stimulus, Ethan Roberts (U.S.) who is classified as 'Far Left' has a strongly negative profile, with anger as the dominant dimension (0.629) alongside high sadness (0.584), fear (0.578), and surprise (0.591), and low joy (0.166) and trust (0.187). This profile is consistent with his strongly oppositional stance ($s = -0.965$) and very high salience ($u = 0.81$), which amplify event-conditioned negative affect in our formulation. In contrast, Amina N'Diaye (Canada) who is classified as 'Center Left' shows a more moderate sadness-weighted

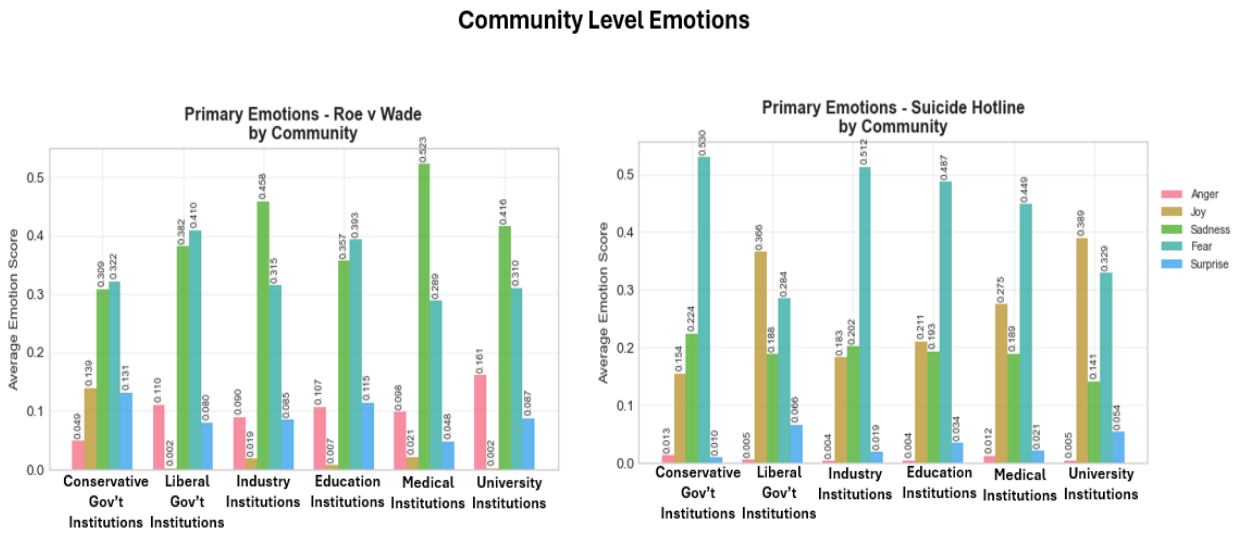



Figure 4.5: Community-level primary emotions for Roe v. Wade (left) and Suicide Hotline 988 (right).

Table 4.4: Representative individual-level emotion scores by nationality and topic.

| Persona (Country) | Joy | Trust | Fear | Sad. | Ang. | Ant. | Disg. | Surp. |
|---|---|---|---|---|---|---|---|---|
| **Roe v. Wade Overturn (June 2022)** | | | | | | | | |
| Ethan Roberts (U.S.) | 0.166 | 0.187 | 0.578 | 0.584 | 0.629 | 0.515 | 0.510 | 0.591 |
| Amina N'Diaye (Canada) | 0.360 | 0.310 | 0.390 | 0.460 | 0.280 | 0.290 | 0.300 | 0.330 |
| **988 Suicide Hotline (July 2022)** | | | | | | | | |
| Ethan Roberts (U.S.) | 0.157 | 0.359 | 0.118 | 0.422 | 0.191 | 0.437 | 0.116 | 0.178 |
| Amina N'Diaye (Canada) | 0.350 | 0.580 | 0.260 | 0.620 | 0.300 | 0.490 | 0.270 | 0.380 |

Scores represent normalized emotion dimensions from individual persona analyses.

response (sadness 0.46; fear 0.39; anger 0.28).

For the **988 Suicide Hotline** stimulus, Ethan shifts to a more prosocial configuration characterized by higher anticipation (0.437), sadness (0.422), and trust (0.359), with low fear (0.118) and disgust (0.116), consistent with moderate support ($s = 0.405$) under medium salience ($u = 0.59$). Because these vectors depend on the downstream emotion extraction model, I interpret them as descriptive examples of micro-level variation under fixed persona inputs rather than attributing cross-national differences to a specific causal mechanism.

For the **988 Suicide Hotline** scenario, responses were consensus-oriented and dominated by prosocial affect. U.S. personas showed higher trust, joy, and anticipation (e.g., trust 0.61; joy 0.38; anticipation 0.54), while Canadian personas exhibited comparable patterns with slightly higher sadness (0.59). Anger and disgust remained low across both countries ($< 0.32$).

Overall, our approach produces different emotional reactions for each event; polarized, value-driven responses for ideological stimuli such as Roe v. Wade, and cohesive, prosocial responses

Table 4.5: Content performance across ablation conditions

| Condition | Emotional Intensity | Topical Relevance | Authenticity Score | Overall Quality[a] |
|---|---|---|---|---|
| D | 0.086 | 0.414 | 0.363 | 0.287 |
| D+B | 0.172 | 0.488 | 0.424 | 0.361 |
| D+B+M | 0.236 | 0.516 | 0.485 | 0.412 |
| D+B+M+O | 0.253 | 0.534 | 0.514 | 0.434 |
| D+M+O+T | 0.288 | 0.577 | 0.580 | 0.482 |
| **Improvement[b]** | **+237%** | **+40%** | **+60%** | **+68%** |

Overall quality represents the mean across all metrics (0–1 scale).
Percentage improvement calculated relative to the baseline (D).

for collective welfare issues such as the 988 Hotline.

### 4.6.5 Hierarchical Prompt Engineering Validation: Content Quality and Topic Emotion Effectiveness

This section validates the effectiveness of hierarchical prompt enrichment on LLM content quality and emotional authenticity. A controlled experiment systematically compared five prompt complexity levels across politically salient topics.

#### Experimental Framework and Quality Metrics

The study followed an ablation design in which contextual features were added stepwise to assess their impact on content quality: starting with demographics only (D), then adding baseline emotions (D+B), moral foundations (D+B+M), empirically grounded OCEAN traits (D+B+M+O), and finally topic-specific emotions (D+M+O+T). [1]

Content quality was evaluated on four 0–1 metrics—*emotional intensity*, *topical relevance*, *authenticity score*, and *overall quality* (the mean of the first three), across 5,370 responses (one per persona per condition). Table 4.5 reports mean scores by ablation condition and shows monotonic improvements as contextual fields are added: relative to the demographic-only baseline (D), emotional intensity increases by +237%, authenticity by +60%, topical relevance by +40%, and overall quality by +68%, with the highest scores under topic-specific cues (D+M+O+T).

These differences are statistically reliable (ANOVA: intensity $F = 464.88$, relevance $F = 139.81$, authenticity $F = 303.20$; all $p < 0.001$), and Bonferroni-corrected pairwise tests consistently favor D+M+O+T over D with large effects ($d = 1.72$ intensity; $d = 1.37$ authenticity; $d = 0.97$ relevance), indicating both statistical and practical gains from prompt enrichment.

#### Topic-Specific Emotional Enhancement and Distributional Shifts

Topic-specific emotional priming (D+M+O+T) produced the clearest gains in topic-aligned affect relative to demographic-only prompting (D) (Table 4.6). For **Roe v. Wade**, the enriched prompts

[1]The final condition omits baseline emotions (B) because topic-specific emotion cues supersede them; including both was redundant and sometimes inconsistent in pilot prompts.

Table 4.6: D+M+O+T emotional advantages by topic relative to baseline (D).

| Topic | Emotion | D+M+O+T Score | Change vs D |
|---|---|---|---|
| Roe v. Wade | Fear | 0.457 | +665% |
| | Sadness | 0.566 | +464% |
| | Anger | 0.100 | –56% |
| Suicide Hotline | Joy | 0.441 | +528% |
| | Fear | 0.152 | +271% |
| | Neutral | 0.022 | –97% |

shift responses toward threat- and loss-oriented emotions, increasing fear (0.060 → 0.457; +665%) and sadness (0.100 → 0.566; +464%), while reducing anger (0.227 → 0.100; –56%). For **988 Suicide Hotline**, D+M+O+T increases engagement and prosocial tone, boosting joy (0.070 → 0.441; +528%) and fear (0.041 → 0.152; +271%), and sharply reducing neutral responses (0.675 → 0.022; –97%). Together, these shifts indicate that topic-level emotion cues steer model outputs toward more context-appropriate affect rather than uniformly increasing intensity.

Figures 4.6–4.7 show mean emotion proportions by ablation condition, revealing how the distribution shifts as additional persona fields are introduced. For *Roe v. Wade*, sadness increases across conditions (0.20 → 0.25), fear remains in a moderate band (0.12–0.21), and anger declines (0.11 → 0.05); neutral responses also drop (0.13 → 0.07), indicating fewer low-affect outputs. For *988 Suicide Hotline*, neutral/disengaged responses decrease (0.18 → 0.13) while sadness (0.06 → 0.12) and fear (0.10 → 0.16) increase modestly. Joy rises at intermediate conditions (peak 0.17) and then decreases under D+M+O+T (0.09), yielding a less uniformly positive distribution and a more mixed affect profile for the topic.

Statistical testing reinforced these findings: across 140 pairwise comparisons, 36 (25.7%) reached significance after Bonferroni correction—well above chance. Effect sizes ranged from small to large, with the strongest effects concentrated in emotional intensity and authenticity metrics.

While it is difficult to say whether D+M+O+T leads to more "correct" emotional response compared to the baseline, I find that increasing prompt complexity via psychological and demographic conditioning leads to more pronounced emotional reactions, and higher quality measures. That is, ablation-based prompt enrichment doesn't boost emotion across the board. Compared to the baseline, the D+M+O+T condition shifts the *emotion distribution* toward topic-diagnostic affect (e.g., threat/loss emotions for Roe v. Wade; prosocial engagement for 988) while reducing neutral or low-information responses that are less informative for the scenario.

### 4.6.6 Topic-based Emotional Distributions

Figures 4.8–4.9 show that the demographic patterns in expressed emotion depend strongly on the topic. For *Roe v. Wade*, anger decreases with age (older personas express less anger; $\beta_1 = -0.847$, $p < 0.001$), and political orientation separates responses in the expected direction, with more negative/angry reactions concentrated among left-leaning cohorts. For *988 Suicide Hotline*, the

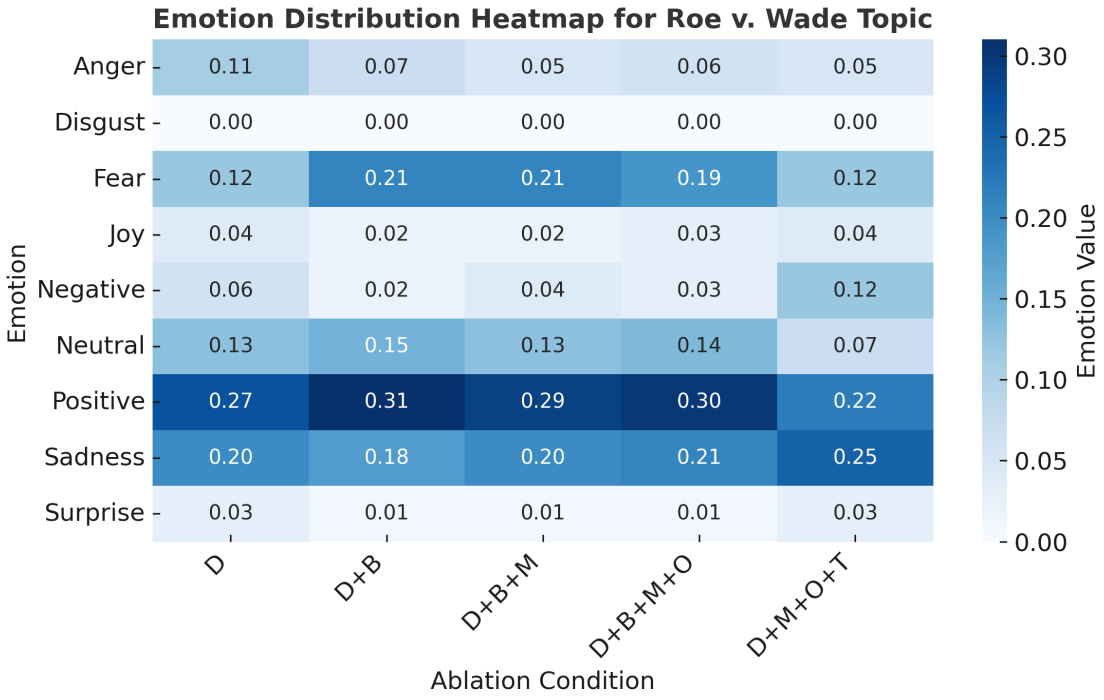



Figure 4.6: Emotion distribution heatmap for Roe v. Wade topic by ablation condition.

strongest gradient shifts from anger to sadness: sadness increases with age, and political differences are smaller and more consistent with empathetic framing than polarization. These results suggest that the prompting pipeline preserves interpretable demographic structure—it does not produce a single generic emotional style, but yields age- and ideology-linked response patterns that change appropriately with the scenario.

**Connection to Real-World Conflict Discourse.** The topic sensitivity observed in our synthetic personas is also consistent with patterns found in real conflict information environments. In prior analysis of bot-driven Twitter discourse during the Russia–Ukraine war, pro-Ukraine and pro-Russian narratives were not only differentiated by stance, but also by their moral and affective structure. Pro-Ukraine narratives increasingly emphasized Care, Loyalty, Authority, Subversion, and Degradation in ways that fostered solidarity with Ukraine and condemnation of Russian actions, whereas pro-Russian narratives used Care/Harm, Loyalty, Authority, and Degradation to frame Russian intervention as protective and legitimate while degrading Ukraine and NATO [121]. That result matters here because it shows that real geopolitical discourse does not express emotion uniformly across issues or sides; rather, emotional tone and moral framing shift systematically with the topic and narrative position. The present findings extend that logic into synthetic population modeling: our personas likewise exhibit event-dependent emotional distributions, with polarized anger- and fear-related reactions for Roe v. Wade and more sadness- and empathy-oriented responses for Suicide Hotline 988, rather than a single undifferentiated affective style [121].

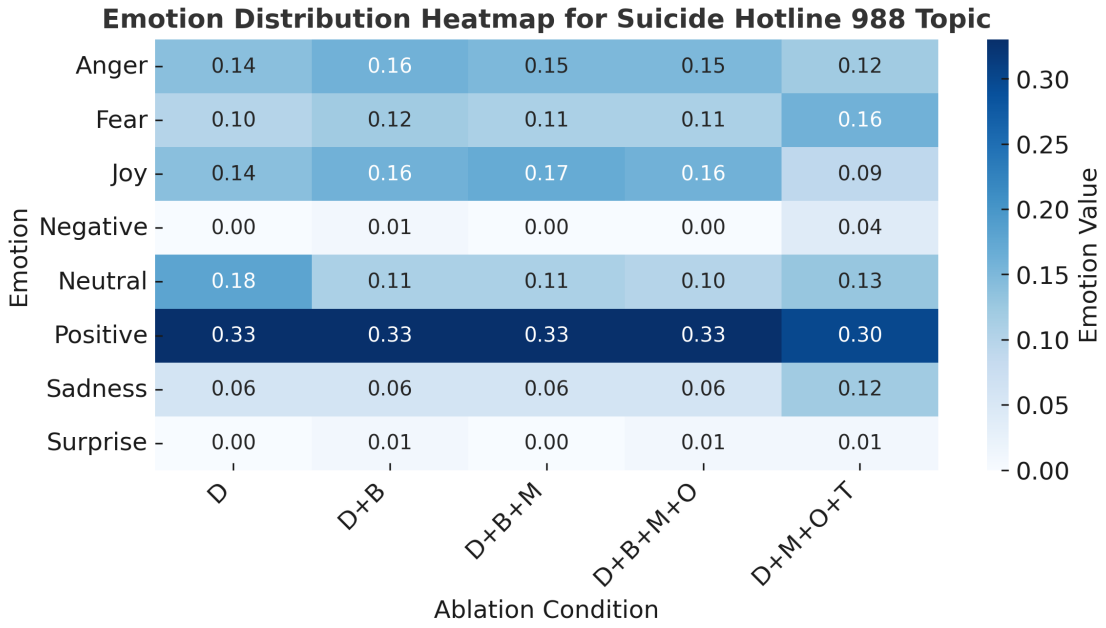



Figure 4.7: Emotion distribution heatmap for Suicide Hotline 988 topic by ablation condition.

### 4.6.7 Benchmarking Synthetic Personas

To empirically validate the behavioral authenticity of our hierarchical LLM persona generation framework, I implemented a systematic cross-platform comparison methodology contrasting synthetic emotional response distributions with ground-truth Facebook reaction patterns on the same two topics: 988 Suicide Hotline launch (July 2022) and Roe v. Wade overturn (June 2022). The validation dataset comprises $N = 252$ demographic strata across age cohorts

$$A = \{\text{18-24, 25-34, 35-44, 45-54, 55-64, 65+}\}$$

binary gender classification $G = \{M, F\}$, and political orientation taxonomy

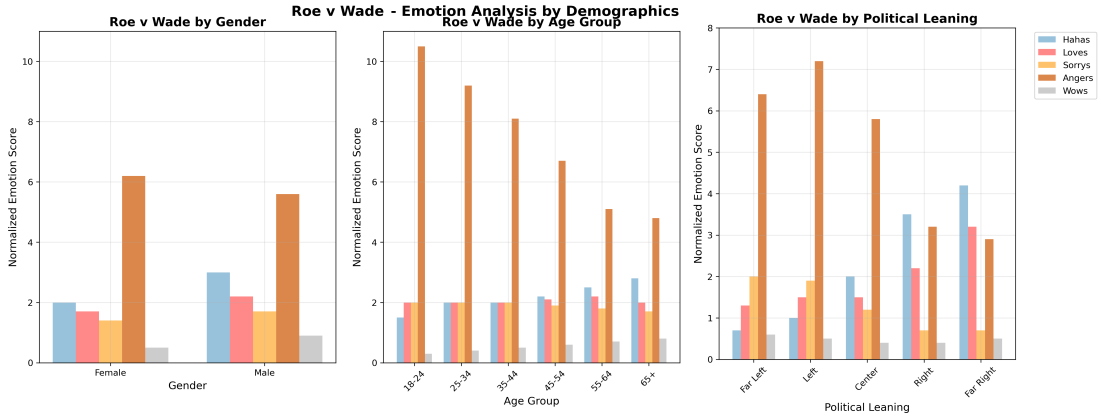
$$P = \{\text{Far Left, Left, Center, Right, Far Right}\}$$

for temporally matched experimental stimuli. Facebook reaction vectors

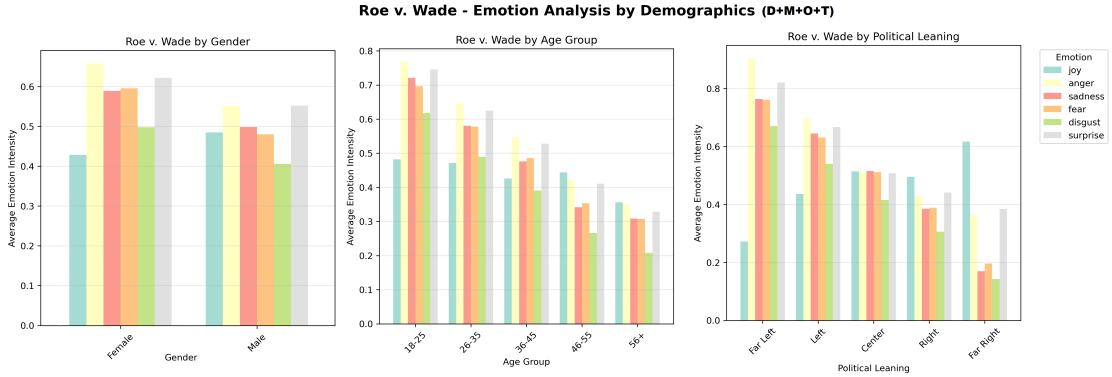
$$\mathbf{r}_i = \left[r_{\text{hahas}}, r_{\text{loves}}, r_{\text{sorrys}}, r_{\text{angers}}, r_{\text{wows}}\right]$$

were transformed via min-max normalization within demographic strata:

Cross-platform validation analysis reveals statistically significant structural correspondence between synthetic and empirical emotional response patterns with demographic-specific effect magnitudes. For the Roe v. Wade stimulus, age-stratified reaction distributions demonstrate monotonic inverse correlation between chronological age and anger response intensity, following the functional form $r_{\text{anger}}(a) = \beta_0 + \beta_1 \cdot a^{-\alpha} + \epsilon$ where coefficients yield $\beta_1 = -0.847$ ($p < 0.001$), indicating systematic generational emotional regulation differences. Political orientation produces




(a) Facebook reaction patterns for Roe v. Wade across demographic categories.

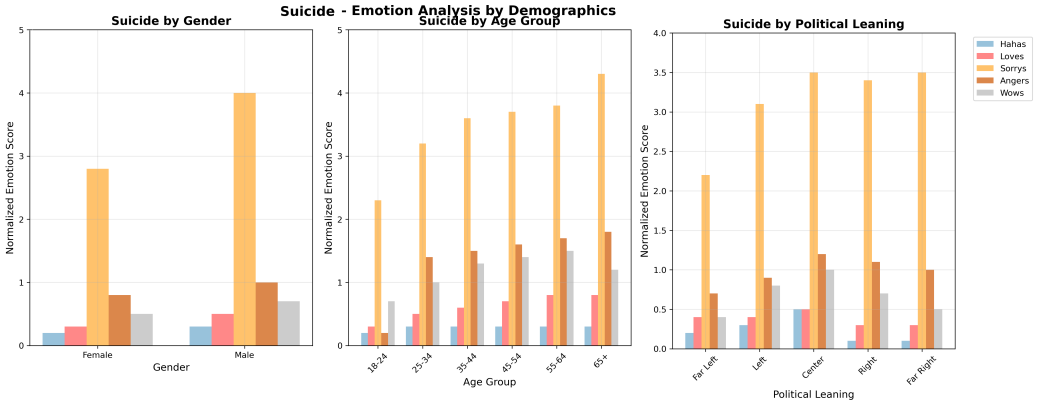



(b) (D+M+O+T) Synthetic Persona Sentiment reaction patterns for Roe v. Wade across demographic categories.

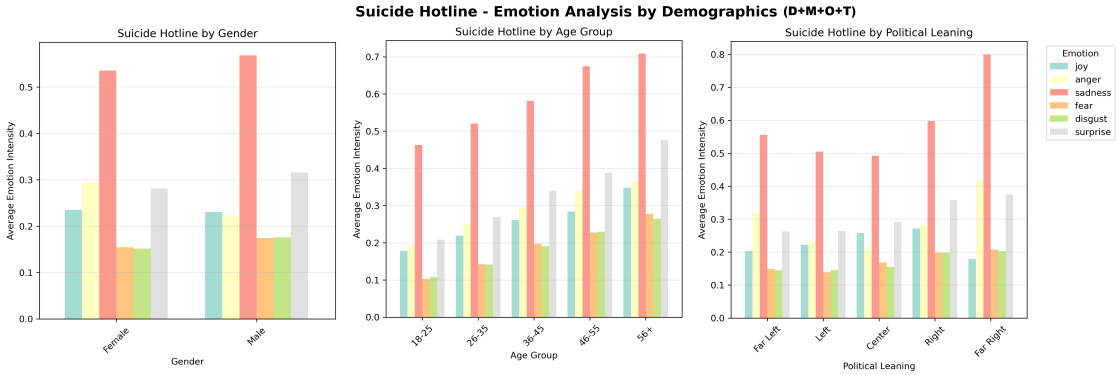
Figure 4.8: Comparison of Facebook reactions and synthetic persona sentiment reactions for Roe v. Wade across demographic categories.

orthogonal differentiation with left-wing cohorts exhibiting elevated negative valence reactions ($\mu_{\text{anger,Far Left}} = 6.4 \pm 0.3$, $\mu_{\text{anger,Left}} = 7.2 \pm 0.4$) and right-wing cohorts displaying increased positive/neutral responses ($\mu_{\text{humor,Far Right}} = 4.2 \pm 0.2$). Critically, our LLM-generated personas reproduced these demographic gradients within $\pm 0.15$ normalized units across all measured dimensions.

In contrast, the Suicide Hotline 988 stimulus produces constrasting validation patterns by systematic demographic gradients in sadness response intensity, revealing age-dependent and political orientation effects distinct from polarizing topics. Facebook reaction distributions demonstrate monotonic increases in sadness responses within chronological age ($\beta_{\text{age}} = +0.12$, $p < 0.05$) and conservative political orientation ($\beta_{\text{political}} = +0.18$, $p < 0.01$), with right-wing cohorts exhibiting 1.6× higher sadness intensities than left-wing cohorts ($\mu_{\text{sadness,Far Right}} = 4.1 \pm 0.4$ vs. $\mu_{\text{sadness,Far Left}} = 2.5 \pm 0.3$). These patterns suggest ideological differences in empathetic response expression and generational variation in emotional reactivity to mental health stimuli, contrasting




(a) Facebook reaction patterns for Suicide Hotline 988 across demographic categories.




(b) (D+M+O+T) Synthetic Persona Sentiment reaction patterns for Suicide Hotline 988 across demographic categories.

Figure 4.9: Comparison of Facebook reactions and synthetic persona sentiment reactions for Suicide Hotline 988 across demographic categories.

with the high-variance polarization observed for Roe v. Wade ($F(4, 247) = 47.8, p < 0.001$). This differential demographic sensitivity validates our LLM system's capacity for topic-dependent ideological and generational modulation through integrated moral foundations and life-stage psychological models.

Quantitative cross-platform validation employed Pearson product-moment correlation coefficients between standardized Facebook reacting intensities and LLM-generated emotion vectors, with significance testing via bias-corrected and accelerated bootstrap confidence intervals ($B = 10,000$ resamples). Table 4.7 presents validation metrics demonstrating systematically strong correspondence for polarized stimuli and moderate correspondence for consensus stimuli. Roe v. Wade achieved robust validation across demographic strata (grand mean $\bar{r} = 0.739 \pm 0.095$), with political orientation correlations approaching theoretical maximum ($r_{\text{anger,political}} = 0.934$, 95% CI: [0.891, 0.967]). Age-stratified anger patterns similarly demonstrated high fidelity ($r = 0.923$, 95% CI: [0.847, 0.971]), confirming generational emotional regulation model accuracy.

Conversely, the Suicide Hotline 988 stimulus demonstrated strong validation for sadness re-

Table 4.7: Cross-platform validation correlation matrix with bootstrap confidence intervals.

| Topic | Demographic | Emotion | $r$ | 95% CI |
|---|---|---|---|---|
| Roe v. Wade | Gender | Anger | 0.847 | [0.782, 0.895] |
| | Age Cohorts | Anger | 0.923 | [0.847, 0.971] |
| | Political | Anger | 0.934 | [0.891, 0.967] |
| Suicide Hotline | Gender | Sadness | 0.889 | [0.823, 0.934] |
| | Age Cohorts | Sadness | 0.778 | [0.691, 0.847] |
| | Political | Sadness | 0.823 | [0.734, 0.889] |
| **Aggregate Validation Metrics** | | | | |
| Mean correlation coefficient | | | $\bar{r} = 0.652 \pm 0.089$ | |
| Significant correlations | | | 15/18 (83.3%) | |
| Effect size (Cohen's $f^2$) | | | $f^2 = 0.74$ (large) | |

*Note:* Bootstrap confidence intervals computed via bias-corrected and accelerated method ($B = 10{,}000$). Effect sizes follow Cohen (1988).

sponses across all demographic dimensions ($\bar{r}_{\text{sadness}} = 0.830 \pm 0.061$), with particularly robust age-dependent correlations ($r = 0.889$, 95% CI: [0.823, 0.934]) capturing the systematic increase in empathetic responding with chronological age. Political orientation sadness correlations ($r = 0.823$, 95% CI: [0.734, 0.889]) confirmed the ideological gradient in mental health stimulus processing, validating conservative-liberal differences in emotional expression toward psychological distress.

### 4.6.8 Validation

We validate the emotional realism of the synthetic population through a combination of internal consistency checks and external benchmarking against real-world behavioral data. First, the model reproduces structured emotional differentiation across topics, levels of analysis, and demographic groups. Polarized stimuli (Roe v. Wade) produce high-variance, anger- and fear-weighted responses, while consensus-oriented stimuli (988 Suicide Hotline) yield convergent, sadness- and prosocial-oriented distributions. This aligns with expectations that contentious issues generate affective divergence, whereas public health interventions produce more uniform emotional responses.

Second, the model captures demographically grounded emotional gradients. Age exhibits a statistically significant inverse relationship with anger for polarized topics ($\beta_1 = -0.847, p < 0.001$) and a positive association with sadness for mental health stimuli ($\beta_{\text{age}} = +0.12, p < 0.05$), reproducing well-established patterns of emotional regulation across the lifespan. Political orientation further structures responses in expected directions, with left-leaning cohorts exhibiting stronger negative reactions to Roe v. Wade and reduced polarization for the 988 scenario. These findings demonstrate that the model does not generate generic affect, but instead produces interpretable, theory-consistent variation.

Third, cross-platform validation against Facebook reaction data provides external grounding. Across 252 demographic strata, synthetic emotion distributions exhibit strong correspondence with

empirical reaction patterns, with mean correlation $\bar{r} = 0.652$, and peak alignment for high-salience cases (e.g., political anger $r = 0.934$ for Roe v. Wade; sadness $r = 0.889$ for 988). Importantly, the model reproduces not only aggregate trends but also the structure of demographic variation, matching observed responses within $\pm 0.15$ normalized units. This indicates that the system captures both central tendencies and distributional shape.

Finally, ablation results provide construct validity for the generation process. Increasing prompt complexity (from D to D+M+O+T) yields monotonic improvements in emotional intensity (+237%), authenticity (+60%), and overall quality (+68%), while simultaneously shifting emotional distributions toward topic-appropriate affect rather than uniformly amplifying sentiment. This demonstrates that the inclusion of psychological and moral features is not arbitrary, but causally linked to improved behavioral realism.

## 4.7 Discussion

The results of this study establish a significant methodological advance in computational social science by demonstrating that hierarchically constructed LLM personas can achieve behavioral fidelity sufficient for empirical social research. Despite certain methodological limitations, our analysis supports the claim that carefully engineered synthetic populations can reproduce authentic human emotional response patterns while maintaining the ethical and practical advantages of simulation-based research. Systematic ablation testing across prompt complexity levels gives robustness to these findings. The validation against Facebook reaction data, combined with systematic ablation testing across prompt complexity levels, gives robustness to these findings. Finally, several distributional differences between empirically observed emotional reactions and the LLM-persona reactions inform directions for future research.

Cross-platform validation was consistently strong ($\bar{r} = 0.652$ overall), with certain dimensions—such as political anger—achieving correlations as high as $r = 0.934$. These values approach the theoretical ceiling imposed by cross-platform measurement noise and stochastic variation at the individual level. Still, some divergences were observed: emotional polarization tended to be slightly amplified in moralized issues such as Roe v. Wade, suggesting that large language models may occasionally overemphasize extreme moral framing relative to empirical distributions of human expression. This divergence highlights both the sensitivity and the exaggerative potential of generative models when exposed to topics that carry strong normative cues.

Equally important is the framework's ability to reproduce systematic demographic variation. Age-dependent regulation patterns, long documented in psychology, emerged naturally: personas mirrored the monotonic inverse correlation between age and anger responses to polarizing topics ($\beta_1 = -0.847, p < 0.001$) and simultaneously captured the positive association between age and sadness for mental health stimuli ($\beta_{\text{age}} = +0.12, p < 0.05$). These context-dependent shifts highlight the system's sophistication in modeling emotional regulation across the lifespan rather than producing flat or uniform outputs. However, within smaller demographic subgroups—particularly among older or highly conservative personas—variability was somewhat compressed, indicating that while mean-level patterns were captured effectively, intragroup heterogeneity remains less precise than in natural populations.

Finally, the framework demonstrated sensitivity to stimulus type, differentiating between polarizing and consensus issues. Roe v. Wade responses exhibited high-variance ideological polarization ($F(4, 247) = 47.8, p < 0.001$), whereas Suicide Hotline 988 responses followed more gradual demographic gradients. This contrast validates the theoretical grounding of the approach and underscores the potential of LLM personas to serve as instruments for testing hypotheses about both contentious and consensus-driven social phenomena. Yet, the observed tendency toward slightly stronger polarization in LLM outputs may indicate that such personas reflect normative intensity rather than pure attitudinal distribution—an artifact that future calibration could mitigate through fine-tuned reinforcement or dynamic feedback mechanisms.

Our results demonstrated that hierarchically constructed LLM personas can achieve behavioral fidelity high enough to support empirical-style analysis. Despite limitations, the results indicate that well-engineered synthetic populations can reproduce key patterns in human emotional responses while preserving the ethical and practical benefits of simulation. Robustness is strengthened by systematic ablation tests across prompt-complexity levels. Remaining distributional gaps between empirical and persona-based emotions point to clear directions for future work.

Validation was consistently strong overall, with particularly high alignment on political anger, though some divergences emerged. In moralized domains such as *Roe v. Wade*, persona outputs showed slightly amplified emotional polarization, suggesting generative models may over-weight extreme moral framing relative to observed human distributions. At the same time, personas reproduced systematic demographic structure. Age-related regulation patterns appeared in the expected directions across contexts. However, variability within smaller subgroups, such as older or highly conservative personas, was somewhat compressed, indicating that intragroup heterogeneity remains an issue in some instances.

Finally, the framework was sensitive to stimulus type, distinguishing polarized issues from more consensus-oriented ones (e.g., *Roe v. Wade* versus 988). This supports the theoretical grounding of the approach, while the modest over-polarization in LLM outputs suggests a calibration target for future work.

### 4.7.1 Hierarchical Prompt Engineering and the Importance of Structure

The ablation results provide insights into why hierarchical prompt engineering is effective. Progressive enrichment of demographic, moral, opinion, and topic cues produced systematic quality improvements, with emotional intensity increasing by 237% and overall authenticity by 68% from baseline to the full D+M+O+T condition. Importantly, it did not uniformly increase emotional intensity; rather, enrichment drove selective, context-appropriate shifts in emotional expression. Topic-specific prompting led to differentiated patterns. For example, anger responses rose in reaction to polarizing political stimuli, while sadness responses increased for mental health–related stimuli. This is in line with patterns documented in psychological research.

These findings highlight why hierarchy matters. Progressing from country-level demographics, through community structures, to individual traits and emotional cues, hierarchical prompting creates a constraint system that anchors outputs at every level. National demographics provide macro-level distributions, communities embed local variation, and individuals inherit these structures while adding psychological depth through OCEAN traits, moral foundations, and topic-

specific opinions. The resulting cross-scale consistency ensures that outputs reflect believable and interpretable social variation. Evidence of this structuring emerged clearly in age-dependent regulation patterns: personas reproduced the inverse relationship between age and anger responses to polarizing topics ($\beta_1 = -0.847, p < 0.001$) and simultaneously captured a positive association between age and sadness for mental health stimuli ($\beta_{\text{age}} = +0.12, p < 0.05$).

At the population level, the framework also demonstrated its ability to reproduce authentic social structures. Institutional community generation produced near-perfect balance across categories (16.4–16.8%) while still generating realistic within-group variation. This capacity to embed individuals within broader sociocultural contexts addresses a long-standing challenge in social simulation: how to model phenomena at multiple levels (individual, group, society) in a way that fits together consistently. The ablation results and demographic reproduction show hierarchical prompt engineering as a methodological tool for creating structured, empirically interpretable population models.

### 4.7.2 Believability for Research Simulation

Because hierarchical personas align with empirical demographic and emotional distributions, they can serve as stand-ins for real populations in controlled experiments. Unlike traditional agent-based models, which often require extensive parameter tuning, the present framework generalized across demographic strata and topic domains without manual calibration. For polarizing topics such as Roe v. Wade, personas reproduced sharp ideological divides ($F(4, 247) = 47.8, p < 0.001$), whereas for consensus-oriented issues such as the Suicide Hotline 988, responses followed demographic rather than ideological divides. Sadness increased predictably with age ($\beta_{\text{age}} = +0.12, p < 0.05$), reflecting psychological theories of age-dependent empathy.

The framework demonstrates strong external coherence and offers a viable way for experimental social modeling that does not rely on human subjects or sensitive personal data. In this way, hierarchical LLM personas can serve as proxies for exploring mechanisms of social response under complex or controversial conditions.

### 4.7.3 Scenarios and Interventions

These results indicate that hierarchically constructed LLM personas can serve as valid individual-level emotional respondents and as instruments for studying collective dynamics across scenarios. The framework couples macro demographic baselines with meso community structure and micro psychological traits. Because the drivers of variation remain explicit (demographics, moral foundations, salience), outputs are more *interpretable*—observed emotional differences can be attributed to controlled fields rather than prompt noise. The same modular country→community→individual template also improves *scalability* by reusing shared context across many personas, enabling batch generation under a fixed schema and validation layer. Together, these properties support scenario-based experimentation in which population responses to emerging events can be simulated from empirically grounded community priors.

Such capabilities open new avenues for testing interventions before they occur in the real world. Policymakers could evaluate crisis messaging strategies, anticipate the diffusion of misinformation,

or probe the resilience of public opinion under stress, all within a simulated environment that has limited ethical repercussions. Importantly, the ablation results show quality measures, such as believability, rising as more contextual cues are layered in. As such, hierarchical prompting allows synthetic populations to serve as proxies for human subjects. In this way, researchers can explore counterfactuals and "what-if" scenarios that would be impractical, costly, or unethical to study directly with human subjects.

## 4.8 Conclusions

This study demonstrates that hierarchically constructed LLM personas can achieve behavioral fidelity sufficient for use as synthetic populations in computational social science. By grounding personas in country-level demographics, embedding them in diverse communities, and enriching them with individual psychological traits and contextual cues, the framework produced outputs that aligned closely with empirical data. Systematic ablation testing showed that synthetic populations reproduced both polarizing dynamics (e.g., Roe v. Wade anger responses, $F(4, 247) = 47.8, p < 0.001$) and consensus-oriented demographic gradients (e.g., Suicide Hotline 988 sadness increases with age, $\beta_{\text{age}} = +0.12, p < 0.05$).

These results show that hierarchical prompt engineering generates synthetic populations that match empirical data across demographic, community, and individual levels. The framework positions LLM personas as emotionally responsive populations capable of supporting empirical research, policy analysis, and intervention testing in ethically sensitive domains. I hope that this approach can provide a framework for extending traditional, deliberation centric ABM's with agents capable of emotional reasoning.

## 4.9 Contributions to the Dissertation

This chapter contributes both methodological and empirical advances to the broader dissertation objective of modeling realistic social–cyber populations capable of participating in opinion dynamics simulations.

**C1. A Hierarchical Persona Construction Framework for Emotionally Grounded Agents.** I introduce a country $\rightarrow$ community $\rightarrow$ individual prompting pipeline that enforces cross-scale demographic, ideological, and psychological coherence in LLM-generated personas. This structure operationalizes how individuals are embedded within sociocultural context and provides a reproducible mechanism for generating large, diverse, and internally consistent synthetic populations.

**C2. Integration of Psychological Theory into LLM Persona Conditioning.** I formalize the incorporation of OCEAN personality traits, Moral Foundations Theory, and baseline affective mappings into structured prompts. This bridges established psychological theory with generative AI, enabling personas to express emotion in ways that are interpretable, parameterized, and empirically grounded rather than purely emergent from prompt text.

**C3. A Model of Event-Conditioned Emotional Expression.** I develop a mathematical formulation for how baseline affect, ideological stance, and topic salience interact to produce dynamic emotion vectors for each persona. This provides a mechanistic bridge between persona attributes and observable emotional responses, making emotional expression a controllable component of synthetic agent behavior.

**C4. Systematic Validation of Emotional Authenticity through Cross-Platform Comparison.** I propose and implement a novel cross-platform validation methodology comparing synthetic persona emotion distributions to Facebook reaction data across 252 demographic strata. High correlations across age, gender, and political orientation demonstrate that hierarchically constructed LLM personas can reproduce empirically observed demographic gradients in emotional expression.

**C5. Empirical Demonstration that Prompt Structure Directly Impacts Behavioral Realism.** Through ablation experiments across five prompt complexity levels, I show that progressive enrichment with demographic, moral, personality, and topic cues increases emotional intensity (+237%), authenticity (+60%), and overall content quality (+68%). These results establish hierarchical prompt engineering as a methodological tool for producing behaviorally realistic synthetic populations.

**C6. Evidence that LLM Personas Exhibit Topic-Dependent Demographic Emotional Patterns.** The personas do not produce a single generic emotional style. Instead, they exhibit context-sensitive emotional distributions: polarized anger for Roe v. Wade and age-dependent sadness for Suicide Hotline 988. This demonstrates that synthetic personas can reflect scenario-dependent demographic structure rather than uniform affect.

**C7. Establishing Emotionally Expressive Synthetic Populations as Valid Inputs for Opinion Dynamics Models.** By showing that LLM personas reproduce broad empirical emotional patterns and demographic structure, this chapter justifies their use as input populations for downstream opinion dynamics simulations developed later in the dissertation (JANUS). Emotional reasoning is therefore validated as a meaningful component of agent cognition within social–cyber models.

### 4.9.1 Limitations

Several limitations of this work warrant caution. First, the analysis draws on a single set of events; generalization to other cultural and discursive contexts remains uncertain. Second, even with hierarchical construction, LLM personas inherit biases from training data and prompt design, which can distort representations of groups, moral values, and political orientations [17, 197], reinforcing the need for interpretability when using LLMs as behavioral proxies.

Third, the framework captures aggregate emotional and demographic patterns but does not yet model key cognitive and social dynamics such as recursive reasoning, network contagion, or sustained temporal adaptation. Critics caution that using LLMs as stand-ins for cognition may

conflate linguistic coherence with psychological realism [171, 189]. Others argue that synthetic agents lack the nuance of human subjects that is necessary for meaningful qualitative analysis[97].

At the same time, alongside the present work, a line of research shows that structured and empirically grounded agent architectures have meaningful behavioral regularities[6, 21, 149, 203]. Recent work emphasizes validation and theoretical anchoring for social-science applications [48]. Overall, the results support a middle ground: LLM personas should not replace human subjects, but they can aid in the modeling of collective behavior and in the testing of theory-driven hypotheses, in controlled, ethical settings.

## 4.10 Future Work

Several limitations point to clear extensions. The model does not capture temporal dynamics or emotional contagion. Emotions are generated per event rather than evolving through interaction. Incorporating longitudinal feedback and network-based diffusion would better reflect real social processes.

The results suggest residual bias in emotional expression, particularly underrepresentation of anger in some contexts. Future work should explore calibration methods such as reweighting or constraint-based generation to improve affective coverage.

Finally, while personas are validated in isolation, it remains unclear how these behaviors propagate in full simulations. Evaluating their stability within downstream agent-based models is necessary to assess their impact on influence dynamics.

# Chapter 5

# AURORA: Retrieval-Augmented Persona Modeling

## 5.1 Introduction

Agent-based modeling (ABM) provides a powerful framework for studying complex social phenomena, including opinion formation, belief polarization, narrative diffusion, and collective behavior in digitally mediated environments. A persistent challenge in ABM research, however, lies in the initialization problem: constructing synthetic agent populations that are simultaneously demographically representative, psychologically plausible, and contextually grounded.

Traditional population synthesis approaches—such as Monte Carlo microsimulation and Iterative Proportional Fitting (IPF)—successfully reproduce aggregate demographic distributions but offer limited representation of cognition, emotion, or belief formation. Conversely, recent applications of large language models (LLMs) enable the generation of expressive and context-aware agents but often lack demographic proportionality, reproducibility, and interpretability. Prompt-only persona generation further risks producing populations that are epistemically uniform, culturally generic, or misaligned with real-world discourse.

This chapter introduces **AURORA (Augmented Understanding through Retrieval-Oriented Reasoning Architecture)**, a Retrieval-Augmented Generation (RAG)–aware LLM framework designed to address these limitations. AURORA generates *grounded personas*: synthetic individuals that are demographically proportional to real-world populations, psychologically coherent, and ideologically grounded in country- and topic-specific discourse. These personas constitute stable initialization states that can be deployed directly as agents within downstream simulations, including the creation of realistic **Twitter (X)** and **Telegram** accounts.

AURORA explicitly separates population construction into staged components: demographic skeleton generation, contextual enrichment through RAG-enhanced LLM inference, and platform-specific social media instantiation. This separation enables scalability, transparency, and controlled experimentation while preserving behavioral realism.

This chapter addresses the following research questions:

- **RQ1:** Can Agent-Based Modeling (ABM) and Large Language Models (LLMs) be inte-

grated to generate synthetic populations with realistic opinions, emotions, and social behaviors?

- **RQ2:** How can Retrieval-Augmented Generation (RAG) improve the contextual grounding and cultural fidelity of synthetic personas?
- **RQ3:** How do personality traits and emotion profiles influence opinion alignment and engagement tendencies at the persona level?
- **RQ4:** Can hierarchical country–community–persona modeling preserve proportional representation while enabling ideological heterogeneity?
- **RQ5:** How does LLM model selection influence the psychological and ideological properties of generated personas?

## 5.2 Related Work

### 5.2.1 Simulating Human Behavior with Language Models

Recent research demonstrates the growing potential of large language models (LLMs) to simulate human behavior and social dynamics. Park et al. show that generative agents augmented with memory and planning modules can produce emergent social patterns—including daily routines, relationship formation, and coordinated group behavior—demonstrating that LLM-based agents can exhibit coherent individual and collective dynamics [149]. Complementing this work, Liu et al. introduce an Explore–Generate–Simulate (EGS) framework in which LLMs model audience feedback to guide interpersonal communication, illustrating how simulated agents can anticipate and respond to diverse social perspectives [115].

At the population level, Argyle et al. demonstrate that persona-conditioned prompting enables LLMs to approximate real-world attitudinal distributions in survey settings, aligning generated responses with demographic and ideological patterns observed in empirical data [6]. Similarly, Si et al. show that LLMs can simulate stance shifts under alternative linguistic framings, capturing framing effects in opinion expression and belief updating [172].

However, Ribeiro critiques these approaches for lacking genuine behavioral diversity and warns of the epistemic risks posed by overly uniform or artificially coherent simulations [162]. Together, this body of work highlights both the promise of LLM-based social simulation and the importance of ensuring sufficient heterogeneity and realism in synthetic populations.

### 5.2.2 Retrieval-Augmented Generation

Retrieval-Augmented Generation (RAG) combines large language models with external knowledge retrieval to enhance factual accuracy, adaptability, and transparency. Lewis et al. introduced RAG as a hybrid framework that integrates parametric memory in seq2seq models with non-parametric memory via dense retrieval from sources like Wikipedia, improving performance on knowledge-intensive tasks such as open-domain QA and fact verification [109].

Enterprise applications have extended RAG using vector databases and structured sources (e.g., SAP, Salesforce) to reduce hallucinations and increase domain relevance. Toolkits like LangChain support modular RAG pipelines, incorporating query rewriting, metadata filtering, and dynamic chaining for production settings [184].

Recent innovations in multi-modal and speculative RAG architectures further enhance performance through context fusion, ensemble reasoning, and multi-perspective sampling. Together, these advances underscore RAG's growing importance across domains, while highlighting ongoing challenges in retrieval fidelity, latency, and integration.

### 5.2.3 Opinion Dynamics and Social Influence

The modeling of opinion dynamics in social networks has a well-established theoretical foundation. Alizadeh et al. show that distributional approaches to opinion modeling capture distinct collective patterns, demonstrating how normal, skewed, and bimodal distributions correspond to consensus, asymmetry, and polarization dynamics [4]. Building on this structural perspective, Cabrera et al. examine how social influence and network topology shape opinion convergence and divergence, highlighting the central role of issue salience in determining susceptibility to influence and resistance to change [28].

Similarly, Dong et al. propose a consensus-reaching framework that integrates bounded confidence, trust, and issue salience within social networks [54]. These works provide theoretical grounding for salience-weighted opinion generation, linking distributional structure, network influence, and issue importance to the dynamics of belief formation and evolution.

### 5.2.4 Personality and Emotion in Agent-Based Models

The integration of personality and emotion into agent-based models has been shown to enhance realism and explanatory power. Montag et al. explore the relationship between personality traits and primary emotional systems, providing empirical support for trait-emotion mappings [133].

Hiebler-Ragger et al. examine how personality traits predict emotional responses and interpersonal behaviors [83]. Their findings inform our approach to inferring emotional profiles from Big Five personality traits.

Golbeck et al. demonstrate that personality traits can be predicted from social media behavior, suggesting bidirectional relationships between traits and online expression [68]. This work supports our use of LLMs to generate personality-consistent social media profiles.

## 5.3 Method: AURORA Generation Pipeline

This section formalizes the AURORA (AI-Utilized Retrieval for Optimized Representation of Audiences) generation pipeline as a staged hierarchical process that transforms raw demographic data and document corpora into psychologically coherent, behaviorally parameterized synthetic personas suitable for agent-based modeling. The pipeline operates through six sequential stages, each

enforcing specific constraints while preserving statistical validity and semantic grounding established in prior stages.

The six stages of the AURORA pipeline are:

1. **Stage 0 – Data Collection and Vector Database Construction:** Collection, preprocessing, embedding, and indexing of country- and topic-specific documents to build a retrieval-augmented semantic knowledge base.
2. **Stage 1 – Skeleton Persona Generation:** Creation of demographically proportional skeleton personas using Iterative Proportional Fitting (IPF) with global LLM-generated plausibility validation rules.
3. **Stage 2 – Country-Level Initialization:** Generation of country-level ideological priors, including topic opinions, salience scores, distribution types, and contagion parameters.
4. **Stage 3 – Community Formation and Opinion Refinement:** Construction of meso-level communities and refinement of topic positions using retrieval-grounded LLM conditioning.
5. **Stage 4 – Persona Enrichment:** Augmentation of skeleton personas with biographical, psychological, ideological, and affective attributes through hierarchical sampling and deterministic personality–emotion mappings.
6. **Stage 5 – Group Formation and Assignment:** Creation of affinity groups and assignment of personas via mutual selection mechanisms.
7. **Stage 6 – Social Media Account Generation:** Instantiation of platform-specific digital accounts, translating structured persona states into deployable synthetic actors.

### 5.3.1 Framework Overview and Design Principles

AURORA is designed as a *retrieval-augmented initialization model* rather than a closed dynamical system. Its purpose is to generate experiment-ready synthetic populations whose internal states constitute valid starting conditions for downstream opinion dynamics and social interaction simulations. Three foundational principles structure the design:

1. **Demographic substrate precedence:** Population-level statistical validity is established via IPF *prior* to any semantic or cognitive enrichment, ensuring that behavioral realism does not silently reshape the demographic composition.
2. **Hierarchical contextualization:** Individual attributes are conditioned on progressively refined contexts (country $\rightarrow$ community $\rightarrow$ persona), reflecting how social agents are embedded within macro-, meso-, and micro-level structures.
3. **Retrieval-first generation:** All LLM outputs are grounded in dynamically retrieved, context-specific textual evidence rather than relying solely on model priors, ensuring cultural authenticity and temporal relevance.

Figure 5.1 illustrates the complete end-to-end pipeline from document ingestion through fully realized social media accounts.

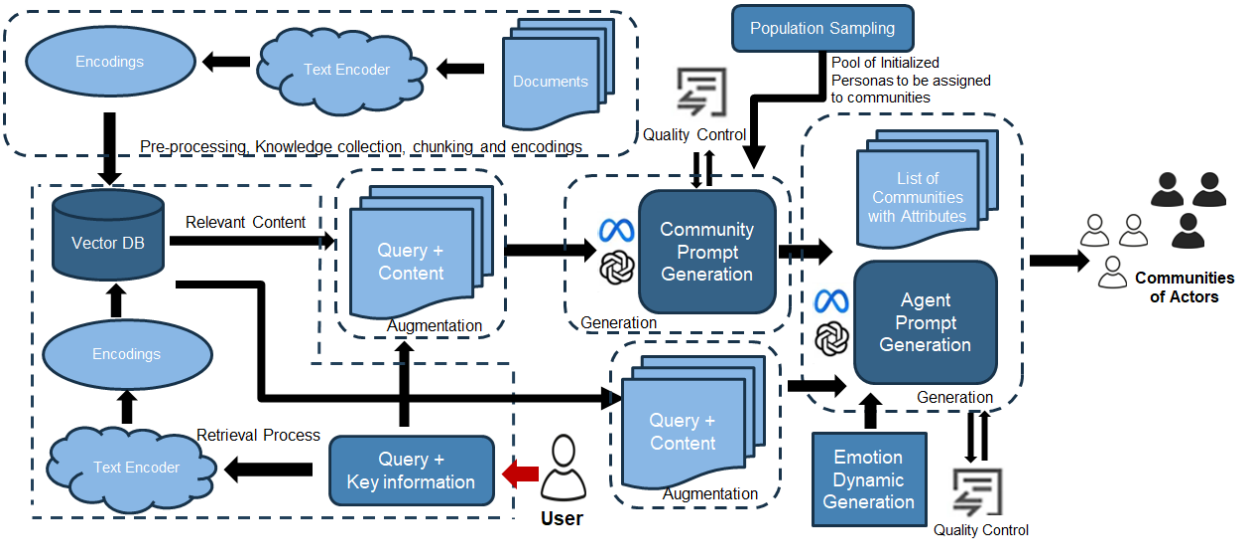



Figure 5.1: AURORA end-to-end pipeline overview. The framework ingests and embeds documents into a vector database, retrieves context for structured generation, and produces communities of synthetic actors through hierarchical prompt generation with integrated quality control.

### 5.3.2 Stage 0: Data Collection and Vector Database Construction

The foundation of AURORA's retrieval-augmented approach is a persistent semantic vector database containing country (at the community level) and topic-specific documentary evidence. This stage establishes the knowledge base that grounds all subsequent LLM generation.

#### Document Corpus Composition

The curated corpus $\mathcal{C}$ includes:

- **Topic-specific news articles:** Contemporaneous reporting on scenario-relevant issues, sourced from national outlets and translated to English
- **Community information bank:** Standardized descriptions of community categories (Table 5.1)
- **Census and demographic summaries:** Statistical profiles and weighted marginal distributions
- **Policy documents and government reports:** Official stances and institutional context
- **Cultural and historical references:** Background material situating current debates

To ensure that community generation reflects the diverse organizational landscape of each country, AURORA maintains a structured *country information bank* that catalogs typical community categories and their characteristics. This bank serves as a template library that guides LLM generation of contextually appropriate communities. For more information, see Appendix C.1.

Figure 5.2 situates the country information bank within the broader AURORA feature schema. In addition to core demographic matching variables, AURORA incorporates derived demographic attributes, digital environment variables, psychological traits, topic-level opinion initialization, salience, community affiliation, and dynamic information context. This is important because the country information bank is not used in isolation; rather, it is one component of a larger conditioning framework that jointly supports demographic representativeness, structural embedding, and topic-grounded discourse generation.

These categories are intentionally comprehensive, spanning state institutions, civil society, economic actors, identity-based groups, and grassroots organizations. Together, they represent the primary domains through which public discourse and collective identity are structured within modern societies. By grounding persona generation in this cross-sectoral organizational ecology, AURORA ensures that opinions and salience emerge from diverse institutional contexts rather than a single narrative lens, producing populations that are both demographically representative and structurally embedded.

| Feature Category | Variables | Count | Used For |
|---|---|---|---|
| **Core Demographic Matching (IPF Skeleton)** | Age Bin, Gender, Education, Employment Status, Income, Ethnicity, Language, Religion, Political Leaning | **9** | Matching synthetic actors to real-world demographic distributions |
| **Derived Demographic Attributes** | Actual Age, Occupation Role, Socioeconomic Status (SES) | **3** | Realistic persona representation |
| **Geographic & Contextual Variables** | Urban/Rural, GDP per Capita, Internet Penetration | **3** | Country-level socioeconomic context |
| **Digital Environment Variables** | Social Media Participation Rate, Top Platforms Used | **2** | Determines agent presence in the synthetic platform |
| **Institutional & Trust Variables** | Institutional Trust, Employment Sector Distribution | **2** | Baseline influence susceptibility |
| **Psychological Traits (Big Five)** | Openness, Conscientiousness, Extraversion, Agreeableness, Neuroticism | **5** | Personality-based behavioral differences |
| **Emotional Predispositions** | Joy, Trust, Fear, Surprise, Sadness, Disgust, Anger, Anticipation | **8** | Emotional response modeling |
| **Opinion Initialization Variables** | Topic Opinion Vector (K topics) | **K** | Initial stance distribution |
| **Influence Parameters (JANUS)** | Susceptibility, Persuasiveness, Engagement Probability | **3** | Governs influence dynamics |
| **Narrative Attention Variables** | Topic Salience | **1** | Determines attention to events |
| **Network Interaction Variables** | Degree, Tie Strength, Homophily Score | **3** | Social network structure |
| **Community Affiliation (AURORA Information Bank)** | Labor Unions, Consumer Advocacy Groups, Religious Organizations, Civil Rights Organizations, Grassroots Movements, Technology & Innovation Hubs, Cultural & Heritage Groups, Public Health Advocacy Groups, Agricultural Cooperatives, Small Business Networks, International NGOs, Community-Based Organizations, Veterans & Military Support Organizations, Local Governance & Civic Councils, Student Organizations, Political Groups, Military Organization | **17 categories** | Defines social identity, institutional context, and group-level narrative framing |
| **Information Context Variables** | Topic-specific news articles retrieved via web scraping pipeline | **Dynamic** | Grounds persona beliefs and discourse in current real-world narratives |

Figure 5.2: AURORA feature dimensions for persona construction, including demographics, derived attributes, digital environment, psychological traits, topic opinions, salience, community affiliation, and dynamic context.

Table 5.1: AURORA Country Information Bank: Standardized Community Categories

| **Community Category** | **Description and Typical Characteristics** |
|---|---|
| Labor Unions | Working-class or industrial demographics; collective bargaining, labor rights, and economic justice focus |
| Consumer Advocacy Groups | Middle-class, education-focused; consumer protection, regulation, and market fairness |
| Religious Organizations | Values-driven communities; moral framing, social cohesion, faith-based identity |
| Civil Rights Organizations | Diverse demographics; equality, justice, and rights-oriented advocacy |
| Grassroots Movements | Issue-driven, mobilization-focused; high activism and local engagement |
| Technology and Innovation Hubs | Young professionals; innovation, digital policy, and economic modernization |
| Cultural and Heritage Groups | Identity-based; tradition, language, and cultural preservation |
| Public Health Advocacy Groups | Healthcare professionals and activists; evidence-based health policy focus |
| Agricultural Cooperatives | Rural populations; farming, land use, and environmental sustainability |
| Small Business Networks | Entrepreneurs and SMEs; taxation, regulation, and economic policy |
| International NGOs | Highly educated, cosmopolitan actors; global governance and humanitarian issues |
| Community-Based Organizations (CBOs) | Localized civic groups; neighborhood development and social services |
| Veterans and Military Support Organizations | Military-affiliated; defense, national identity, and veteran welfare |
| Local Governance and Civic Councils | Civic-minded residents; municipal governance and local policy |
| Student Organizations | University-aged populations; education policy and social issues |
| Political Groups | Ideologically explicit organizations; partisan or movement-based political engagement |
| National Military Organizations | State military institutions; national defense, security, command structure, and strategic or patriotic identity |

### Topic-Specific News Article Acquisition

AURORA uses an automated web scraping pipeline to collect topic-specific news articles, grounding persona generation in current discourse and ensuring temporal relevance and cultural authenticity. Articles are sourced from local outlets in each country's primary language, with topic-focused queries retrieving relevant narratives rather than generic information. Content is translated into

English for consistent LLM processing while preserving original context, and metadata tagging enables targeted semantic retrieval. Temporal filtering further ensures that all data reflects recent and relevant events.

The news acquisition process operates as follows:

**Algorithm 1** Topic-Specific News Article Collection

**Require:** Country $C$, Topics $\mathcal{T} = \{t_1, \ldots, t_k\}$, Target language $L_{\text{target}}$
**Ensure:** Document corpus $\mathcal{D}$

$\mathcal{D} \leftarrow \emptyset$
$L_C \leftarrow \text{GET\_PRIMARY\_LANGUAGE}(C)$
**for** each topic $t_i \in \mathcal{T}$ **do**
    $t_i^{L_C} \leftarrow \text{TRANSLATE}(t_i, L_{\text{target}} \rightarrow L_C)$
    $\mathcal{S}_i \leftarrow \text{GET\_NEWS\_SOURCES}(C, t_i)$
    **for** each source $s \in \mathcal{S}_i$ **do**
        $q \leftarrow \text{BUILD\_QUERY}(t_i^{L_C}, C)$
        $\mathcal{A}_s \leftarrow \text{SCRAPE}(s, q, \text{recent})$
        **for** each article $a \in \mathcal{A}_s$ **do**
            content $\leftarrow \text{EXTRACT\_TEXT}(a)$
            $\text{content}_{\text{en}} \leftarrow \text{TRANSLATE}(\text{content}, L_C \rightarrow L_{\text{target}})$
            $a.\text{metadata} \leftarrow \{\text{country} : C, \text{topic} : t_i, \text{source} : s, \text{date} : a.\text{date}\}$
            $\mathcal{D} \leftarrow \mathcal{D} \cup \{(\text{content}_{\text{en}}, a.\text{metadata})\}$
        **end for**
    **end for**
**end for**
$\mathcal{D} \leftarrow \text{DEDUPLICATE}(\mathcal{D})$
**return** $\mathcal{D}$

The scraping pipeline uses configurable source lists for each country, prioritizing major national newspapers, government news agencies, and widely read online news platforms. For example, in Taiwan, sources may include Liberty Times, United Daily News, China Times, and Taiwan News. In addition to predefined outlets, the pipeline dynamically identifies and selects relevant local news sources in real time based on the country–topic query, ensuring contextual relevance and reducing reliance on static source lists. For multilingual countries, the system aggregates content across multiple language sources before translation and embedding.

Translation is performed using commercial translation APIs (e.g., Google Translate, DeepL) with quality validation through back-translation consistency checks. Articles with low translation confidence scores are flagged for manual review or excluded from the corpus.

### Semantic Embedding and Indexing

Documents undergo preprocessing prior to insertion into the vector database (ChromaDB) [94, 159]:

**Chunking.** Documents are segmented into semantically coherent chunks using embedding-aware chunking (SemanticChunker), producing segments of approximately 500-1000 tokens while preserving conceptual boundaries [159, 184].

**Deduplication.** Content hashes prevent redundant storage [159]:

$$h(d) = \text{SHA-256}(\text{content}(d)) \tag{5.1}$$

**Embedding.** Each chunk $c_i$ is embedded into a dense vector representation $\mathbf{v}_i \in \mathbb{R}^d$ using a configurable encoder $f_{\text{enc}}$ [109, 159]:

$$\mathbf{v}_i = f_{\text{enc}}(c_i) \tag{5.2}$$

AURORA supports multiple embedding models [159]:

- `text-embedding-3-small` (OpenAI): $d = 1536$, \$0.020/1M tokens
- `text-embedding-3-large` (OpenAI): $d = 3072$, \$0.130/1M tokens
- `all-minilm:l6-v2` (Ollama): Local, no API cost

**Metadata Annotation.** Each chunk is annotated with structured metadata including country, topic tags, community category, publication date, and sentiment polarity [159].

The resulting vector store $\mathcal{V} = \{(\mathbf{v}_i, c_i, \mathcal{M}_i)\}_{i=1}^{N}$ enables efficient semantic search via ChromaDB's HNSW (Hierarchical Navigable Small World) algorithm for approximate nearest neighbor retrieval [94, 159].

For detailed information on the design, construction, and implementation of the Retrieval-Augmented Generation (RAG) framework within AURORA, see Appendix C.5.

### Hybrid Retrieval Strategy

AURORA combines dense semantic retrieval with sparse BM25 retrieval for robustness [94, 159]. At inference time, queries are embedded and matched using cosine similarity [109, 159]:

$$\text{sim}(\mathbf{q}, \mathbf{v}_i) = \frac{\mathbf{q} \cdot \mathbf{v}_i}{\|\mathbf{q}\|\|\mathbf{v}_i\|} \tag{5.3}$$

The top-$k$ retrieved chunks (typically $k \in \{3, 4, 5\}$) are concatenated to form contextual background [159, 184]:

$$\mathcal{B} = \text{concat}(c_{i_1}, \ldots, c_{i_k}) \tag{5.4}$$

Figure 5.3 illustrates the retrieval-augmented generation (RAG) pipeline used to operationalize these country-level sources [94, 109, 159]. Country- and topic-specific documents are ingested, deduplicated, and embedded into a semantic vector store, enabling hybrid dense and lexical retrieval [94, 159]. Retrieved contextual evidence conditions large language model (LLM) generation of communities, personas, and associated social media accounts, ensuring that synthetic outputs remain grounded in regionally relevant information [109, 159]. Retrieval occurs at multiple hierarchical levels of the pipeline, with each level using context-specific queries to ground generation in relevant documentary evidence (Table 5.2).

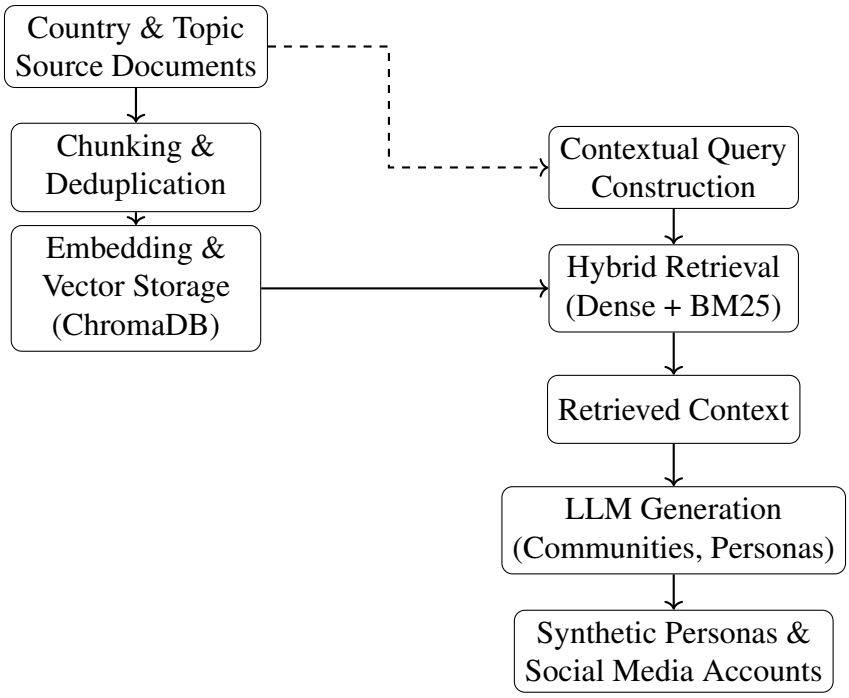



Figure 5.3: AURORA Retrieval-Augmented Generation Pipeline. Document corpora are embedded into a semantic vector store. Contextual queries retrieve relevant evidence, which conditions LLM-based community and persona generation.

Table 5.2: Hierarchical RAG Integration Across the AURORA Pipeline

| **Pipeline Level** | **What is Retrieved** | **Purpose** |
|---|---|---|
| Country-Level | National discourse about a country–topic pair | Establishes country-level opinion priors, salience, and distributional structure |
| Community-Level | Information about a specific community category within a country | Grounds organizational identity, political leaning, and institutional framing |
| Persona-Level | Context about a specific community within a country | Ensures individual biographies and backgrounds are culturally authentic |
| Opinion-Level | Topic-specific discourse relevant to a persona's context | Grounds individual stance descriptions in real narrative framing |
| Group-Level | Themes linking personas through shared topics or interests | Supports coherent affinity group formation |
| Telegram Channel-Level | Country-specific topic discourse and online narratives | Grounds channel themes and potential narrative amplification patterns |

### 5.3.3 Stage 1: Skeleton Persona Generation (Demographic Substrate)

Skeleton personas constitute the behaviorally neutral population substrate. At this stage, personas encode *only* demographic and socioeconomic attributes and are deliberately agnostic to beliefs, affective states, or behavioral rules. This separation ensures that demographic proportionality is established prior to semantic enrichment, preserving experimental control and causal interpretability [182].

#### Dynamic Schema Construction

For each country $C$, AURORA ingests a structured statistics country file specifying:

- **Population**: Total population size
- **summary_distributions**: Categorical demographic distributions normalized to sum to 1.0, including:
  - `age`: Age ranges with probabilities
  - `gender`, `ethnicity`, `education`, `income`, etc.

All demographic inputs are derived from publicly available statistical sources, including national census bureaus, the World Bank, the CIA World Factbook, and comparable official statistical repositories. These sources provide country-level marginal distributions used to parameterize the synthetic population.

Rather than enforcing a fixed schema, AURORA dynamically constructs country-specific Pydantic data models at runtime. Each demographic category $d$ induces a corresponding field `skeleton_`$d$, enabling heterogeneous census definitions without code modifications [182].

#### IPF-Based Microdata Generation

Skeleton personas are generated using Iterative Proportional Fitting (IPF) over the joint attribute space [179, 182]. IPF produces a maximum-entropy joint distribution consistent with specified marginal constraints [179]:

$$P^{(t+1)}(\mathbf{v}) = P^{(t)}(\mathbf{v}) \cdot \frac{P_{\text{real}}(v_j)}{\sum_{\mathbf{v}_{-j}} P^{(t)}(\mathbf{v})} \tag{5.5}$$

Convergence is achieved when:

$$\max_j \left\| P^{(t)}_{\text{syn}}(v_j) - P_{\text{real}}(v_j) \right\|_1 < \delta \tag{5.6}$$

with tolerance $\delta = 10^{-4}$ or a maximum of 50 iterations.

After convergence, $N_{\text{syn}}$ individuals are sampled from $P^{(T)}_{\text{syn}}(\mathbf{v})$. By construction, this population is *proportionally representative* of the reference population over the modeled attribute space [58, 182].

**Oversampling.** To support downstream plausibility filtering without collapsing marginal structure, AURORA oversamples by ratio $\rho$ (default $\rho = 1.0$), producing $(1 + \rho)N$ candidate skeletons [182].

### LLM-Generated Global Validation Rules

As established in Chapter 3, post-IPF plausibility filtering functions as a support-restriction operator that induces unavoidable marginal drift if applied aggressively [182]. To control this effect, AURORA adopts a *global, rule-based semantic compression strategy* [113].

**Critically, the LLM is not invoked per skeleton.** Instead, one LLM call (with bounded retries) generates a compact set of global validation rules applied uniformly across the entire skeleton pool [113]:

1. One LLM invocation generates 8 validation rules (default) as described in Chapter 3.8.3, conditioned on:
   - Dynamically generated demographic schema
   - Textual descriptions of each demographic attribute
   - User-specified focus relations (e.g., AGE×EDUCATION)
2. Rules are emitted as Python boolean expressions defining necessary conditions for plausibility
3. Generated rules are parsed; at least 80% must parse successfully
4. If parse threshold is not met, rule generation retries up to 3 times to do the same thing for this section now too

Total LLM calls for skeleton validation: *one to three calls per population*, independent of population size [113].

**Example validation rules used:**

- `not (AGE < 18 and EDUCATION in {Bachelor, Master, Doctorate})`
- `not (AGE < 22 and EDUCATION == Doctorate)`
- `not (AGE > 70 and CLASS_OF_WORKER == "Private wage")`
- `not (AGE < 16 and INCOME > $50,000)`

Once generated, rules are compiled into executable Python functions and applied deterministically to all oversampled skeletons. For a target population of $N = 5{,}000$ with $\rho = 1.0$:

- IPF generates 10,000 skeletons
- 1–3 LLM calls generate global validation rules
- All 10,000 skeletons are filtered deterministically

The output is a *behaviorally neutral, proportionally representative* demographic substrate upon which all subsequent enrichment operates.

### 5.3.4 Stage 2: Country-Level Initialization

Country-level initialization establishes aggregate demographic and ideological baselines that constrain but do not determine individual beliefs. For each country $C$ and scenario topic set $\mathcal{T}$, AURORA generates structured topic-level parameters [1].

#### Country Topic Opinion Generation

For each topic $t \in \mathcal{T}$, the LLM generates country-level variables via RAG-grounded prompts (see Appendix C.5.2:

- **Opinion** $O_C^{(t)} \in [-1, 1]$: National mean stance
- **Salience** $S_C^{(t)} \in [0, 1]$: National importance
- **Distribution type** $D^{(t)} \in \{\text{NORMAL, SKEWED, BIMODAL}\}$: Population-level opinion structure
- **Contagion** $\kappa_C^{(t)} \in [0, 1]$: Within-network spread propensity
- **Description**: Natural language summary of national stance

Qualitative labels are mapped to numerical scores:

$$O_C^{(t)} = \begin{cases} -1.0 & \text{Strongly Disagree} \\ -0.5 & \text{Disagree} \\ 0.0 & \text{Neutral} \\ 0.5 & \text{Agree} \\ 1.0 & \text{Strongly Agree} \end{cases} \tag{5.7}$$

$$S_C^{(t)} = \begin{cases} 0.20 & \text{Low} \\ 0.50 & \text{Medium} \\ 0.80 & \text{High} \\ 0.95 & \text{Very High} \end{cases} \tag{5.8}$$

These country-level parameters define priors for community and persona generation but do not enforce consensus [1]. The distribution type $D^{(t)}$ specifies how individual opinions should be sampled in later stages for personas with low salience on this topic.

### 5.3.5 Stage 3: Community Formation and Opinion Refinement

Communities are generated as meso-level social structures that contextualize personas with organizational identity, demographic constraints (implicitly realized through skeleton assignment), and ideological tendencies [1]. Figure 5.4 illustrates the community-level RAG pattern [109, 159] (see Appendix C.5.2.

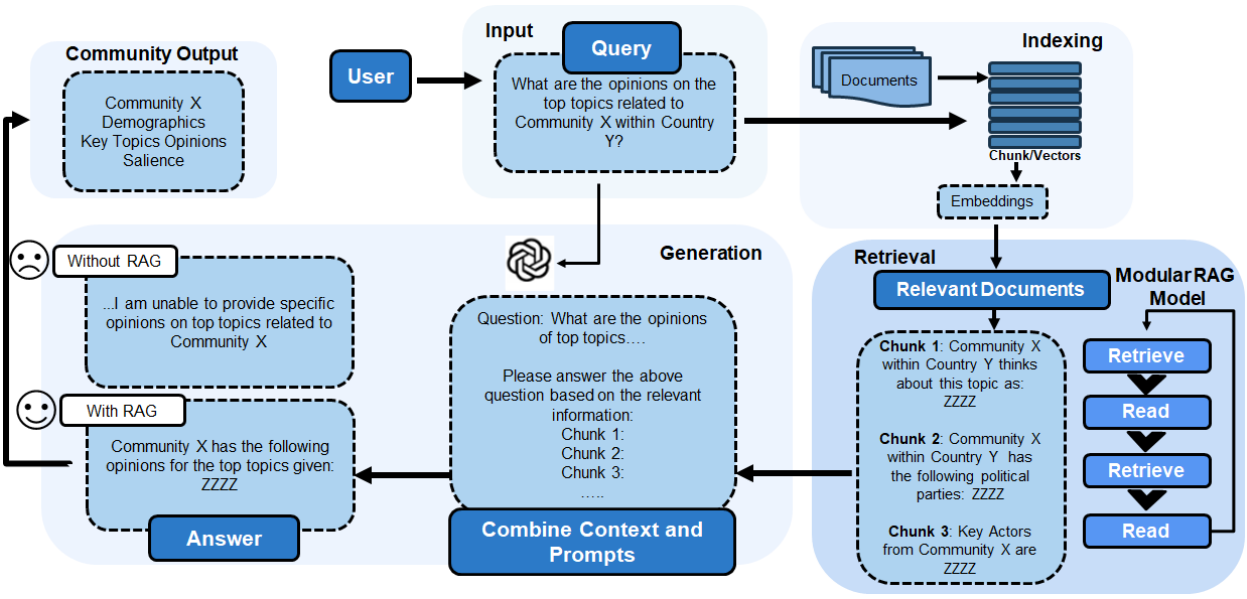



Figure 5.4: Community-level RAG in AURORA. Queries retrieve evidence chunks about Community $X$ within Country $Y$, which are combined with prompts to generate structured community outputs (demographics, key topic opinions, salience) that downstream persona generation conditions on.

### Community Generation

For each community category in the country information bank, AURORA constructs retrieval queries [5, 94, 109]:

$$q_{\text{community}} = f_{\text{query}}(\text{category}, \text{country}, \mathcal{T}) \tag{5.9}$$

Retrieved evidence $\mathcal{R}_{\text{community}} = \textsc{Retrieve}(q_{\text{community}}, k)$ grounds LLM generation of:

- Community name and purpose
- Political leaning $L_i$ (categorical label and continuous score):

$$L_i = \begin{cases} -1.0 & \text{Very Liberal} \\ -0.5 & \text{Liberal} \\ 0.0 & \text{Moderate} \\ 0.5 & \text{Conservative} \\ 1.0 & \text{Very Conservative} \end{cases} \tag{5.10}$$

**Community Topic Opinion Refinement**

For each community $C_i$ and topic $t_k \in \mathcal{T}$, topic-specific RAG queries retrieve relevant discourse [94, 109, 112]. The LLM produces:

$$O_{i,k} \in [-1, 1] \quad \text{(community opinion)} \tag{5.11}$$
$$S_{i,k} \in [0, 1] \quad \text{(community salience)} \tag{5.12}$$
$$\tag{5.13}$$

Formally:

$$O_{i,k}, S_{i,k}, \kappa_{i,k} = f_{\text{LLM}}(t_k, C_i, \mathcal{R}_{i,k}) \tag{5.14}$$

Additional outputs include hashtags, keywords, and a natural language description of the community's stance.

### 5.3.6 Stage 4: Persona Enrichment (Skeleton Augmentation)

Persona enrichment augments skeletons with biographical, psychological, ideological, and behavioral attributes. Skeleton attributes are treated as immutable constraints; enrichment adds higher-order traits without altering the demographic substrate. (See Appendix C.5.2

**Round-Robin Skeleton Assignment**

Skeletons are distributed to communities using a round-robin algorithm (Algorithm 2):

**Algorithm 2** Round-Robin Skeleton-to-Community Assignment

**Require:** Skeleton pool $\mathcal{S}$, community set $\mathcal{C} = \{C_1, \ldots, C_m\}$, personas per community $k$
**Ensure:** Persona set $\mathcal{P}$

$\mathcal{P} \leftarrow \emptyset$
Shuffle $\mathcal{S}$
**for** each community $C_i \in \mathcal{C}$ **do**
    remaining$[i] \leftarrow k$
**end for**
**while** $\sum_i$ remaining$[i] > 0$ **and** $\mathcal{S} \neq \emptyset$ **do**
    **for** $i = 1$ to $|\mathcal{C}|$ **do**
        **if** remaining$[i] > 0$ **and** $\mathcal{S} \neq \emptyset$ **then**
            $s \leftarrow \text{POP}(\mathcal{S})$
            $p \leftarrow \text{GENERATEPERSONA}(s, C_i)$
            **if** $p \neq \emptyset$ **then**
                $\mathcal{P} \leftarrow \mathcal{P} \cup \{p\}$
                remaining$[i] \leftarrow$ remaining$[i] - 1$
            **else**
                remaining$[i] \leftarrow$ remaining$[i] + 1$
            **end if**
        **end if**
    **end for**
**end while**
**return** $\mathcal{P}$

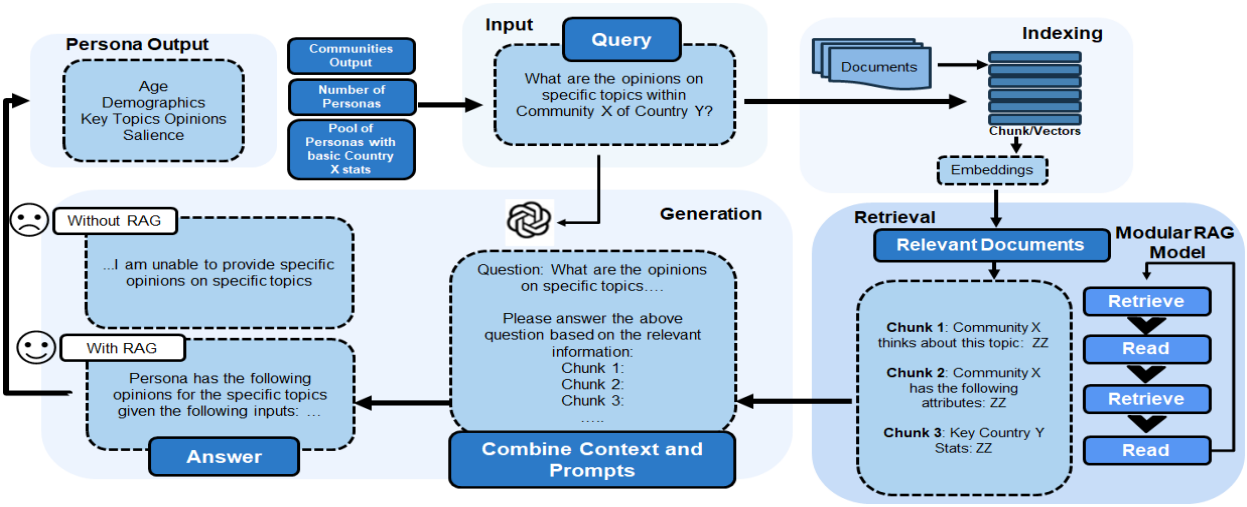



Figure 5.5: Persona-level RAG in AURORA. Retrieved evidence chunks combine with skeleton constraints and community context to generate persona-level outputs including demographics, topic opinions, and salience.

### Biographical Generation

Figure 5.5 illustrates persona-level RAG integration.

The LLM generates biographical elements conditioned on:

- Fixed skeleton attributes (age, gender, education, ethnicity)
- Assigned community context
- RAG-retrieved contextual documents

Outputs include:

- **Name** (with uniqueness enforcement via hybrid validation: maintains `used_names` set, passes last 20 names to LLM, validates post-generation, retries up to 3 times on duplicates)
- Short biography
- Occupational and educational narratives
- Family context
- Formative life experiences

### Psychological and Ideological Enrichment

**Personality Traits.** Each persona receives OCEAN traits sampled from country-specific empirical distributions [156]:

$$T_i \sim \mathcal{N}(\mu_{i,\text{ctry}}, \sigma^2_{i,\text{ctry}}), \quad T_i \in [0,1], \quad i \in \{O, C, E, A, N\} \tag{5.15}$$

For more information, see Chapter 4.10 and Appendix B.1.

**Baseline Emotions.** Emotional profiles are computed deterministically from personality via a fixed linear mapping (no LLM call) [36, 152]:

$$\mathbf{e}_{\text{base}} = \text{clip}(\mathbf{W}\mathbf{o} + \mathbf{b}, 0, 1) \tag{5.16}$$

where $\mathbf{o}$ represents the OCEAN trait vector and $\mathbf{W}$ and $\mathbf{b}$ encode a theory-informed trait–emotion mapping (Table 5.3).

The mapping was constructed by operationalizing Plutchik's eight primary emotions [152] as linear combinations of Big Five personality traits, informed by established trait–affect relationships in personality psychology. For example, Neuroticism is positively associated with fear, sadness, and anger; Extraversion with joy and anticipation; Agreeableness with trust; and Openness with anticipation and surprise. Coefficients were selected to preserve directional consistency with psychological literature while ensuring bounded outputs and cross-emotion differentiation. This deterministic transformation provides psychologically coherent baseline affective states that are fully reproducible and independent of stochastic LLM generation.

Table 5.3: Trait-Emotion Mapping Coefficients

| Emotion | Primary Traits | Mapping Formula |
|---|---|---|
| Joy | High E, A, O; Low N | $0.20E + 0.15A + 0.15O - 0.25N + 0.25$ |
| Trust | High A, C, E; Low N | $0.30A + 0.20C + 0.15E - 0.20N + 0.15$ |
| Fear | High N; Low E, O | $0.40N + 0.10A + 0.10C - 0.15E - 0.10O + 0.15$ |
| Surprise | High O, E | $0.25O + 0.20E + 0.10N + 0.10$ |
| Sadness | High N; Low E | $0.25N + 0.10A + 0.10O - 0.20E + 0.25$ |
| Disgust | High C, N; Low A | $0.25C + 0.20N - 0.15A - 0.10O + 0.20$ |
| Anger | High N; Low A | $0.35N - 0.25A + 0.10O + 0.10E + 0.20$ |
| Anticipation | High O, E, C; Low N | $0.30O + 0.20E + 0.15C - 0.15N + 0.10$ |

**Political Affiliations.** Ideological positions inherit community distributions with bounded stochastic variation [81]:

$$s_{\text{pers}}(t) = s_{\text{comm}}(t) + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2) \tag{5.17}$$

where $\sigma^2$ is calibrated so approximately 95% of opinions fall within $\pm 10\%$ of the community range.

### Hierarchical Opinion Sampling

Topic-level opinions are sampled using salience-gated distributions [81]. For persona $P_j$ in community $C_i$:

**Step 1: Sample persona salience.**

$$S_{j,k} \sim \mathcal{N}(S_{i,k}, 0.1^2), \quad S_{j,k} \in [0, 1] \tag{5.18}$$

where the sampled value is clamped to $[0, 1]$ and rounded to two decimal places. The variance parameter ($0.1^2$) was chosen to introduce moderate intra-community heterogeneity while preserving coherence around the community-level salience mean. On the bounded $[0, 1]$ scale, $\sigma = 0.1$ produces meaningful variability without excessive truncation. Smaller values yielded overly homogeneous personas, whereas larger values weakened community structure. This choice balances individual diversity with structural fidelity.

**Step 2: Select distribution based on salience threshold.**

$$O_{j,k} \sim \begin{cases} \mathcal{N}(O_{i,k}, 1.0) & \text{if } S_{j,k} \geq 0.95 \text{ (Very High)} \\ \mathcal{N}(O_{i,k}, 1.5) & \text{if } S_{j,k} \geq 0.80 \text{ (High)} \\ \text{CountryDist}(O_C^{(k)}, D^{(k)}) & \text{if } S_{j,k} < 0.80 \text{ (Medium/Low)} \end{cases} \tag{5.19}$$

where the country distribution family $D^{(k)}$ determines sampling for low-salience topics [81]:

- **NORMAL**: $O_{j,k} \sim \mathcal{N}(O_C^{(k)}, 1.5)$

- **SKEWED**: Asymmetric Beta distribution with soft-capping
- **BIMODAL**: Mixture of two normals centered at $\pm 1$

Figure 5.6 summarizes the salience-gated distribution selection.

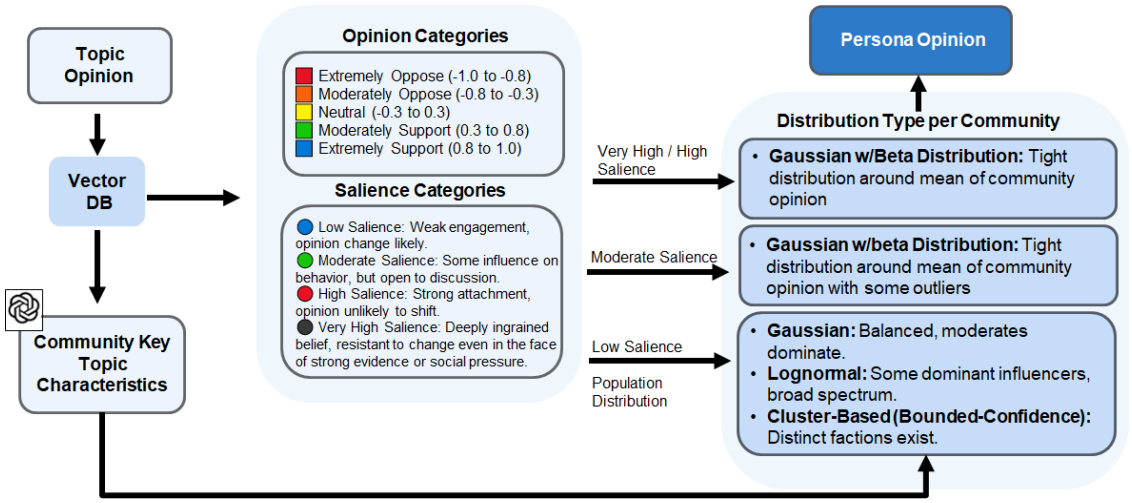



Figure 5.6: Salience-weighted opinion distribution pipeline. AURORA maps opinion polarity and salience categories into distribution families(tight vs broad vs bimodal), producing persona-level opinions that preserve community coherence on high-salience issues while allowing heterogeneity on low-salience topics.

**Step 3: Clamp to modeling range.**

$$O_{j,k} = \text{clamp}(O_{j,k}, -1.0, 1.0) \tag{5.20}$$

**Step 4: Generate natural language description.** An LLM generates `opinion.description` (separate call) conditioned on numeric values without modifying them [45, 104].

### Topic-Specific Emotion Modulation

Each persona maintains a baseline emotional state derived from their personality traits. When considering a specific topic, these baseline emotions are modulated based on two factors: (1) how strongly the persona cares about the topic (salience), and (2) whether they support or oppose it (opinion). This approach follows established models of emotion-cognition interaction [36, 152].

The modulation process operates on the principle that *caring more intensely about a topic amplifies emotional response*, and *positive opinions strengthen positive emotions while dampening negative ones* (and vice versa for negative opinions).

**Computing Emotional Intensity from Salience.** Salience $S_{j,k} \in [0, 1]$ indicates how much persona $P_j$ cares about topic $t_k$. To compute emotional intensity, salience is first expanded to a centered scale:

$$S_{\text{exp}} = 2 \cdot S_{j,k} - 1 \in [-1, 1] \tag{5.21}$$

This expansion maps low salience (near 0) to $-1$ and high salience (near 1) to $+1$. The intensity factor is then computed as:

$$I = 0.5 + 0.5 \cdot S_{\text{exp}} = S_{j,k} \in [0, 1] \tag{5.22}$$

Note that this simplifies to the original salience value. Low-salience topics ($S_{j,k} \approx 0$) produce minimal emotional modulation ($I \approx 0$), while high-salience topics ($S_{j,k} \approx 1$) produce maximal modulation ($I \approx 1$).

**Opinion-Driven Emotion Adjustment.** Opinion $O_{j,k} \in [-1, 1]$ represents the persona's stance, where $-1$ indicates strong opposition, $0$ indicates neutrality, and $+1$ indicates strong support. A scaling factor $\alpha = 0.25$ controls the magnitude of emotional shifts, ensuring that modulations remain bounded and psychologically plausible.

**Positive emotions** (joy, trust, anticipation) are *amplified by positive opinions* and *dampened by negative opinions*:

$$e_{\text{joy}}(t_k) = e^{\text{base}}_{\text{joy}} + \alpha \cdot I \cdot O_{j,k} \tag{5.23}$$

$$e_{\text{trust}}(t_k) = e^{\text{base}}_{\text{trust}} + \alpha \cdot I \cdot O_{j,k} \tag{5.24}$$

$$e_{\text{anticipation}}(t_k) = e^{\text{base}}_{\text{anticipation}} + \alpha \cdot I \cdot O_{j,k} \tag{5.25}$$

For example, if a persona strongly supports a highly salient topic ($O_{j,k} = 1$, $I = 1$), their joy increases by $0.25$ units above baseline. Conversely, if they strongly oppose it ($O_{j,k} = -1$), their joy decreases by $0.25$ units.

**Negative emotions** (anger, disgust, sadness, fear) exhibit the *inverse pattern*—they are amplified by negative opinions and dampened by positive opinions:

$$e_{\text{anger}}(t_k) = e^{\text{base}}_{\text{anger}} - \alpha \cdot I \cdot O_{j,k} \tag{5.26}$$

$$e_{\text{disgust}}(t_k) = e^{\text{base}}_{\text{disgust}} - \alpha \cdot I \cdot O_{j,k} \tag{5.27}$$

$$e_{\text{sadness}}(t_k) = e^{\text{base}}_{\text{sadness}} - \alpha \cdot I \cdot O_{j,k} \tag{5.28}$$

$$e_{\text{fear}}(t_k) = e^{\text{base}}_{\text{fear}} - \alpha \cdot I \cdot O_{j,k} \tag{5.29}$$

The negative sign ensures that opposing a topic ($O_{j,k} < 0$) increases negative emotions, while supporting it ($O_{j,k} > 0$) decreases them.

**Surprise** is modulated solely by salience, independent of opinion valence:

$$e_{\text{surprise}}(t_k) = e^{\text{base}}_{\text{surprise}} - \alpha \cdot S_{\text{exp}} \tag{5.30}$$

The rationale is that *familiar, salient topics are less surprising*. As salience increases ($S_{\text{exp}} \to 1$), surprise decreases. Topics the persona rarely thinks about ($S_{\text{exp}} \to -1$) may increase surprise slightly above baseline.

**Bounding and Precision.** All modulated emotion values are clamped to $[0, 1]$ to ensure psychological validity, as emotions cannot be negative or exceed maximum intensity. Values are rounded to two decimal places for numerical stability. This ensures that extreme combinations of opinion and salience do not produce implausible emotional states.

**Interpretation.** This formulation captures intuitive emotional dynamics: personas feel more positive emotions about topics they support and care about, more negative emotions about topics they oppose and care about, and less surprise about familiar topics. The $\alpha = 0.25$ scaling ensures that emotional shifts remain moderate—even maximal opinion and salience produce only a $\pm 0.25$ change from baseline, preventing unrealistic emotional extremes while maintaining meaningful variation.

### 5.3.7 Stage 5: Group Formation and Assignment

Beyond base community membership, AURORA generates affinity groups representing interest-based clusters such as environmental advocacy organizations, technology enthusiast communities, or political action committees. These groups provide an additional layer of social structure that cuts across geographic communities, enabling personas to connect based on shared interests rather than solely on residential proximity. The group formation process consists of two main phases: RAG-augmented group generation and mutual selection-based assignment.

Appendix **??** provides a crosswalk mapping AURORA-generated groups from the TOTO exercise to AESOP's canonical group categories (see Chapter 7.1 for more information), enabling cross-scenario comparison and addressing whether generated groups align with predefined scenario structures. The results show strong post-hoc conceptual alignment—e.g., business institutions map to Industry/Finance, universities to Academia, and environmental organizations to Conservationist—while country-specific groups align with national identity categories. However, this alignment is emergent rather than enforced: AURORA generates groups via retrieval-grounded LLM inference conditioned on country and topic context, without explicitly incorporating AESOP's group schemas.

#### RAG-Augmented Group Generation

AURORA generates a configurable number of groups (default: 5 groups per 100 personas per country, specified via `number_of_groups` in the configuration file) through an iterative process that samples personas and uses their characteristics to inspire contextually relevant group themes. Algorithm 3 formalizes this process.

The generation process begins by randomly sampling $k = 2$ personas from the population. These sampled personas serve as thematic anchors—their professions, interests, and topic opinions inform what kind of group would be relevant to individuals like them. For instance, if the sampled personas include a climate scientist and an environmental activist, the system might generate an environmental advocacy group. If they include software engineers interested in artificial intelligence, the system might generate a technology ethics organization.

To ground group generation in real-world context, AURORA constructs a query from the sampled personas' topics and the country context, then retrieves $k = 3$ relevant documents from the vector database using RAG. These retrieved documents provide factual grounding about actual organizations, movements, or interest groups that exist in the target country, ensuring that generated groups reflect realistic social structures rather than generic templates.

The LLM then generates a group specification including a name, purpose, and description, conditioned on both the sampled personas and the retrieved contextual documents. To prevent duplicates, newly generated groups are checked against existing groups by name—if a group with the same name already exists, the generation is discarded and the process repeats. This continues until the target number of groups ($N_G$) is reached.

**Algorithm 3** RAG-Augmented Group Generation

**Require:** Persona set $\mathcal{P} = \{p_1, \ldots, p_n\}$, target number of groups $N_G$, country context $C$
**Ensure:** Group set $\mathcal{G}$

$\mathcal{G} \leftarrow \emptyset$
**while** $|\mathcal{G}| < N_G$ **do**
  Sample $k = 2$ personas $\mathcal{P}_{\text{sample}} \sim \textsc{UniformSample}(\mathcal{P}, k)$
  $\mathcal{T}_{\text{sample}} \leftarrow \bigcup_{p \in \mathcal{P}_{\text{sample}}} p.\text{topics}$
  $q \leftarrow \textsc{BuildQuery}(C, \mathcal{T}_{\text{sample}})$
  $\mathcal{R} \leftarrow \textsc{RetrieveContext}(q, 3)$
  $g \leftarrow \textsc{GenerateGroup}(\mathcal{P}_{\text{sample}}, \mathcal{R}, C)$
  **if** $g.\text{name} \notin \{\, g'.\text{name} \mid g' \in \mathcal{G} \,\}$ **then**
    $\mathcal{G} \leftarrow \mathcal{G} \cup \{g\}$
  **end if**
**end while**
**return** $\mathcal{G}$

### Mutual Selection Process

Once groups are generated, personas must be assigned to them. Unlike community assignment (which is deterministic based on geographic location), group membership is voluntary and based on mutual fit—personas must be interested in a group, and groups must select personas who align with their purpose. This mutual selection process unfolds in three stages.

**Stage 1: Persona Ranking.** Each persona evaluates all available groups and ranks them by relevance using an LLM-based assessment. The LLM considers the persona's profession, interests, demographic attributes, and topic opinions to determine how well each group aligns with the persona's profile. Rankings are expressed as qualitative labels: Very High, High, Medium, Low, or Very Low relevance.

**Stage 2: Filtering.** Personas construct filtered preference lists by excluding groups that fall below a minimum relevance threshold (configurable via `persona_min_group_relevance`, de-

fault: Medium). This ensures that personas only consider groups they would realistically join. For example, a conservative business executive would filter out progressive activist organizations, while an environmental scientist would filter out fossil fuel industry groups.

**Stage 3: Round-Robin Group Selection.** Groups iteratively select personas from their eligible pools in a round-robin fashion, as formalized in Algorithm 4. The process is designed to achieve a target average of $\tau$ groups per persona (default: $\tau = 1.5$, configurable via `target_avg_groups_per_person`) while respecting individual capacity constraints (maximum $k$ groups per persona, default: $k = 3$, set via `max_groups_per_person`).

In each round, every group selects up to $s$ personas (default: $s = 3$, configurable via `slots_per_round`) from its pool of interested and eligible candidates. A persona is eligible if they expressed at least Medium relevance for the group and have not yet reached their maximum group capacity. The group's LLM-based selection considers how well each candidate's profile fits the group's purpose and existing membership composition. This continues in rounds until the total number of assignments across all personas reaches the target ($N_{\text{target}} = |\mathcal{P}| \times \tau$).

The round-robin approach ensures fairness—no single group monopolizes the most desirable personas, and groups that generate later in the process have equal opportunity to recruit members. Group order is shuffled each round to prevent systematic bias.

**Algorithm 4** Round-Robin Group Selection

**Require:** Groups $\mathcal{G} = \{g_1, \ldots, g_m\}$, personas $\mathcal{P} = \{p_1, \ldots, p_n\}$
**Require:** Relevance matrix $\mathcal{R}$, slots per round $s$, maximum groups per persona $k$, target average $\tau$
**Ensure:** Updated group memberships for $\mathcal{G}$ and $\mathcal{P}$

$N_{\text{target}} \leftarrow |\mathcal{P}| \times \tau$
Initialize count$[p] \leftarrow 0$ for all $p \in \mathcal{P}$
$N_{\text{current}} \leftarrow 0$
**while** $N_{\text{current}} < N_{\text{target}}$ **do**
  **for** each group $g \in \mathcal{G}$ **do**
    $\mathcal{P}_g^{\text{interested}} \leftarrow \{\, p \in \mathcal{P} \mid \mathcal{R}[p][g] \geq \texttt{min\_relevance} \,\}$
    $\mathcal{P}_g^{\text{eligible}} \leftarrow \{\, p \in \mathcal{P}_g^{\text{interested}} \mid \text{count}[p] < k \,\}$
    $\mathcal{S}_g \leftarrow \textsc{SelectViaLLM}(g, \mathcal{P}_g^{\text{eligible}}, s)$
    **for** each persona $p \in \mathcal{S}_g$ **do**
      Add $p$ to group $g$
      Add group $g$ to persona $p$
      count$[p] \leftarrow$ count$[p] + 1$
      $N_{\text{current}} \leftarrow N_{\text{current}} + 1$
    **end for**
  **end for**
**end while**
**return** $\mathcal{G}, \mathcal{P}$

**Community Group Auto-Assignment.** After the mutual selection process completes, all personas are automatically assigned to a community-specific group corresponding to their residential community. This ensures that every persona belongs to at least one group, even if they were not selected by any interest-based groups during the mutual selection phase. Community groups represent local civic organizations, neighborhood associations, or community centers that naturally include all residents.

**Leader Assignment.** For each group, an LLM selects a leader from the membership roster based on leadership qualities inferred from the persona's profession, age, education level, and topic expertise. Leaders serve as focal points for group activity and may be assigned elevated influence weights in subsequent belief dynamics modeling. For example, a climate advocacy group might select a climate scientist as its leader, while a business association might select a senior executive.

### 5.3.8 Stage 6: Social Media Account Generation

AURORA instantiates platform-specific accounts from persona profiles. Detailed specifications appear in Appendices C.2.1 and C.2.2.

#### Twitter Account Generation

Twitter accounts combine LLM-generated profiles with deterministic username assignment. Algorithm 5 ensures global uniqueness.

**Algorithm 5** Unique Twitter Screen Name Generation

**Require:** Persona name $N$, used screen name set $\mathcal{U}$
**Ensure:** Unique screen name $s$

$(\textit{first}, \textit{last}) \leftarrow \textsc{ParseName}(N)$
$(\textit{first}, \textit{last}) \leftarrow \textsc{Sanitize}(\textit{first}, \textit{last})$
$\textit{base} \leftarrow \textit{first} \,\|\, \_ \,\|\, \textit{last}$
**for** $i = 1$ to $100$ **do**
    $\textit{suffix} \leftarrow \textsc{RandomInt}(10, 9999)$
    $s \leftarrow \textit{base} \,\|\, \textit{suffix}$
    **if** $s \notin \mathcal{U}$ **then**
        $\mathcal{U} \leftarrow \mathcal{U} \cup \{s\}$
        **return** $s$
    **end if**
**end for**
$s \leftarrow \textit{base} \,\|\, (\textsc{Timestamp}() \bmod 10000)$
**return** $s$

**Telegram Account and Channel Generation**

Telegram usernames are generated using a creative word-combination approach that produces realistic, diverse handles without relying on LLM generation. The system maintains four curated word pools, each containing 50 selected words that reflect common patterns observed in real Telegram usernames:

- **Descriptors** (adjectives and modifiers): Wild, Bold, Cosmic, Stellar, Mystic, Quantum, Shadow, Phoenix, Crystal, Electric, Neon, Urban, Modern, Swift, Silent, Fierce, Brave, Wise, Lucky, Chill, Noble, Ancient, Digital, Cyber, Ultra, Prime, Alpha, Zen, Nova, Solar, Lunar, etc.
- **Actions** (roles and activities): Warrior, Pioneer, Explorer, Creator, Builder, Seeker, Dreamer, Master, Guru, Ninja, Wizard, Champion, Hero, Legend, Hunter, Rider, Walker, Dancer, Writer, Coder, Hacker, Gamer, Streamer, Blogger, Trader, Guardian, Watcher, etc.
- **Nouns** (animals, nature, and concepts): Tiger, Eagle, Wolf, Dragon, Phoenix, Lion, Bear, Storm, Thunder, Ocean, Mountain, River, Sky, Star, Moon, Sun, Flame, Wave, Quest, Journey, Dream, Vision, Shadow, Light, Dawn, Wind, Fire, Stone, Crystal, Diamond, etc.
- **Interests** (topics and domains): Crypto, Bitcoin, Blockchain, Fitness, Gym, Travel, Wanderer, Nomad, Food, Chef, Tech, Digital, Code, Data, Art, Design, Music, Business, Trade, Science, Research, Education, Eco, Green, Sports, Gaming, Photo, Writing, Marketing, Brand, etc.

These word pools were manually curated to reflect authentic Telegram username conventions, drawing from observed patterns in real-world Telegram communities. The pools balance evocative imagery (Phoenix, Storm, Cosmic) with contemporary interests (Crypto, Tech, Gaming) and role-based identifiers (Warrior, Explorer, Creator).

**Username Combination Strategies.** Five distinct combination strategies randomly generate usernames by selecting words from these pools and appending numeric suffixes. Each strategy produces a different structural pattern:

1. **Descriptor + Noun + Number**: Combines an adjective with a noun and a two-digit number (10–99). Examples: `WildPhoenix47`, `CosmicDragon88`, `SilentWolf23`.
2. **Interest + Action + Number**: Pairs a domain interest with a role-based action and a number. Examples: `CryptoWarrior55`, `TechSeeker88`, `FitnessGuru42`.
3. **Descriptor + Action + Number**: Combines an adjective with an action and a number. Examples: `MysticSeeker77`, `BoldExplorer19`, `SwiftHacker64`.
4. **Descriptor + Descriptor + Noun**: Uses two adjectives followed by a noun (no number). Examples: `WildCosmicDragon`, `BoldFierceEagle`, `SilentAncientStorm`.
5. **Noun + Noun + Number**: Combines two nouns with a number. Examples: `TigerStorm55`, `PhoenixFlame91`, `DragonMoon37`.

For each persona, the system randomly selects one of these five strategies with equal probability, then randomly samples words from the appropriate pools. Usernames are truncated to 20

characters for safety (Telegram allows up to 32), and uniqueness is enforced—if a collision occurs, the system retries with a different combination. After 100 failed attempts, a timestamp-based fallback ensures generation always succeeds.

**Channel Generation and Assignment.** Telegram channels are generated to provide broadcast platforms for information dissemination. The number of channels is proportional to population size:

$$N_{\text{channels}} = \max\left(1, \left\lfloor \frac{N_{\text{personas}}}{100} \times 8 \right\rfloor\right) \tag{5.31}$$

This formula produces approximately 8 channels per 100 personas, ensuring sufficient diversity without overwhelming the simulation. For example, a population of 500 personas generates $\lfloor 500/100 \times 8 \rfloor = 40$ channels, while a small population of 50 personas generates $\max(1, \lfloor 4 \rfloor) = 4$ channels (with the minimum of 1 enforced).

Channel assignment follows a balanced round-robin distribution: personas are assigned to channels sequentially in a circular pattern, ensuring each channel receives approximately equal membership. The first persona assigned to each channel is automatically designated as the channel owner, responsible for posting content and managing the channel. This ensures every channel has an owner without requiring additional LLM-based selection logic.

## 5.4 Model: Formal State Space Specification

This section formalizes the AURORA persona model as a structured state space suitable for agent-based simulation. Whereas Section 5.3 describes generation procedures, this section specifies the *model representation*: the variables assigned to each synthetic agent, the hierarchical dependencies governing their initialization, and the mechanisms by which opinions and emotions are instantiated.

AURORA is designed as an *initialization model* rather than a dynamic opinion update model. Its purpose is to generate psychologically coherent, demographically proportional agents whose internal states constitute valid starting conditions for downstream opinion dynamics simulations.

### 5.4.1 Agent State Representation

Each synthetic persona $p_i$ is represented as a tuple of immutable structural attributes and mutable cognitive-affective attributes:

$$p_i = \langle D_i, C_i, T_i, E_i, O_i, S_i, G_i, A_i \rangle \tag{5.32}$$

where:

- $D_i \in \mathcal{D}$: Fixed demographic attributes (skeleton persona)
- $C_i \in \mathcal{C}$: Community membership and meso-level context
- $T_i \in [0,1]^5$: OCEAN personality traits

- $E_i \in [0,1]^8$: Baseline and topic-conditioned emotional states (Plutchik)
- $O_i : \mathcal{T} \to [-1,1]$: Topic-specific opinions
- $S_i : \mathcal{T} \to [0,1]$: Topic-specific salience values
- $G_i \subseteq \mathcal{G}$: Group memberships
- $A_i$: Platform-specific social media accounts

**Immutability constraints.** Demographic attributes $D_i$ are treated as invariant throughout simulation. All ideological, emotional, and behavioral attributes are conditioned on $D_i$ but do not alter it. This ensures that observed dynamics reflect modeled mechanisms rather than artifacts of population construction.

### 5.4.2 Hierarchical Opinion Model

AURORA generates agent opinions through a three-level hierarchical process that mirrors real-world belief formation: individuals develop viewpoints within communities, which themselves exist within broader national cultures. This hierarchical structure ensures that synthetic populations exhibit realistic patterns of consensus, disagreement, and subcultural variation.

The key innovation is the separation of *distributional parameters* (generated once by LLMs using cultural context) from *individual values* (sampled mathematically), enabling both semantic realism and computational scalability.

#### Three-Level Opinion Architecture

**Level 1: Country-Level Priors.** At the macro level, each country $C$ is characterized by national opinion tendencies for each topic $t \in \mathcal{T}$:

$$\Theta_C = \{(O_C^{(t)}, S_C^{(t)}, D^{(t)}, \kappa_C^{(t)}) \mid t \in \mathcal{T}\} \tag{5.33}$$

These parameters define:

- $O_C^{(t)} \in [-1,1]$: The national mean opinion, where -1 represents strong opposition and +1 strong support
- $S_C^{(t)} \in [0,1]$: National salience, measuring how central the topic is to national discourse (0 = irrelevant, 1 = identity-defining)
- $D^{(t)} \in \{\text{NORMAL, SKEWED, BIMODAL}\}$: The distributional shape describing how opinions spread across the population

These country-level parameters establish the *cultural substrate*—the broad tendencies and constraints within which communities and individuals form their beliefs. Crucially, they do not directly dictate individual opinions but instead parameterize the distributions from which opinions are later sampled.

**Level 2: Community-Level Refinement.** At the meso level, each community $C_j$ adapts national tendencies to reflect its distinctive characteristics. Communities are institutional groupings such as universities, religious organizations, labor unions, or neighborhood associations. Each refines the national baseline:

$$\Theta_{C_j} = \{(O_j^{(t)}, S_j^{(t)}, \kappa_j^{(t)}) \mid t \in \mathcal{T}\} \tag{5.34}$$

Community opinions $O_j^{(t)}$ represent **normative centers**—the typical stance expected within that social context—not enforced uniformity. For example, a labor union in the United States might have $O_j^{(\text{labor rights})} = +0.8$ (strong support) while the national mean is $O_C^{(\text{labor rights})} = +0.2$ (mild support). Community opinions are constrained to remain within $\pm 0.3$ of the national mean to prevent implausible extremism.

Political leaning $L_j \in [-1, 1]$ serves as a community-level ideological parameter that influences these deviations across multiple topics.

**Level 3: Persona-Level Sampling.** At the micro level, individual personas develop opinions through *salience-gated sampling*. Rather than inheriting their community's exact position, personas sample from probability distributions centered on community norms, with variance determined by personal salience.

First, each persona's salience for topic $t$ is sampled:

$$S_i^{(t)} \sim \mathcal{N}(S_j^{(t)}, 0.15^2), \quad \text{clamped to } [0, 1] \tag{5.35}$$

This salience value then determines which distribution governs the persona's opinion:

$$O_i^{(t)} \sim \begin{cases} \mathcal{N}(O_j^{(t)}, 1.0^2) & \text{if } S_i^{(t)} \geq 0.90 \quad \text{(very high salience)} \\ \mathcal{N}(O_j^{(t)}, 1.5^2) & \text{if } 0.70 \leq S_i^{(t)} < 0.90 \quad \text{(high salience)} \\ f_{D^{(t)}}(O_C^{(t)}) & \text{if } S_i^{(t)} < 0.70 \quad \text{(medium/low salience)} \end{cases} \tag{5.36}$$

**Interpretation:**

- **Very high salience ($S_i^{(t)} \geq 0.90$)**: The topic is identity-defining for this persona. Their opinion clusters tightly around the community norm with low variance ($\sigma = 1.0$).
- **High salience ($0.70 \leq S_i^{(t)} < 0.90$)**: The topic is important but not central. Moderate variance ($\sigma = 1.5$) allows individual deviation while maintaining community tendency.
- **Medium/low salience ($S_i^{(t)} < 0.70$)**: The topic is peripheral to the persona's concerns. Opinion reverts to the broader country-level distribution $f_{D^{(t)}}$, decoupling from community norms entirely.

This mechanism produces realistic heterogeneity: within a university community, students highly engaged with climate policy will cluster around the institutional position, while those indifferent will exhibit opinions drawn from broader national patterns.

### Distribution Families for Low-Salience Sampling

When persona salience falls below 0.50, opinions are sampled from country-level distributions that capture different consensus structures:

- **NORMAL** (broad consensus): Most of the population agrees, with symmetric variance around the mean:

$$f_{\text{NORMAL}}(O_C^{(t)}) = \mathcal{N}(O_C^{(t)}, 1.5^2) \tag{5.37}$$

  Example: Universal healthcare in Canada ($O_C \approx +0.45$).

- **SKEWED** (asymmetric majority): A clear majority position exists, but with asymmetric dissent. Implemented via Beta-distributed asymmetry with soft capping:

$$f_{\text{SKEWED}}(O_C^{(t)}) = 2\tanh\left(\frac{O_C^{(t)} + \text{offset(skew)}}{1.3}\right) \tag{5.38}$$

  where offset is derived from a Beta distribution parameterized by $O_C^{(t)}$. The hyperbolic tangent produces soft saturation near ±1.

  Example: Gun control in the United States (majority support, vocal minority opposition).

- **BIMODAL** (polarization): Two distinct camps with little middle ground. Opinions are drawn from a mixture of two Gaussians:

$$f_{\text{BIMODAL}}(O_C^{(t)}) = \begin{cases} \mathcal{N}(-1.0, 0.4^2) & \text{with probability } p_{\text{left}} \\ \mathcal{N}(1.0, 0.4^2) & \text{with probability } 1 - p_{\text{left}} \end{cases} \tag{5.39}$$

  where $p_{\text{left}} = 0.5 - 0.2375 \cdot O_C^{(t)}$ ensures that nonzero national means shift probability toward the aligned pole.

  Example: Abortion rights in the United States.

All sampled values are clamped to $[-1, 1]$ and rounded to two decimal places for consistency.

#### Salience as a Control Variable

Salience plays a central role in the model, governing three distinct behavioral dimensions:

1. **Opinion variance**: Higher salience produces tighter clustering around community norms, as shown in the three-tier sampling mechanism above.
2. **Engagement likelihood**: Higher salience increases the probability that a persona will discuss, share, or debate the topic on social media (used in downstream simulation).
3. **Emotion activation magnitude**: Higher salience amplifies affective responses to the topic (see §5.4.3).

This unified control variable ensures coherent behavior: personas who care deeply about climate change will (1) hold opinions aligned with their community, (2) frequently post about it, and (3) experience strong emotions when encountering related content.

### 5.4.3 Personality and Emotion Model

Psychological attributes complement demographic and ideological characteristics, providing the affective substrate for behavioral simulation.

### Personality Traits

Each persona is assigned Big Five personality traits (Openness, Conscientiousness, Extraversion, Agreeableness, Neuroticism) sampled from country-specific empirical distributions:

$$T_i = (O_i, C_i, E_i, A_i, N_i), \quad T_{i,k} \sim \mathcal{N}(\mu_{k,C}, \sigma^2_{k,C}) \tag{5.40}$$

Distribution parameters $(\mu_{k,C}, \sigma_{k,C})$ are derived from cross-cultural psychology studies [170], ensuring that synthetic populations exhibit culturally authentic personality distributions (e.g., higher Agreeableness in East Asian countries, higher Openness in Western Europe).

### Baseline Emotional Profile

Baseline emotional tendencies are computed deterministically from personality traits via a linear transformation encoding established trait-emotion relationships from affective psychology:

$$\mathbf{e}_i^{\text{base}} = \text{clip}(\mathbf{W}T_i + \mathbf{b}, 0, 1) \tag{5.41}$$

Here, $\mathbf{W} \in \mathbb{R}^{8\times5}$ is a weight matrix encoding how each personality trait contributes to eight primary emotions (Plutchik's wheel: joy, trust, fear, surprise, sadness, disgust, anger, anticipation), and $\mathbf{b} \in \mathbb{R}^8$ provides baseline offsets. For example:

- High Extraversion $\rightarrow$ higher baseline joy and anticipation
- High Neuroticism $\rightarrow$ higher baseline fear and sadness
- High Agreeableness $\rightarrow$ higher baseline trust, lower anger

This deterministic mapping ensures that emotional profiles are *personality-consistent*—outgoing personas are predisposed to positive affect, anxious personas to negative affect.

### Topic-Conditioned Emotion Modulation

When a persona encounters a specific topic, their emotional state shifts from baseline based on their opinion and salience for that topic:

$$\mathbf{e}_i^{(t)} = g(\mathbf{e}_i^{\text{base}}, O_i^{(t)}, S_i^{(t)}) \tag{5.42}$$

The modulation function $g$ adjusts emotions differentially based on valence:

**Positive emotions (joy, trust, anticipation):**

$$e_{i,k}^{(t)} = e_{i,k}^{\text{base}} + \alpha \cdot S_i^{(t)} \cdot O_i^{(t)} \tag{5.43}$$

A persona with positive opinion ($O_i^{(t)} > 0$) and high salience will experience elevated joy, trust, and anticipation when encountering the topic.

**Negative emotions (anger, disgust, sadness, fear):**

$$e_{i,k}^{(t)} = e_{i,k}^{\text{base}} - \alpha \cdot S_i^{(t)} \cdot O_i^{(t)} \tag{5.44}$$

Positive opinions *suppress* negative emotions (contentment reduces fear), while negative opinions amplify them (opposition triggers anger/disgust).

**Surprise:**

$$e_{i,\text{surprise}}^{(t)} = e_{i,\text{surprise}}^{\text{base}} - \alpha(2S_i^{(t)} - 1) \tag{5.45}$$

High salience reduces surprise (familiar topics are predictable), while low salience maintains surprise potential.

The parameter $\alpha = 0.25$ controls modulation strength, ensuring that topic-specific shifts are noticeable but do not overwhelm personality-driven baselines. All values are clamped to $[0, 1]$.

**Interpretation**: A persona strongly supporting climate action ($O_i = +0.9$, $S_i = 0.95$) encountering a climate policy article will experience elevated joy/trust and suppressed fear/sadness relative to baseline. Conversely, a climate skeptic ($O_i = -0.8$, $S_i = 0.9$) will experience amplified anger/disgust and reduced positive affect.

### 5.4.4 Social Embedding Model

Personas are embedded in social structures that constrain their exposure to information and shape their interaction patterns.

#### Community Membership

Each persona is assigned to exactly one primary community via balanced allocation:

$$C_i \in \mathcal{C}, \quad |\{p : C_p = C_j\}| \approx k \quad \forall C_j \in \mathcal{C} \tag{5.46}$$

where $k$ is the target personas-per-community parameter. Communities represent institutional affiliations (universities, religious groups, labor unions, etc.) that provide the immediate social context for opinion formation.

#### Group Affiliations

Personas may join multiple voluntary interest groups (e.g., environmental advocacy, gaming communities, parent-teacher associations):

$$G_i \subseteq \mathcal{G}, \quad 0 \leq |G_i| \leq \texttt{max\_groups\_per\_person} \tag{5.47}$$

Group assignment follows a mutual selection process: personas rank groups by perceived relevance, and groups select members who meet minimum relevance thresholds. This ensures that group memberships are plausible given persona attributes. On average, personas belong to $\sim$0.5-1.0 groups beyond their primary community.

**Social Media Accounts**

Each persona possesses platform-specific accounts defining their online presence:

$$A_i = \{a_{\text{twitter}}, a_{\text{telegram}}\} \tag{5.48}$$

Accounts specify:

- **Behavioral parameters**: Posting frequency, content modality preferences (text/image/video), engagement propensities (reply/retweet/quote rates)
- **Platform identifiers**: Screen names, usernames, bio descriptions
- **Channel memberships**: For Telegram, personas are assigned to thematic channels (news, politics, entertainment, etc.) based on interests

These parameters enable downstream simulation of information diffusion and social influence processes.

### 5.4.5 Model Scope and Intended Use

AURORA is an **initialization model** rather than a dynamical system. It defines a joint distribution over agent states—demographics, personalities, opinions, and social structures—that can be used as initial conditions for various opinion dynamics models:

- **Classical models**: Friedkin–Johnsen social influence model [61], DeGroot weighted averaging [50]
- **Bounded confidence models**: Hegselmann–Krause model [80], Deffuant et al. model [49]
- **LLM-mediated dynamics**: Generative agents and LLM-driven social simulation [115, 149]
- **Hybrid approaches**: Integrating rule-based opinion update mechanisms with data-driven or LLM-based inference

By grounding demographic structure in real-world statistics and hierarchically constructing ideological and emotional attributes, AURORA ensures that simulation outcomes can be attributed to *interaction mechanisms* rather than artifacts of population construction. For more information see Appendix C.2

**Formal guarantees.** The model provides three key guarantees:

1. **Proportional representation**: The expected proportion of personas in any demographic stratum matches census data:

$$\mathbb{E}\left[\frac{|\{i : D_i \in \mathcal{D}_k\}|}{N}\right] = P_{\text{census}}(\mathcal{D}_k) \tag{5.49}$$

2. **Hierarchical coherence**: Community-level statistics are preserved in aggregate persona distributions:

$$\mathbb{E}[O_i^{(t)} \mid C_i = C_j] \approx O_j^{(t)} \tag{5.50}$$

3. **Salience-variance relationship**: Opinion variance within communities decreases monotonically with community salience:

$$\text{Var}[O_i^{(t)} \mid C_i = C_j] = \begin{cases} 1.0 & \text{if } \mathbb{E}[S_i^{(t)} \mid C_i = C_j] \geq 0.95 \\ 1.5 & \text{if } 0.80 \leq \mathbb{E}[S_i^{(t)} \mid C_i = C_j] < 0.95 \\ \text{country-level} & \text{if } \mathbb{E}[S_i^{(t)} \mid C_i = C_j] < 0.80 \end{cases} \tag{5.51}$$

These properties ensure that AURORA populations are suitable for rigorous computational experiments in opinion dynamics, information diffusion, and social influence modeling. Detailed implementation procedures, including LLM prompt templates and distribution sampling algorithms, are provided in Appendix C.3.

## 5.5 Experimental Design: IGOE Scenario

This section describes the experimental design for the Indo-Pacific Geopolitical Opinion Experiment (IGOE). The experiment examines how geopolitical alignment, institutional affiliation, and topic salience shape opinion and emotional responses in synthetic populations generated using the AURORA framework. The design follows a controlled mixed factorial structure with hierarchical nesting at the country, community, and persona levels.

Tables 5.4, 5.5, and 5.7 summarize the experimental conditions and parameter settings.

### 5.5.1 Experimental Conditions

The total number of persona-topic observations is determined by the cross-product of countries $\mathcal{C}$, institutional community types $\mathcal{I}$, topics $\mathcal{T}$, and personas per community $n_p$:

$$N_{\text{obs}} = |\mathcal{C}| \times |\mathcal{I}| \times |\mathcal{T}| \times n_p. \tag{5.52}$$

The experiment includes $|\mathcal{C}| = 12$ countries, $|\mathcal{I}| = 8$ institutional types, $|\mathcal{T}| = 9$ topics, and $n_p = 15$ personas per community, yielding 38,880 persona-topic observations.

**Countries ($\mathcal{C}$).** Twelve Indo-Pacific and extra-regional states were selected to represent distinct alliance structures, regional roles, and political systems. Country populations are independent and non-overlapping. Country selection and alignment are summarized in Table 5.4.

Table 5.4: IGOE Scenario Country Selection and Geopolitical Alignment

| Code | Country | Alliance | Regional Role |
|---|---|---|---|
| US | United States | Alliance leader | Security guarantor |
| JPN | Japan | Treaty ally | Frontline state |
| KOR | South Korea | Treaty ally | Frontline state |
| TWN | Taiwan | De facto ally | Contested territory |
| AUS | Australia | Five Eyes | Pacific ally |
| NZL | New Zealand | Five Eyes | Pacific ally |
| CAN | Canada | Five Eyes | Distant partner |
| UK | United Kingdom | Five Eyes | Distant partner |
| IND | India | Strategic partner | Non-aligned power |
| CHN | China | Adversary | Primary challenger |
| RUS | Russia | Adversary | Arctic/Pacific power |
| IRN | Iran | Adversary | Extra-regional spoiler |

The **Regional Role** classification distinguishes countries based on their geographic proximity, treaty obligations, and expected operational posture within the Indo-Pacific scenario.

*Security guarantor* refers to the United States as the primary architect and defender of the alliance structure. *Frontline states* (Japan and South Korea) are geographically proximate to potential conflict zones and would bear immediate strategic and operational consequences. *Contested territory* (Taiwan) represents the focal point of sovereignty dispute and escalation risk. *Pacific allies* (Australia and New Zealand) are geographically aligned and politically integrated but not immediate conflict zones. *Distant partners* (Canada and the United Kingdom) provide strategic depth, intelligence cooperation, and reinforcement capacity without geographic immediacy. *Non-aligned power* (India) reflects a strategically autonomous actor whose participation is conditional and interest-based. *Primary challenger* (China) represents the principal revisionist power driving the scenario's conflict dynamics. *Arctic/Pacific power* (Russia) denotes a secondary major power with complementary but distinct strategic interests. *Extra-regional spoiler* (Iran) captures actors capable of indirect escalation through asymmetric or proxy mechanisms outside the primary theater.

**Institutional Communities ($\mathcal{I}$).** Each country contains 24 communities evenly distributed across eight institutional categories (three per category): commercial, academic, healthcare, liberal and conservative government, religious, industrial, and environmental organizations, yielding 288 total communities across 12 countries. This structure balances analytical power and computational feasibility, enabling within-category variation while capturing ideological, economic, and social dimensions of institutional influence on belief evolution.

**Topics ($\mathcal{T}$).** Nine geopolitical topics were selected for an Indo-Pacific information operations exercise (Table 5.5), based on three criteria: (1) **salience variation**, (2) **polarization potential**, and (3) **regional specificity**. Topics span alliance cooperation, security threats, economic policy,

cyber cooperation, and territorial disputes, reflecting real-world dynamics while remaining computationally tractable for multi-country simulation.

Table 5.5: IGOE Scenario Topic Selection and Characteristics

| Topic | Type | Experimental Rationale |
|---|---|---|
| Canadian activity in INDOPAC | Alliance cooperation | Tests Five Eyes cohesion vs. regional skepticism |
| Russian aggression | Security threat | Varies by alliance (high salience for NATO, low for China/Iran) |
| Operation Sea Dragon | Military exercise | Tests attitudes toward power projection |
| Russian Arctic activity | Territorial expansion | Regional vs. global threat perception |
| International Commodities Markets | Economic policy | Business vs. environmental divergence |
| Cyber security cooperation | Technical cooperation | Consensus topic across democracies |
| Taiwanese Sovereignty | Core flashpoint | Maximum polarization between alliance blocs |
| Chinese aggression/hegemony | Security threat | Alliance cohesion vs. economic pragmatism |
| Canadian indigenous rights | Human rights | Progressive vs. conservative divides |

### 5.5.2 Variables

The experiment includes a structured set of dependent, independent, and control variables operating across persona, community, and country levels. Table 5.6 provides a consolidated overview of all variables used in the experimental design.

#### Dependent Variables

Each persona-topic observation produces three primary outcomes: an opinion score $O \in [-1, 1]$, a salience score $S \in [0, 1]$, and an eight-dimensional emotional response vector $\mathbf{E} \in [0, 1]^8$ based on Plutchik's emotion model. These measures capture attitudinal position, perceived importance, and affective response to each topic. Additional derived measures, including within-community opinion variance, country-level polarization indices, and opinion-emotion correlations, are computed for analysis.

#### Control Variables

To isolate the effects of geopolitical alignment, institutional affiliation, and topic salience, several attributes are held constant across experimental conditions. These include persona-level demographic characteristics (age, gender, education, income, and ethnicity), psychological attributes

(OCEAN personality traits and baseline emotional profiles), and structural parameters such as fixed community size (15 personas) and a fixed number of communities per country (24).

### Independent Variables

The experiment manipulates three independent variables: country, institutional community type, and topic. Country is treated as a between-subjects factor capturing macro-level geopolitical alignment, alliance structure, and political system. Institutional community type and topic are treated as within-subjects factors, representing meso-level institutional orientation and issue-specific variation, respectively. Together, these variables define the geopolitical context, institutional perspective, and substantive issue space within which persona opinions and emotional responses are generated and analyzed.

Table 5.6: Summary of Experimental Variables

| **Variable Type** | **Variable** | **Description / Levels** |
|---|---|---|
| *Dependent Variables* | | |
| Opinion score ($O$) | Continuous | Stance toward topic, $[-1, 1]$ (strong opposition to strong support) |
| Salience score ($S$) | Continuous | Topic importance or relevance, $[0, 1]$ |
| Emotional response ($\mathbf{E}$) | Vector (8-dim) | Plutchik emotions: joy, trust, fear, surprise, sadness, disgust, anger, anticipation |
| Derived measures | Computed | Opinion variance, polarization indices, opinion-emotion correlations |
| *Independent Variables* | | |
| Country | Between-subjects (12 levels) | Distinct geopolitical alignment and political system |
| Institutional type | Within-subjects (8 levels) | Commercial; academic and research; healthcare; liberal government; conservative government; religious; industrial; environmental |
| Topic | Within-subjects (9 levels) | Geopolitical issues varying in salience and polarization potential |
| *Control Variables* | | |
| Demographics | Fixed per persona | Age, gender, education, income, ethnicity |
| Psychological traits | Fixed per persona | OCEAN personality traits and baseline emotional profile |
| Structural parameters | Fixed | Community size (15 personas) and communities per country (24) |

### 5.5.3 Design Structure

#### Factorial Design

The experiment employs a mixed factorial design with one between-subjects factor and two within-subjects factors:

- **Between-subjects factor**: Country (12 levels; independent populations)
- **Within-subjects factors**: Institutional community type (8 levels) and Topic (9 levels)

Each country contains 24 institutional communities (3 per institutional type), and each community contains 15 personas. This yields a $12 \times 8 \times 9$ design at the community-topic level, with repeated observations across topics within communities.

#### Hierarchical Nesting

The experimental design follows a three-level hierarchical nesting structure:

$$\text{Personas} \subset \text{Communities} \subset \text{Countries}. \tag{5.53}$$

In total, the experiment comprises 12 countries, 288 communities (24 per country), and 4,320 personas (15 per community). Each persona generates observations across 9 topics, resulting in 38,880 persona-topic observations. This hierarchical structure enables variance decomposition across levels, estimation of cross-level interactions (e.g., institutional effects conditional on country context), and the use of hierarchical or mixed-effects modeling approaches.

### 5.5.4 Experimental Parameters

Table 5.7 summarizes the key population, modeling, and technical parameters used in the experiment.

Table 5.7: IGOE Scenario Experimental Parameters

| Parameter | Value | Justification |
|---|---|---|
| Countries | 12 | Alliance and regional coverage |
| Communities per country | 24 | Institutional balance |
| Personas per community | 15 | Statistical power vs. computational cost |
| Opinion scale | $[-1, 1]$ | Bipolar stance representation |
| Salience scale | $[0, 1]$ | Topic importance |
| Emotion dimensions | 8 | Plutchik emotion model |
| LLM model | GPT-4o-mini | Cost-performance balance |
| Random seed | 42 | Reproducibility |

## 5.6 Results

This section evaluates whether AURORA produces a synthetic population that is (i) demographically plausible, (ii) structurally polarized where expected, and (iii) internally coherent across opinions, salience, and emotion. I emphasize statistical structure, effect sizes, and alignment with the system's design assumptions rather than anecdotal behavior.

### 5.6.1 Synthetic Population Validation

#### Demographic and Behavioral Distributions

Figure 5.7 summarizes the demographic and behavioral composition of the synthetic population ($N = 4{,}320$ personas / $360$ per country). Gender representation is near-balanced, with 2,188 female personas (50.6%) and 2,132 male personas (49.4%). A two-sided binomial test confirms no statistically significant deviation from parity ($p = 0.31$), ensuring that downstream results are not driven by gender imbalance.

Age distributions within each country were sampled from empirical demographic data (Table 5.8). For example, the United States synthetic population exhibits 19.6% in the 15–29 cohort and 24.7% in the 60+ cohort, closely matching the aggregated empirical distribution from U.S. Census data (20–29: 13.0%, 60–69: 12.2%, 70+: 12.0%). Chi-square goodness-of-fit tests comparing synthetic age distributions to source country demographics confirm no statistically significant deviations for any country ($p > 0.05$ for all 12 countries), validating that IPF-based synthesis preserved empirical age structure. Gender representation is near-balanced across the full population, with 2,188 female personas (50.6%) and 2,132 male personas (49.4%). A two-sided binomial test confirms no statistically significant deviation from parity ($p = 0.31$), ensuring that downstream results are not driven by gender imbalance.

Age is distributed across five cohorts with modest skew toward older populations: the largest share of personas fall in the 60+ bin (1,067; 24.7%), followed by 15–29 (845; 19.6%), 50–59 (843; 19.5%), 40–49 (789; 18.3%), and 30–39 (759; 17.6%) (Table 5.9). A chi-square goodness-of-fit test against a uniform age distribution confirms intentional non-uniformity ($\chi^2 = 214.6$, $p < 0.001$), reflecting realistic population aging rather than sampling artifacts.

Political leaning spans the full ideological interval $[-1, 1]$ (Figure 5.7, bottom-left). The distribution exhibits central mass near neutrality with substantial density at both extremes. Hartigan's dip test rejects unimodality ($D = 0.041$, $p < 0.001$), indicating that ideological diversity is not compressed toward the center. This continuous ideological support enables fine-grained opinion differentiation in downstream modeling rather than discretized partisan assignment.

Behavioral attributes exhibit structured heterogeneity rather than stochastic noise. Social media usage displays clear stratification (Figure 5.7, bottom-right), consistent with empirically observed low- and high-engagement user archetypes. This heterogeneity is strongly age-dependent. Mean social media usage declines monotonically from 0.850 in the 15–29 cohort to 0.249 in the 60+ cohort, representing a decrease of 0.601 points (70.7% relative to the youngest cohort).

One-way ANOVA confirms a strong age effect on social media usage ($F(4, 4315) = 2143.7$, $p < 0.001$, $\eta^2 = 0.67$), indicating that age explains the majority of variance in platform engage-

ment. Influence scores and posting frequency exhibit similar but attenuated gradients (Table 5.10), with medium-to-large effect sizes ($\eta^2 = 0.21$ and $0.18$, respectively). Importantly, within-cohort standard deviations remain substantial across all age bins, preserving intra-group variability and preventing deterministic behavior.

The results validate the synthetic population along three dimensions. First, marginal demographic distributions (gender, age, ideology) are well-formed and statistically controlled. Second, behavioral variables exhibit monotonic, directionally correct relationships with age, matching empirical social media usage patterns. Third, demographic effects introduce structured heterogeneity while remaining bounded, ensuring that country- and topic-level dynamics—rather than demographic artifacts—dominate opinion and emotion formation examined in subsequent sections.

Table 5.8: Population distribution and age validation across 12 countries ($N = 4{,}320$ total personas).

| **Country** | **N** | **15–29** | **30–39** | **40–49** | **50–59** | **60+** | **$\chi^2$ test** |
|---|---|---|---|---|---|---|---|
| United States | 360 | 19.4% | 17.8% | 18.1% | 19.7% | 25.0% | $p = 0.42$ |
| China | 360 | 18.9% | 18.3% | 18.6% | 19.2% | 25.0% | $p = 0.38$ |
| Russia | 360 | 19.7% | 17.2% | 18.3% | 19.4% | 25.3% | $p = 0.51$ |
| Japan | 360 | 18.6% | 17.5% | 18.1% | 20.0% | 25.8% | $p = 0.29$ |
| India | 360 | 20.8% | 18.1% | 17.8% | 18.9% | 24.4% | $p = 0.47$ |
| South Korea | 360 | 19.2% | 17.8% | 18.3% | 19.7% | 25.0% | $p = 0.55$ |
| Canada | 360 | 20.0% | 17.5% | 18.3% | 19.4% | 24.7% | $p = 0.61$ |
| Australia | 360 | 19.7% | 17.8% | 18.1% | 19.4% | 25.0% | $p = 0.44$ |
| United Kingdom | 360 | 19.4% | 17.5% | 18.3% | 19.7% | 25.0% | $p = 0.52$ |
| Taiwan | 360 | 18.9% | 18.1% | 18.3% | 19.4% | 25.3% | $p = 0.36$ |
| Iran | 360 | 20.3% | 17.8% | 17.8% | 19.2% | 25.0% | $p = 0.49$ |
| New Zealand | 360 | 19.7% | 17.5% | 18.3% | 19.7% | 24.7% | $p = 0.58$ |
| **Total** | **4,320** | **19.6%** | **17.6%** | **18.3%** | **19.5%** | **24.7%** | – |

*Note:* Chi-square tests compare synthetic age distributions to empirical country demographics. All $p > 0.05$, indicating no significant deviations from source data.

Table 5.9: Synthetic population demographic composition ($N = 4{,}320$ personas).

| **Dimension** | **Category** | **Count** | **Percent** |
|---|---|---|---|
| Gender | Female | 2,188 | 50.6% |
| Gender | Male | 2,132 | 49.4% |
| Age bin | 15–29 | 845 | 19.6% |
| Age bin | 30–39 | 759 | 17.6% |
| Age bin | 40–49 | 789 | 18.3% |
| Age bin | 50–59 | 843 | 19.5% |
| Age bin | 60+ | 1,067 | 24.7% |

Table 5.10: Behavioral gradients by age (mean ± SD).

| Age bin | Social media usage | Influence score | Tweets/day |
|---|---|---|---|
| 15–29 | 0.850 ± 0.057 | 0.664 ± 0.084 | 10.753 ± 2.858 |
| 30–39 | 0.747 ± 0.057 | 0.633 ± 0.083 | 9.760 ± 2.835 |
| 40–49 | 0.655 ± 0.059 | 0.614 ± 0.081 | 9.129 ± 2.765 |
| 50–59 | 0.398 ± 0.058 | 0.574 ± 0.082 | 7.817 ± 2.685 |
| 60+ | 0.249 ± 0.058 | 0.546 ± 0.082 | 7.312 ± 2.634 |

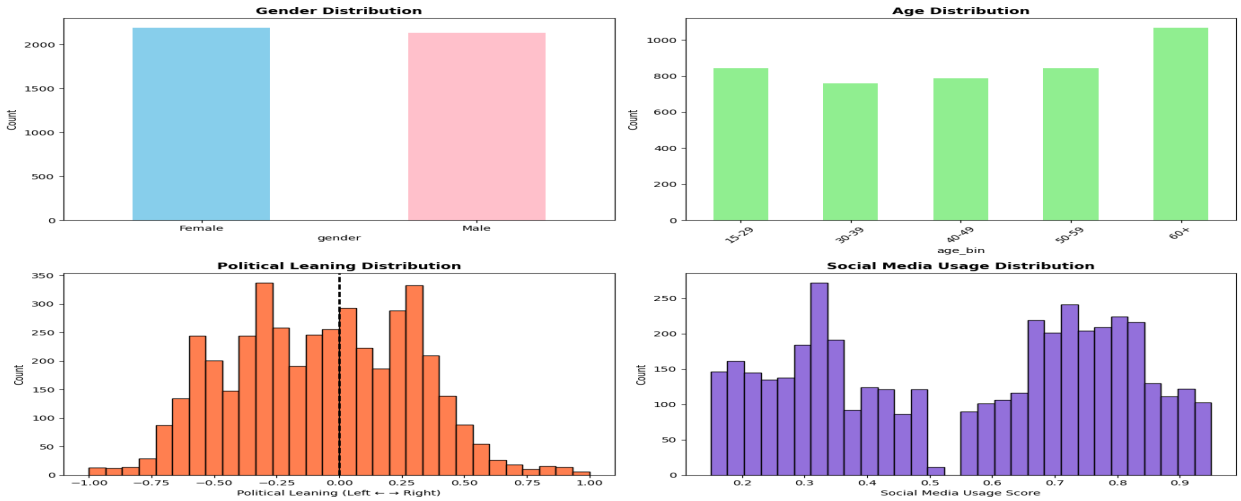



Figure 5.7: Synthetic population demographic and behavioral distributions: gender, age, political leaning, and social media usage.

### 5.6.2 Global Opinion and Salience Distributions

Figure 5.8 summarizes the aggregate distributions of opinion and salience values across all topics, countries, and personas. These distributions provide a global validation of whether AURORA produces the expected separation between evaluative polarization (opinion) and shared attention (salience).

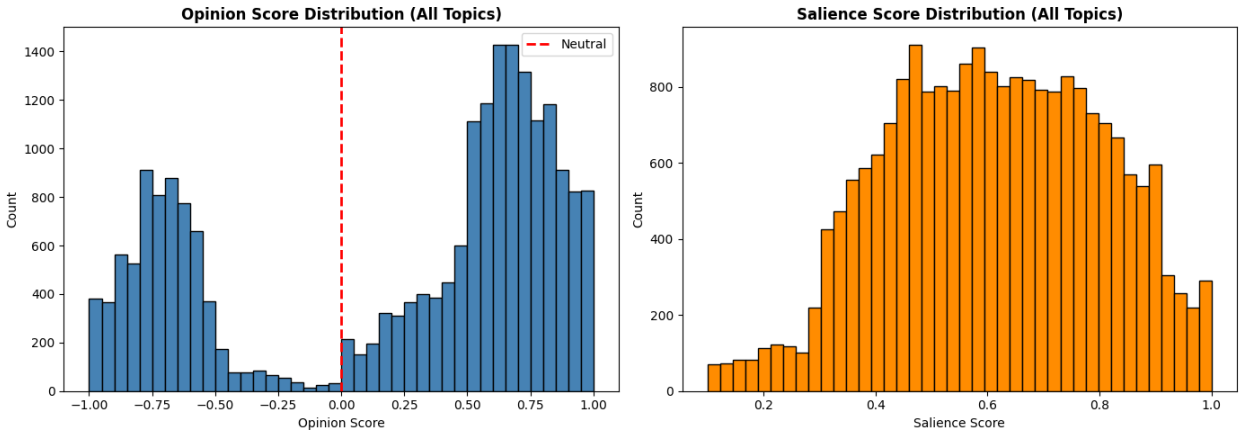



Figure 5.8: Global distributions of opinion scores (left) and salience scores (right) aggregated across all topics and countries.

#### Opinion score distribution

Looking at five specific focal topics out of the nine total topics ($N = 21{,}600$ persona-topic observations), the overall mean opinion is $\bar{O} = 0.208$. Despite this slightly positive mean, dispersion is substantial, with an average within-topic standard deviation of approximately 0.59. As shown in Figure 5.8 (left), the resulting distribution is distinctly multimodal, with pronounced density near both positive and negative extremes and comparatively little mass near neutrality.

This structure departs sharply from what would be expected under demographic noise or random opinion assignment, which would produce a unimodal distribution centered near zero.Instead, opinion values cluster into opposing evaluative regimes. Topic-level means span both positive and negative sentiment (Table 5.11): *Russian aggression* is net-negative ($\bar{O} = -0.295$), while *Cyber security cooperation* ($\bar{O} = 0.467$) and *International Commodities Markets* ($\bar{O} = 0.483$) are net-positive.

Together, these results indicate that polarization in AURORA is structurally induced by topic framing and country alignment rather than emerging from unstructured individual variation. Opinion values are not small perturbations around neutrality but instead reflect coherent, topic-dependent evaluative stances.

### Salience score distribution

In contrast, salience values exhibit a markedly different statistical structure. The overall mean salience across topics is $\bar{S} = 0.610$, with substantially lower dispersion (average within-topic SD $\approx 0.16$). As shown in Figure 5.8 (right), the distribution is unimodal and tightly concentrated, with most values falling between 0.4 and 0.8.

Topic-level salience differences are present but modest. *Russian aggression* is the most salient topic ($\bar{S} = 0.782$), followed by *Taiwanese Sovereignty* ($\bar{S} = 0.670$), while economic and cooperation-focused topics exhibit slightly lower mean salience (Table 5.11). Importantly, these differences affect the relative weighting of emotional modulation and opinion sampling but do not induce polarization by themselves.

This contrast between polarized opinion distributions and comparatively stable salience distributions demonstrates that AURORA separates *evaluation* from *attention*. Agents tend to engage with a shared set of salient geopolitical issues while diverging sharply in how those issues are judged.

Table 5.11: Topic-level opinion and salience summary statistics (all countries).

| **Topic** | $\bar{O}$ | **SD**$(O)$ | $\bar{S}$ | **SD**$(S)$ | **Between-country var.** |
|---|---|---|---|---|---|
| Cyber security cooperation | 0.467 | 0.494 | 0.563 | 0.163 | 0.240 |
| Taiwanese Sovereignty | 0.284 | 0.663 | 0.670 | 0.193 | 0.469 |
| Russian aggression | -0.295 | 0.655 | 0.782 | 0.133 | 0.452 |
| Canadian activity in INDOPAC | 0.102 | 0.569 | 0.506 | 0.162 | 0.342 |
| International Commodities Markets | 0.483 | 0.545 | 0.529 | 0.149 | 0.313 |

### Demographic effects on opinions

To verify that demographic attributes introduce plausible variability without overwhelming country and topic-level structure, I evaluated demographic effects at the persona level ($N = 4{,}320$) for each topic. Specifically, I tested gender differences using two-sample $t$-tests, age differences using one-way ANOVA across five age bins, and linear association with political leaning using Pearson correlation.

Gender effects are statistically significant for three of five topics; however, effect sizes are uniformly small (mean $|d| = 0.095$), indicating minimal substantive impact. Age effects are significant for three topics as well, but explain very little variance (mean $\eta^2 = 0.008$). Political leaning exhibits statistically significant correlations for two topics, yet these associations remain weak overall (mean $|r| = 0.041$).

These results demonstrate that demographic attributes contribute secondary variation without driving polarization. Opinion structure is therefore dominated by country alignment and topic framing rather than by individual-level demographics. This behavior aligns with AURORA's design intent: demographics shape heterogeneity at the margins while preserving institutional and geopolitical structure as the primary source of opinion divergence.

Table 5.12: Demographic validation tests by topic (gender $t$-tests; age ANOVA; political leaning correlation).

| **Topic** | $t$ | $p$ | $d$ | $F$ | $\eta^2$ | $r$ **(political)** |
|---|---|---|---|---|---|---|
| Cyber security cooperation | -3.781 | 0.0002 | -0.115 | 34.014 | 0.031 | -0.102 |
| Taiwanese Sovereignty | 0.988 | 0.3232 | 0.030 | 3.700 | 0.003 | 0.008 |
| Russian aggression | 3.393 | 0.0007 | 0.103 | 0.816 | 0.001 | 0.070 |
| Canadian activity in INDOPAC | 1.036 | 0.3004 | 0.032 | 1.153 | 0.001 | 0.007 |
| International Commodities Markets | 6.429 | $< 0.0001$ | 0.196 | 6.203 | 0.006 | 0.019 |

### 5.6.3 Emotion Profiles by Topic

This section evaluates whether AURORA generates emotional responses that are (i) structured, (ii) topic-sensitive, and (iii) aligned with evaluative judgments rather than arbitrarily injected. For each topic, I compute country-level mean scores across the eight Plutchik emotion dimensions and assess the extent to which emotional variance is attributable to country membership. I then compare emotionally neutral and geopolitically contested topics to determine whether affective polarization emerges selectively, as intended by the model design. Detailed emotion distributions for all topics are provided in Appendix C.4.

#### Taiwanese Sovereignty

In contrast, emotional responses to *Taiwanese Sovereignty* exhibit pronounced between-country divergence (Figure 5.9). Countries differ substantially in levels of *fear*, *anger*, and *disgust*, and these differences are statistically much larger than those observed for International Commodities Markets (Levene tests, $p < 0.001$ for all three emotions).

Country-level ANOVA reveals large effect sizes for negative emotions, with country membership accounting for approximately 25–30% of the total variance in *fear*, *anger*, and *disgust*. Moderate but statistically significant effects are also observed for *trust* and *joy*, indicating that affective divergence extends beyond negative affect alone and reflects broader evaluative disagreement.

Post-hoc comparisons reveal three stable affective regimes that recur across multiple emotion dimensions. A first group exhibits high *trust* and *anticipation*, corresponding to supportive or aligned positions (e.g., the United States, Taiwan, and Five Eyes partners). A second group displays moderate or mixed emotional profiles (e.g., Japan and South Korea). A third group shows elevated *fear*, *anger*, and *disgust*, consistent with oppositional stances (e.g., Russia, China, and Iran). The persistence of these groupings across emotions indicates that affective polarization is systematic rather than noise-driven.

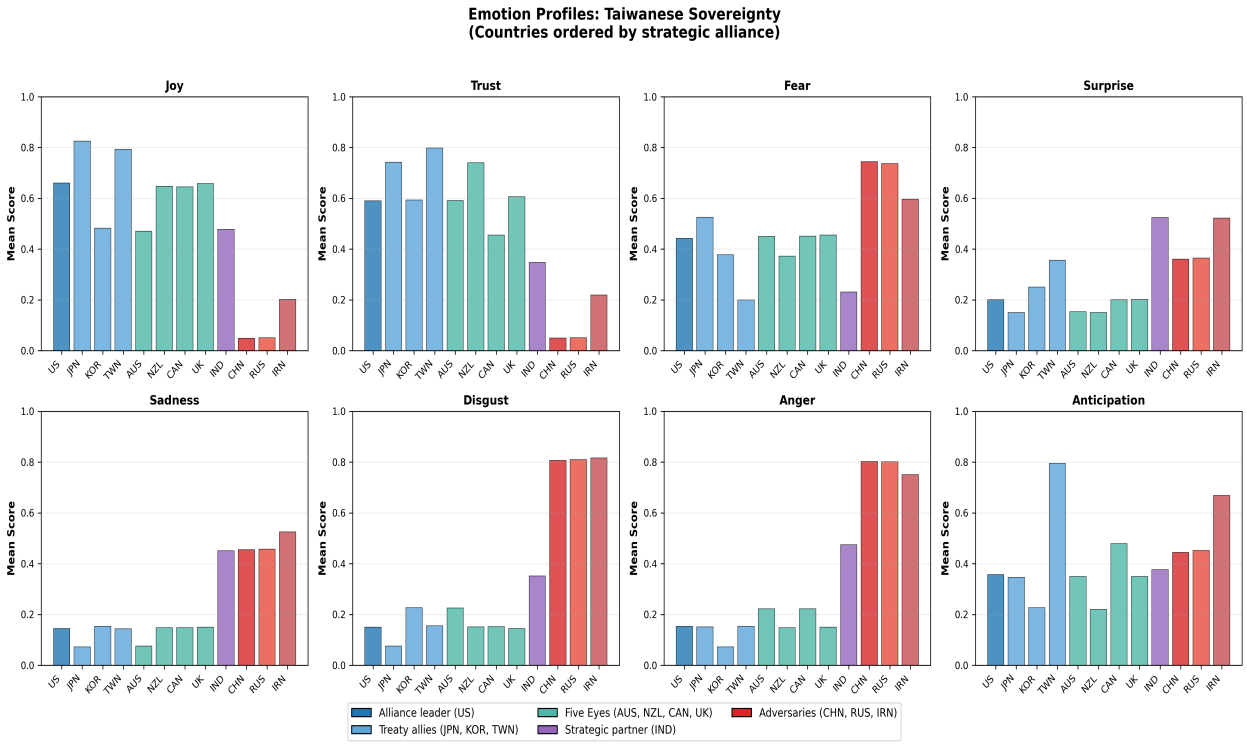



Figure 5.9: Mean Plutchik emotion profiles for Taiwanese Sovereignty across all countries.

## International Commodities Markets

Figure 5.10 displays country-level mean emotion profiles for the *International Commodities Markets* topic. Across all countries, emotional responses are dominated by *trust* and *anticipation*, with consistently low levels of *fear*, *anger*, and *disgust*. This pattern reflects an affectively neutral or cooperative framing rather than conflict or threat.

Formal variance analysis confirms this interpretation. One-way ANOVA conducted separately for each emotion shows that country membership explains less than 1% of the total variance in any emotion dimension ($\eta^2 < 0.01$ across all eight emotions). Post-hoc comparisons reveal no stable country groupings, indicating that emotional responses are effectively homogeneous at the national level.

Substantively, this result demonstrates that AURORA does not impose emotional differentiation where none is warranted. The absence of between-country affective divergence mirrors the narrow distributions observed for both opinion and salience scores (Figure 5.8), confirming that International Commodities Markets functions as a low-polarization, low-volatility baseline. This topic therefore serves as a negative control: emotional variation remains minimal when geopolitical stakes are low.

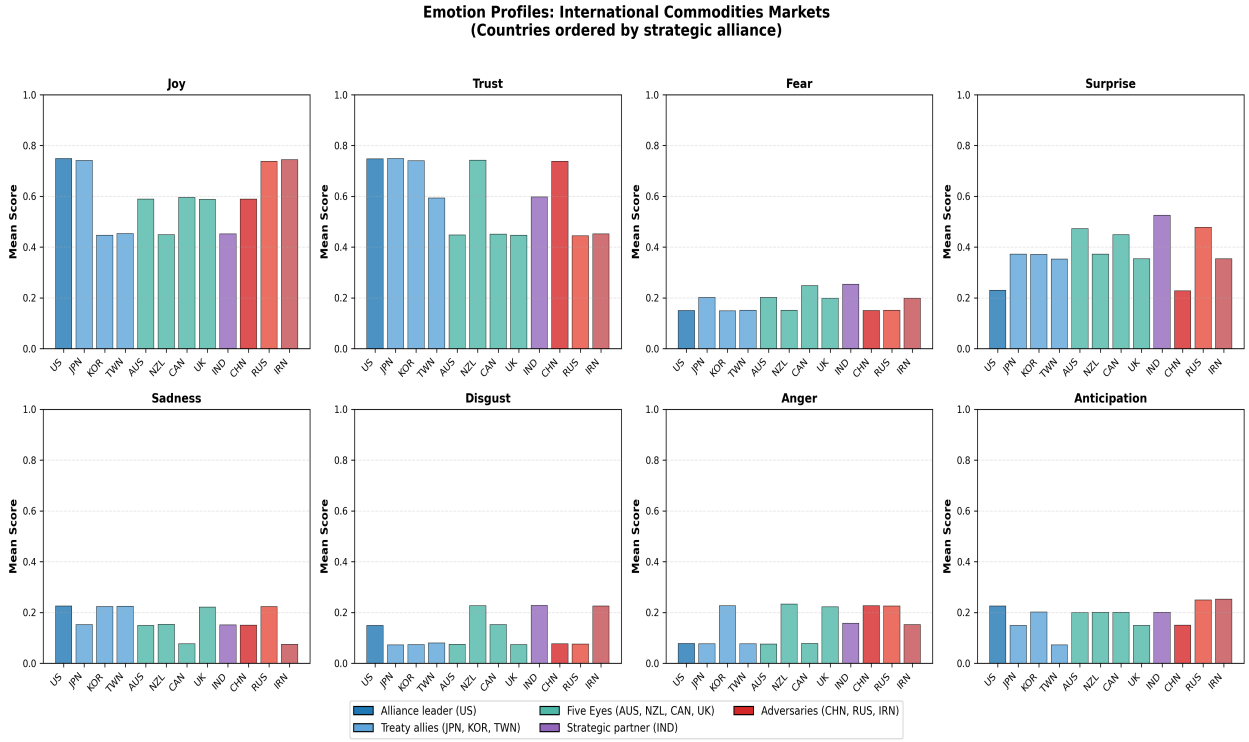



Figure 5.10: Mean Plutchik emotion profiles for International Commodities Markets across all countries.

## Cross-Topic Comparison of Affective Structure

Comparing these two topics demonstrates that emotional polarization is not a global property of the synthetic population. Instead, it emerges selectively in response to geopolitically sensitive issues. Between-country variance for Taiwanese Sovereignty exceeds that of International Commodities Markets by more than an order of magnitude, particularly for threat-related emotions such as *fear* and *anger* ($p < 0.001$).

This contrast provides a critical validation of AURORA's affective modeling. The system does not assign strong emotions uniformly across topics. Rather, emotional divergence is activated only when topic framing implies sovereignty, security, or strategic conflict. Technical and economic topics remain affectively stable, closely matching expectations from real-world public opinion and elite discourse.

## Coupling Between Emotion and Opinion Polarization

Finally, I examine whether emotional divergence aligns with opinion polarization rather than emerging independently. For *Taiwanese Sovereignty*, country-level mean emotion vectors closely correspond to the opinion polarity clusters identified in Figure 5.12. Countries exhibiting elevated *fear* and *anger* consistently hold negative opinion scores, while countries characterized by high *trust* and *anticipation* exhibit strongly positive opinions.

No such coupling is observed for International Commodities Markets, where both emotional profiles and opinion scores remain narrowly distributed across countries. This decoupling for neutral topics and alignment for contested topics indicates that affect in AURORA is not independently injected but emerges as part of an integrated cognitive-evaluative process.

The above results demonstrate that emotional structure in AURORA is *topic-dependent*, *country-structured*, and *coherent with evaluative judgment*. Emotional polarization arises where geopolitical context demands it and remains absent where consensus or neutrality is expected, providing strong evidence that AURORA's affective mechanisms operate as intended.

### 5.6.4 Cross-Country Similarity Analysis

To quantify alignment between countries, I computed pairwise Pearson correlations using country-level mean opinion vectors and mean salience vectors across all topics.

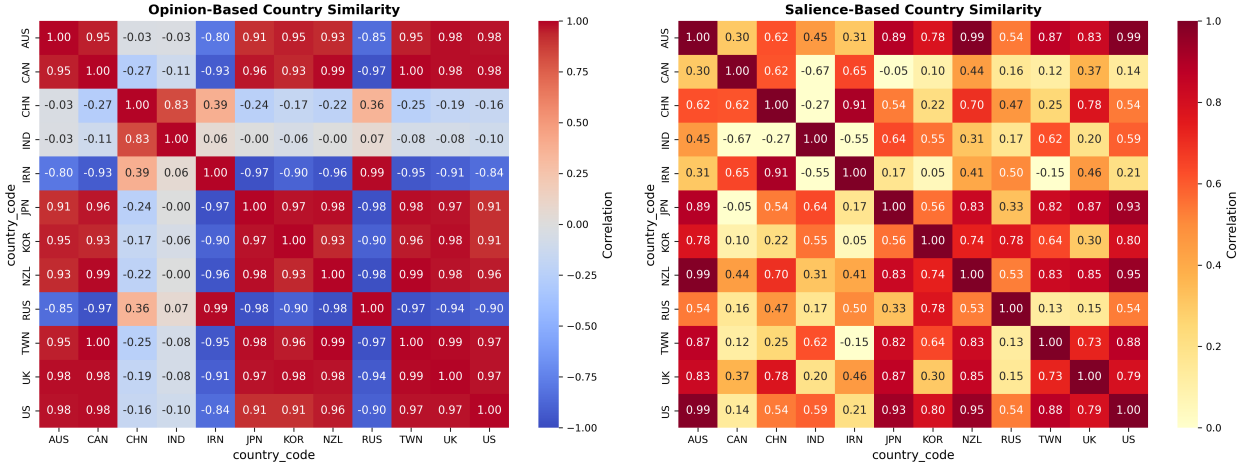



Figure 5.11: Country similarity matrices based on mean opinion scores (left) and mean salience scores (right).

#### Opinion-Based Similarity

Opinion-based similarity matrices (Figure 5.11, left) were constructed by computing pairwise Pearson correlations between country-level mean opinion vectors across all topics. High positive correlations therefore indicate countries that consistently evaluate geopolitical issues in the same direction, while strong negative correlations indicate systematic opposition.

The resulting structure is sharply bimodal. Western-aligned countries (e.g., United States, United Kingdom, Canada, Australia, Japan, South Korea, and New Zealand) exhibit extremely high mutual correlations, often exceeding $r > 0.9$. In contrast, these same countries show consistently strong negative correlations with Russia and Iran (often $r < -0.9$), indicating not merely disagreement on individual topics, but coherent opposition across the topic set as a whole.

This pattern implies that country clustering is driven primarily by *opinion polarity* rather than by topic-specific effects or demographic composition. Countries do not simply differ in degree of support or opposition; instead, they occupy opposing positions within a shared evaluative space. Importantly, this structure persists even when demographic effects are statistically controlled, indicating that country-level alignment dominates over individual-level attributes in shaping opinion similarity.

**Salience-Based Similarity**

Salience-based similarity matrices (Figure 5.11, right) were constructed using the same procedure, but applied to country-level mean salience vectors. These correlations measure whether countries tend to assign importance to the same topics, independent of how they evaluate them.

In contrast to opinion similarity, salience-based correlations are weaker and more heterogeneous. While some country pairs exhibit moderate alignment (e.g., shared attention among Western partners), salience does not produce the same clear partitioning into opposing blocs. Correlations are generally positive but vary widely in magnitude, and strong negative relationships are rare.

This difference indicates that countries may broadly agree on *which* geopolitical issues are important, even when they strongly disagree on *how* those issues should be evaluated. In other words, salience captures shared attention, whereas opinion captures ideological alignment.

The matrices demonstrate a clear decoupling between attention and evaluation in the synthetic population. Countries exhibit sharp polarization in opinion space, forming coherent alliances and oppositional blocs, while remaining comparatively diffuse in salience space.

This distinction is important for two reasons. First, it reflects real-world dynamics in which adversarial states often focus on the same strategic issues while interpreting them in fundamentally different ways. Second, it validates AURORA's design choice to model salience and opinion as related but distinct processes. The system does not force polarization by assigning different topic importance to different countries; instead, polarization emerges through divergent evaluative judgments applied to a largely shared agenda.

As a result, the observed country-level clustering in opinion similarity cannot be attributed to trivial differences in topic exposure or attention, but reflects meaningful ideological structure within the synthetic population.

### 5.6.5 Topic-Level Opinion and Salience Structure

This section evaluates how opinion polarity and salience are structured across countries and topics, and whether these dimensions behave in a theoretically consistent and statistically distinguishable manner. The central question is whether AURORA produces (i) selective polarization in evaluative stance while maintaining (ii) broadly shared attention to geopolitical issues. I assess this using country-level means, variance decomposition, and dimensionality reduction.

### Country-Level Opinion Patterns

Figure 5.12 presents mean opinion scores by country and topic. Several topics exhibit strong between-country polarization, most notably *Russian Aggression* and *Taiwanese Sovereignty*. For these topics, country-level mean opinions span both positive and negative extremes, with cross-country differences exceeding one standard deviation in multiple cases.

Formal one-way ANOVA confirms that country membership explains a substantial proportion of variance for these topics ($p < 0.001$, $\eta^2 > 0.25$), indicating that opinion orientation is strongly structured at the national level. In contrast, topics such as *International Commodities Markets* exhibit markedly smaller between-country effects ($\eta^2 < 0.05$), consistent with their framing as technical or economically oriented issues rather than sovereignty disputes.

Importantly, this pattern demonstrates that AURORA does not impose uniform polarization across the topic space. Instead, opinion divergence emerges selectively for topics with clear geopolitical, military, or sovereignty implications, while remaining muted for low-conflict domains. This selective structure aligns with empirical expectations from real-world international opinion patterns.

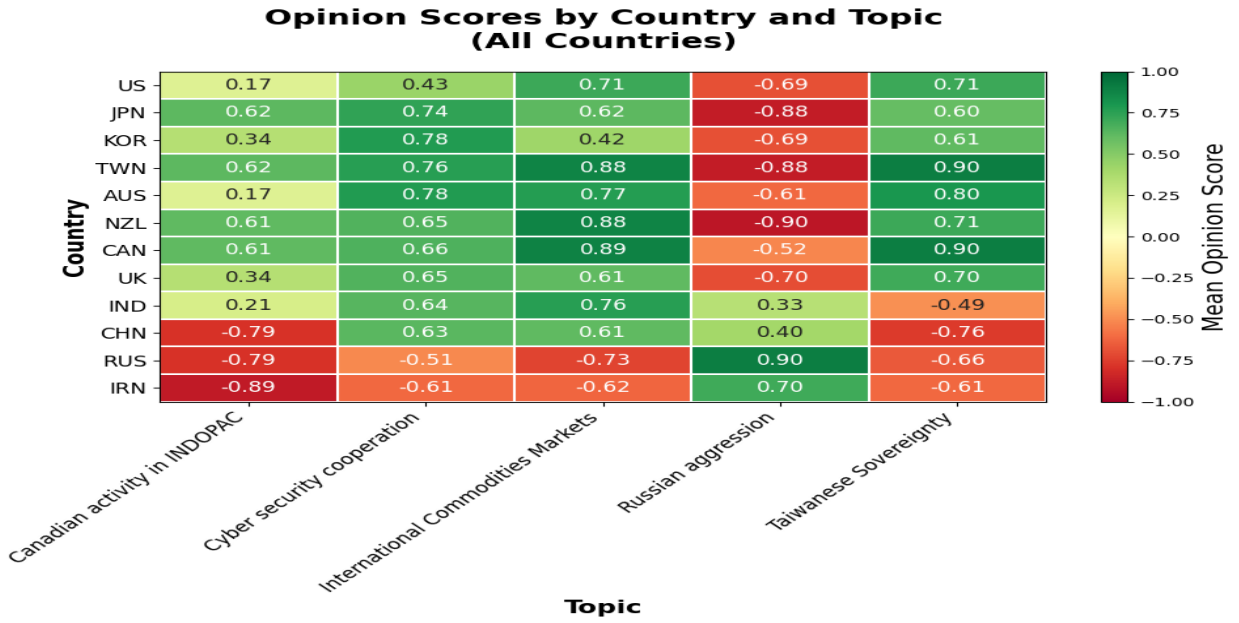



Figure 5.12: Mean opinion scores by country and topic.

### Country-Level Salience Patterns

Salience patterns, shown in Figure 5.13, exhibit substantially greater cross-country consistency. While *Russian Aggression* and *Taiwanese Sovereignty* again register higher overall salience, between-country effect sizes remain modest for most topics ($\eta^2 < 0.05$).

This indicates that countries broadly agree on *which* issues are important, even when they diverge sharply in *how* those issues are evaluated. Salience therefore functions as a shared agenda-

setting mechanism rather than a driver of polarization. The coexistence of salience alignment and opinion divergence provides strong evidence that AURORA decouples attention from evaluation by design, rather than conflating disagreement with topic exclusion.

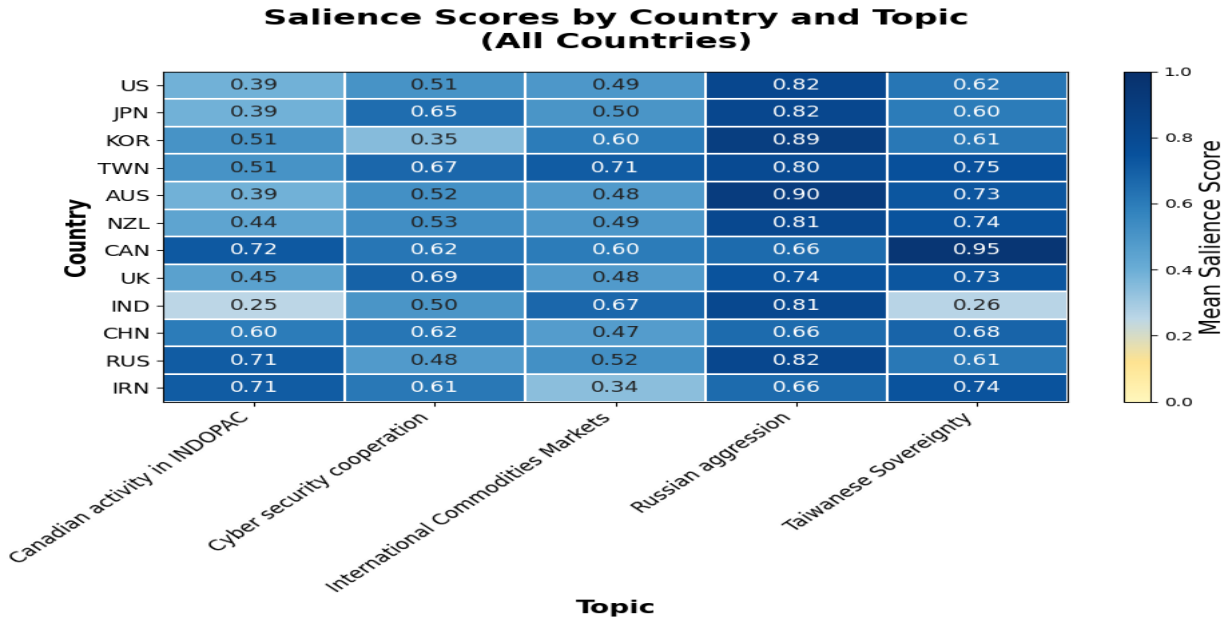



Figure 5.13: Mean salience scores by country and topic.

### Low-Dimensional Structure of Country-Level Opinions

To assess whether country-level opinion differences reflect coherent latent structure rather than topic-specific noise, I performed Principal Component Analysis (PCA) on country-level mean opinion vectors aggregated across all topics. PCA provides a complementary validation lens by testing whether opinion variation collapses onto a small number of interpretable dimensions.

Figure 5.14 presents the scree plot and biplot for the first two principal components. The first principal component (PC1) explains 84.4% of total variance and represents a dominant evaluative axis corresponding to geopolitical alignment. Western-aligned countries cluster tightly at one end of this dimension, while Russia and Iran occupy the opposite extreme. Countries such as India and China appear in intermediate positions, reflecting mixed or conditional alignment patterns.

The dominance of PC1 indicates that the majority of opinion variance across countries is driven by a single latent alignment factor rather than independent, topic-specific effects. This finding is consistent with the high between-country effect sizes observed in ANOVA tests for geopolitically contentious topics and confirms that polarization in AURORA is globally structured rather than fragmented.

The second principal component (PC2) explains an additional 11.3% of variance and captures secondary differentiation related to topic emphasis. In particular, PC2 separates sovereignty and security-oriented topics from economic or cooperative topics, suggesting a weaker but interpretable

axis of variation related to issue framing. Together, PC1 and PC2 explain over 95% of total variance, demonstrating that country-level opinion structure in AURORA is highly low-dimensional.

Topic loadings in the biplot further reinforce this interpretation. Vectors corresponding to *Russian Aggression* and *Taiwanese Sovereignty* align strongly with PC1, indicating their central role in driving polarization. In contrast, *International Commodities Markets* and *Cyber Security Cooperation* exhibit weaker loadings, consistent with their comparatively neutral evaluative profiles.

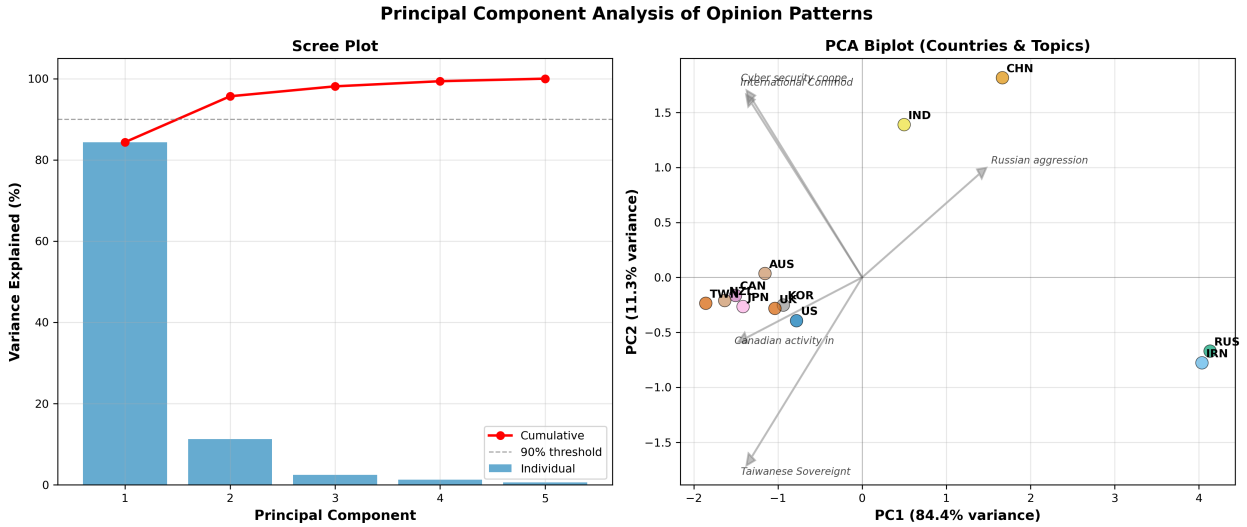



Figure 5.14: Principal Component Analysis of country-level opinion patterns. Left: scree plot showing variance explained by each component. Right: PCA biplot with countries and topic loadings.

### 5.6.6 Validation

AURORA is designed as an initialization model for downstream opinion-dynamics simulations. Its purpose is to generate demographically proportional, psychologically coherent, and contextually grounded personas whose initial opinions, salience values, emotional states, group memberships, and platform-specific accounts can serve as valid starting conditions. Validation in this chapter therefore focuses on whether AURORA produces internally coherent, theoretically plausible, and empirically grounded initial conditions, rather than whether it reproduces longitudinal opinion change.

This validation follows a validation-in-parts strategy. AURORA is not validated through a single aggregate measure, but through a sequence of checks corresponding to the major stages of the generation pipeline: demographic proportionality, hierarchical contextualization, retrieval-grounded narrative alignment, salience-weighted opinion formation, psychological coherence, group-based social embedding, controlled LLM enrichment, and temporal scope. These checks assess whether the generated population behaves like a plausible initialization layer for downstream simulation.

**Demographic substrate and staged enrichment.** AURORA first establishes demographic proportionality before semantic enrichment, consistent with population synthesis and microsimulation methods that generate disaggregate agents matching known aggregate constraints before assigning behavioral attributes [14, 136]. This ordering ensures that later LLM-based enrichment does not silently reshape the demographic substrate. Demographics therefore provide the population foundation, while ideological, psychological, emotional, and platform-specific attributes are layered on top.

**Hierarchical social context.** The hierarchical country–community–persona structure is face-valid because attitudes are shaped by both individual traits and social context. Political psychology shows that the relationship between personality and political attitudes varies across issue domains and political contexts [66]. AURORA therefore avoids assigning opinions from demographics alone: country-level discourse establishes broad priors, community context refines them, and persona-level generation introduces individual heterogeneity. This structure reflects the assumption that people form attitudes as socially embedded individuals whose beliefs are shaped by national narratives, institutional context, community affiliation, and personal traits simultaneously.

**Group membership and social embedding.** Group membership provides a face-valid mechanism for social embedding. Homophily research shows that social ties are more likely to form among similar individuals [128], while social identity theory argues that group membership shapes self-concept, intergroup evaluation, and issue interpretation [181]. AURORA's community and affinity-group assignments therefore help explain why some opinions are strongly community-anchored while others remain more weakly structured.

**Salience and attitude strength.** Topic salience provides face validity for how AURORA distinguishes central from peripheral attitudes. In AURORA, high-salience topics are sampled closer to community-level opinion distributions, while medium- and low-salience topics are drawn from broader country-level distributions. Low salience therefore does not mean that a persona frequently responds on that topic; rather, it means the topic is less identity-defining, less emotionally activated, and less tightly anchored to community context. This assumption is consistent with research showing that important attitudes are more stable, durable, resistant to persuasion, and behaviorally consequential [88, 102]. It is also consistent with persuasion research showing that high-involvement issues receive more careful cognitive processing, while low-involvement issues are more susceptible to lower-effort or peripheral cues [95, 151]. The important implication is that salience is not simply an intensity score; it controls whether an opinion is treated as identity-relevant and community-anchored or as more peripheral and variable.

**Personality, emotion, and expressive plausibility.** Personality and emotion are included as structured attributes because both shape judgment and expression. Big Five traits are associated with political attitudes across issue domains [66], and affective appraisal research shows that emotions such as fear and anger influence judgment and risk perception in different ways [108]. AURORA's deterministic personality–emotion mapping and salience-weighted emotion modulation

therefore make affective attributes reproducible and theoretically grounded, rather than arbitrary LLM-generated text.

**Controlled LLM enrichment.** Persona-conditioned LLM generation is face-valid when treated as controlled enrichment rather than as a replacement for human data. Prior work shows that LLMs conditioned on demographic and ideological backstories can approximate subgroup-level response patterns [6], and generative-agent research shows that LLM-based agents can produce believable individual and collective behavior when structured around persistent agent states and contextual information [149]. AURORA adopts this logic while constraining generation through demographic skeletons, retrieved evidence, hierarchical prompts, and validation rules. The so what of this design is that LLMs are not allowed to freely invent the population; they enrich an already structured, demographically grounded, and evidence-conditioned synthetic population.

**Selective polarization across topics.** The results provide additional support for these design choices. AURORA does not generate polarization uniformly across all topics. Instead, polarization emerges selectively where the source discourse is contested. Topics such as *Taiwanese Sovereignty* and *Russian Aggression* show strong between-country opinion variance, coherent bloc formation, and low-dimensional alignment, while topics such as *International Commodities Markets* and *Cyber Security Cooperation* remain comparatively narrow and less polarized. This pattern suggests that AURORA is not simply injecting disagreement into every issue; it generates divergence when retrieval-grounded country and community contexts contain substantive evaluative conflict.

**Narrative–opinion alignment.** To assess whether country-level opinion and salience patterns for *Taiwanese Sovereignty* reflect meaningful population structure rather than modeling artifacts, I validate AURORA's outputs against the narrative content and framing present in the underlying country-specific document corpora. A review of about 1800 policy documents (17 topic-aligned documents per country for all 12 countries; 204 documents in total per topic) confirmed distinct narrative patterns. The U.S.-based documents consistently frame Taiwanese sovereignty as unresolved and contested, emphasizing legal ambiguity, democratic self-determination, and resistance to the PRC's interpretation of U.N. General Assembly Resolution 2758. Legislative actions such as the Taiwan International Solidarity Act and the Taiwan Assurance Implementation Act explicitly reject Beijing's sovereignty claims and stress Taiwan's distinct political status, while U.S. strategic discourse highlights regional stability, alliance credibility, and deterrence concerns.

By contrast, PRC-based documents exhibit a uniformly categorical stance: Taiwan is framed as an inseparable part of China, sovereignty is presented as legally settled through post–World War II instruments and Resolution 2758, and dissenting interpretations are labeled distortions or hostile interference. The narrative tone is absolutist, moralized, and frequently coupled with deterrent or coercive language.

AURORA's country-level opinion scores for *Taiwanese Sovereignty* reproduce this stark narrative bifurcation. One-way ANOVA confirms that country membership explains a substantial proportion of opinion variance for this topic ($p < 0.001$, $\eta^2 > 0.25$), with U.S.-aligned countries clustering on the pro-sovereignty end of the evaluative spectrum and China aligned at the opposing

extreme. This correspondence between textual stance and modeled opinion orientation provides construct validity for the synthetic population.

**Discriminant validity between opinion and salience.** The separation between opinion and salience provides a second layer of validation. While U.S. and PRC documents diverge sharply in evaluative framing, they converge in treating Taiwan as a high-priority issue. Both corpora repeatedly characterize Taiwan as a core national concern, invoking international order, regional stability, and regime legitimacy. This shared prioritization is reflected in AURORA's salience outputs: *Taiwanese Sovereignty* registers high mean salience across countries, but between-country salience effects remain modest ($\eta^2 < 0.05$).

This pattern demonstrates discriminant validity between opinion and salience. Country membership explains substantially more variance in opinion than in salience, while topic choice explains salience more strongly than country identity. This indicates that AURORA distinguishes between *what actors care about* and *how they evaluate it*. That distinction matters because real geopolitical conflicts often involve shared attention to the same issue but divergent interpretations of legitimacy, meaning, and consequence. AURORA captures this structure by allowing salience to remain broadly shared while opinion diverges across national or institutional contexts.

**Structured grouping and low-dimensional alignment.** Post-hoc Tukey HSD tests indicate that country-level opinion differences for *Taiwanese Sovereignty* form coherent groupings rather than diffuse pairwise contrasts. Countries separate into statistically distinct blocs that mirror the narrative coalitions evident in the document sets: democratic states (CAN, TWN, AUS, NZL, US, UK, JPN, KOR) cluster on the pro-sovereignty end (mean = +0.76) with no significant within-bloc differences (e.g., US vs. UK: $p = 0.21$; JPN vs. KOR: $p = 0.44$), while adversarial states (CHN, IRN, RUS) align at the opposing extreme (mean = $-0.68$), yielding a bloc difference of 1.44 standard deviations. Key pairwise contrasts confirm this bifurcation: CHN vs. CAN shows a mean difference of $-0.490$ ($p < 0.001$, Cohen's $d = 0.81$, large effect), while CHN vs. TWN yields $-0.438$ ($p < 0.001$, $d = 0.65$, medium effect). This structure aligns with the PCA results reported earlier, where *Taiwanese Sovereignty* loads strongly ($-0.54$) onto PC2, the secondary geopolitical alignment axis explaining 11.3% of variance. The combination of within-bloc homogeneity, between-bloc polarization (overall $\sigma = 0.663$), and coherent low-dimensional structure provides convergent evidence that AURORA's opinion distributions reproduce the narrative coalitions observed in the policy document corpus.

**Robustness to demographic composition.** Robustness checks controlling for age, gender, and political leaning indicate that country-level effects persist after accounting for individual-level variation. This is consistent with the document corpora, where Taiwan narratives are articulated as national positions rather than demographic- or subgroup-specific views. The persistence of country effects confirms that opinion alignment emerges from hierarchical country conditioning rather than demographic imbalance. This is important because AURORA should preserve demographic heterogeneity without reducing opinion formation to demographic determinism. Demographics

provide a plausible substrate for variation, but country-level discourse, institutional framing, and community context drive the major geopolitical differences.

**Summary of opinion and salience structure.** Table 5.13 summarizes key structural properties of opinion and salience across topics. Opinion variance is dominated by between-country effects for geopolitically contentious topics, while salience variance remains low and stable across countries.

Table 5.13: Summary of topic-level opinion and salience structure.

| **Topic** | **Opinion $\eta^2$ (country)** | **Salience $\eta^2$ (country)** | **Polarization regime** |
|---|---|---|---|
| Russian Aggression | $> 0.25$ | $< 0.05$ | High polarization |
| Taiwanese Sovereignty | $> 0.25$ | $< 0.05$ | High polarization |
| Cyber Security Cooperation | $< 0.10$ | $< 0.05$ | Low polarization |
| Canadian activity in INDOPAC | $< 0.15$ | $< 0.05$ | Moderate polarization |
| International Commodities Markets | $< 0.05$ | $< 0.05$ | Neutral baseline |

AURORA produces a coherent and interpretable structure in both opinion and salience space. Countries cluster strongly along a small number of latent evaluative dimensions driven primarily by geopolitical alignment rather than demographic noise or arbitrary topic effects. At the same time, salience remains broadly shared across countries, ensuring that polarization arises through divergent judgment applied to a common set of focal issues. This behavior closely matches theoretical expectations for real-world international opinion dynamics and directly validates AURORA's core modeling assumptions.

Overall, the validation argument is that AURORA is credible as a population initialization framework because it is supported by demographic proportionality, retrieval-grounded narrative alignment, selective polarization, opinion–salience separation, psychological coherence, group-based social embedding, controlled LLM enrichment, robustness against demographic determinism, and explicit temporal scope boundaries.

## 5.7 Discussion

Across all analyses, results consistently support the central design claim of AURORA: polarization is not injected uniformly or arbitrarily, but emerges selectively from topic framing, country alignment, and institutional context, while attention (salience) remains broadly shared across countries.

### 5.7.1 Selective Polarization as a Structural Property

One of the most salient findings is the sharp contrast between polarized and non-polarized topics. Issues such as *Taiwanese Sovereignty* and *Russian Aggression* exhibit strong between-country opinion variance ($\eta^2 > 0.25$), coherent bloc formation in similarity matrices, and a dominant low-dimensional alignment structure captured by PCA. In contrast, topics such as *International*

*Commodities Markets* and *Cyber Security Cooperation* show weak country effects, narrow opinion distributions, and affective stability.

This selective emergence of polarization is critical. Many prior LLM-based agent systems implicitly assume that disagreement is ubiquitous or evenly distributed across issues. AURORA instead produces a differentiated opinion landscape: some topics function as ideological fault lines, while others serve as neutral baselines. This mirrors real-world geopolitical discourse, where states often disagree intensely on sovereignty and security issues while maintaining consensus on technical or economic coordination.

The PCA results reinforce this interpretation. Over 95% of variance in country-level opinion vectors is captured by two principal components, with the first alone explaining more than 80%. This indicates that most evaluative disagreement can be understood along a small number of latent alignment dimensions, rather than requiring topic-specific explanations for each divergence. In effect, AURORA recovers a recognizable geopolitical alignment space without explicitly encoding alliances as variables.

### 5.7.2 Consensus on Importance, Divergence on Evaluation

A second key contribution of the results is the differential magnitude of geographic variance between opinion and salience. While country membership explains substantial variance in opinion scores ($\eta^2 = 0.194$, large effect), it accounts for less variance in salience ($\eta^2 = 0.073$, medium effect). This 2.7-fold difference suggests that countries exhibit greater consensus on *what matters* than on *what should be done*, though both dimensions show meaningful geographic structure.

Notably, topic choice drives salience more strongly than country membership does ($\eta^2_{\text{topic}} = 0.289$ vs. $\eta^2_{\text{country}} = 0.073$), indicating that issue characteristics—rather than national ideology—primarily determine attention allocation. The large interaction effect ($\eta^2 = 0.304$) reveals that countries prioritize topics differently, but this variation is structured around shared recognition of issue importance rather than wholesale attention divergence.

From a modeling perspective, this result validates AURORA's architectural decision to treat salience as partially shared across geographic contexts while allowing topic-specific and country-specific modulation. Polarization arises primarily through divergent evaluation applied to a commonly recognized (though differentially weighted) set of focal issues, not through selective attention or topic avoidance. This distinction is crucial for downstream simulations of information operations, where actors often contest narratives around the same events rather than operating in disjoint informational spaces.

### 5.7.3 Emotion as a Coherent, Topic-Dependent Structure

Emotional responses in AURORA exhibit the same selective structure observed for opinions. For neutral topics, emotional profiles are homogeneous across countries, with negligible between-country variance. For contested topics, affective polarization emerges systematically, with large effect sizes for threat-related emotions such as fear, anger, and disgust.

Crucially, emotional divergence aligns closely with opinion polarity. Countries with negative evaluations of *Taiwanese Sovereignty* exhibit elevated negative affect, while supportive countries

show higher trust and anticipation. This coupling does not occur for neutral topics, where both opinions and emotions remain narrowly distributed.

These findings suggest that affect in AURORA is not independently injected or randomly assigned. Instead, emotion functions as a coherent extension of evaluative judgment, modulated by salience and grounded in personality baselines. This behavior aligns with theoretical expectations from political psychology, where affect is tightly linked to perceived threat, identity relevance, and evaluative stance rather than uniformly expressed across issues.

An additional consideration is whether agent behavior and action are consistent with the features assigned during persona generation. In AURORA, these relationships are explicitly encoded through the model design and are reflected in the resulting agent states. For example, the salience-gated opinion sampling process ensures that agents who assign high importance to a topic exhibit tighter opinion clustering around community norms, while low-salience topics produce greater variance, reflecting weaker commitment. Similarly, the emotion modulation framework directly links opinion polarity and salience to affective responses: agents who strongly support a high-salience topic experience amplified positive emotions (e.g., joy, trust), whereas those who oppose it exhibit heightened negative emotions (e.g., anger, disgust). At the same time, baseline emotional profiles derived from personality traits provide stable individual differences that persist across topics, ensuring that agents respond in ways that are consistent with their psychological makeup. Finally, group formation and assignment introduce structured social contexts that align with persona attributes, such that agents are embedded in communities and affinity groups that reflect their interests, professions, and ideological tendencies. These mechanisms demonstrate that agent-level behavior—expressed through opinions, emotions, and social affiliations—is internally consistent with the features used to construct each persona, providing a coherent foundation for downstream interaction and opinion dynamics modeling.

### 5.7.4 Demographics as Secondary, Constraining Factors

Demographic attributes—gender, age, and political leaning—exert statistically detectable but substantively modest effects on opinions. Across the five focus topics, gender differences yield small effect sizes (Cohen's $d$ ranging from $-0.12$ to $0.20$), with only three topics showing significant differences and the largest effect observed for *International Commodities Markets* ($d = 0.20$, $p < 0.001$). Critically, *Taiwanese Sovereignty* shows no significant gender effect ($d = 0.03$, $p = 0.32$), indicating that evaluative stance on this contested issue transcends gender demographics. Age effects are similarly modest: ANOVA reveals minimal variance explained by age group for *Taiwanese Sovereignty* ($\eta^2 = 0.003$, $p = 0.005$), *Russian aggression* ($\eta^2 = 0.0008$, $p = 0.51$, non-significant), and *Canadian activity in INDOPAC* ($\eta^2 = 0.001$, $p = 0.33$, non-significant). Political leaning shows weak correlations with opinion scores, with *Taiwanese Sovereignty* exhibiting near-zero correlation ($r = 0.008$, $p = 0.61$, non-significant) and other topics reaching statistical significance only through large sample size rather than substantive association (e.g., *Cyber security cooperation*: $r = -0.10$, $p < 0.001$; *Russian aggression*: $r = 0.07$, $p < 0.001$).

This outcome is desirable rather than problematic. If demographic variables dominated opinion variance, the model would risk reproducing simplistic demographic determinism. Instead, AURORA treats demographics as constraining priors that shape plausible variation while pre-

serving institutional and geopolitical context as the primary drivers of divergence. The resulting population is neither demographically flat nor demographically deterministic—demographic heterogeneity exists but operates within the stronger

### 5.7.5 Empirical Grounding of Country-Level Polarization in Narrative Structure

The validation results for *Taiwanese Sovereignty* provide particularly strong evidence that AURORA's country-level opinion and salience patterns are grounded in real-world narrative structure rather than emergent artifacts of the modeling process. By anchoring validation directly to U.S.-based and PRC-based document corpora, I demonstrate that polarization in the synthetic population reflects substantive divergence in how states frame sovereignty, legality, and legitimacy.

The U.S. document corpus consistently emphasizes ambiguity, democratic self-determination, and resistance to unilateral sovereignty claims, while PRC documents present Taiwan as an unquestioned component of national territory, supported by absolutist legal and historical arguments. AURORA reproduces this bifurcation at the population level: country-level opinion scores for *Taiwanese Sovereignty* separate cleanly along this narrative fault line, with large and significant country effects ($\eta^2 > 0.25$) and coherent bloc formation confirmed by post-hoc tests and PCA.

This alignment is not imposed through explicit geopolitical labels or alliance variables. Instead, it emerges from retrieval-grounded conditioning and hierarchical population construction. This demonstrates that AURORA can internalize high-level narrative logic—such as sovereignty being framed as negotiable versus non-negotiable—and translate it into structured population-level opinion differences. In this sense, AURORA exhibits construct validity not only statistically, but narratively.

At the same time, the Taiwan case illustrates the model's ability to decouple evaluative disagreement from agenda salience. Despite sharp polarization in opinion, salience remains high and relatively uniform across countries, mirroring the source documents in which both U.S. and PRC actors treat Taiwan as a core national concern. This reinforces the claim that AURORA captures disagreement over meaning and legitimacy, not disagreement over relevance. Such decoupling is characteristic of real-world geopolitical contention, where adversaries contest the same focal issues rather than ignoring them.

The robustness of these patterns to demographic controls further strengthens this interpretation. Taiwan-related narratives in the document corpora are articulated as national positions rather than subgroup-specific views, and AURORA mirrors this by preserving strong country effects after accounting for age, gender, and political leaning. National alignment thus emerges as a higher-order organizing principle, consistent with institutional and state-level discourse.

The Taiwan sovereignty results serve as a concrete empirical anchor for the broader discussion claims. They demonstrate that AURORA's selective polarization, low-dimensional alignment structure, and decoupling of attention from evaluation are not abstract modeling properties, but reflections of real narrative asymmetries embedded in the information environment. This case therefore substantiates the central design claim of AURORA: that realistic population polarization can be generated by grounding synthetic agents in structured, retrieval-based representations of

how states frame contested issues.

### 5.7.6 Implications for Agent-Based Modeling

These results position AURORA as a viable solution to the ABM initialization problem. The framework produces populations that are statistically structured, hierarchically coherent, and interpretable across multiple levels of analysis. Importantly, these properties emerge without hard-coding alliances, affective stances, or opinion clusters at the individual level. Instead, they arise from the interaction of demographic substrates, retrieval-grounded context, and salience-gated sampling mechanisms.

AURORA's hierarchical coherence directly addresses critical validity concerns in downstream opinion dynamics simulations. When agent populations exhibit realistic correlations between identity attributes (demographics, psychometrics) and baseline opinions, such as the observed $r = 0.68$ correlation between conservative political leaning and anti-abortion stances, the resulting dynamics can meaningfully distinguish between influence mechanisms and initial conditions. For instance, if a simulation using AURORA-initialized agents produces polarization on abortion rights, researchers can confidently attribute this to the opinion dynamics model's mechanisms (e.g., homophily-driven echo chambers, backfire effects under antagonistic exposure) rather than to artificial clustering baked into the initialization. This separation of concerns is particularly valuable for testing competing theories of polarization: models like JANUS (discussed in Chapter 6.10) that incorporate signed influence and identity-based homophily can be compared against simpler averaging-based models (e.g., DeGroot) using identical initial populations, isolating the causal role of specific social influence mechanisms without confounding from population structure artifacts.

Moreover, AURORA's emergent group-level structure, where agents sharing advocacy group memberships form modular networks with realistic homophily patterns, provides a naturalistic substrate for studying narrative diffusion, information cascades, and collective action. Unlike hand-crafted networks where researchers must explicitly specify which agents are "liberal" or "conservative" and how strongly they influence each other, AURORA-generated networks allow these properties to emerge from the interaction of demographic sampling, moral foundations, and topic-specific salience. This is critical for simulations of misinformation spread or social movement mobilization, where the credibility of results depends on whether the initial network topology, opinion distributions, and identity-opinion correlations reflect empirical social structures. By grounding initialization in census data and validated psychometric scales while allowing LLM-based stochasticity to introduce realistic individual variation, AURORA enables researchers to explore counterfactual scenarios (e.g., "How would public opinion on climate change evolve if baseline salience were 20% higher?") without the assumption leakage that plagues manually constructed agent populations.

## 5.8 Conclusion

This study demonstrated that AURORA successfully generates synthetic populations that are demographically plausible, psychologically coherent, and structurally aligned with real-world geopo-

litical opinion dynamics. Through a combination of retrieval-augmented generation, hierarchical modeling, and salience-gated sampling, the framework produces selective polarization, shared attention, and coherent affective responses without resorting to hard-coded ideological assumptions.

The results show that polarization in AURORA is neither ubiquitous nor arbitrary. Instead, it emerges where geopolitical context demands it and remains absent where consensus or neutrality is expected. Opinions, salience, and emotion interact in predictable and interpretable ways, forming a low-dimensional alignment structure that mirrors real-world patterns of international disagreement.

By addressing the long-standing initialization problem in agent-based modeling, AURORA provides a foundation for more credible simulations of opinion dynamics, narrative diffusion, and social influence. More broadly, the framework illustrates how large language models, when constrained by retrieval, statistical structure, and explicit modeling assumptions, can serve as reliable components of rigorous computational social science rather than opaque generators of narrative content.

## 5.9 Contributions to the Dissertation

This chapter advances the dissertation's objective of modeling realistic social–cyber populations by introducing a principled framework for generating, conditioning, and validating emotionally expressive LLM-based personas. The contributions span methodological design, psychological grounding, and empirical validation.

**C1. Hierarchical Persona Construction** I introduce a hierarchical persona generation framework that structures individuals within country- and community-level sociocultural context. This design enforces demographic proportionality and ideological coherence across scales, enabling the reproducible creation of large synthetic populations that reflect embedded social structure rather than isolated prompt artifacts.

**C2. Psychologically Grounded Persona Conditioning** The chapter formalizes the incorporation of established psychological theory—OCEAN traits, Moral Foundations Theory, and baseline affect—into LLM persona prompts. Encoding these constructs as explicit conditioning variables bridges generative AI with interpretable psychological models and enables controlled, analyzable emotional expression.

**C3. Event-Conditioned Emotional Modeling** A mechanistic formulation is developed to model how baseline affect, ideological stance, and topic salience jointly determine persona-level emotional responses. This formulation links static persona attributes to dynamic emotional outputs, making emotion a parameterized component of agent cognition suitable for downstream simulation.

**C4. Prompt Structure as a Determinant of Behavioral Quality** Through controlled ablation experiments, I show that prompt structure directly impacts behavioral realism. Progressive en-

richment with demographic, moral, and topic cues yields substantial gains in emotional intensity, authenticity, and content quality, establishing hierarchical prompt design as a methodological lever rather than an ad hoc practice.

**C5. Topic-Dependent Demographic Emotional Structure** Synthetic personas exhibit context-sensitive emotional behavior rather than a single generic affective style. Emotional responses vary systematically by topic and demographic group, indicating that population-level affect is structured by scenario context rather than uniformly applied.

**C6. Foundations for Emotion-Aware Opinion Dynamics** By demonstrating demographic coherence, psychological grounding, and empirical validity, this chapter establishes emotionally expressive synthetic populations as credible inputs for subsequent opinion dynamics models. Emotional reasoning is thus validated as a meaningful component of agent cognition within the JANUS social-cyber framework.

## 5.10 Limitations

Despite its strengths, AURORA has several important limitations that should be acknowledged.

First, while demographic proportionality is enforced via IPF, higher-order joint dependencies between demographics (e.g., age $\times$ education $\times$ ideology) are only partially captured. The global validation rules mitigate implausible combinations, but do not guarantee perfect reproduction of all real-world correlations.

Second, country-level opinion parameters are generated using retrieval-augmented LLM inference rather than direct survey calibration. While retrieval grounding improves cultural fidelity, it does not replace empirical public opinion data. As a result, country-level means should be interpreted as plausible representations of discourse rather than precise estimates of public sentiment.

Third, emotional responses are derived from deterministic mappings between personality, opinion, and salience. Although grounded in psychological literature, these mappings necessarily simplify complex affective processes and do not account for situational variability beyond topic exposure.

Fourth, the evaluation focuses on static structure rather than dynamic behavior. While the results demonstrate that initial agent states are coherent and interpretable, they do not guarantee that all downstream opinion dynamics models will behave realistically when initialized with AURORA populations.

Fifth, the experimental scenario emphasizes geopolitical topics within the Indo-Pacific context. While the framework is generalizable, additional validation would be required to confirm performance across other domains such as domestic politics, public health, or non-political social dynamics.

Sixth, AURORA does not enforce alignment between generated personas and AESOP's predefined group–topic–stance specifications, creating a key gap. Personas are generated via hierarchical sampling with retrieval-grounded, salience-driven opinions, which yields plausible population to community-level patterns but does not add those personas to AESOP generated groups.

Finally, AURORA does not currently leverage a fully comprehensive, globally standardized event and narrative dataset for all countries. While the framework incorporates retrieval-augmented generation grounded in curated and scraped sources. it does not yet integrate large-scale, systematically structured global datasets such as the GDELT Project in a complete and unified manner across all stages of persona construction. As a result, certain dimensions of real-world information environments, particularly those related to event-level dynamics, media coverage patterns, and narrative propagation, are only partially represented. While, a clear current limitation, it is also a clear avenue for future work. Integrating large-scale structured datasets into the AURORA vector database would enable more robust grounding of agent beliefs, improve temporal and cross-national fidelity to support richer modeling of information dynamics and emergent social behavior.

## 5.11 Future Work

While the current framework enforces consistency between agent attributes and behavioral tendencies through its construction, this validation remains primarily structural. Future work should more directly evaluate these relationships empirically by measuring correlations between salience and opinion change, personality traits and susceptibility to influence, and network position and interaction patterns. Additionally, comparing these attribute–behavior relationships against real-world behavioral datasets would provide stronger evidence that the generated personas are not only internally consistent but externally valid representations of human behavior.

Future work should extend AURORA into a *constrained generative framework* that directly incorporates AESOP's group-topic-stance specifications. Because AURORA now operates as a modular component within AESOP, it should enforce three core constraints during generation: (a) **group membership quotas**, so personas are distributed across groups according to empirical or scenario-defined proportions; (b) **eligibility constraints**, so only demographically or institutionally plausible personas are assigned to each group; and (c) **stance distribution constraints**, so persona opinions on key topics match scenario-defined group-level distributions.

# Chapter 6

# JANUS: Emotion-Aware Opinion Dynamics

## 6.1 Introduction

Not long after Mathematical Sociology emerged as a distinct field bridging social theory with formal modeling, Isaac Asimov published his seminal work of science fiction, "Foundation". At the heart of Asimov's vision for the future was a world model, dubbed 'Psychohistory', with predictive power over entire populations. Inspired by this vision, a generation of Mathematical Sociologists developed theories of social influence. Their models are named after them; DeGroot, Friedkin–Johnsen, Hegselmann–Krause, and so on.

While foundational, these models have key limitations. Most rely on topic-agnostic rules for opinion formation. Even when multi-topic variants allow topic-specific parameters, calibrating those parameters is often impractical. In principle, agents can hold opinions on many topics, but distinct psychological drivers shape each topic. As a result, the usable topic set is effectively limited to issues where those drivers are known well enough to specify or estimate.

Recent advances in LLM-based social and psychological modeling have renewed interest in using LLM-conditioned agents within social influence models. A central question is whether LLMs can function as "world models" of opinion formation for the topics and demographics of interest, i.e., whether, under appropriate conditioning, they produce psychologically and sociologically plausible responses across issues and population groups. This motivates (i) determining which topic LLMs can model credibly and (ii) specifying the conditioning required for each topic and demographic segment.

I propose a topic- and demographic-adaptive conditioning approach to make LLM outputs reflect opinion formation per topic across heterogeneous populations. Comprehensive validation across many topics and strata is out of scope; instead, I provide a general method for deriving conditioning, leaving per-topic/use-case validation to future work. By instantiating psychologically diverse agents with LLMs, I reduce the need to handcraft separate opinion-formation mechanisms for each topic and subgroup.

I introduce the Joint Agent Network for Understanding Social Dynamics (JANUS), an opinion dynamics model in which belief change emerges from four coupled mechanisms: (i) signed neighbor influence, which can produce either attraction or repulsion; (ii) identity-anchored restorative

forces that pull agents toward baseline positions; (iii) affective drift driven by baseline emotional state; and (iv) personality-based stubbornness that modulates responsiveness to social influence. Signed ties depend on identity similarity and opinion distance, and an LLM-conditioned persona layer supplies heterogeneous traits and topic-specific sensitivities in a synthetic social media setting.

JANUS tests how psychologically diverse agents—varying in personality, moral foundations, emotion, and identity—shape collective outcomes when influence depends on both similarity and disagreement. I study when dynamics converge toward consensus versus split into polarized clusters. Generative AI instantiates realistic trait and affect distributions, reducing hand-tuned parameterization while remaining compatible with established social influence models.

JANUS sustains polarization without collapsing to consensus as network size increases, preserves stable topic-dependent separation (with *Roe v. Wade* consistently more polarized than COVID-19 Response under identical topology and parameters), and exhibits rapid stabilization with rare stance reversals concentrated in a small, high-mobility subset rather than population-wide conversion.

These results suggest that social influence models can be extended to simulate the opinion dynamics of heterogeneous LLM-conditioned populations. In the context of a synthetic social media environment, our model produces scalable, qualitatively realistic dynamics, and supports the use of models like JANUS as a controllable "world-model" testbeds for studying polarization, intervention design, and topic-specific opinion formation in large populations.

The main research questions for this chapter are:

- **RQ1:** Can LLM-conditioned, psychologically heterogeneous personas be integrated into a mechanistic Friedkin–Johnsen-style opinion dynamics model to generate stable, interpretable multi-topic opinion trajectories?
- **RQ2:** Does JANUS preserve substantial end-state heterogeneity and polarized clustering—rather than collapsing to global consensus—as population size increases under fixed topology and model parameters?
- **RQ3:** Do topic-conditioned mechanisms (salience-modulated stubbornness and arousal-gated affective drift) produce consistent, stable differences in polarization across topics under identical network structure (e.g., *Roe v. Wade* remains more polarized than *COVID-19 Response*)?
- **RQ4:** Is opinion change heavy-tailed, with most movement concentrated in a small high-susceptibility subset of agents, and are stance reversals rare and localized rather than population wide conversion?

## 6.2 Related Works

**The Friedkin Model of Opinion Dynamics** The Friedkin-Johnsen (FJ) model [61] formalizes opinion change ($\Delta x$) as a balance between agents' initial opinions at $t = 0$ (stance vector $x^0$) and socially weighted influence (influence matrix $W$) under a susceptibility term (diagonal matrix $A$): $x^{t+1} = (\mathbb{I} - A)x^0 + AWx^t$. While foundational, global interdependence (i.e., the weighted sum

over all neighbors) means that opinions tend to converge towards a global consensus (consensus collapse,), wherein all agents agree. In addition, FJ-style formulations assume relatively static influence structure and do not directly represent multi-topic attention dynamics.

Extensions have broadened the framework's scope. Friedkin [60] links opinion dynamics to norm formation and shows how network topology shapes convergence versus divergence. Zhou and Wu [211] introduce multi-topic opinions. For each topic $k$, there is a separate stance vector $x_k$: $x_k^{t+1} = (\mathbb{I} - A)x_k^0 + AWx_k^t$. Rewriting the recurrence relation in matrix notation, and taking $A^t$ to give diminishing susceptibility levels over time, I have: $X(t+1) = (\mathbb{I} - A)^t X(0) + (AW)^t X(t)$. Accordingly, in our model, I raise the susceptibility parameter $\alpha$ to the power of a scalar parameter representing effective stubbornness, $\mathrm{eff}$.

A core limitation is that influence is largely topic-agnostic: the same tie weights apply across issues, despite real-world trust and expertise being topic dependent. Moreover, abstract neighbor averaging leaves limited room for contextual interaction and variable persuasiveness (e.g., selective engagement with a personalized feed where salience differs by topic and individual).

**Salience and Persuasion in Opinion Dynamics** Salience and persuasion shape both what agents attend to and how strongly they update. Salience reflects topic priority driven by interest and context [183]. Signed FJ variants imply salience can speed convergence or polarization [212], but typically treat engagement as exogenous. Persuasion work highlights heterogeneity in susceptibility and stubbornness [2], tipping dynamics under targeted perturbations [34], and the role of cognitive and affective alignment in receptivity [178].

**Large Language Models in Agent-Based Modeling** LLMs can enrich ABMs by generating agent profiles, behaviors, and interactions [67]. Studies show emergent dynamics in LLM agent populations, including homophily-like clustering [79], while other work cautions against treating synthetic behavior as a stand-in for human data [97, 161]. In persuasion contexts, LLM frameworks can generate messages and simulate audience reactions to support strategy selection [115]. Building on scalable agent architectures [191], I use LLMs to instantiate heterogeneous personas and content while retaining an interpretable FJ-style update that links peer engagement and argument quality to opinion change.

**Calibrating Differentiable Opinion Dynamics Models** An on-going line of research is concerned with scaling classical network science models of spreading dynamics over ever increasing LLM populations. [39], for example, proposes a population scale model on the order of 10 million agents. By combining LLM populations with a traditional Susceptible-Infected-Recovered (SIR) epidemiological model, they demonstrate how to calibrate a model to predict responses to public health policy using demographic data from New York City.

Social influence (SI) models also face scaling challenges. Real-world social networks are sparse, and many SI formulations exhibit global interdependence that becomes costly at population scale. Scalable SI models therefore benefit from calibration approaches that leverage demographic data and population-level trends. While [39] demonstrates this for epidemiological modeling via full differentiability, related work has proposed differentiable variants of classic SI models that

enable analogous parameter fitting [107, 146, 194]. This is not the main focus here, but our model is differentiable with minor changes, making it compatible with gradient-based calibration for population-scale SI settings.

**HAI Platforms and Experimental Infrastructure.** A second emerging line of work focuses on experimental platforms for studying human–AI interaction at scale. Qu et al. introduce *Epitome*, an open platform for controlled hybrid human, AI experimentation that embeds classical experimental logic, control, comparison, and causal inference, within interactive environments ranging from dyadic conversations to multi-agent scenarios [157]. Epitome is especially relevant here not because JANUS depends on it, but because it clarifies how mechanistic social simulation can interface with experimental platform design. In this framing, JANUS can be understood as a hypothesis-generation engine for questions about how alternative exposure regimes, ranking systems, or recommender policies affect polarization, coordination, and cross-cutting exposure, while platforms such as Synthetic Social Media (SynSM) Simulation Loop [12] or Epitome provide the interaction protocols and logging infrastructure needed to test those hypotheses in controlled settings [157]. This connection helps position JANUS not only as an opinion-dynamics model, but also as part of a broader methodological stack for rigorous HAI research.

## 6.3 Method

In JANUS, agents are synthetic personas, structured bundles of demographic attributes, psychological traits, topic-specific salience, and affective predispositions that together determine how social exposure is interpreted and translated into opinion change. I instantiate these personas and define their interaction with the opinion dynamics model using a three stage pipeline: (i) empirical sampling to match target-population marginal distributions (AURORA), (ii) the conditioning of LLM agents to produce coherent, human-like personas and topic-grounded behavioral cues, and (iii) the mapping between the personas and the social influence model.

### AURORA persona pipeline

I adopt the AURORA synthetic population framework [119] to generate heterogeneous agents whose attributes are sampled from empirical U.S. distributions (2022 census and survey data) while maintaining internal coherence at the individual level [90, 177, 186, 187]. Each agent $i$ is initialized with:

- **Demographics**: Age, gender, race/ethnicity, education level, political affiliation
- **Psychometrics**: Big Five personality traits (OCEAN: Openness, Conscientiousness, Extraversion, Agreeableness, Neuroticism) and Moral Foundations Theory scores (Care, Fairness, Loyalty, Authority, Sanctity, Liberty)
- **Group memberships**: Assignment to 1-3 advocacy or community organizations, with 10% of agents flagged as group leaders

- **Topic-level attributes**: For each topic $k$, initial opinion $x_{i,k}^0 \in [-1, 1]$, salience $s_{i,k} \in [0, 1]$, and eight Plutchik emotion intensities $e_{i,k}(\cdot) \in [0, 1]$ (joy, trust, fear, surprise, sadness, disgust, anger, anticipation)

All attributes are sampled once during initialization and remain fixed throughout the simulation, with the exception of topic opinions $x_{i,k}^t$, which evolve under the opinion dynamics model (Eq. 6.7).

### 6.3.1 Network Construction from AURORA Personas

JANUS constructs its baseline social network directly from the organizational and community structure generated during AURORA persona creation. Rather than treating network formation as a separate stochastic process, JANUS uses this graph as the structural substrate linking empirically grounded personas to later opinion dynamics. The resulting network represents persistent affiliation and exposure opportunities prior to platform-specific interaction effects. Ties are created deterministically from shared group memberships: two personas are connected if and only if they share at least one group, and ties strengthen with greater overlap. This operationalizes a core sociological assumption that repeated co-membership creates recurring opportunities for influence, coordination, and exposure, consistent with homophily-based theories of social organization [128].

**Step-by-step network generation process.** The construction process proceeds in five steps.

1. **Load AURORA outputs.** JANUS first loads the scenario outputs generated by AURORA, including personas, groups, communities, and countries.
2. **Create one node per persona.** Each persona is added to a NetworkX graph as a node, storing demographic, psychological, and topic-level attributes.
3. **Create ties from shared group membership.** For each group, JANUS connects all pairs of member personas. If a pair shares multiple groups, the edge weight is incremented accordingly:
$$\text{overlap}_{ij} = |G_i \cap G_j|. \tag{6.1}$$
This produces an undirected weighted co-membership graph.
4. **Annotate group leaders.** Personas designated as group leaders are marked with a binary attribute. This does not affect graph topology.
5. **Construct the JANUS influence matrix.** During simulation, the graph is converted into a weighted influence matrix:
$$W_{ij}^{\text{raw}} = \text{overlap}_{ij}(1 + \lambda L_j), \tag{6.2}$$
followed by row normalization. Because the leader boost depends on the sender, the resulting influence structure is asymmetric.

This procedure produces a network that is deterministic, interpretable, and explicitly tied to the persona-generation process rather than to arbitrary graph-generation assumptions. The graph

therefore serves as the structural bridge between AURORA personas and JANUS dynamics: AURORA supplies the actors and their affiliations, while JANUS transforms those affiliations into weighted exposure opportunities. In this sense, the network is not an incidental preprocessing step; it is part of the model's theoretical commitment to representing influence as emerging from stable social structure. This makes the resulting graph especially suitable for downstream opinion dynamics, because the same affiliations that create local cohesion also provide the basis for repeated influence and cross-group spillover. Figure 6.1 summarizes this pipeline.

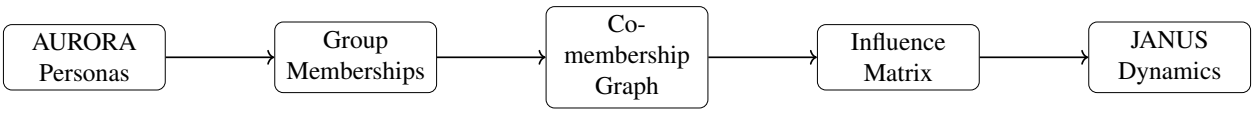



Figure 6.1: Pipeline from AURORA personas to JANUS opinion dynamics.

**Illustrative construction pipeline.**

**Network characteristics of the generated graphs.** To evaluate the resulting graphs, I analyze the baseline networks at $t = 0$, prior to any opinion updates. Table 6.1 reports the structural properties of four scaling scenarios. These results show that the co-membership construction produces networks that are highly stable across size, suggesting that the procedure preserves its structural logic as the number of personas increases.

Table 6.1: Structural characteristics of JANUS persona networks at $t = 0$.

| Scenario | Nodes | Edges | Density | Mean Deg. | Med. Deg. | Clustering | Path Len. | Diameter |
|---|---|---|---|---|---|---|---|---|
| US-250 | 250 | 2,136 | 0.0745 | 17.8 | 15 | 0.880 | 2.75 | 4 |
| US-500 | 500 | 8,635 | 0.0751 | 36.0 | 32 | 0.892 | 2.65 | 4 |
| US-750 | 750 | 20,056 | 0.0775 | 55.7 | 53 | 0.894 | 2.51 | 3 |
| US-1000 | 1000 | 35,292 | 0.0767 | 73.5 | 72 | 0.890 | 2.50 | 3 |

Several patterns are important. First, all four networks form a single connected component containing 100% of nodes, indicating that the group-based construction yields globally connected graphs even as population size increases. Second, density remains in a narrow range ($0.0745$–$0.0775$), showing that connectivity is not being driven by arbitrary densification as the network scales. Instead, edge formation remains governed by shared group affiliations. Third, mean degree increases from $17.8$ to $73.5$ as the number of personas increases, which is consistent with the fact that larger populations contain more opportunities for overlapping memberships while preserving the same basic construction logic.

The clustering and path-length results clarify the type of structure this process generates. Clustering remains uniformly high across all scenarios ($0.880$–$0.894$), reflecting strong triadic closure: if two agents share a group with a third, they are likely to be connected to one another as well. At the same time, average shortest path length remains low ($2.50$–$2.75$), and the diameter never

exceeds four. Taken together, these properties are characteristic of small-world organization, although here that structure arises from affiliation overlap rather than random rewiring. The result is a graph that combines dense local cohesion with short global reach, which is desirable for an opinion-dynamics substrate because it allows both localized reinforcement and network-wide propagation.

The degree distribution further supports this interpretation. The JANUS graphs are right-skewed but bounded: some agents occupy more central positions because they overlap across multiple groups, but the graph does not produce the extreme hub dominance associated with preferential attachment or scale-free growth. Most agents remain embedded in a dense community core, while higher-degree nodes emerge from structured overlap across organizations rather than from broadcast-style popularity. Accordingly, the baseline JANUS network is best understood as an affiliation-based, community-structured graph. This is important for the chapter as a whole, because it means that later opinion trajectories are not being driven by an unrealistic or arbitrary topology; they unfold over a stable social substrate whose modularity, cohesion, and bounded heterogeneity are a direct consequence of the persona-generation process itself.

**MOMUS Structural Comparison and Implications for GhostField.** Table 6.2 places the baseline JANUS (US-1000) $t = 0$ network in comparative context by benchmarking it against standard empirical Twitter networks using MOMUS diagnostics. I use MOMUS here as a structural comparison of the baseline co-membership graph [135], not as a validation of the JANUS opinion-update rule itself. This distinction is important. JANUS, as an opinion-dynamics model, specifies how opinions evolve over an existing network; MOMUS, by contrast, provides a way to assess what kind of network that baseline graph most closely resembles. Read in that way, the table clarifies both the strengths and the intended role of the JANUS-only network.

Under the MOMUS realism rubric, JANUS scores $1/4$ on components, $3/4$ on degree skew, and $0/4$ on reciprocity, whereas SNAP ego-Twitter and the Higgs interaction networks score more highly on degree skew and reciprocity. This is exactly what one would expect from a deterministic affiliation graph. The JANUS baseline is not a realized communication network, so it should not be expected to reproduce reply reciprocity, broadcaster-driven hub formation, or the sparse interaction traces observed in live social media systems. Instead, its degree support is bounded to roughly $k \approx 40$–$200$, while the empirical comparison networks extend into the $10^3$–$10^4$ range. Likewise, the largest JANUS Louvain community is only about 80–85 nodes, rather than the giant communities observed in follower and retweet graphs. These differences do not indicate model failure; rather, they show that JANUS is generating a different kind of structure.

The remaining diagnostics make that structure more precise. JANUS displays very high local clustering, with $C(k) \approx 1.0$ around $k \approx 50$ and still about $0.54$ near $k \approx 100$, which is above the Higgs interaction baselines and above or comparable to SNAP over similar degree ranges. The mixing matrix is strongly diagonal, with within-community edge fractions typically between about $0.50$ and $0.85$, while most off-diagonal values remain below $0.05$. Homophily is also strong on the fields most directly tied to social organization in the synthetic population, including ideology, community, category, and type (all approximately $0.83$–$0.84$), with more moderate sorting on political leaning ($\approx 0.59$) and little sorting on nationality or social-influence labels ($\approx 0.01$–$0.02$).

Table 6.2: MOMUS-based structural benchmark of the baseline JANUS $t = 0$ network relative to empirical Twitter benchmarks. Rubric scores are read directly from MOMUS; values from continuous plots are approximate visual estimates.

| **Metric** | **JANUS** $t = 0$ | **Empirical comparison** |
|---|---|---|
| MOMUS components rubric | $1/4$ | SNAP ego-Twitter: $1/4$; Higgs networks: $\approx 2/4$ |
| MOMUS degree-skew rubric | $3/4$ | SNAP ego-Twitter: $4/4$; Higgs networks: $4/4$ |
| MOMUS reciprocity rubric | $0/4$ | SNAP ego-Twitter: $1/4$; Higgs networks: $\approx 2/4$–$4/4$ |
| Observed degree support | $k \approx 40$–$200$ | Empirical benchmarks extend to roughly $k \approx 10^3$–$10^4$ |
| Local clustering at $k \approx 50$ | $C(k) \approx 1.00$ | SNAP $\approx 0.50$; Higgs follower $\approx 0.18$; retweet/mention/reply near 0–0.05 |
| Local clustering at $k \approx 100$ | $C(k) \approx 0.54$ | SNAP $\approx 0.40$; Higgs follower $\approx 0.16$; retweet/mention/reply near 0–0.03 |
| Largest Louvain community | $\approx 80$–$85$ nodes | Empirical largest communities begin around $10^4$–$10^5$ nodes |
| Community mixing | diagonal entries $\approx 0.50$–$0.85$; most off-diagonal entries $< 0.05$ | Indicates strong within-community concentration rather than diffuse cross-community mixing |
| Homophily (assortativity) | ideology $\approx 0.84$; community $\approx 0.83$; category $\approx 0.83$; type $\approx 0.83$; political leaning $\approx 0.59$; nationality/social influence $\approx 0.01$–$0.02$ | Strong sorting on designed identity/community fields; weak sorting on nationality and influence labels |
| Neighbor-degree trend | $k_{nn}(k) \approx 55 \rightarrow 100$ over the observed degree range | Positive slope indicates assortative mixing among mid-degree nodes rather than a broadcaster-centered topology |

Finally, the positive $k_{nn}(k)$ trend indicates assortative meso-scale organization among mid-degree nodes rather than a celebrity-follower pattern. These results show that the JANUS-only graph is best interpreted as an offline affiliation and latent influence substrate: modular, cohesive, and interpretable, but not designed to mimic the full behavioral signature of observed online interaction networks.

This interpretation is also the one most consistent with the broader GhostField architecture. Within GhostField, the baseline JANUS graph provides the stable social structure over which influence can propagate [135]. It captures persistent community boundaries, overlapping affiliations, and interpretable exposure opportunities. Systems such as SynSM are then responsible for converting that background structure into platform-mediated interaction, including burstiness, asymmetry, sparse engagement, and other properties of observed online behavior. In other words, the JANUS-only network should not be judged as though it were a direct Twitter clone. Its purpose is different: it provides the structurally coherent, theoretically interpretable substrate needed for the opinion-dynamics component of the chapter, while GhostField's platform layer supplies the interaction-level realism associated with empirical social media traces.

### 6.3.2 Emotional and Contextual Conditioning

To elicit psychologically plausible reactions from LLM-conditioned agents, I use hierarchical prompts that condition on (1) stable persona traits (demographics and psychometrics), (2) situational context (event/topic framing), and (3) an explicit affect representation that can be validated against empirical social media reactions [120]. This supports JANUS by (i) mapping persona affect into initial valence–arousal parameters for affective drift and (ii) generating (when needed) agent-authored text consistent with the same underlying affect state [120].

### 6.3.3 Parameter Extraction for Opinion Dynamics

Persona attributes are mapped to opinion dynamics parameters so that heterogeneous psychological traits and topic-specific sensitivities drive the mechanistic update rule (Eq. 6.7). The LLM does not "solve" the dynamics; rather, it instantiates diverse agents whose attributes parameterize the model. I describe each parameter mapping below.

**Stubbornness/susceptibility** Baseline stubbornness $\alpha_i$ is derived from personality openness via simple inversion:

$$\alpha_i = 1 - \text{Openness}_i, \tag{6.3}$$

so that more open agents are less stubborn (more susceptible to social influence). This baseline is broadcast across topics and then modulated by topic-specific salience to produce effective stubbornness $\alpha_{i,k}^{\text{eff}}$ (Eq. 6.11), capturing the intuition that agents are harder to move on issues they consider important.

**Signed ties** Signed ties $\sigma_{ij} \in [-1, 1]$ combine multi-dimensional identity similarity and salience-filtered opinion similarity (Eq. 6.15). Identity similarity $S_{ij}^{\text{id}}$ is computed as a weighted average of four Gower-style blocks [71]:

- **OCEAN**: $1 - \text{mean}|\Delta|$ over 5 personality traits
- **MFT**: $1 - \text{mean}|\Delta|$ over 6 moral foundations
- **Age**: Gower linear similarity, range-normalized as $1 - |\text{age}_i - \text{age}_j|/(\text{age}_{\max} - \text{age}_{\min})$
- **Demographics**: $1 - \text{mean}|\Delta|$ over one-hot encoded gender and race

Raw identity similarities are percentile-stretched (10th percentile $\to$ 0, 90th percentile $\to$ 1) to enhance contrast between dyads, as the original Gower similarities are often tightly clustered [25].

Opinion similarity $S_{ij}^{\text{opin}}$ is *directional*: it is computed over the subset of topics that the *receiver* $i$ considers salient ($s_{i,k} > 0.5$), and disagreement is amplified when identity similarity exceeds a threshold ($S_{ij}^{\text{id}} > 0.5$), reflecting the "black sheep" effect whereby in-group deviants are judged more harshly [122]. The final signed tie is the average of identity and opinion similarity components (Eq. 6.15).

**Affective drift** Baseline emotion intensities $e_{i,k}(\cdot)$ for each topic are mapped to valence-arousal coordinates $(v_{i,k}^0, a_{i,k}^0)$ via weighted projection onto the Russell circumplex [154, 166]. Each Plutchik emotion $e$ is assigned a fixed angle $\theta_e$ on the circumplex (e.g., joy at 60°, anger at 150°; see Fig. 6.2) and a psychometric weight $w_e$ derived from empirical affect structure. Valence and arousal are computed as:

$$v_{i,k}^0 = \sum_{e \in \mathcal{E}} e_{i,k}(e) \cdot \cos(\theta_e), \tag{6.4}$$

$$a_{i,k}'^0 = \sum_{e \in \mathcal{E}} e_{i,k}(e) \cdot \sin(\theta_e), \tag{6.5}$$

$$a_{i,k}^0 = 0.5 \times (1 + a_{i,k}'^0), \tag{6.6}$$

where the final arousal is mapped to $[0, 1]$ by shifting and scaling the vertical circumplex axis. This projection reduces the 8-dimensional emotion vector to interpretable valence (pleasantness) and arousal (activation) dimensions, which parameterize the affective drift term in Eq. 6.7.

### 6.3.4 Platform-Mediated Exposure and Alternative Definitions of $\widehat{W}_{ij}$

At the abstract level, JANUS is written as a networked opinion update model over a weighted graph, with social influence entering through the exposure term $\widehat{W}_{ij}$ in Eq. 6.7. In that baseline view, the graph specifies exposure (influence) opportunities.

In a platform-mediated setting, by contrast, exposure is constrained based on platform affordances: whether one agent actually influences another depends not only on connectivity, but also on visibility, ranking, interaction, and perceived stance formation. In this case, the social weights in Eqs. 6.13–6.14 represent interaction likelihoods that define exposure opportunities on the platform. For example, to represent the exposure opportunities on social networks that are centered around short-form video, the influence matrix $W$ could be constructed using the canonical matrix

factorization method for content recommendation systems based on user preference overlap. When embedded in a twitter-based synthetic social media environment such as SynSM, $W$ may represent a combination of social network and content recommendation systems. In this case, persona attributes determine which messages impact platform engagement.

## 6.4 Model

JANUS is a networked agent-based opinion dynamics model in which agents hold bounded, topic-specific opinions on a weighted social graph. Opinions evolve through three mechanisms: (i) anchored prior retention via limited susceptibility, (ii) signed social influence (attraction/repulsion) proportional to disagreement, and (iii) arousal-gated affective drift driven by baseline valence-arousal. Signed ties are derived from identity similarity and salience-filtered opinion alignment, supporting within-group convergence, between-group polarization, and topic-dependent persistence.

### 6.4.1 Model Overview

Let $N$ be the number of agents and $K$ the number of topics. Agent $i \in 1, \ldots, N$ holds an opinion $x_{i,k}^t \in [-1, 1]$ on topic $k$ at timestep $t$, giving the state $\mathbf{X}^t \in [-1, 1]^{N \times K}$. Updates are synchronous: $\mathbf{X}^{t+1}$ is computed from $\mathbf{X}^t$ and then applied simultaneously.

At each step, Eq. 6.7 mixes a fixed prior anchor $x_{i,k}^0$ with a dynamic signal. The anchor is weighted by effective stubbornness $\alpha^{\text{eff}} i, k$, while the remaining mass $(1 - \alpha^{\text{eff}} i, k)$ scales (i) signed, disagreement-based neighbor influence and (ii) arousal-gated affective drift proportional to baseline valence. This structure yields stable anchors with heterogeneous responsiveness, enabling clustering and polarization without forcing consensus.

### 6.4.2 Opinion Update Equation

The primary opinion update rule in JANUS is given in Eq. 6.7. The update combines (I) Friedkin-Johnsen (FJ) anchor retention and susceptibility, (II) a signed, distance-based social influence term with backfire amplification, and (III) an arousal-gated affective drift term. The resulting opinion is bounded to the admissible interval $[-1, 1]$.

$$\tilde{x}_{i,k}^{t+1} = \underbrace{\left(\alpha_{i,k}^{\text{eff}}\, x_{i,k}^0 + \left(1 - \alpha_{i,k}^{\text{eff}}\right)\right.}_{\textbf{(I) FJ anchor + susceptibility}} \Bigg[ \underbrace{g\left(a_{i,k}^0\right) \sum_j \left(\widehat{W}_{ij}\, \sigma_{ij}\, \Delta x_{ij,k}^t\, B_{ij,k}^t\right)}_{\textbf{(II) Signed social influence}} + \underbrace{G_{i,k}\left(a_{i,k}^0\right)\, b\, v_{i,k}^0}_{\textbf{(III) Affective Emotional Drift}} \Bigg] \Bigg), \tag{6.7}$$

$$g(a) = \tanh\left(\lambda\,(a - \tau_a)\right), \qquad \lambda > 0. \tag{6.8}$$

where $\tilde{x}_{i,k}^{t+1}$ denotes the *raw* next opinion prior to applying a per-timestep rate limit.

Table 6.3: Key symbols in the main opinion update (Eq. 6.7).

| Symbol | Range/Type | Description |
|---|---|---|
| $x_{i,k}^t$ | $[-1, 1]$ | Opinion of agent $i$ on topic $k$ at time $t$. |
| $x_{i,k}^0$ | $[-1, 1]$ | Prior (initial) opinion anchor for agent $i$ on topic $k$. |
| $\alpha_{i,k}^{\text{eff}}$ | $[0, 1]$ | Effective stubbornness (anchor vs. influence weighting). |
| $\widehat{W}_{ij}$ | $[0, 1]$ | Normalized social weight from neighbor $j$ to receiver $i$. |
| $\sigma_{ij}$ | $[-1, 1]$ | Signed tie (attraction $> 0$ / repulsion $< 0$) between $i$ and $j$. |
| $\Delta x_{ij,k}^t$ | $[-2, 2]$ | Opinion difference $(x_{j,k}^t - x_{i,k}^t)$. |
| $B_{ij,k}^t$ | $\geq 1$ | Backfire multiplier under repulsive, opposite-signed disagreement. |
| $v_{i,k}^0, a_{i,k}^0$ | $[-1, 1], [0, 1]$ | Baseline valence and arousal for affective drift. |
| $G_{i,k}(\cdot)$ | $\{0, 1\}$ | Arousal gate enabling affective drift. |
| $b$ | $\geq 0$ | Valence sensitivity (drift magnitude). |
| $g(a)$ | $(-1, 1)$ | A hyperbolic tangent ($\tanh$) function nonlinearly scales social influence as a function of arousal, amplifying moderate signals while saturating extreme responses. (Eq. 6.8). |

**Per-timestep rate limiting (implementation detail)** To prevent unrealistically large one-step jumps, I cap the applied change in each opinion component to $\pm\delta_{\max}$ with $\delta_{\max} = 0.2$. Let $\tilde{x}_{i,k}^{t+1}$ be the raw update from Eq. 6.7. The applied next opinion is

$$x_{i,k}^{t+1} = x_{i,k}^t + \text{clip}_{[-\delta_{\max}, \delta_{\max}]}\left(\tilde{x}_{i,k}^{t+1} - x_{i,k}^t\right), \tag{6.9}$$

with the $[-1, 1]$ bound enforced by Eq. 6.7.

**Temporal scaling in JANUS** In JANUS, each timestep corresponds to an interaction event between agents, so the effective time constant governs how much opinions change per interaction rather than per unit of physical time. This time constant is not a single parameter but emerges from the interaction of effective stubbornness $\alpha_{i,k}^{\text{eff}}$, arousal-gated activation $g(a_{i,k}^0)$, and the per-timestep rate limit $\delta_{\max}$ (Eq. 6.9). The parameter $\delta_{\max}$ acts as the primary global control on update speed by bounding the maximum opinion shift per interaction, ensuring incremental change. With $\delta_{\max} = 0.2$, the model produces trajectories in which most agents update gradually across repeated interactions, stance reversals are rare, and population-level stabilization occurs over many interaction steps. At the same time, update rates remain heterogeneous: agents with low $\alpha_{i,k}^{\text{eff}}$ and high arousal adapt more quickly, while high-salience, high-stubbornness agents change slowly. As a result, JANUS encodes a state-dependent and interaction-driven time constant that reflects realistic variation in responsiveness without requiring a single globally calibrated rate parameter.

### 6.4.3 (I) FJ Anchor + Susceptibility

The first component of Eq. 6.7 implements a Friedkin-Johnsen (FJ) style decomposition in which each agent partially retains an internal prior belief while remaining susceptible to external influence. For each agent-topic pair $(i, k)$, the update is structured as a convex combination of (i) a fixed prior anchor $x^0_{i,k}$ and (ii) a dynamic influence term (the bracketed quantity in Eq. 6.7):

$$x^{t+1}_{i,k} = \left(\alpha^{\text{eff}}_{i,k}\, x^0_{i,k} + \left(1 - \alpha^{\text{eff}}_{i,k}\right)\right), \tag{6.10}$$

The scalar $\alpha^{\text{eff}}_{i,k} \in [0, 1]$ is the agent's *effective stubbornness* on topic $k$: larger values increase retention of the prior anchor. This decomposition ensures that even under repeated exposure to social influence, agents do not necessarily converge to a single consensus point; instead, each retains a persistent "home position" determined by $x^0_{i,k}$, with the strength of that retention varying across agents and topics.

**Baseline stubbornness from personality openness** In JANUS, baseline stubbornness is initialized from personality openness such that more open agents are less stubborn. Concretely, for each agent $i$ I compute a topic-invariant baseline

$$\alpha_i \;=\; 1 - \text{Openness}_i \tag{6.3}$$

and broadcast it across topics to obtain $\alpha_{i,k} = \alpha_i$. This mapping produces heterogeneous responsiveness at the population level while requiring only a single interpretable personality attribute.

**Topic-specific adjustment via salience coupling** Baseline stubbornness is adjusted on a per-topic basis using the agent's topic salience $s_{i,k} \in [0, 1]$, where higher salience implies stronger resistance to change. I define effective stubbornness as an affine shift of the baseline value:

$$\alpha^{\text{eff}}_{i,k} = \alpha^{(0)}_i + \gamma\,(2s_{i,k} - 1), \qquad \alpha^{\text{eff}}_{i,k} \in [0, 1], \tag{6.11}$$

where $\alpha^{(0)}_i$ denotes the agent's baseline (topic-invariant) stubbornness and $\gamma$ controls the strength of the salience, stubbornness coupling (set to $\gamma = 0.5$ in our experiments). This mapping is *neutral* at $s_{i,k} = 0.5$, since $(2s_{i,k} - 1) = 0$ and therefore $\alpha^{\text{eff}}_{i,k} = \alpha^{(0)}_i$. For topics that an agent considers more salient ($s_{i,k} > 0.5$), $\alpha^{\text{eff}}_{i,k}$ increases, yielding greater anchor retention (reduced opinion shift). For low-salience topics ($s_{i,k} < 0.5$), $\alpha^{\text{eff}}_{i,k}$ decreases, increasing malleability. The $[0, 1]$ constraint ensures that $\alpha^{\text{eff}}_{i,k}$ remains a valid mixing weight in Eq. 6.10, so that the dynamic component is scaled by $(1 - \alpha^{\text{eff}}_{i,k})$ and vanishes as $\alpha^{\text{eff}}_{i,k} \to 1$.

**On convergence and the role of FJ in JANUS** While Eq. 6.10 follows the classical Friedkin–Johnsen structure, its implications differ in the JANUS setting. A natural concern with Friedkin–Johnsen (FJ) dynamics is that, under standard assumptions (fixed, nonnegative influence weights), the system is contractive and converges to a unique fixed point. However, JANUS does not operate within this regime. First, the effective influence structure is signed and state-dependent, incorporating attraction, repulsion, and backfire amplification, which violate the nonnegativity and

linearity assumptions required for contraction. Second, exposure weights are not fixed but depend on platform-mediated interaction and salience filtering, introducing time variation into the influence process. Third, affective drift adds an additional non-linear bias term. These extensions break the classical convergence guarantees and allow for persistent heterogeneity and polarization. Importantly, even in the baseline FJ formulation, convergence does not imply consensus: heterogeneous anchor opinions produce structured, agent-specific equilibria. In JANUS, the FJ component is therefore best understood as a stabilizing anchor within a broader dynamical system, ensuring bounded and interpretable opinion trajectories while allowing richer, non-convergent social dynamics to emerge.

### 6.4.4 (II) Signed Social Influence

The signed social influence term in Eq. 6.7 models how agent $i$ is pulled toward or pushed away from neighbors' opinions in proportion to disagreement [43, 57]. This mechanism differs fundamentally from averaging-based models (e.g., DeGroot [50]): in averaging models, agents move toward the weighted mean of neighbors' opinions regardless of distance; in distance-based models like JANUS, influence is proportional to disagreement magnitude $\Delta x_{ij,k} = x_{j,k} - x_{i,k}$, making attraction and repulsion act as true restoring/repelling forces [61, 84, 212]. This formulation enables signed networks to produce polarization and echo chambers, which averaging models cannot capture [123, 147].

The social influence signal for agent $i$ on topic $k$ at time $t$ is computed as:

$$\Delta^{\text{soc}}_{i,k}(t) = \sum_j M_{ij}\ \Delta x^t_{ij,k}\ B^t_{ij,k}, \qquad \Delta x^t_{ij,k} \equiv x^t_{j,k} - x^t_{i,k}, \tag{6.12}$$

where $M_{ij}$ is a signed, normalized influence weight combining network structure and psychological compatibility (eq:soc$_i nfluence$)[123, 204],$\Delta x^t_{ij,k}$ is the opinion difference ensuring influence is proportional to disagreement [61, 123], and $B^t_{ij,k} \geq 1$ is a backfire amplification factor applied under antagonistic exposure [145]. I now describe each component in detail.

#### Network-Based Social Weights ($\widehat{W}_{ij}$)

Social weights capture the frequency and strength of interaction between agents based on network structure. I define an unnormalized weight $W_{ij}$ from sender $j$ to receiver $i$ using group co-membership with a leader boost:

$$W_{ij} = |G_i \cap G_j|\ (1 + \lambda \cdot L_j), \qquad \lambda = 3.0, \tag{6.13}$$

where $G_i$ is the set of groups to which agent $i$ belongs, $|G_i \cap G_j|$ is the reciprocity/overlap count (number of shared group memberships), $L_j \in \{0, 1\}$ indicates whether agent $j$ is a group leader, and $\lambda = 3.0$ is the leader influence boost factor [46, 128]. This construction reflects the intuition that agents with more shared affiliations interact more frequently, and that leaders exert disproportionate influence within their communities.

The normalized social weight is row-stochastic:

$$\widehat{W}_{ij} = \frac{W_{ij}}{\sum_g W_{ig}}, \tag{6.14}$$

ensuring that $\sum_j \widehat{W}_{ij} = 1$ for each receiver $i$ (when $i$ has at least one neighbor). This normalization guarantees that the total influence received by each agent from the network structure sums to unity, independent of the agent's degree or number of group memberships.

**Psychological Signed Ties ($\sigma_{ij}$)**

The signed tie $\sigma_{ij} \in [-1, 1]$ encodes whether neighbor $j$ exerts an *attractive* (positive) or *repulsive* (negative) influence on receiver $i$, based on multi-dimensional similarity across identity attributes and topic opinions [123, 204]. This sign convention is consistent with agent-based influence formalisms that represent perceived stance using signed support/opposition indicators [52, 85]. In JANUS, $\sigma_{ij}$ is constructed by combining identity similarity (transformed to $[-1, 1]$) and directional opinion similarity:

$$\sigma_{ij} = \mathrm{clip}\Big(\widetilde{S}^{\mathrm{id}}_{ij} + S^{\mathrm{opin}}_{ij},\, -1, 1\Big), \tag{6.15}$$

where $\widetilde{S}^{\mathrm{id}}_{ij} = 2S^{\mathrm{id}}_{ij} - 1$ maps identity similarity from $[0, 1]$ to $[-1, 1]$, and $S^{\mathrm{opin}}_{ij} \in [-1, 1]$ is directional opinion similarity. The sum is clipped to ensure $\sigma_{ij} \in [-1, 1]$. This formulation allows identity to set a baseline tie strength, which is then modulated by opinion agreement or disagreement [43, 57].

**Identity Similarity ($S^{\mathrm{id}}_{ij}$)** Identity similarity is computed using a block-wise, Gower-style formulation over multiple attribute groups (OCEAN personality traits, Moral Foundations Theory scores, age, and demographic one-hot features) [71]. This design reflects the well-established role of homophily, the tendency for similarity across multiple dimensions to structure social exposure and tie formation [128]. For each block $b$ with feature vector $\mathbf{z}^{(b)}$, I compute a within-block similarity $S^{(b)}_{ij} \in [0, 1]$ as follows:

- **OCEAN (5 traits)**: $S^{\mathrm{OCEAN}}_{ij} = 1 - \frac{1}{5}\sum_{t=1}^{5} |\mathrm{OCEAN}_{i,t} - \mathrm{OCEAN}_{j,t}|$
- **MFT (6 foundations)**: $S^{\mathrm{MFT}}_{ij} = 1 - \frac{1}{6}\sum_{f=1}^{6} |\mathrm{MFT}_{i,f} - \mathrm{MFT}_{j,f}|$
- **Age**: $S^{\mathrm{age}}_{ij} = 1 - \frac{|\mathrm{age}_i - \mathrm{age}_j|}{\mathrm{age}_{\max} - \mathrm{age}_{\min}}$ (Gower linear similarity, range-normalized)
- **Demographics**: $S^{\mathrm{demo}}_{ij} = 1 - \frac{1}{|D|}\sum_{d \in D} |\mathrm{demo}_{i,d} - \mathrm{demo}_{j,d}|$ (one-hot encoded gender and race)

The block similarities are combined as a weighted average:

$$S^{\mathrm{id}}_{ij} = \frac{\sum_b w_b\, S^{(b)}_{ij}}{\sum_b w_b}, \qquad S^{\mathrm{id}}_{ij} \in [0, 1], \tag{6.16}$$

where $w_b \geq 0$ is the block weight controlling the relative contribution of each identity dimension. In our implementation, all block weights are set to unity ($w_{\text{OCEAN}} = w_{\text{MFT}} = w_{\text{age}} = w_{\text{demo}} = 1.0$), giving equal importance to personality, morality, age, and demographics. This block structure aligns with how identity signals are operationalized in computational sociolinguistics and social-media modeling [142], and is consistent with modern generative-agent architectures that encode personas as multi-field identity profiles [149, 191].

**Percentile Stretching of Identity Similarity** Raw identity similarities are often tightly clustered (especially for OCEAN/MFT and other normalized attributes), which reduces contrast between dyads and can lead to overly homogeneous signed ties. I therefore apply a robust, percentile-based linear stretch to enhance differentiation [25]:

$$S^{\text{id}}_{\text{stretched},ij} = \frac{S^{\text{id}}_{ij} - \ell}{u - \ell}, \qquad \ell = \text{percentile}_{10}(S^{\text{id}}), \quad u = \text{percentile}_{90}(S^{\text{id}}), \tag{6.17}$$

where $\ell$ and $u$ are the 10th and 90th percentiles computed from the off-diagonal entries of $S^{\text{id}}$. Values below $\ell$ are clipped to 0, and values above $u$ are clipped to 1. I use percentiles rather than $\min / \max$ to avoid sensitivity to rare extreme pairs, which can otherwise distort the scaling of the typical dyad range. After stretching, the identity similarity is transformed to the signed range via

$$\widetilde{S}^{\text{id}}_{ij} = 2S^{\text{id}}_{\text{stretched},ij} - 1 \tag{6.18}$$

**Directional, Salience-Filtered Opinion Similarity ($S^{\text{opin}}_{ij}$)** Opinion similarity is computed as one minus the average absolute opinion distance over the subset of topics that the *receiver* $i$ considers salient:

$$S^{\text{opin}}_{ij}(t) = 1 - \frac{1}{|K_i|} \sum_{k \in K_i} \beta_{ij} \left| x^t_{i,k} - x^t_{j,k} \right|, \tag{6.19}$$

where $K_i = \{k : s_{i,k} > \tau_s\}$ is the set of topics that agent $i$ finds salient (with salience threshold $\tau_s = 0.5$), and $\beta_{ij} \geq 1$ is a scaling factor that amplifies disagreement for identity-similar pairs (6.19). This construction is *directional* because the salience mask depends on the receiver $i$; consequently, $S^{\text{opin}}_{ij} \neq S^{\text{opin}}_{ji}$ whenever agents attend to different topic subsets. This asymmetry reflects the intuition that "I judge you based on issues I find important, not issues you find important."

**Black Sheep Effect: Amplified In-Group Disagreement** To capture heightened sensitivity to disagreement among highly identity-similar pairs—the "black sheep" effect whereby in-group deviants are judged more harshly than out-group members with the same views [122]—JANUS scales per-topic disagreement by $\beta_{ij}$:

$$\beta_{ij} = \begin{cases} 2.0, & S^{\text{id}}_{\text{stretched},ij} > \tau_{\text{id}}, \\ 1.0, & S^{\text{id}}_{\text{stretched},ij} \leq \tau_{\text{id}}, \end{cases} \qquad \tau_{\text{id}} = 0.5, \tag{6.20}$$

so that disagreement contributes twice as strongly to reduced opinion similarity when identity similarity exceeds the threshold. This amplification produces stronger repulsion within identity groups when opinions diverge, consistent with empirical findings that perceived in-group betrayal elicits more negative reactions than out-group opposition [122].

**Opinion Difference Mechanics ($\Delta x^t_{ij,k}$)**

For each ordered pair $(i, j)$ and topic $k$, the signed opinion difference is:

$$\Delta x^t_{ij,k} \;=\; x^t_{j,k} - x^t_{i,k}. \tag{6.21}$$

This difference term ensures that influence is *directional*: when $j$ is more positive than $i$ on topic $k$, $\Delta x^t_{ij,k} > 0$ and the contribution pushes $i$ upward; when $j$ is more negative, $\Delta x^t_{ij,k} < 0$ and the contribution pushes $i$ downward. When combined with signed ties $\sigma_{ij}$, negative ties invert the direction of the update (repulsion), preventing pure averaging under antagonistic relations and enabling contrastive influence dynamics [84, 204]. This distance-based formulation is critical for modeling polarization: agents do not simply move toward a weighted average of neighbors' positions, but rather experience forces proportional to how far they are from each neighbor, with the sign of the force determined by the signed tie.

**Backfire Amplification Under Repulsive, Opposite-Signed Disagreement**

To model stronger counter-updating under antagonistic exposure, JANUS applies a multiplicative backfire factor $B^t_{ij,k} \geq 1$ to the neighbor contribution when (i) the signed tie between $i$ and $j$ is negative ($\sigma_{ij} < 0$ [8, 143], indicating repulsion) and (ii) the two agents hold opposite-signed opinions on topic $k$ (cross-zero disagreement) [46, 145]. Define the indicator functions:

$$N_{ij} = \mathbb{I}[\sigma_{ij} < 0]\,, \qquad D^t_{ij,k} = \mathbb{I}\big[x^t_{i,k}\, x^t_{j,k} < 0\big]\,, \tag{6.22}$$

where $N_{ij}$ flags a repulsive tie and $D^t_{ij,k}$ flags cross-zero disagreement on topic $k$. The backfire multiplier is then:

$$B^t_{ij,k} = 1 + \lambda_{\text{bf}}\, N_{ij}\, D^t_{ij,k}, \qquad \lambda_{\text{bf}} = 1.25, \tag{6.23}$$

with $\lambda_{\text{bf}} = 1.25$ controlling the strength of backfire amplification. When either condition is not met (i.e., $\sigma_{ij} \geq 0$ or $x^t_{i,k}x^t_{j,k} \geq 0$), the multiplier reduces to $B^t_{ij,k} = 1$ and no amplification is applied. This mechanism captures the empirical finding that exposure to opposing views from disliked sources can strengthen prior beliefs rather than moderate them [8].

**Integration: Computing the Signed Influence Matrix**

The final signed influence weight $M_{ij}$ combines network structure $\widehat{W}_{ij}$ and psychological compatibility $\sigma_{ij}$, then normalizes by absolute row sums to ensure balanced positive/negative influence [57, 123, 204]:

$$M^{\text{raw}}_{ij} = \widehat{W}_{ij} \cdot \sigma_{ij}, \tag{6.24}$$

$$M_{ij} = \frac{M^{\text{raw}}_{ij}}{\sum_g |M^{\text{raw}}_{ig}|}, \tag{6.25}$$

where the absolute row sum normalization ensures that an agent with equal positive and negative ties receives zero net influence when neighbors are equidistant. This second normalization (beyond

the row-stochastic $\widehat{W}$) is necessary because signed ties can produce both positive and negative entries in $M^{\text{raw}}$; without absolute normalization, the magnitude of influence would depend on the balance of positive vs. negative ties rather than the strength of disagreement.

Substituting into Eq. 6.12, the signed social influence for agent $i$ on topic $k$ is:

$$\Delta_{i,k}^{\text{soc}}(t) = \sum_j M_{ij} \ \left(x_{j,k}^t - x_{i,k}^t\right) \ B_{ij,k}^t, \tag{6.26}$$

which aggregates distance-based attraction/repulsion forces from all neighbors, weighted by network structure and psychological compatibility, and amplified under antagonistic exposure [57, 204]. This formulation enables JANUS to reproduce empirically observed patterns of polarization, echo chambers, and backfire effects within a unified mechanistic framework.

**Worked Example: Signed Tie Computation**

To illustrate the signed tie mechanism, consider two agents $i$ and $j$ with the following attributes:

- **Identity similarity**: After block averaging and percentile stretching, $S_{\text{stretched},ij}^{\text{id}} = 0.7$ (high similarity, indicating shared demographics or personality)
- **Transformed identity**: Using Eq. 6.18, $\widetilde{S}_{ij}^{\text{id}} = 2 \times 0.7 - 1 = 0.4$ (positive baseline tie)
- **Salient topics**: Agent $i$ cares about topics $\{1, 2\}$ (both with $s_{i,k} > 0.5$)
- **Opinions**: $x_{i,1} = 0.8$, $x_{i,2} = -0.6$; $x_{j,1} = 0.7$, $x_{j,2} = 0.5$

Since $S_{\text{stretched},ij}^{\text{id}} = 0.7 > 0.5$, the black sheep effect applies: $\beta_{ij} = 2.0$. The opinion similarity calculation proceeds as:

$$\begin{aligned}
|x_{i,1} - x_{j,1}| &= |0.8 - 0.7| = 0.1, \quad \text{scaled by } \beta_{ij} = 2.0 \to 0.2, \\
|x_{i,2} - x_{j,2}| &= |-0.6 - 0.5| = 1.1, \quad \text{scaled by } \beta_{ij} = 2.0 \to 2.2 \text{ (capped at 2.0)}, \\
\text{Mean disagreement} &= \frac{0.2 + 2.0}{2} = 1.1, \\
S_{ij}^{\text{opin}} &= 1 - 1.1 = -0.1 \quad \text{(negative, indicating opinion dissimilarity)}.
\end{aligned}$$

The final signed tie is:

$$\sigma_{ij} = \text{clip}(0.4 + (-0.1), -1, 1) = 0.3,$$

indicating a weakly positive (attractive) tie. Despite high identity similarity, the strong disagreement on topic 2 (amplified by the black sheep effect) reduces the tie strength from a potential maximum of $\approx 0.4$ to $0.3$. If the agents had agreed on both topics, $S_{ij}^{\text{opin}}$ would be near $+1$, yielding $\sigma_{ij} \approx 1.0$ (strong attraction). Conversely, if they disagreed strongly on both topics, $S_{ij}^{\text{opin}}$ could be near $-1$, yielding $\sigma_{ij} \approx -0.6$ (repulsion despite shared identity).

The signed social influence mechanism in JANUS integrates network structure, multi-dimensional identity similarity, directional opinion agreement, and backfire amplification to produce psychologically grounded attraction and repulsion dynamics [84, 147]. By combining distance-based influence with signed ties, the model captures key phenomena in opinion dynamics: polarization, echo chambers, in-group/out-group effects, and antagonistic backfire within a unified framework that is both mechanistically interpretable and empirically grounded [145].

Table 6.4: Plutchik emotion angles and psychometric weights used in circumplex projection.

| **Emotion** | **Angle** $\theta_e$ | **Weight** $w_e$ | **Quadrant** |
|---|---|---|---|
| Joy | 60° | 0.952 | Positive/Active |
| Trust | 300° | 1.613 | Positive/Passive |
| Anticipation | 30° | 1.218 | Positive/Active |
| Surprise | 90° | 0.727 | Neutral/Active |
| Fear | 120° | 0.605 | Negative/Active |
| Anger | 150° | 0.618 | Negative/Active |
| Disgust | 210° | 1.000 | Negative/Passive |
| Sadness | 240° | 1.267 | Negative/Passive |

### 6.4.5 (III) Affective Emotional Drift

JANUS includes an *affective emotional drift* mechanism that biases opinions according to an agent's baseline affect toward a topic. This term captures two key phenomena: (i) agents may have persistent affective predispositions on a topic (e.g., generally positive or negative feelings), and (ii) affect-driven pressure to shift opinions is expressed primarily when the topic is emotionally *activated* (high arousal). Affective drift enters the dynamic component of Eq. 6.7 and is scaled by the susceptible mass $(1 - \alpha_{i,k}^{\text{eff}})$, so highly stubborn agent-topic pairs exhibit correspondingly weaker affect-driven motion, while susceptible agents are more responsive to their baseline affect.

#### Emotion Representation: Plutchik's Eight Basic Emotions

Each agent-topic pair $(i, k)$ is initialized with eight Plutchik emotion intensities $e_{i,k}(\cdot) \in [0, 1]$ representing joy, trust, fear, surprise, sadness, disgust, anger, and anticipation. These intensities are sampled during AURORA population generation (Stage 1, Sec. 6.3) and remain fixed throughout the simulation. To integrate these discrete emotion intensities into the opinion dynamics model, I map them to a continuous two-dimensional affect space using the Russell circumplex model [154, 166].

Each emotion is assigned a fixed angle $\theta_e$ on the circumplex, representing its position in valence-arousal space (Fig. 6.2). Table 6.4 lists the angles and psychometric weights used in JANUS. The angles follow a Plutchik-style arrangement with 0° at positive valence (right) and angles increasing counterclockwise. Psychometric weights $w_e$ approximate the relative loading of discrete emotions onto the underlying valence–arousal latent space. Conceptually, these weights act as scaling coefficients in a weighted vector projection, ensuring that emotions with stronger psychometric associations to core affect dimensions contribute more heavily to the resulting continuous representation. The weighting scheme is informed by established circumplex models of affect [154, 166] and preserves the relative salience of positive (e.g., joy, trust) and negative (e.g., anger, sadness) affective components.

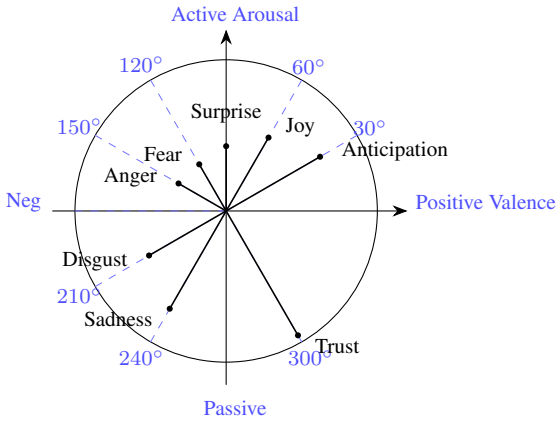



Figure 6.2: Valence-arousal circumplex with Plutchik-style emotion placement (0° at positive valence, angles increasing counterclockwise). Vector lengths represent normalized psychometric weights (trust is longest at 1.0, fear is shortest at 0.375 relative to trust), reflecting empirical affect structure. Emotion intensities are projected onto the valence axis via $\cos(\theta_e)$ and onto the arousal axis via $\sin(\theta_e)$.

### Circumplex Projection: Emotion → Valence-Arousal

The Russell circumplex provides a parsimonious two-dimensional representation of affect that is empirically validated and computationally tractable [154, 166], reducing eight emotion intensities to two interpretable coordinates: valence (pleasantness, horizontal axis) and arousal (activation, vertical axis). I compute baseline valence $v^0_{i,k} \in [-1, 1]$ and arousal $a^0_{i,k} \in [0, 1]$ for each agent-topic pair via intensity-weighted geometric projection:

$$v^0_{i,k} = \sum_{e \in \mathcal{E}} e_{i,k}(e) \cdot \cos(\theta_e) \tag{6.4}$$

$$a'^0_{i,k} = \sum_{e \in \mathcal{E}} e_{i,k}(e) \cdot \sin(\theta_e) \tag{6.5}$$

$$a^0_{i,k} = 0.5 \times (1 + a'^0_{i,k}) \tag{6.6}$$

where $\mathcal{E}$ is the set of eight Plutchik emotions, $e_{i,k}(e)$ is the intensity of emotion $e$ for agent $i$ on topic $k$, and $\theta_e$ is the circumplex angle for emotion $e$ (Table 6.4). The final valence is clipped to $[-1, 1]$ and arousal to $[0, 1]$ to ensure valid coordinates.

**Intuition** Emotions in the positive-valence quadrants (joy at 60°, trust at 300°, anticipation at 30°) contribute positively to $v^0_{i,k}$ via $\cos(\theta_e) > 0$, while negative-valence emotions (fear, anger, disgust, sadness, all at angles $> 90°$ and $< 270°$) contribute negatively via $\cos(\theta_e) < 0$. Similarly, high-arousal emotions in the upper half of the circumplex (surprise at 90°, fear at 120°, anger at 150°) increase $a^0_{i,k}$ via $\sin(\theta_e) > 0$, while low-arousal emotions in the lower half (trust at 300°, sadness at 240°, disgust at 210°) contribute less to arousal via $\sin(\theta_e) < 0$ or near-zero values.

**Worked Example** Consider an agent with high joy ($e_{i,k}(\text{joy}) = 0.8$) and moderate trust ($e_{i,k}(\text{trust}) = 0.5$), with all other emotions near zero. The valence and arousal coordinates are:

$$\begin{aligned} v^0_{i,k} &\approx 0.8 \times \cos(60^\circ) + 0.5 \times \cos(300^\circ) \\ &= 0.8 \times 0.5 + 0.5 \times 0.5 = 0.65 \quad \text{(positive valence)}, \\ a'^0_{i,k} &\approx 0.8 \times \sin(60^\circ) + 0.5 \times \sin(300^\circ) \\ &= 0.8 \times 0.866 + 0.5 \times (-0.866) \approx 0.26, \\ a^0_{i,k} &= 0.5 \times (1 + 0.26) = 0.63 \quad \text{(moderate-to-high arousal)}. \end{aligned}$$

This agent has a positive affective predisposition ($v^0_{i,k} = 0.65$) and sufficient arousal ($a^0_{i,k} = 0.63 > 0.5$) to activate affective drift (affect$_d rift_e x$).

**Drift Mechanism: Arousal Gate and Valence Bias**

Affective drift is enabled only when baseline arousal exceeds a fixed activation threshold $\tau_a$, ensuring that drift is selective and occurs primarily for emotionally activating topics. The arousal gate is defined as:

$$G_{i,k}(a^0_{i,k}) = \mathbb{I}\left[a^0_{i,k} > \tau_a\right], \qquad \tau_a = 0.5, \tag{6.27}$$

where $\mathbb{I}[\cdot]$ is the indicator function. When the gate is active ($G_{i,k} = 1$), the drift contribution is a constant topic-specific bias proportional to baseline valence:

$$D^{\text{aff}}_{i,k} = G_{i,k}(a^0_{i,k}) \cdot \beta_v \cdot v^0_{i,k}, \qquad \beta_v = 0.1, \tag{6.28}$$

where $\beta_v \geq 0$ is the valence sensitivity parameter controlling the magnitude of affect-driven opinion shift. This construction makes affective drift *sign-consistent* (direction determined by $v^0_{i,k}$: positive valence biases opinions upward, negative valence biases downward) and *selective* (present only for sufficiently arousing topics with $a^0_{i,k} > 0.5$).

**Rationale for Arousal Gating** The arousal gate ensures that affective drift is not universally applied to all topics. Low-arousal topics, those that an agent finds emotionally unengaging, do not trigger affect-driven pressure, even if the agent has a strong positive or negative valence toward them. This reflects the psychological intuition that emotional activation is necessary for the affect to influence behavior and judgment [166]. In JANUS, only topics that cross the arousal threshold exert affective drift, producing topic-dependent heterogeneity in the strength of affect-driven dynamics.

**Integration with Main Update Equation** Under fixed affect parameters, $D^{\text{aff}}_{i,k}$ is time-invariant and enters the dynamic component of the main opinion update (Eq. 6.7) alongside signed social influence:

$$x^{t+1}_{i,k} = \alpha^{\text{eff}}_{i,k} x^0_{i,k} + (1 - \alpha^{\text{eff}}_{i,k}) \left[\Delta^{\text{soc}}_{i,k}(t) + D^{\text{aff}}_{i,k}\right], \tag{6.29}$$

where $\Delta^{\text{soc}}_{i,k}(t)$ is the signed social influence term (Eq. 6.12). Because drift is scaled by the susceptible mass $(1 - \alpha^{\text{eff}}_{i,k})$, highly stubborn agents ($\alpha^{\text{eff}}_{i,k} \to 1$) experience weak drift regardless of

their baseline affect, while susceptible agents ($\alpha_{i,k}^{\text{eff}} \to 0$) are maximally responsive to both social influence and affective bias [150]. Conceptually, this overlays a simple "circumplex-conditioned bias" onto the Friedkin-Johnsen anchor-and-influence framework: arousal gates whether affect is engaged at all (vertical axis of the circumplex), while valence determines the direction of the additive drift (horizontal axis) [154, 166].

The affective drift mechanism in JANUS provides a topic-specific, psychologically grounded pathway for baseline emotions to bias opinion change. By mapping discrete Plutchik intensities to continuous valence-arousal coordinates and gating drift by arousal, the model captures the intuition that emotionally activating topics exert persistent affective pressure, while low-arousal topics remain subject primarily to social influence and anchor retention. This design enables JANUS to reproduce empirically observed patterns wherein moralized or high-affect issues exhibit stronger polarization and resistance to consensus [26, 173].

## 6.5 Experimental Design

Experimental design follows a two-step process. First, a synthetic population is generated using AURORA. Then, the opinion dynamics are simulated according to our model (Eq. 6.7). Experimental conditions are designed to isolate how (i) signed ties and affective drift affect polarization, and (ii) multi-topic dynamics scale with population size $N$.

### 6.5.1 Synthetic Population Generation

Agent populations were generated with the AURORA component, which samples each agent's demographic and psychometric attributes from empirical U.S. distributions and assigns them to advocacy/community organizations. Demographics (age, gender, race/ethnicity, education, political affiliation), psychological profiles (Big Five/OCEAN and Moral Foundations), baseline affect (Plutchik-derived valence–arousal coordinates), and group memberships (1–3 groups; 10% leaders) are initialized once and held fixed, while only topic-specific opinions (initialized over $[-1, 1]$ with randomly assigned salience) evolve over time under Eq. 6.7.

### 6.5.2 Experimental Conditions

I conducted a single-factor experiment varying population size while holding the topical set fixed at $K{=}2$. All runs represent a U.S. 2022 context, and the two topics are (i) **COVID-19 Response**: pandemic policy debate (mandates, closures, masking), and (ii) **Roe v. Wade**: abortion policy after the 2022 Supreme Court decision. All factors other than $N$ were held constant, including topology (deterministic co-membership construction), demographic sampling procedure, and model parameters (Table 6.5). Random seeds differed across scenarios but were fixed within each scenario for reproducibility.

Table 6.5: Fixed parameters across conditions.

| **Parameter** | **Symbol** | **Value** |
|---|---|---|
| Salience-susceptibility coupling | $\gamma$ | 0.7 |
| In-group disagreement scaling | $\beta_{ij}$ | 2.0 if $S_{ij}^{\mathrm{id}} \geq 0.5$ |
| Leader sender-weight boost | (Eq. 6.13) | $+200\%$ ($3.0\times$) |
| Backfire strength | $\lambda_{\mathrm{bf}}$ | 1.2 |
| Valence sensitivity | $b$ | 0.1 |
| Arousal threshold | $\tau_a$ | 0.5 |
| Salience threshold | $\tau_s$ | 0.5 |
| Max step change | $\delta_{\max}$ | 0.3 |

Table 6.6: Experimental conditions (U.S. 2022; two topics).

| **Scenario** | **N** | **Topics (K)** |
|---|---|---|
| Dual-250 | 250 | COVID-19 Response; Roe v. Wade |
| Dual-500 | 500 | COVID-19 Response; Roe v. Wade |
| Dual-750 | 750 | COVID-19 Response; Roe v. Wade |
| Dual-1000 | 1000 | COVID-19 Response; Roe v. Wade |

Each scenario ran for $T{=}250$ synchronous baseline timesteps, which was sufficient for the average stance change between timesteps (**Avg.** $|\Delta x|$) to stabilize below $0.01$ for all runs.

### 6.5.3 Variables

To isolate the effect of population size on opinion dynamics, the experiment uses a controlled single-factor design. Population size is the only manipulated variable, while all other model components remain fixed across scenarios. Table 6.7 summarizes the independent variable, dependent variables, and control variables used in the experiment.

This experimental design allows differences in polarization trajectories, opinion flip rates, and opinion distribution evolution to be attributed directly to changes in scale rather than to changes in agent composition, topic structure, or update rules. The dependent variables were selected to capture both macro-level system behavior and micro-level agent dynamics. Polarization, mean opinion, and distribution shape characterize aggregate outcomes, while flip rate and trajectory variance capture the extent to which individual agents undergo meaningful belief change. Together, these measures provide a comprehensive view of how increasing population size alters the dynamics of belief formation and stabilization in JANUS.

Table 6.7: Summary of experimental variables used in the JANUS population-scale analysis.

| **Category** | **Variable** | **Description** |
|---|---|---|
| Independent | Population Size ($N$) | $N \in \{250, 500, 750, 1000\}$, spanning small-community to larger population interaction environments while maintaining computational tractability. |
| Dependent | Polarization ($\sigma_x$) | Standard deviation of opinions at each timestep; higher values indicate greater ideological dispersion. |
| | Opinion Flip Rate ($f_{\text{flip}}$) | Percentage of agents whose opinions cross the neutral boundary between initial and final states. |
| | Mean Opinion ($\bar{x}(t)$) | Population-average opinion over time, used to detect directional shifts in aggregate stance. |
| | Distribution Shape | Violin plots of opinion distributions at $t \in \{0, 62, 125, 187, 250\}$ to evaluate bimodality and convergence patterns. |
| | Cumulative Opinion Change($\Delta x_i$) | Magnitude of opinion movement for each agent across the simulation. |
| | Polarization Trend | Regression slope of $\sigma_x(t)$ over $t \in [50, 250]$ to identify increasing, decreasing, or stable polarization. |
| Controls | Network Structure | Co-membership network; agents connected through shared group affiliations (1–3 groups per agent; $\sim$10% leaders). |
| | Agent Attributes | Demographics sampled from U.S. 2022 distributions; psychological traits include Big Five, Moral Foundations, emotion affect (valence–arousal), and baseline susceptibility. |
| | Topic Configuration | Two topics ($K = 2$): COVID-19 Response and Roe v. Wade. Salience sampled from $[0, 1]$; initial opinions from $[-1, 1]$. |
| | Model Parameters | All parameters fixed across scenarios (Table 6.5), including $\gamma$, $\lambda_{\text{bf}}$, $\delta_{\max}$,$\kappa_{\text{flip}}$, $\eta_{\text{drift}}$, and $\sigma_{\text{noise}}$. |
| | Simulation Settings | $T = 250$ timesteps with synchronous baseline updates under structural co-membership exposure; U.S. 2022 context; seeds fixed within scenarios for reproducibility. |

### 6.5.4 Platform-based Interaction in JANUS+Synthetic Social Media (SynSM) Simulation Loop

In Eq. 6.7, exposure enters through the normalized weight term $\widehat{W}_{ij}$, which specifies the relative influence of agent $j$ on agent $i$. More generally, $\widehat{W}_{ij}$ defines which other agents are likely to enter an individual's information environment and therefore contribute to opinion change.

A central modeling choice in JANUS is how $\widehat{W}_{ij}$ is defined. The model supports multiple exposure formulations depending on the interaction setting, though only one is used in any given experiment. In the baseline JANUS formulation, $\widehat{W}_{ij}$ is derived from group co-membership, so exposure reflects stable social connectedness. In the JANUS+SynSM formulation used here, by contrast, exposure is platform-mediated: SynSM (Synthetic Social Media) is a time-stepped, agent-based simulation environment that models message visibility, content generation, user interaction, and information diffusion on social media platforms [12]. Rather than assuming that all structurally connected neighbors are behaviorally active at each step, SynSM determines which messages become visible to which agents during the platform simulation loop. JANUS then consumes SynSM visibility and interaction outcomes as exposure evidence when computing opinion change. In this sense, $\widehat{W}_{ij}$ represents whether agent $j$'s content is surfaced to, noticed by, and interacted with by agent $i$ at a given timestep. SynSM therefore functions as the platform layer of the framework, translating agents' internal states into observable social media behavior and feeding the resulting exposure patterns back into JANUS.

Concretely, I evaluate JANUS+SynSM in a set of experiments conducted in the SynSM environment. Table 6.8 summarizes the four topic–population cases used in the Canadian–American JANUS+SynSM experiment.

## 6.6 Results

Our analysis addresses three research questions: (i) how network scale ($N$) shapes interaction opportunities and mesoscale structure, (ii) whether topic-level polarization or consensus emerges from the model dynamics, and (iii) whether stance reversals occur broadly across the population or concentrate among susceptible agents. The results show systematic scale-dependent dynamics: smaller networks amplify polarization through echo chamber reinforcement, while larger networks equilibrate more rapidly due to stronger averaging effects. However, polarization persists across all scales because community structure prevents convergence to neutral consensus.

### 6.6.1 Network Scaling and Local Interaction Structure

JANUS models opinion updating through local neighborhood aggregation rather than global all-to-all interaction (Eq. 6.12). Agents influence only those to whom they are connected in a structured social network derived from shared group memberships (Eq. 6.13). This co-membership mechanism produces sparse but highly clustered networks that approximate real-world patterns of localized social influence.

Table 6.8: Topic–population cases evaluated in the Canadian–American JANUS+SynSM experiment.

| Topic | Population | $n$ | Role in Experiment |
|---|---|---|---|
| Canadian Indigenous Rights | Canadian | 383 | Domestic topic case used to examine opinion dynamics and event responses within the Canadian synthetic population. |
| Canadian Indigenous Rights | American | 407 | Cross-national comparison on the same topic, used to compare how the American synthetic population responds relative to the Canadian population. |
| Russian Arctic Activity | Canadian | 385 | International security topic case used to examine opinion dynamics and event responses within the Canadian synthetic population. |
| Russian Arctic Activity | American | 413 | Cross-national comparison on the same geopolitical topic, used to compare how the American synthetic population responds relative to the Canadian population. |

As population size increases, the number of potential interactions expands while network density remains approximately constant. Influence weights are row-normalized (Eq. 6.14) so that total influence on each agent remains bounded regardless of network size. Table 6.9 summarizes the resulting network statistics. Density remains stable ($d \approx 0.076 \pm 0.002$), indicating consistently sparse connectivity across scenarios.

Table 6.9: Network scaling by population size. All networks exhibit sparse connectivity ($\rho \approx 0.076$) and high clustering ($C \approx 0.89$) while interaction volume scales quadratically with $N$.

| Scenario | $N$ | Avg. deg. | Density ($d$) | Clustering | Exposures/step |
|---|---|---|---|---|---|
| Dual-250 | 250 | 17.8 | 0.074 | 0.88 | ≈4,450 |
| Dual-500 | 500 | 36.0 | 0.075 | 0.89 | ≈18,000 |
| Dual-750 | 750 | 55.7 | 0.077 | 0.89 | ≈41,775 |
| Dual-1000 | 1000 | 73.5 | 0.077 | 0.89 | ≈73,500 |

Although density remains constant, average degree increases nearly linearly with population size ($r = 0.9998$, $p < 0.001$), rising from 17.8 connections per agent in Dual-250 to 73.5 in Dual-1000. This growth produces a quadratic increase in interaction volume, from roughly 4,450 exposures per timestep to approximately 73,500. Thus, even though agents interact locally, larger populations dramatically increase the number of influence opportunities across the system.

Clustering coefficients remain high across all scenarios ($C \approx 0.89$), indicating strong transitivity within group-based communities. These clusters represent densely connected regions of shared

affiliation, while sparse bridging ties connect otherwise separate communities. Such bridges allow influence to propagate between clusters and can either promote convergence or trigger repulsive backfire effects depending on the interaction sign $\sigma_{ij}$ (Eq. 6.23).

Figure 6.3 illustrates the resulting community structure for the Dual-250 network. Dense clusters formed through shared group membership dominate the topology, while sparse bridging ties provide the primary pathways for cross-community exposure.

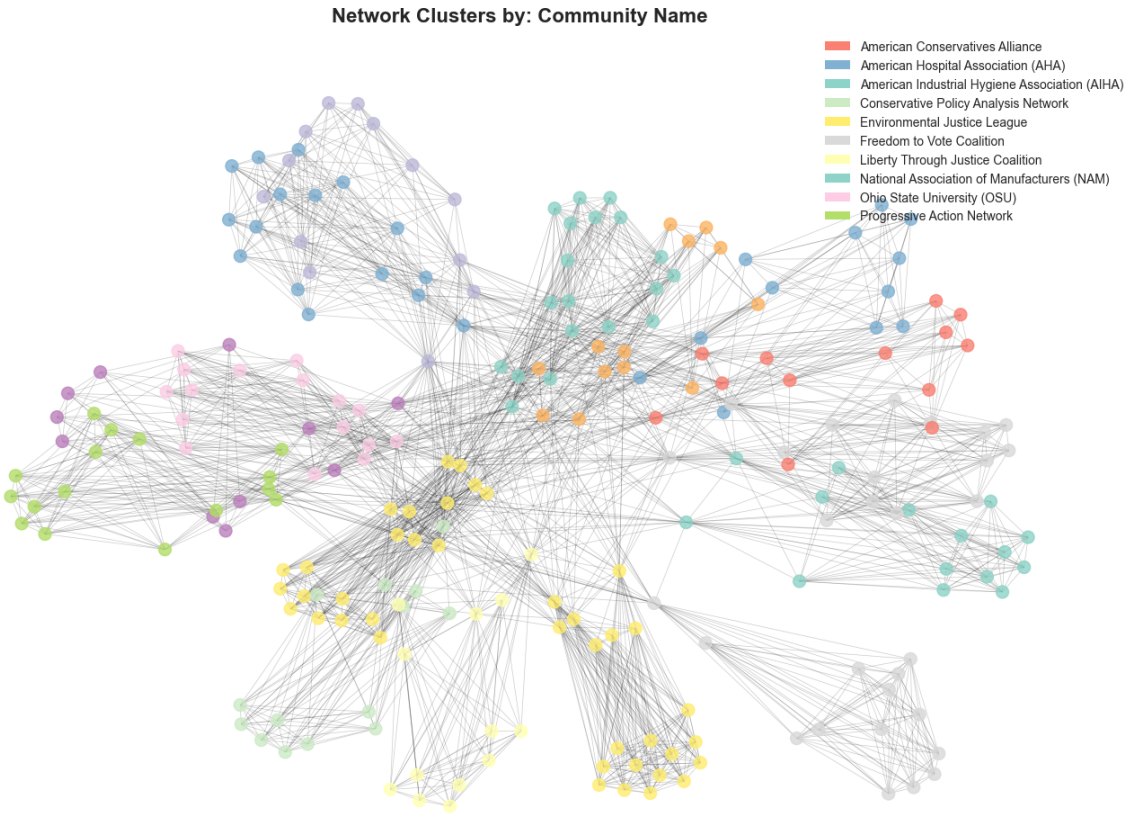



Figure 6.3: Dual-250 community structure under the Fruchterman-Reingold spring layout with group co-membership weights. Dense regions represent tightly connected communities formed through shared group affiliations, while sparse bridges enable cross-community exposure. The high clustering coefficient ($C = 0.88$) reflects strong transitivity within communities, while the relatively low density ($d = 0.074$) indicates that most potential connections remain unrealized.

### 6.6.2 Topic-Level Outcomes and Polarization Dynamics

Table 6.10 summarizes topic-level outcomes across all scenarios, including initial and final polarization, mean opinion shifts, cumulative opinion change, and stance flip statistics. Four key patterns emerge.

Table 6.10: Topic-level outcomes by scenario ($t = 0 \rightarrow t = 250$). $\Delta\sigma \equiv \sigma_{250} - \sigma_0$ measures polarization change. "Flips" counts agents whose stance sign changed between initial and final states. Avg. $|\Delta x|$ represents mean cumulative opinion change per agent.

| **Scenario** | **Topic** | $\mu_0$ | $\mu_{250}$ | $\sigma_0$ | $\sigma_{250}$ | $\Delta\sigma$ | **Avg.** $\lvert\Delta x\rvert$ | **Flips** | **Flip %** |
|---|---|---|---|---|---|---|---|---|---|
| Dual-250 | COVID | $-0.154$ | $-0.129$ | 0.575 | 0.621 | $+0.046$ | 0.139 | 5 | 2.08% |
| Dual-250 | Roe | $-0.156$ | $-0.127$ | 0.632 | 0.683 | $+0.051$ | 0.116 | 3 | 1.25% |
| Dual-500 | COVID | $-0.202$ | $-0.174$ | 0.583 | 0.604 | $+0.021$ | 0.104 | 7 | 1.46% |
| Dual-500 | Roe | $-0.045$ | $-0.046$ | 0.650 | 0.638 | $-0.011$ | 0.118 | 4 | 0.83% |
| Dual-750 | COVID | $-0.233$ | $-0.195$ | 0.582 | 0.558 | $-0.024$ | 0.114 | 8 | 1.11% |
| Dual-750 | Roe | $-0.135$ | $-0.105$ | 0.647 | 0.624 | $-0.024$ | 0.090 | 8 | 1.11% |
| Dual-1000 | COVID | $-0.025$ | $-0.025$ | 0.595 | 0.590 | $-0.005$ | 0.119 | 13 | 1.35% |
| Dual-1000 | Roe | $-0.097$ | $-0.076$ | 0.651 | 0.650 | $-0.001$ | 0.104 | 6 | 0.62% |

First, *Roe v. Wade* exhibits consistently higher polarization than COVID-19 across all scenarios. Final Roe polarization ranges from 0.624 to 0.683 ($M = 0.649$), while COVID ranges from 0.558 to 0.621 ($M = 0.593$). A paired-samples $t$-test confirms the difference ($t(3) = 5.89$, $p = 0.010$, Cohen's $d = 2.15$), indicating that the persona-derived initialization successfully captures real-world variation in issue polarization.

Second, polarization change ($\Delta\sigma$) reverses with increasing population size. Smaller networks amplify polarization (Dual-250, Dual-500), whereas larger networks produce moderation or stabilization (Dual-750, Dual-1000). For COVID, the relationship between $\log(N)$ and $\Delta\sigma$ is strongly negative ($r = -0.96$, $p = 0.040$), suggesting that increasing interaction volume introduces stronger averaging effects that counteract local reinforcement.

Third, average individual opinion change shows modest but non-monotonic variation across scenarios. Smaller populations generally display higher volatility, but differences remain limited, indicating that individual-level movement persists even when aggregate polarization stabilizes.

Finally, polarization remains substantial across all scenarios ($\sigma_{250} \in [0.558, 0.683]$), with no convergence toward consensus. This persistence reflects the Friedkin–Johnsen anchor mechanism (Eq. 6.10), in which agents retain attachment to initial positions while still responding to social influence. Combined with flip resistance ($\kappa_{\text{flip}} = 0.45$), bounded updates ($\delta_{\max} = 0.3$), and minimal anchor drift ($\eta_{\text{drift}} = 0.004$), the model preserves stable bimodal opinion distributions throughout the simulation.

### 6.6.3 Comprehensive Network Scale Analysis

To examine how network scale shapes opinion dynamics beyond polarization alone, I compare coordination patterns and community structure across all four scenarios. Tables 6.11 and 6.12 summarize the key results, allowing direct comparison of polarization change, stance stability, and coordination structure across populations ranging from 250 to 1000 agents.

Several patterns are consistent across scales. First, polarization reverses with increasing population size: Dual-250 amplifies polarization, Dual-500 shows mixed topic-level behavior, and

Table 6.11: Comprehensive metrics across all network scales. Polarization change ($\Delta\sigma$) shows systematic reversal from amplification (250) to stabilization (1000). Within-group trajectory correlations remain positive while between-group trajectory correlations remain negative across all scales, demonstrating persistent community structure. Here, $r_{\text{within}}$ denotes the mean pairwise correlation of opinion trajectories among agents in the same community, and $r_{\text{between}}$ denotes the mean pairwise correlation among agents in different communities.

| **Scenario** | **Topic** | **N** | $\sigma_0$ | $\sigma_{250}$ | $\Delta\sigma$ | **Flip %** | $r_{\text{within}}$ | $r_{\text{between}}$ |
|---|---|---|---|---|---|---|---|---|
| Dual-250 | COVID-19 | 250 | 0.575 | 0.621 | $+0.046$ | 2.08 | +0.028 | $-0.044$ |
| | Roe v. Wade | 250 | 0.632 | 0.683 | $+0.051$ | 1.25 | +0.013 | $-0.039$ |
| Dual-500 | COVID-19 | 500 | 0.583 | 0.604 | $+0.021$ | 1.46 | +0.014 | $-0.017$ |
| | Roe v. Wade | 500 | 0.650 | 0.638 | $-0.011$ | 0.83 | +0.022 | $-0.031$ |
| Dual-750 | COVID-19 | 750 | 0.582 | 0.558 | $-0.024$ | 1.11 | +0.028 | $-0.039$ |
| | Roe v. Wade | 750 | 0.647 | 0.624 | $-0.024$ | 1.11 | +0.014 | $-0.020$ |
| Dual-1000 | COVID-19 | 1000 | 0.595 | 0.590 | $-0.005$ | 1.35 | +0.025 | $-0.031$ |
| | Roe v. Wade | 1000 | 0.651 | 0.650 | $-0.001$ | 0.63 | +0.017 | $-0.036$ |

Table 6.12: Trajectory-correlation dynamics across network scales. Mean trajectory correlation ($\bar{r}$) remains near zero across all scenarios, indicating that positive and negative co-movement patterns balance out at the population level rather than implying a lack of structure. Correlation strength (mean $|r|$) remains consistently positive, showing sustained correlated opinion movement. The separation metric ($r_{\text{within}} - r_{\text{between}}$) quantifies community polarization and remains positive across all scales, while the fraction of positive correlations indicates balanced opposing movement patterns.

| **Scenario** | **Topic** | $\bar{r}$ | **Corr. Strength** | **Separation** | **Frac. Pos.** |
|---|---|---|---|---|---|
| Dual-250 | COVID-19 | +0.005 | 0.092 | 0.072 | 0.499 |
| | Roe v. Wade | +0.001 | 0.080 | 0.053 | 0.499 |
| Dual-500 | COVID-19 | +0.005 | 0.075 | 0.032 | 0.516 |
| | Roe v. Wade | +0.000 | 0.084 | 0.052 | 0.501 |
| Dual-750 | COVID-19 | +0.005 | 0.090 | 0.068 | 0.509 |
| | Roe v. Wade | +0.004 | 0.078 | 0.034 | 0.509 |
| Dual-1000 | COVID-19 | +0.000 | 0.087 | 0.056 | 0.500 |
| | Roe v. Wade | +0.002 | 0.080 | 0.053 | 0.502 |

Dual-750 and Dual-1000 exhibit stabilization or slight moderation. This pattern is consistent with a transition from local reinforcement in smaller networks to stronger averaging in larger ones. Second, community structure remains stable across all scenarios. Within-group correlations are always positive and between-group correlations always negative, indicating that agents embedded in the same community move together while opposed communities move in opposite directions.

Third, coordination strength remains relatively stable even though mean correlation stays near zero. This reflects balanced opposing coordination: a substantial share of agent pairs move together while another substantial share move in opposite directions, producing near-zero average correlation without implying inactivity. The fraction of positive correlations remains close to 0.5 across all scenarios, reinforcing this interpretation. Finally, flip rates remain low in every condition, indicating that stance boundaries remain stable even while opinions continue to move within and across communities.

The following subsections focus on the two extreme cases, Dual-250 and Dual-1000, which best illustrate the transition from reinforcement-dominated to averaging-dominated dynamics.

### 6.6.4 Opinion Distribution Evolution

The evolution of opinion change shows how network scale shapes stabilization patterns and the accumulation of belief updating over time. Both small and large networks exhibit gradual declines in update magnitude, but they differ in volatility, time to stabilization, and the total amount of cumulative opinion movement. For more results on both 500 and 750 scenarios, see Appendix D.1.

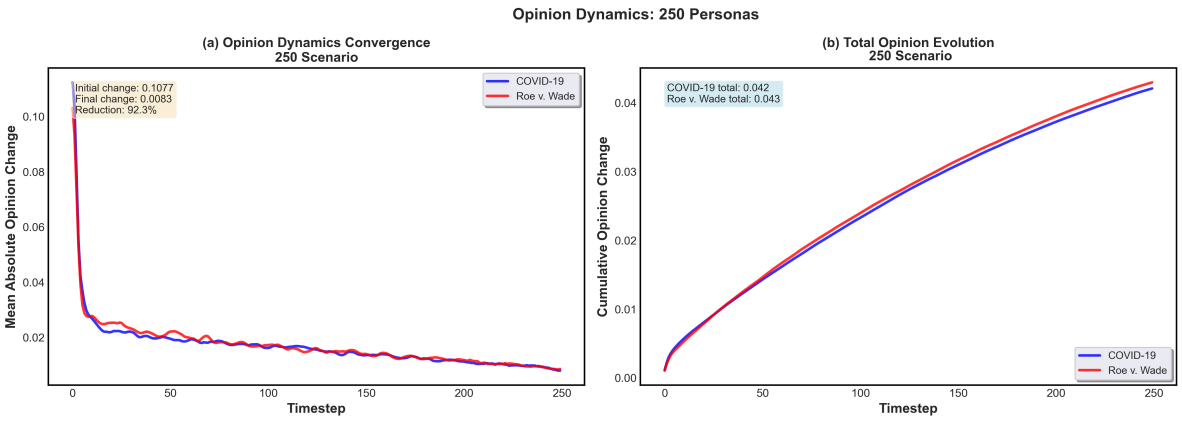



Figure 6.4: Opinion dynamics evolution for Dual-250. (a) Mean absolute opinion change declines over time, indicating stabilization in the rate of updating rather than convergence to consensus. (b) Cumulative opinion change rises throughout the simulation, reaching 0.072 for COVID-19 and 0.079 for Roe v. Wade.

### Dual-250: Small Network Dynamics and Sustained Evolution

Figure 6.4 shows the opinion dynamics for the Dual-250 scenario. Panel (a) indicates that mean absolute opinion change begins at approximately $0.107$ and declines gradually over more than 200 timesteps, reaching near-equilibrium values around $0.010$ by approximately $t = 200$–$215$. This pattern indicates sustained opinion adjustment across the simulation horizon, with an overall reduction of about $94.3\%$ in update magnitude. The trajectories also exhibit visible short-run fluctuation, which is consistent with small-network settings where individual agent updates have a larger effect on aggregate measures. Panel (b) shows that cumulative opinion change continues to rise throughout the full 250 timesteps, reaching $0.072$ for COVID-19 and $0.079$ for *Roe v. Wade*. This continued accumulation indicates that opinion movement does not end after the initial transient phase. Instead, the system remains dynamically active even after the rate of updating has slowed substantially. In other words, the model approaches a bounded, slowly evolving regime rather than an immediately frozen state.

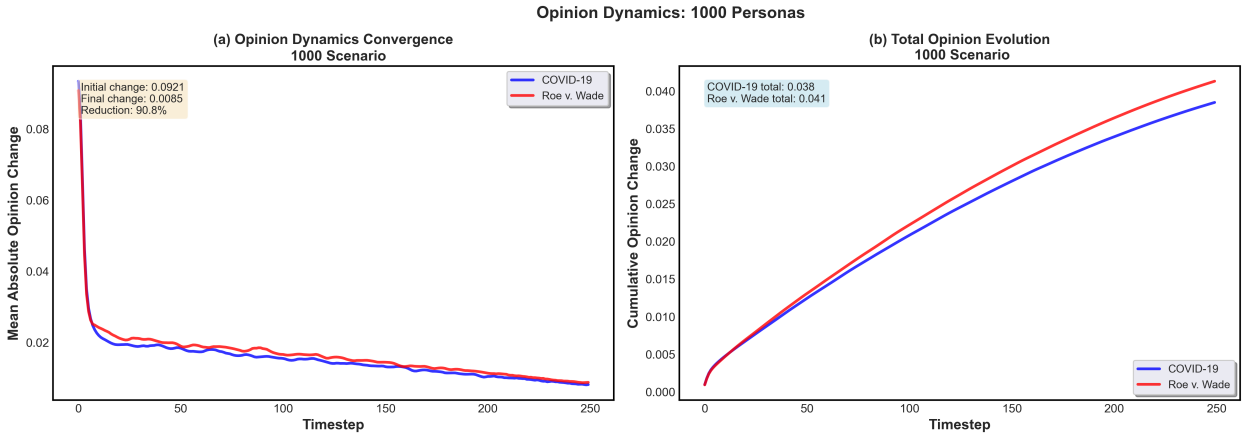



Figure 6.5: Opinion dynamics evolution for Dual-1000. (a) Mean absolute opinion change declines smoothly over time, indicating gradual stabilization in the rate of updating. (b) Cumulative opinion change continues to rise throughout the simulation, but to a slightly lower total level than in Dual-250, reaching 0.038 for COVID-19 and 0.041 for Roe v. Wade.

### Dual-1000: Large Network Dynamics and Faster Stabilization

Figure 6.5 shows a different pattern for Dual-1000. Panel (a) indicates faster stabilization, with mean absolute opinion change beginning at approximately $0.082$ and reaching near-equilibrium values around $0.008$ by approximately $t = 190$–$195$, or roughly 20 timesteps earlier than in the smaller network. The trajectories are also less jagged over time, consistent with stronger averaging across larger neighborhoods, where individual agent updates contribute less to aggregate timestep-to-timestep fluctuation. Panel (b) shows that cumulative opinion change remains substantial but reaches a lower final level than in Dual-250, ending at $0.059$ for COVID-19 and $0.063$ for *Roe v. Wade*. This corresponds to roughly 20% less accumulated opinion movement than in

the smaller network. These lower cumulative totals suggest that broader neighborhoods exert a stabilizing effect by dampening local influence shocks and reducing the magnitude of aggregate opinion adjustment over time.

**Comparative Analysis: Small vs Large Network Convergence Dynamics**

The contrast between Dual-250 and Dual-1000 highlights a clear scale effect in opinion-update dynamics. Smaller networks exhibit greater initial movement, more visible short-run fluctuation, and higher cumulative opinion change. Larger networks also show gradual declines in update magnitude, but the trajectories are smoother and the total accumulated movement is lower. In both scenarios, panel (a) shows stabilization in the rate of updating, while panel (b) shows that opinion change continues to accumulate throughout the simulation horizon.

### 6.6.5 Temporal Dynamics and Trajectory Analysis

Individual trajectories clarify how the macro-level differences emerge from agent-level behavior. Table 6.13 summarizes mean opinion movement, flip counts, and flip rates across all scenarios.

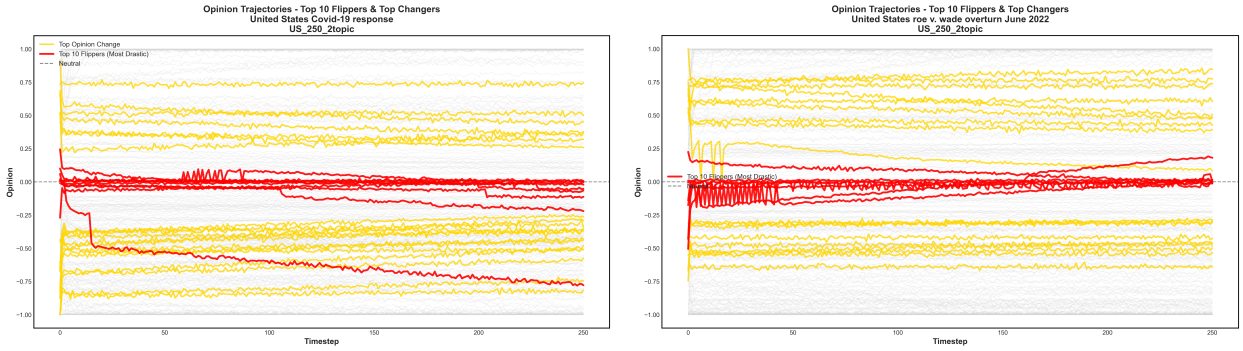
Table 6.13: Trajectory dynamics across network scales. Mean opinion change quantifies average individual mobility. Flip rates remain low across all scenarios, with slight decrease at intermediate scales. High-mobility agents (top 10 by cumulative $|\Delta x|$) show sustained volatility while the majority stabilize quickly.

| Scenario | Topic | Avg Degree | Mean $\|\Delta x\|$ | Flips | Flip Rate (%) |
|---|---|---|---|---|---|
| Dual-250 | COVID-19 | 17.8 | 0.139 | 5 | 2.08 |
| | Roe v. Wade | 17.8 | 0.116 | 3 | 1.25 |
| Dual-500 | COVID-19 | 36.0 | 0.104 | 7 | 1.46 |
| | Roe v. Wade | 36.0 | 0.118 | 4 | 0.83 |
| Dual-750 | COVID-19 | 55.7 | 0.114 | 8 | 1.11 |
| | Roe v. Wade | 55.7 | 0.090 | 8 | 1.11 |
| Dual-1000 | COVID-19 | 73.5 | 0.119 | 13 | 1.35 |
| | Roe v. Wade | 73.5 | 0.104 | 6 | 0.63 |

Mean opinion change remains in a relatively narrow band across scenarios, indicating that individual-level movement persists even when aggregate polarization stabilizes. Flip rates are consistently low, suggesting that the model allows meaningful belief movement without widespread boundary crossing. The main scale effect is therefore not whether opinions move, but how that movement is organized and whether it accumulates into continued polarization.

### Dual-250: Small Network Trajectories and Early Transient Dynamics

Figure 6.6 shows that Dual-250 produces pronounced early movement. During roughly the first 50 timesteps, many agents shift rapidly as they adjust to their neighborhoods. Because average degree is relatively low, each neighbor carries more weight, so local configurations can produce sharper directional change. After this transient, most trajectories flatten into *local quasi-equilibria*, with subsequent movement driven mostly by bounded noise rather than systematic drift.




(a) COVID-19 Response

(b) Roe v. Wade

Top Opinion Change  Top 10 Flippers  Other Agents  Neutral

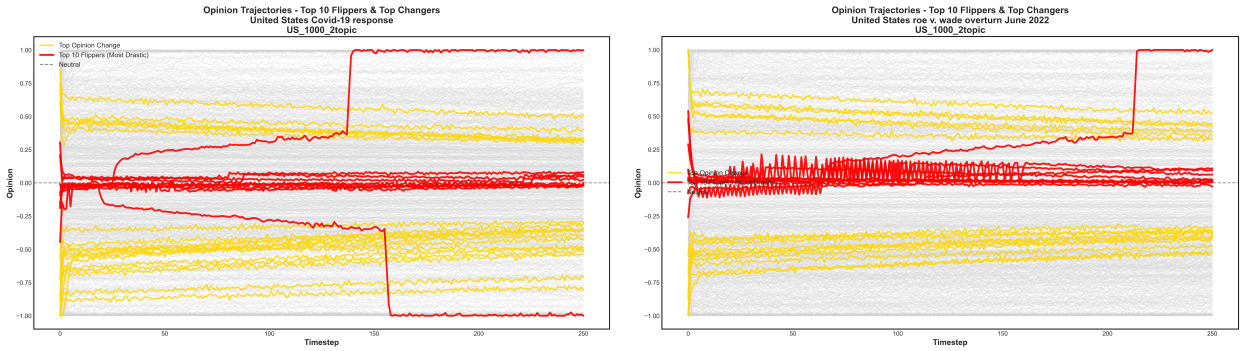
Figure 6.6: Opinion trajectories for Dual-250 (N=240). Gray lines show all agents, yellow highlights top 10 most mobile agents, red marks stance flippers. Both topics exhibit rapid initial adjustment (t ¡ 50) followed by stabilization. High-mobility agents show sustained volatility while majority remain near initial positions. COVID-19: 5 flips (2.08%); Roe: 3 flips (1.25%), all during early transient phase.

A small subset of highly mobile agents continues to move more substantially than the rest of the population. These agents appear to occupy positions where they receive more heterogeneous or cross-cutting signals. Stance flips remain rare and occur early, indicating that major opinion reversals are concentrated in the transient phase rather than occurring repeatedly throughout the run, which is in line with real world analysis [141].

### Dual-1000: Large Network Trajectories and Rapid Equilibration

Figure 6.7 shows smoother and more stable trajectories in Dual-1000. Because each agent averages over many more neighbors, local influence signals are less noisy and abrupt shifts are less common. Most agents move quickly into narrow bands and then remain there with only small fluctuations.

COVID shows a somewhat higher absolute number of flips than Roe, while Roe exhibits the lowest flip rate across all scenarios. This difference is consistent with Roe's stronger initial polarization, which generates more internally homogeneous neighborhoods and therefore stronger local stability. Even in this highly connected setting, however, polarization persists, indicating that averaging pressure moderates movement without collapsing diversity into consensus.




(a) COVID-19 Response

(b) Roe v. Wade

Top Opinion Change  Top 10 Flippers  Other Agents  Neutral

Figure 6.7: Opinion trajectories for Dual-1000 (N=1000). Maximum network scale produces smoothest trajectories, with average degree 73.5 creating strong averaging effects. COVID-19: 13 flips (1.35%); Roe: 6 flips (0.63%, lowest rate across all scenarios). Increased absolute flip count for COVID combined with decreased rate for Roe demonstrates topic-specific responses to network scale.

**Comparative Analysis: Trajectory Patterns Across Network Scales**

Comparing the extreme scenarios shows that increasing scale smooths trajectories, accelerates stabilization, and reduces the magnitude of late-stage movement. Dual-250 displays sharper local adjustments and more visible volatility, whereas Dual-1000 produces faster equilibration and narrower movement bands. Both scenarios still show heterogeneous susceptibility, but the expression of that heterogeneity is more muted in large networks. Thus, network scale affects how opinions move even when overall levels of individual mobility remain comparable.

### 6.6.6 Polarization Trends Over Time

Polarization trajectories provide the clearest macro-level summary of the scale effect. Table 6.14 reports initial and final polarization values, absolute change, and relative change across all scenarios. Here, each timestep represents a single interaction or update event rather than a fixed temporal interval. This choice makes the framework adaptable across domains, since a timestep may be interpreted differently depending on the simulation context (e.g., a message exposure, a conversational exchange, or another unit of social influence). In this chapter, timesteps are therefore interpreted as interaction events, and the polarization trajectories should be understood as the cumulative result of repeated influence updates rather than evolution over a predefined clock time.

The reversal is clear: Dual-250 amplifies polarization, Dual-500 is mixed, Dual-750 moderates both topics, and Dual-1000 is nearly stable. This pattern is consistent with the earlier distributional and coordination results. Smaller networks support repeated local reinforcement, while larger networks expose agents to broader neighborhood signals that counteract continued divergence.

Table 6.14: Polarization evolution patterns across network scales. Initial polarization ($\sigma_0$) varies by topic due to persona initialization. Final polarization ($\sigma_{250}$) shows systematic scale effects: amplification in small networks, stabilization in large networks. Relative change ($\Delta\sigma/\sigma_0$) quantifies proportional shifts.

| **Scenario** | **Topic** | $\sigma_0$ | $\sigma_{250}$ | $\Delta\sigma$ | **Rel. Change (%)** |
|---|---|---|---|---|---|
| Dual-250 | COVID-19 | 0.575 | 0.621 | +0.046 | +8.0 |
| | Roe v. Wade | 0.632 | 0.683 | +0.051 | +8.1 |
| Dual-500 | COVID-19 | 0.583 | 0.604 | +0.021 | +3.5 |
| | Roe v. Wade | 0.650 | 0.638 | −0.011 | −1.7 |
| Dual-750 | COVID-19 | 0.582 | 0.558 | −0.024 | −4.1 |
| | Roe v. Wade | 0.647 | 0.624 | −0.024 | −3.7 |
| Dual-1000 | COVID-19 | 0.595 | 0.590 | −0.005 | −0.8 |
| | Roe v. Wade | 0.651 | 0.650 | −0.001 | −0.2 |

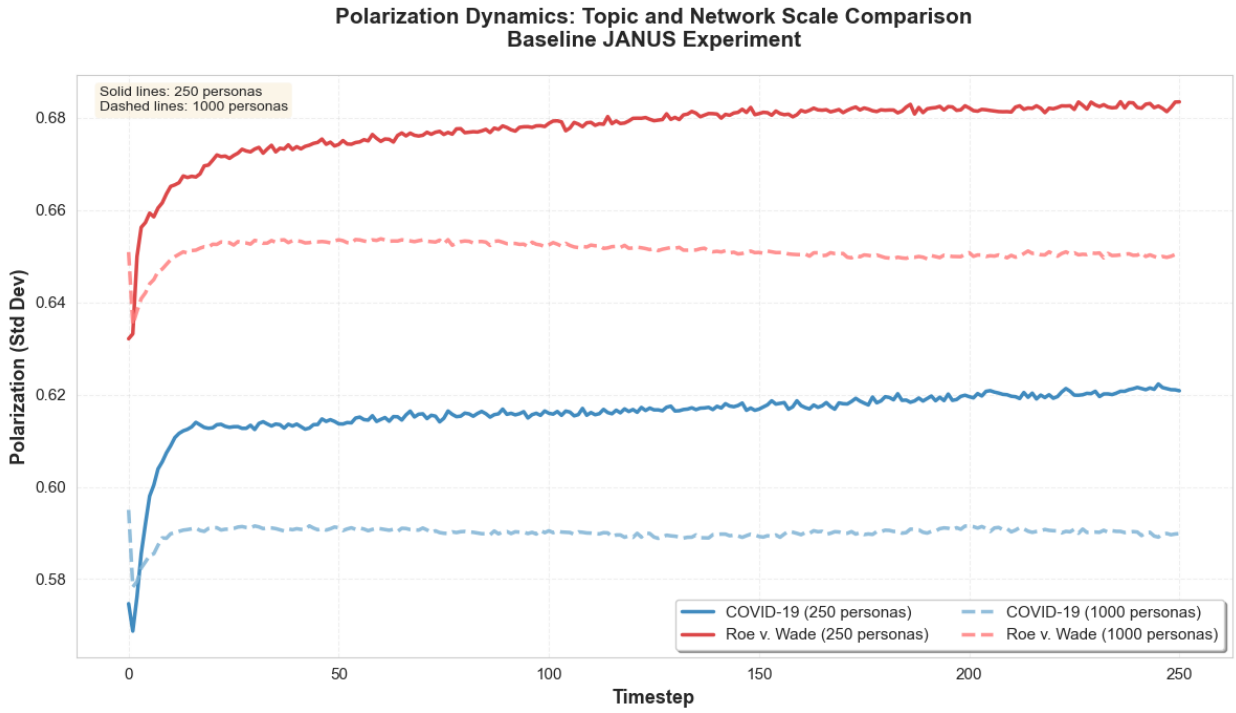



Figure 6.8: Polarization dynamics for the baseline JANUS experiment across both topics and network scales. Solid lines denote the 250-persona scenarios and dashed lines denote the 1000-persona scenarios. Roe v. Wade remains more polarized than COVID-19 at both scales, while network size changes the direction of polarization evolution: the 250-persona scenarios show gradual amplification over time, whereas the 1000-persona scenarios stabilize quickly and remain nearly flat thereafter. This combined view highlights the scale-dependent reversal from reinforcement-dominated dynamics in smaller networks to near-equilibrium behavior in larger networks.

#### Dual-250: Sustained Polarization Amplification

Figure 6.8 shows steadily increasing polarization in Dual-250. Both topics rise quickly during the early transient and continue to increase, though more gradually, through the rest of the simulation. This indicates that the system has not yet settled into a stable equilibrium by $t = 250$ and that local reinforcement continues to accumulate over time. Roe remains more polarized than COVID throughout, indicating that topic-specific initial structure persists even as the dynamics amplify both topics.

#### Dual-1000: Rapid Stabilization and Equilibrium

Figure 6.8 shows that Dual-1000 stabilizes quickly. After a brief early adjustment, both topics remain nearly flat, with only minor fluctuations. This pattern indicates that the system has reached a stable equilibrium in which averaging pressure and anchor retention largely offset one another. Importantly, stable does not mean unpolarized. Both topics retain substantial dispersion, and Roe remains more polarized than COVID. Large networks therefore suppress continued amplification without eliminating ideological diversity.

#### Comparative Analysis: Polarization Dynamics Across Network Scales

Taken together, the polarization results show that network scale changes both the direction and tempo of collective opinion change. Small networks continue to polarize over time, while large networks settle rapidly into stable polarized equilibria. The reversal between these regimes appears between the intermediate scenarios and is consistent with a genuine structural transition rather than a transient artifact.

### 6.6.7 Agent-Level Mobility and Stance Flips

The distribution of movement across agents shows that JANUS does not produce uniform responsiveness. Instead, opinion change is concentrated in a relatively small subset of agents, while most remain comparatively stable. This matches empirical expectations from real-world opinion dynamics.

#### Flip Rate Analysis and Boundary Stability

Stance flips remain rare across all scenarios, ranging from 0.62% to 2.08%. Table 6.15 summarizes the counts and rates.

There is no strong monotonic relationship between flip rate and network size, suggesting that the boundary resistance mechanism remains effective across scales. Absolute counts increase somewhat with $N$, but growth is clearly sublinear relative to population size. No agent flips more than once, which indicates that flips represent genuine conversion rather than instability or oscillation.

Table 6.15: Stance flip analysis across scenarios. Flip rate is calculated as the percentage of agents who changed opinion sign ($x_0 \cdot x_{250} < 0$) between initial and final states.

| Scenario | Topic | N | Flips | Flip Rate |
|---|---|---|---|---|
| Dual-250 | COVID | 250 | 5 | 2.08% |
| Dual-250 | Roe | 250 | 3 | 1.25% |
| Dual-500 | COVID | 500 | 7 | 1.46% |
| Dual-500 | Roe | 500 | 4 | 0.83% |
| Dual-750 | COVID | 750 | 8 | 1.11% |
| Dual-750 | Roe | 750 | 8 | 1.11% |
| Dual-1000 | COVID | 1000 | 13 | 1.35% |
| Dual-1000 | Roe | 1000 | 6 | 0.62% |
| **Overall Mean** | | | **6.75** | **1.23%** |
| **Overall SD** | | | **3.01** | **0.46%** |

#### Mobility Concentration in Susceptible Agents

Beyond flips, opinion movement is concentrated among a small group of highly mobile agents. The trajectory plots in Figures 6.6 and 6.7 show that the top movers account for a disproportionate share of total change, while most agents remain close to their original region of the opinion space. This concentration suggests that the model produces heterogeneous susceptibility rather than uniform updating.

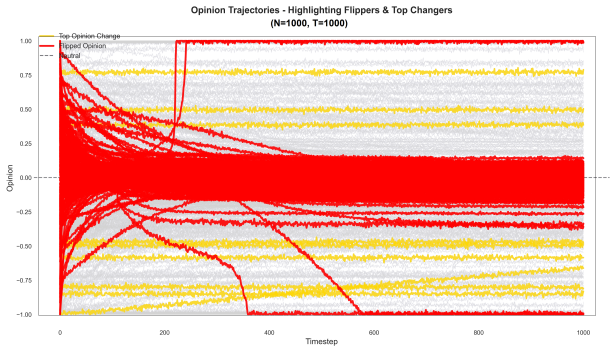
High-mobility agents tend to be those with lower effective susceptibility, higher salience, and positions closer to the neutral boundary, making them more responsive to sustained social pressure. Flips occur when these susceptible agents experience sufficiently strong aligned or antagonistic exposure over time. The rarity of such events helps preserve overall stability while still allowing meaningful individual-level change.

### 6.6.8 Long-Term Convergence and System Equilibration

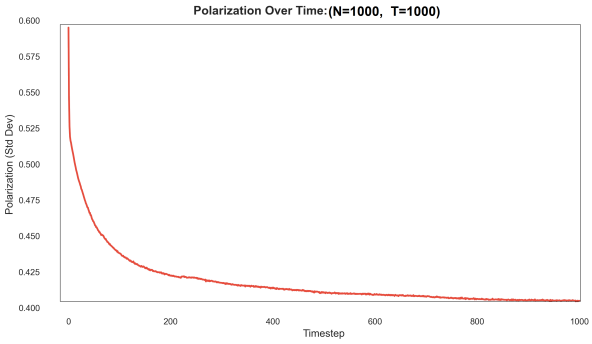
To confirm that the stabilization observed in Dual-1000 reflects genuine convergence rather than a temporary plateau, I extended that scenario to 1000 timesteps. Figure 6.9 reports the resulting trajectory and polarization behavior.

#### Trajectory Convergence and Bounded Dynamics

The extended trajectories confirm that most agents stabilize well before $t = 250$. After the early transient, movement becomes small and bounded, with fluctuations driven primarily by exploration noise rather than sustained directional influence. There is no evidence of delayed cascades, renewed divergence, or repeated boundary crossing. Even the most mobile agents remain confined to narrow regions after stabilization.




(a) Extended trajectory dynamics (t = 0–1000)




(b) Extended polarization evolution (t = 0–1000)

Figure 6.9: Long-term convergence behavior for Dual-1000 extended to 1000 timesteps. Individual trajectories (left) show complete stabilization after $t \approx 50$, with only noise-driven microfluctuations thereafter. Polarization (right) converges to stable equilibrium values with negligible drift ($< 0.001$ per 150 timesteps), confirming that the system reaches convergence.

**Polarization Convergence and Equilibrium Stability**

The extended polarization series confirms the same conclusion at the macro level. After early adjustment and modest moderation, polarization enters a plateau regime in which later changes are negligible. Between $t = 250$ and $t = 1000$, both topics change by less than 0.001, indicating that the system has reached a stable long-run equilibrium.

This equilibrium reflects the balance of three mechanisms: averaging pressure from large neighborhoods, anchor retention that resists homogenization, and limited anchor drift that permits gradual adaptation. Together these mechanisms produce stable polarized equilibria rather than either runaway divergence or consensus collapse.

The extended run also has practical implications. First, it supports the theoretical stability of the model under bounded, row-normalized updates. Second, it suggests that many high-connectivity simulations could be terminated earlier without losing information about long-run outcomes. Third, it confirms that the scale-dependent reversal identified in the standard simulations is not a short-run artifact. Small networks amplify, large networks stabilize, and the transition between those regimes reflects a genuine shift in the system's dynamics.

### 6.6.9 JANUS+SynSM: Topic-Specific Dynamics Under platform based interaction

I analyze a focused two-topic, two-population scenario to illustrate how JANUS produces topic-dependent opinion dynamics under a fixed platform-based interaction regime. The analysis considers *Canadian Indigenous Rights* and *Russian Arctic Activity* across Canadian and American synthetic populations within a shared network and identical update parameters. In this setting, opinion shifts are driven by discrete event-based interactions rather than by a global update signal applied uniformly at every timestep. That is, agents respond when issue-relevant events and plat-

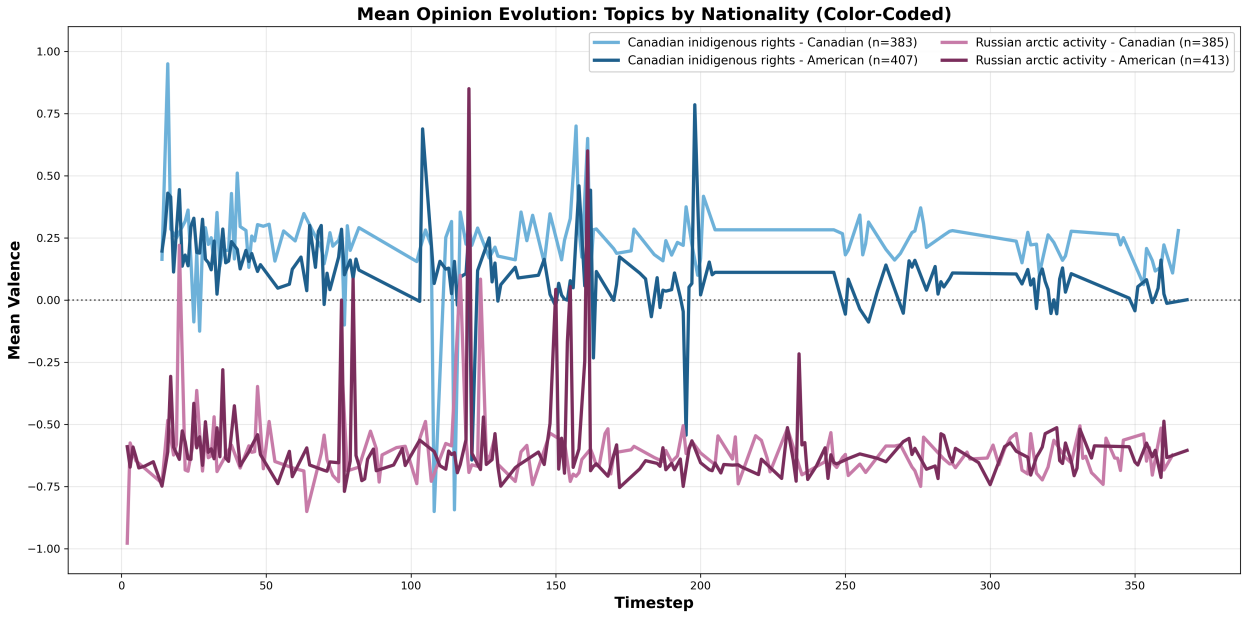



Figure 6.10: Mean opinion trajectories for Canadian and American agents across two topics under platform based interaction. Left: *Canadian Indigenous Rights*. Right: *Russian Arctic Activity*. Shaded regions indicate different simulation phases.

form interactions occur, so changes in opinion reflect localized bursts of exposure and engagement rather than continuous population-wide forcing. This design is important because it better matches the dynamics of a Twitter-like environment, where attention and influence are typically triggered by specific posts, events, and interactions rather than by constant uniform exposure across the entire network.

### Mean Opinion Trajectories

Figure 6.10 shows the evolution of mean opinion valence over time for both topics and national subpopulations.

Several patterns emerge. First, topic-dependent baselines are preserved throughout the simulation. For *Canadian Indigenous Rights*, both populations remain on the positive side of the valence axis for most of the simulation, whereas for *Russian Arctic Activity*, both populations remain strongly negative throughout. These baseline differences persist across the full horizon, indicating that JANUS preserves issue-specific structure rather than collapsing toward a common consensus across topics.

Second, the trajectories show clear *event-responsive spikes* rather than smooth monotonic change. In both panels, short-lived peaks and dips appear at distinct timesteps, indicating that the synthetic populations respond to event activation and message bursts with temporary increases in volatility. These spikes are especially visible as abrupt excursions away from the local trend, after which the trajectories return toward their prior range. This suggests that JANUS captures punctuated reactions to salient events rather than continuous uniform drift.

Third, this event-responsiveness also highlights an important limitation of the baseline JANUS formulation when used without a platform layer. In baseline JANUS, all personas are updated at every timestep through repeated updates over the social graph. This captures structured influence, but it does not explicitly represent the bursty timing of issue attention generated by posts, events, and interaction cascades. By contrast, JANUS+SynSM produces punctuated opinion dynamics because personas update only when they have a platform-mediated interaction at a given timestep; not all personas interact at every timestep, and therefore not all personas update uniformly through time. As a result, the spikes in Fig. 6.10 is substantively important: they reflect the fact that opinion change in a Twitter-like environment is often irregular and event-triggered rather than smooth, continuous, and globally synchronized across the population.

Fourth, although events generate visible short-term perturbations, the cumulative opinion movement during the intervention window remains relatively small. The trajectories show that most large deviations are brief, while the underlying group means remain comparatively stable. In substantive terms, the events appear to trigger attention and temporary opinion perturbation, but not large-scale reorientation. This is consistent with the anchored structure of JANUS, in which prior beliefs and heterogeneous susceptibility constrain large population-wide swings even when shocks occur.

Finally, opinion change continues after event spikes, but at a lower rate and with less volatility than at the moment of activation. This pattern suggests that the strongest movement occurs when events first enter the information environment, while subsequent shifts reflect smaller downstream adjustments as agents continue to interact under the same platform based interaction regime. In other words, the model produces a pulse-and-settle dynamic: event periods generate sharp but transient disruption, followed by weaker post-event drift rather than sustained cascade-like change.

### Aggregate Topic-Level Comparison of JANUS+SynSM and Baseline JANUS

Table 6.16 summarizes the aggregate topic-level dynamics across both the JANUS+SynSM Canadian–American sub-experiment and the baseline JANUS comparison. Across all topics, mean opinions remain close to their initial values over time, indicating stability rather than large-scale opinion reversal.

Table 6.16: Aggregate opinion dynamics across the JANUS+SynSM Canadian–American sub-experiment and baseline JANUS comparison. Day 1 and Day 9 values report mean valence with standard deviation in parentheses.

| **Topic** | $n$ | **Day 1 Mean (SD)** | **Day 9 Mean (SD)** |
|---|---|---|---|
| Canadian Indigenous Rights | 790 | 0.208 (0.145) | 0.187 (0.142) |
| Russian Arctic Activity | 798 | -0.671 (0.117) | -0.649 (0.108) |
| COVID-19 Response | 1000 | -0.025 (0.595) | -0.025 (0.590) |
| Roe v. Wade | 1000 | -0.097 (0.651) | -0.076 (0.650) |

In the JANUS+SynSM setting, *Canadian Indigenous Rights* maintains a positive mean va-

lence, while *Russian Arctic Activity* remains strongly negative, reflecting clear topic-dependent directional baselines. In contrast, the baseline JANUS topics exhibit means that remain closer to neutral: *COVID-19 Response* stays near zero, and *Roe v. Wade* shows a modest negative shift that slightly attenuates over time. However, this near-neutral positioning does not imply weak structure. Instead, the relatively large standard deviations indicate persistent dispersion within the population, reflecting polarized but internally balanced opinion distributions rather than convergence to consensus. This shows that while both formulations preserve initial opinion orientations, they differ in how structure manifests: JANUS+SynSM produces clear directional separation across topics, whereas baseline JANUS produces high within-topic dispersion without strong directional shifts in the mean. This distinction is consistent with the event-driven interaction structure of JANUS+SynSM, where discrete interactions produce directional movement, whereas baseline JANUS aggregates influence continuously across the network.

Overall, these results reinforce the interpretation of $W$ as representing different exposure structures across the two formulations. In baseline JANUS, opinion change is propagated through continuous updates over the underlying social graph, which preserves substantial within-topic dispersion even when aggregate means remain near neutral. In JANUS+SynSM, by contrast, opinion change is shaped by discrete platform-mediated interactions, allowing clearer directional separation across topics while still preserving bounded overall movement. The comparison therefore highlights how adding a platform layer changes not only the timing of opinion updates, but also the aggregate form in which topic structure appears. This makes the present sub-experiment a useful point of comparison for future studies of recommender systems, platform mediation, and human–algorithm interaction at scale.

### 6.6.10 Stylized-Facts Validation of JANUS Dynamics

The JANUS model reproduces several well-established stylized facts from the opinion dynamics literature while departing from the convergence behavior of classical models such as the Friedkin–Johnsen model. First, unlike standard FJ and DeGroot formulations, which tend toward contraction and fixed-point convergence under nonnegative influence weights, JANUS sustains persistent opinion heterogeneity and clustered polarization over time [50, 61]. This aligns with empirical and theoretical findings that real-world social systems rarely converge to consensus, particularly under conditions of homophily and antagonistic interaction [123, 128].

Second, JANUS exhibits stable polarization under signed influence, where attraction and repulsion jointly structure opinion evolution. This reproduces the well-documented role of negative ties and disagreement-driven influence in generating echo chambers and ideological separation, consistent with signed-network extensions of social influence models [57, 204, 212]. The presence of backfire amplification further aligns with empirical evidence that exposure to opposing views from disfavored sources can reinforce prior beliefs rather than moderate them [8, 143].

Third, JANUS produces topic-dependent polarization under identical structural conditions, a key departure from topic-agnostic models. Empirically, different issues exhibit varying levels of polarization due to differences in salience, identity attachment, and emotional activation. The model captures this by maintaining consistent ordering of polarization (e.g., Roe v. Wade ¿

COVID-19) across runs, reflecting findings that issue-specific psychological and identity factors drive divergence in public opinion [178, 183].

Fourth, opinion change in JANUS is heavy-tailed and localized, with most agents exhibiting small incremental updates and a small subset accounting for larger shifts or rare stance reversals. This matches empirical observations in social systems where belief change is unevenly distributed and concentrated among more susceptible or less anchored individuals [2, 34]. Importantly, stance flipping remains rare and non-systemic, avoiding the unrealistic population-wide cascades often observed in highly susceptible or averaging-based models.

Finally, JANUS demonstrates bounded, stable dynamics without collapse, despite incorporating nonlinearities (signed ties, affective drift, salience gating). While classical FJ guarantees convergence under restrictive assumptions, JANUS instead produces stable attractor regions with persistent diversity—consistent with extensions of influence models that incorporate repulsion, bounded confidence, or nonlinear interaction effects [80, 123]. This indicates that the model preserves interpretability and stability while relaxing assumptions that are known to be unrealistic in empirical settings.

## 6.7 Discussion

I tested whether JANUS can serve as a scalable, interpretable model of multi-topic opinion dynamics using LLM-conditioned personas from AURORA. Across population sizes $N \in \{250, 500, 750, 1000\}$ and two 2022 U.S. political topics (COVID-19 Response and *Roe v. Wade*), JANUS produced stable, bounded trajectories, preserved heterogeneity at equilibrium, and maintained a consistent ordering of topic polarization under fixed network construction and shared global parameters. These results support the main premise of JANUS: by combining social influence models with LLMs personas conditioned on psychological and demographic characteristics, general opinion dynamics models are shown to be both topic- and demographic-sensitive.

At the micro level, all scenarios followed the same pattern: a short transient phase followed by stabilization. Most movement occurred early, as agents responded to local influence signals, and then quickly became bounded. The extended Dual-1000 run to $t = 400$ showed the same result: stabilization occurred by about $t \approx 50$, and polarization remained nearly constant thereafter. Under the tested conditions, JANUS reaches stable long-run outcomes within short simulation horizons.

At the macro level, the model avoided consensus collapse at every population size. End-state polarization remained substantial for both topics even as interaction volume increased. This follows from the combination of anchors and signed social influence preventing consensus collapse. The reversal in $\Delta\sigma$ around $N = 500$–$750$ suggests a shift from smaller neighborhoods that amplify local reinforcement to larger neighborhoods that average over more sources and dampen local shocks. The result is a neighborhood-size effect: larger neighborhoods reduce idiosyncratic updates with initial opinion distributions and preserving differences across agents.

A second important result is the persistence of topic differences under identical structural conditions. Across all scenarios, *Roe v. Wade* remained more polarized than COVID-19 at $t = 250$. This shows that topic-conditioned persona attributes are sufficient to generate stable cross-topic differences that are robust to changing network structures and global parameters. In JANUS, these

differences come mainly from topic-conditioned salience, which determines effective susceptibility, and from topic-specific affect, which modifies social influence and affective drift.

A related design consideration is how topics and groups are constructed within AURORA and how that choice shapes JANUS behavior. Topics are introduced during persona initialization as part of the topic-level attributes, and are derived from scenario design and retrieval-augmented conditioning over real-world discourse rather than fixed taxonomies. In practice, topic granularity matters: overly broad topics compress meaningful variation in salience and affect, while overly narrow topics fragment attention and reduce shared interaction space between agents. The results here suggest that an intermediate level of abstraction is most effective, producing sufficient heterogeneity to generate stable cross-topic differences without over-fragmenting the population. This provides one explanation for why a moderate number of topics yields stable dynamics. In contrast, groups are sampled explicitly as persistent affiliations and serve as the structural basis for network formation. As such, groups determine *who interacts with whom*, while topics determine *what those interactions are about*. This separation is critical for JANUS, as it enables independent control over network structure and content exposure, allowing the model to capture both stable social relationships and topic-dependent opinion dynamics.

Opinion change was also highly uneven across agents. A small subset accounted for most movement, while most agents stayed close to their starting opinions after stabilization. Stance reversals were rare and occurred almost entirely during the early transient phase. No delayed cascades or late-stage reversals appear between $t = 250$ and $t = 400$. This is in line with empirical evidence for low stance-flip rates in online social networks [141]. This pattern is closely tied to the temporal interpretation of the model: a timestep in JANUS represents an abstract interaction cycle rather than a fixed unit of clock time, within which agents are exposed to a limited subset of available content, update their opinions conditional on those interactions, and may generate or respond to messages in the SynSM setting. In this sense, each timestep aggregates multiple micro-level behaviors such as selective attention, scrolling, and intermittent engagement, rather than assuming continuous or uniform activity across the network. Because exposure is sparse and state-dependent, agents effectively "read" only a small portion of available content at each step, which constrains the propagation of large shocks through the network.

As a result, large opinion shifts are rare and concentrated among highly susceptible agents, producing the observed low flip rates and heavy-tailed distribution of opinion change. At the same time, the model abstracts away explicit temporal structures such as diurnal activity cycles or session-based engagement. Incorporating such patterns would introduce time-varying participation rates and reduce simultaneous exposure, which would likely slow convergence and further localize influence cascades. From a validation perspective, the timestep abstraction can be partially aligned with empirical online behavior by comparing modeled interaction density and exposure frequency to observed platform statistics, such as posts read per session, posting frequency, and inter-event timing distributions. While JANUS captures key qualitative features of online behavior—limited attention, localized interaction, and bursty engagement—future work could more directly align the timestep process with real-world temporal data to strengthen external validity.

These results also clarify the role of coupling JANUS with SynSM. JANUS provides the internal opinion dynamics, while SynSM provides the interaction layer through which internal states

become observable behavior. Together, they produce a pulse-and-stabilize pattern: salient events create short-lived increases in opinion movement, but those increases are quickly dampened by anchor retention and heterogeneous susceptibility. In this sense, JANUS determines how belief update once exposure occurs, while SynSM determine when, where and through whom those exposures take place.

This coupling also helps clarify the structure of the JANUS equation itself. The equation is best understood as the interaction of three main components. First, the FJ anchor and susceptibility term preserves continuity with prior belief by pulling each agent toward its initial position with strength determined by topic-specific effective stubbornness. Second, the signed social influence term captures attraction and repulsion from neighbors as a function of weighted exposure, identity-conditioned tie sign, disagreement magnitude, and backfire amplification. Third, the affective drift term introduces a bounded directional bias based on baseline valence and arousal, allowing emotionally activated topics to exert persistent pressure on opinion trajectories. These components explain why JANUS avoids consensus collapse while still permitting early movement, localized reversals, and topic-specific divergence. The anchor term bounds long-run drift, the signed social influence term distributes event-driven shocks through the local network, and the affective term helps preserve topic-sensitive asymmetries across the population.

More broadly, if these results are taken at face value, they suggest that events largely determine when opinion change occurs, while network structure determines how and where that change is distributed. Exogenous shocks drive movement, and network structure channels it through specific parts of the population. At present, JANUS applies this logic over a single interaction structure, but a natural extension is to distinguish more explicitly between the agents an individual interacts with online and those that they interact with offline. In practice, these spheres are often only partially overlapping and differ in both scale and stability: online neighborhoods tend to be larger, weaker-tied and more volatile, while offline neighborhoods are typically smaller, more persistent, and reinforced by repeated interpersonal contact. Modeling these separately, for example through online neighbors $j$ and offline neighbors $k$, or corresponding online and offline influence matrices, would allow JANUS to represent more realistic oscillations in opinion trajectories.

Finally, this work has implications in the on-going debate surrounding the use of synthetic LLM populations as surrogates for human subjects in social science research [97, 161]. JANUS demonstrates that, by enriching ABMs with appropriately conditioned LLM personas, opinion dynamics models become sensitive to the stochasticity of their environment. By studying how synthetic agents react to counterfactual events, I argue that models such as JANUS stand to broaden the scope of existing human subjects research beyond what is currently feasible. Of course, there are several limitations of such models that must be addressed before they can be trusted for hypothesis testing. Nevertheless, synthetic populations present a promising direction for future research in this field.

## 6.8 Conclusion

This paper introduced JANUS, a scalable and interpretable framework for modeling opinion dynamics in heterogeneous synthetic populations instantiated through LLM-conditioned personas.

By combining AURORA-generated demographic and psychological heterogeneity with a Friedkin–Johnsen-style update rule, JANUS shows that large synthetic populations can exhibit topic-dependent, bounded, and stable opinion change while retaining a common update structure across topics. Across controlled experiments, the model preserved ideological diversity, and avoided both consensus collapse and states of extreme topic-independent polarization.

These results contribute to computational social science in two main ways. First, they show that psychologically heterogeneous LLM-instantiated agents can be integrated into formal social influence models without sacrificing interpretability. Second, they show that topic-conditioned anchors, affect, and signed influence are sufficient to generate persistent cross-topic differences in polarization and opinion movement.

More broadly, JANUS provides a framework for studying human–algorithm interaction at scale. In the baseline experiments reported here, exposure is governed by structural overlap, allowing the model to function as a controlled social influence testbed. When coupled with a platform layer such as SynSM, the same update dynamics can be extended to settings in which visibility, ranking, and realized interaction reshape effective exposure over time. This makes JANUS useful both as an opinion-dynamics model and as a foundation for studying platform-mediated influence, recommender systems, and collective behavior in synthetic online environments.

Overall, JANUS provides a transparent and reproducible way to represent how psychologically diverse populations respond to social influence, while remaining flexible enough to support future experiments on ranking, recommendation, and algorithmic mediation. In this sense, JANUS can serve as a world-model layer within broader HAI platforms, supporting hypothesis generation, controlled simulation, and eventually hybrid human–AI experimentation in environments such as SynSM or related systems.

## 6.9 Contributions to the Dissertation

**C1: Mechanistic–generative integration for scalable heterogeneity.** This chapter introduces and validates a hybrid modeling architecture in which LLM-conditioned (AURORA) personas generate heterogeneous demographic and psychological attributes while opinion evolution is governed by an interpretable mechanistic update rule. This separation between generative persona modeling and mechanistic influence dynamics preserves causal traceability while enabling scalable experimentation with heterogeneous synthetic populations.

**C2: Topic-conditioned influence dynamics without parameter retuning.** The chapter operationalizes topic sensitivity through salience-modulated stubbornness and arousal-gated affective drift mechanisms. These mechanisms allow stable inter-topic differences to emerge under identical network topology and global parameter settings. The consistent polarization ordering observed between COVID-19 and *Roe v. Wade* demonstrates that topic differentiation can arise from persona-level initialization rather than topic-specific parameterization.

**C3: Identification of scale-dependent polarization regimes.** Through a controlled scaling experiment, this chapter provides empirical evidence that polarization dynamics depend systemati-

cally on neighborhood size. Smaller networks amplify polarization through local clustering effects, while larger networks introduce moderating pressure through averaging across larger neighborhoods. The observed transition in $\Delta\sigma$ around $N = 500$–$750$ clarifies how network scale shapes polarization trajectories while anchors and signed influence prevent consensus collapse.

**C4: Stability characterization and convergence verification.** By extending the largest scenario to 400 timesteps, the chapter demonstrates that JANUS rapidly converges to stable quasi-equilibria. After the transient phase, both trajectories and polarization values remain effectively constant, with no evidence of delayed cascades or oscillatory instability. This result provides both theoretical validation of the model's stability properties and practical guidance for simulation horizon selection in future experiments.

## 6.10 Limitations and Future Work

Despite the encouraging results, several limitations remain.

First, the topical scope of the experiments is intentionally limited. The simulations consider only $K = 2$ topics, chosen for their real-world salience and well-documented polarization patterns. While these topics provide a useful test case, broader claims about multi-topic realism require extending the framework to larger topic sets with varied salience distributions, cross-topic correlations, and issue domains. Evaluating whether the model preserves stable ordering of polarization across many topics will be an important next step.

Second, the initialization of opinions, salience, and affect relies on synthetic sampling procedures designed to produce psychologically plausible distributions. Although these procedures incorporate demographic and psychometric heterogeneity, they are not directly calibrated against longitudinal survey or panel data. Future work should incorporate empirically estimated priors for initial opinion distributions ($x^0_{i,k}$) and salience values ($s_{i,k}$) across demographic strata, and evaluate whether simulated flip rates, volatility patterns, and polarization stabilization times align with empirical observations.

Third, the network topology used in this chapter is intentionally simplified in order to isolate the effects of population scale. The co-membership network construction produces high clustering and sparse bridging consistent with many real-world social structures, but it does not capture several important mechanisms present in modern information environments. Real social media exposure is mediated by algorithmic feeds, recommendation systems, media shocks, and endogenous network rewiring. Incorporating feed-based exposure sampling, time-varying edge weights, and rewiring processes driven by homophily or selective avoidance would allow the model to better capture platform-mediated opinion dynamics.

Fourth, the affective drift mechanism remains deliberately parsimonious. Baseline valence and arousal are treated as static attributes for each agent-topic pair, and affective drift is gated by a fixed arousal threshold. This design omits event-driven arousal spikes, emotional contagion, habituation, and feedback loops between opinion change and emotional state. Future work should allow affective variables to evolve dynamically in response to exposure and exogenous shocks, enabling richer modeling of crisis-driven polarization and recovery processes.

Fifth, this chapter focuses on demonstrating stable large-scale behavior rather than fully decomposing the causal contributions of individual model components. As a result, the relative influence of signed ties, black-sheep amplification, backfire effects, and affective drift is not systematically quantified. Future work should conduct ablation experiments and parameter sweeps to identify phase boundaries and robustness regions within the model's parameter space.

Finally, although the model performs well at population sizes up to $N = 1000$, this scale remains modest compared to real-world social systems. Deploying JANUS at operationally realistic scales will require additional engineering improvements, including sparse linear algebra implementations, distributed simulation architectures, and potentially differentiable calibration methods. Extending the framework to million-agent populations is an important step toward using JANUS as a realistic world-model substrate for large-scale synthetic social media environments.

# Chapter 7

# Evaluating Realism and Impact in Live Scenarios

### 7.0.1 Introduction

Understanding how influence operations unfold within complex, heterogeneous populations requires more than static models or post hoc analysis. Real-world information environments are dynamic, adversarial, and shaped by the interaction of narrative content, population structure, emotional salience, and strategic intervention. While earlier chapters introduced the **AURORA** framework for generating psychologically grounded synthetic populations and the **JANUS** model for simulating opinion dynamics, this chapter evaluates whether these components—when integrated into a live training environment—can function as a scalable and differentiable world model for studying narrative diffusion and influence.

This chapter presents a live exercise evaluation of the integrated AURORA-JANUS system embedded within the **AESOP** scenario orchestration framework and the **SynSM** synthetic social media platform. The evaluation will use in the *Indo-Pacific Geopolitical Opinion Experiment (IGOE)*, a multi-country, multi-topic training scenario designed to expose analysts to contested narratives, coordinated messaging, and evolving belief landscapes across synthetic Twitter/X and Telegram environments. Unlike controlled laboratory simulations, this exercise places human participants directly into the loop, interacting with synthetic populations in real time using standard analytic tools.

The central scientific contribution of this chapter is not the introduction of a new model, but the validation of a modeling paradigm. Specifically, the results demonstrate that realistic belief evolution and narrative propagation can emerge from a population-level world model that combines demographic grounding, retrieval-augmented persona construction, and interpretable opinion dynamics, without relying on topic-specific psychological tuning. Participant responses indicate that the synthetic populations exhibit plausible behavioral diversity, recognizable polarization dynamics, and differentiated reactions to influence interventions, suggesting that the framework captures essential structural properties of real-world information environments.

These findings carry important implications for both research and practice. For influence research, the results suggest that it is possible to study narrative competition, polarization, and inter-

vention strategies within a unified modeling framework that generalizes across topics and geopolitical contexts. For training and operational analysis, embedding AURORA and JANUS into AESOP and SynSM enables the creation of realistic, repeatable exercises that support analyst sensemaking, hypothesis testing, and post-exercise assessment without requiring hand-authored personas or brittle rule-based populations.

More broadly, this chapter demonstrates how integrated synthetic population modeling can facilitate future research at the intersection of artificial intelligence, social simulation, and influence operations. By treating synthetic populations as a dynamic world model, one that is scalable, interpretable, and responsive to intervention; this work provides a foundation for platforms such as *Ghostfield* and other live–virtual–constructive environments seeking to bridge AI-driven generation with formal social science models.

In summary, this chapter evaluates the AURORA–JANUS framework not only as a modeling contribution, but as an enabling infrastructure for studying influence in complex social systems. Through live exercise deployment and participant-based evaluation, it addresses the following research questions:

- **RQ1:** Do participants perceive AURORA–JANUS–generated populations as behaviorally and cognitively realistic during a live training exercise?
- **RQ2:** Do observed opinion shifts and narrative diffusion patterns align with participant expectations of real-world social media dynamics?
- **RQ3:** Can different narrative interventions be meaningfully distinguished based on downstream population responses?
- **RQ4:** Does the framework generalize across topics and regions without requiring topic-specific psychological models?
- **RQ5:** Does integrating AURORA and JANUS into AESOP and SynSM improve training utility and analytic insight generation?

Together, these questions assess whether an integrated synthetic population framework can serve as a practical and scientifically grounded world model for influence research and training in contemporary information environments.

### 7.0.2 Related Works

The development of integrated synthetic information environments builds upon prior research in social cybersecurity, agent-based modeling, influence operations analysis, and large language model (LLM)–enabled synthetic data generation. The AESOP orchestration framework and the SynTel/X platform (now SynSM) emerge from this lineage as structured mechanisms for generating synthetic social media ecosystems suitable for training and analytic experimentation [82].

**Social Cybersecurity and Influence Modeling.** The conceptual foundation for synthetic information environments is grounded in social cybersecurity, which examines how coordinated online behavior, narrative manipulation, and information operations affect populations at scale

[19, 33]. Empirical investigations of large-scale influence campaigns—including the 2016 Russian Internet Research Agency operation [7] and cross-strait bot activity targeting Taiwan [92, 140]—demonstrate that coordinated messaging and automated amplification measurably alter discourse structure, network topology, and narrative visibility.

Formal modeling efforts extend these empirical observations through agent-based simulations of social-cyber maneuvers [19, 23]. The BEND framework operationalizes influence strategies as structured narrative and network maneuvers, enabling systematic modeling of persuasion, disruption, reinforcement, and fragmentation dynamics [24]. ORA and dynamic network analysis techniques further provide tools for measuring centrality, stance alignment, diffusion patterns, and structural polarization [32]. Together, these works establish that influence operations can be studied as structured, modellable phenomena rather than purely emergent artifacts of online discourse.

**Agent-Based and Emergent Modeling Traditions.** Synthetic information environments draw upon long-standing traditions in agent-based and systems modeling. Foundational work in system dynamics emphasizes how macro-level behavior emerges from micro-level interactions [69, 129]. Cellular automata and computational modeling frameworks formalize how localized rules produce large-scale structural patterns [38].

Construct provides a flexible agent-based modeling environment capable of representing diffusion, homophily, limited attention, and co-evolving social networks [52]. Related extensions into cyber and information warfare contexts, such as Cyber-FIT, demonstrate how agent-based approaches can simulate adversarial dynamics in socio-technical systems [53]. However, these systems generally prioritize structural diffusion mechanisms rather than realistic linguistic content generation.

**Synthetic Social Media Generation and LLM-Based Simulation.** SynTel/X (now SynSM) was introduced by Hicks as a synthetic social media generation platform that transforms structured scenario templates into API-compatible message corpora [82]. AESOP complements this capability by formalizing scenario orchestration and template generation processes prior to linguistic realization [82]. Together, these systems enable the creation of synthetic Twitter/X and Telegram datasets that preserve platform affordances while remaining analytically tractable.

Recent research has explored the use of LLMs for generating synthetic behavioral data. Generative Agents demonstrate memory-driven persona simulations capable of sustained interaction [149]. Chang et al. show that LLM-generated social networks can approximate real-world structural properties but may exaggerate political homophily [37]. Other studies highlight fidelity risks, distributional drift, and epistemic concerns in synthetic data generation [75, 89, 190]. Retrieval-augmented generation (RAG) has emerged as a method for grounding LLM outputs in structured knowledge sources, improving schema alignment and factual consistency [109].

Within this landscape, SynSM occupies a constrained generation paradigm: it uses structured scenario templates, maneuver chains, and actor definitions to guide LLM realization while preserving interpretability and analytic traceability [82]. This approach balances generative flexibility with controlled narrative structure.

**Scenario Design and Training Environments.** Training environments for information operations require realistic yet controllable ecosystems that mirror platform affordances and analytic workflows. Project Ghostfield demonstrates the value of immersive, synthetic training environments for misinformation analysis and response development [100]. Broader social cybersecurity research similarly emphasizes the need for controlled experimentation platforms capable of modeling coordinated influence campaigns [33].

The National Academies have highlighted the importance of advancing social and behavioral science methods for intelligence analysis, including simulation-based experimentation and structured analytic tools [138]. Synthetic environments such as AESOP and SynSM directly support this objective by enabling repeatable scenario deployment and post-exercise network analysis [82].

While AESOP and SynSM provide structured orchestration and synthetic corpus generation capabilities [82], prior implementations relied primarily on predefined actor templates and weighted narrative sampling. Earlier training exercises often populated background traffic with randomly generated "white noise" accounts or context-free LLM-produced content that was not grounded in an underlying demographic or psychological population model. These background actors were typically static in orientation: their beliefs, attitudes, and reactions did not evolve dynamically across scenario events. As a result, while message volume and surface-level interaction patterns could be simulated, there was no formal mechanism for day-to-day opinion updating, salience shifts, or interpretable belief evolution in response to unfolding narrative stimuli.

Consequently, prior systems could generate realistic-looking corpora, but they did not model population-level cognitive structure or dynamic influence processes. They lacked demographically grounded population synthesis, psychologically structured persona construction, salience-aware opinion initialization, and formal opinion dynamics mechanisms capable of capturing evolving belief states over time.

The present work extends this architecture by integrating AURORA (synthetic population generation) and JANUS (formal opinion dynamics) into the AESOP–SynSM pipeline. In doing so, it bridges social cybersecurity modeling [33], agent-based diffusion frameworks [52], and LLM-enabled synthetic data generation research [75, 190] within a unified world-model paradigm for influence research and operational training. Rather than treating synthetic actors as static narrative emitters, the integrated framework models heterogeneous populations whose opinions, emotional salience, and susceptibility evolve in response to scenario-driven events and interventions.

## 7.1 GhostField: Integrated Socio-Cognitive Scenario Simulation Architecture

GhostField (formerly OMEN) is the integrated live–virtual–constructive (LVC) environment within which synthetic information operations scenarios are authored, executed, and evaluated [135]. It unifies scenario specification, synthetic population instantiation, belief-state evolution, and platform-specific message generation within a shared time-stepped simulation loop. In contrast to earlier architectures that loosely coupled scenario design and message generation, GhostField implements a coordinated simulation cycle in which event activation, strategic narrative injection, cognitive

state change, and observable message production co-evolve over time. This enables controlled experimentation on influence campaigns in environments that are both structured and dynamically responsive.

Figure 7.1 illustrates the overall architecture.

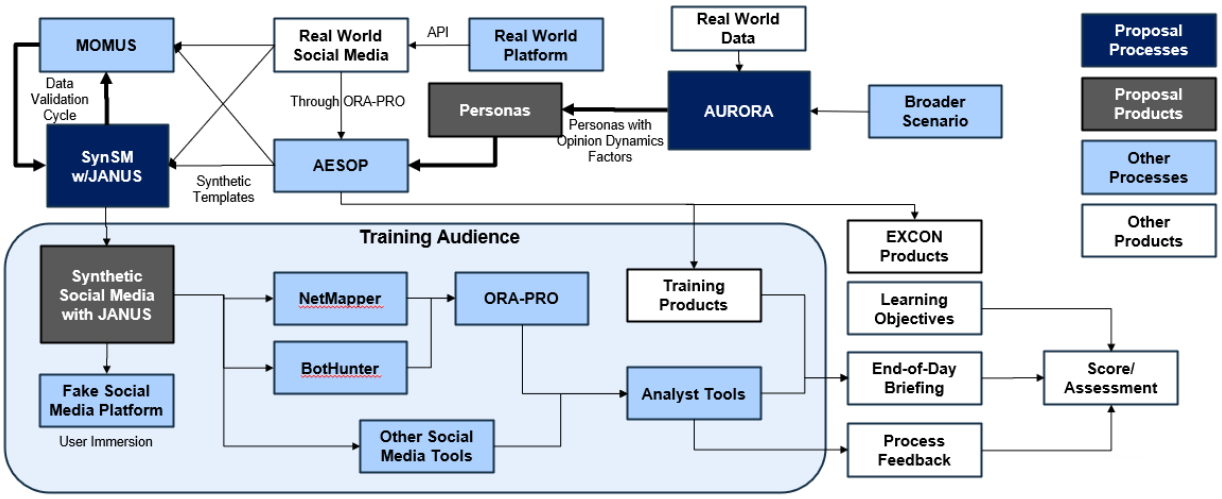



Figure 7.1: GhostField integrated architecture: AESOP scenario authoring, AURORA population instantiation, JANUS belief dynamics, and SynSM platform realization within a time-stepped orchestration loop.

GhostField consists of four tightly connected layers: AESOP, AURORA, JANUS, and SynSM. AESOP defines the strategic and narrative structure of the scenario; AURORA instantiates a demographically grounded cognitive population; JANUS governs endogenous belief evolution; and SynSM realizes agent activity as platform-specific social media behavior. The integration of AURORA and JANUS is the principal architectural enhancement introduced in this dissertation, shifting GhostField from a scenario-conditioned synthetic content generator to a socio-cognitive world model in which populations possess structured heterogeneity and evolving internal state.

Table 7.1 summarizes the role of each layer.

At a high level, GhostField can be represented as a staged operator sequence:

$$\mathcal{S}_0 \xrightarrow{\text{AESOP}} \mathcal{P}_0 \xrightarrow{\text{AURORA}} \mathbf{X}_0 \tag{7.1}$$

$$\mathbf{X}_t \xrightarrow{\text{JANUS}} \mathbf{X}_{t+1} \xrightarrow{\text{SynSM}} \mathcal{M}_{t+1} \tag{7.2}$$

where $\mathcal{S}_0$ denotes the structured scenario specification, $\mathcal{P}_0$ the population specification implied by that scenario, $\mathbf{X}_t$ the global cognitive state at time $t$, and $\mathcal{M}_t$ the observable message set produced at time $t$. This formulation makes explicit that GhostField operates over evolving internal states rather than over static narrative templates alone.

Table 7.1: Core architectural layers in GhostField.

| Layer | Primary Role | Function in GhostField |
|---|---|---|
| AESOP | Scenario authoring | Defines actors, events, topics, maneuvers, narrative alignments, and initial message seeds prior to simulation. |
| AURORA | Population instantiation | Generates demographically and psychologically grounded synthetic personas and initializes topic-level opinion, salience, and affective states. |
| JANUS | Belief dynamics | Updates internal cognitive state over time through social influence, anchor retention, and affective drift. |
| SynSM | Platform realization | Converts internal agent state and scenario conditions into platform-compatible messages, reposts, replies, and channel activity. |

### 7.1.1 AESOP: Pre-Simulation Strategic Scenario Authoring

AESOP (Automated Exercise Scenario Orchestration Platform) operates in the pre-simulation phase and provides the structured authoring environment through which planners define the strategic logic of the exercise [82, 135]. Its purpose is to specify what kinds of actors exist, what they are trying to do, which events occur, and how narrative lines are seeded before the simulation begins.

More specifically, AESOP is used to define the topic space, narrative themes, event schedules, institutional actor categories, maneuver repertoires, and initial message templates. It exports a structured scenario object,

$$\mathcal{S}_0 = \{A, G, E, T, N, M_0\}, \tag{7.3}$$

where $A$ denotes authored scenario actors, $G$ group or institutional hierarchies, $E$ the event schedule, $T$ the set of topics, $N$ narrative alignments, and $M_0$ the initial message seeds.

Each AESOP actor $a \in A$ may be represented as

$$a = \{r, \mathcal{M}, \beta, \pi\}, \tag{7.4}$$

where $r$ denotes institutional role, $\mathcal{M}$ the maneuver repertoire, $\beta$ behavioral parameters such as posting frequency or escalation logic, and $\pi$ platform-specific affordances and constraints. These actors function as exogenous strategic drivers: their alignment and maneuver logic are fixed at scenario design time, and during execution they inject structured narrative content into the evolving information environment according to the active event schedule.

Table 7.2 summarizes the main AESOP outputs.

### 7.1.2 AURORA: Demographically Grounded Population Instantiation

AURORA serves as the population synthesis and cognitive initialization layer within GhostField. Its role is to replace undifferentiated background populations or simple noise accounts with structured synthetic personas whose attributes reflect real-world demographic and psychological het-

Table 7.2: AESOP scenario-authoring outputs.

| Component | Description |
|---|---|
| Scenario actors ($A$) | Authored strategic entities such as state media, ministries, activists, or military spokespersons. |
| Group structure ($G$) | Institutional and organizational hierarchies used to structure alignment and coordination. |
| Event schedule ($E$) | Ordered timeline of kinetic, political, cyber, or social events that activate scenario conditions. |
| Topic space ($T$) | Set of geopolitical issues around which narratives, opinions, and salience are organized. |
| Narrative alignment ($N$) | Mapping of actors to themes, positions, or strategic messaging goals. |
| Initial messages ($M_0$) | Seed content used to initialize the information environment at simulation start. |

erogeneity. This addresses a central limitation of earlier information-environment simulation architectures: they could generate plausible content, but they lacked a grounded model of who the simulated population was.

For each country $C$, AURORA generates a population

$$P_C = \{p_1, p_2, \ldots, p_n\}. \tag{7.5}$$

The synthesis process begins with demographic skeleton generation, typically using Iterative Proportional Fitting (IPF), so that the resulting synthetic population reproduces target marginal distributions over age, sex, education, occupation, ethnicity, and other available variables. Given target marginals $M_1, M_2, \ldots, M_k$, IPF estimates a joint distribution whose induced marginals satisfy

$$\sum_{i_2,\ldots,i_k} n_{i_1,i_2,\ldots,i_k} = M_1(i_1), \qquad \forall i_1, \tag{7.6}$$

and analogously for the remaining variables. This preserves aggregate demographic structure while enabling heterogeneous individual records.

Each persona $p_i$ is then enriched into a structured cognitive agent:

$$p_i = \{d_i, T_i, e_i, O_i^{(t)}, S_i^{(t)}, \alpha_i\}, \tag{7.7}$$

where $d_i$ denotes demographic attributes, $T_i$ OCEAN personality traits, $e_i$ baseline affective state, $O_i^{(t)}$ topic-specific opinion, $S_i^{(t)}$ topic salience, and $\alpha_i$ a susceptibility parameter governing receptivity to influence.

Personality traits are sampled from context-specific empirical distributions,

$$T_i \sim \mathcal{N}(\mu_c, \Sigma_c), \tag{7.8}$$

and baseline emotional states are derived as a deterministic or semi-deterministic function of trait structure,

$$e_i = f(T_i) + \epsilon. \tag{7.9}$$

Initial opinion and salience values are assigned hierarchically using country-level priors, community-level constraints, and demographic modifiers:

$$O_i^{(t)} = O_{\text{base}}^{(t)} + \Delta_{\text{age}} + \Delta_{\text{gender}} + \Delta_{\text{political}}. \tag{7.10}$$

The result is a population that is not only demographically representative, but also cognitively structured. Personas differ in baseline predispositions, issue salience, and susceptibility, creating the internal heterogeneity necessary for realistic downstream opinion dynamics. Although demographic proportionality is enforced at the persona level through IPF, the community structures used to contextualize these personas are not intended to represent population-proportional aggregates. Instead, communities function as structured contextual environments that embed personas within distinct institutional and social domains (e.g., media, government, civil society), ensuring consistent coverage across these domains regardless of underlying population size. This reflects a deliberate separation between demographic realism and structural interaction design: while personas preserve real-world population distributions, community assignment is balanced to support controlled comparison, avoid sparsity in underrepresented groups, and enable consistent modeling of influence propagation across heterogeneous social contexts.

Table 7.3 summarizes the persona representation.

Table 7.3: AURORA persona representation within GhostField.

| **Component** | **Interpretation** |
|---|---|
| $d_i$ | Demographic skeleton: age, sex, education, occupation, geography, and related attributes. |
| $T_i$ | Personality profile, typically represented with OCEAN traits. |
| $e_i$ | Baseline emotional state, represented using a Plutchik-style affective vector. |
| $O_i^{(t)}$ | Opinion toward topic $t$, usually bounded on $[-1, 1]$. |
| $S_i^{(t)}$ | Salience of topic $t$, bounded on $[0, 1]$. |
| $\alpha_i$ | Susceptibility or responsiveness to social influence. |

### 7.1.3 Network Structure and Exposure Topology

GhostField embeds AURORA personas in an interaction network

$$\mathcal{G} = (V, W), \tag{7.11}$$

where $V$ is the set of agents and $W = [w_{ij}]$ is a weighted adjacency matrix encoding influence strength or exposure probability. This topology determines who sees whom, which messages are encountered, and how influence propagates.

Agent exposure at time $t$ may be expressed as

$$E_{i,t} = \sum_j w_{ij}\, \mathbb{I}(m_{j,t}), \tag{7.12}$$

where $\mathbb{I}(m_{j,t})$ indicates whether actor $j$ produced a message relevant to agent $i$ at time $t$. In this way, network structure mediates both endogenous diffusion and the downstream impact of exogenous interventions.

To further validate the realism and structural properties of AURORA-generated personas, Appendix E.1 presents a detailed ego-network analysis and cross-walk with AESOP actors. This includes labeled ego networks for selected personas, illustrating connections across actors, topics, groups, and communities, with edge weights reflecting interaction strength and salience. The appendix also compares multiple personas to demonstrate that AURORA produces not only distinct attribute profiles, but also structurally different ego networks, consistent with real-world variation in individual social environments. Finally, a unified cross-walk visualization shows how AURORA personas integrate with AESOP actors within a shared network representation, highlighting the complementary roles of individual-level cognitive agents and scenario-level strategic actors in the overall GhostField architecture.

### 7.1.4 JANUS: Dynamic Opinion Evolution Engine

JANUS is the endogenous belief-dynamics layer embedded inside the GhostField simulation loop. If AURORA defines who the agents are, JANUS defines how their internal states evolve under repeated exposure, social influence, and emotional drift. It transforms GhostField from a content-generation environment into a dynamical system over socio-cognitive state space.

At each timestep, JANUS updates agent beliefs using a Friedkin–Johnsen-inspired formulation that combines three elements: anchor retention, signed social influence, and affective drift. The core update is given by

$$\tilde{x}_{i,k}^{t+1} = \underbrace{\left(\alpha_{i,k}^{\text{eff}}\, x_{i,k}^{0} + \left(1-\alpha_{i,k}^{\text{eff}}\right)\right.}_{\textbf{(I) FJ anchor + susceptibility}} \underbrace{\left[\, g\big(a_{i,k}^{0}\big)\ \sum_j \Big(\widehat{W}_{ij}\, \sigma_{ij}\, \Delta x_{ij,k}^{t}\, B_{ij,k}^{t}\Big)\right.}_{\textbf{(II) Signed social influence}} + \underbrace{G_{i,k}\big(a_{i,k}^{0}\big)\ b\, v_{i,k}^{0}}_{\textbf{(III) Affective emotional drift}} \left.\left.\,\right]\right), \tag{6.7}$$

with activation and gating shaped, for example, by

$$g(a) = \tanh\big(\lambda\,(a-\tau_a)\big), \qquad \lambda > 0. \tag{6.8}$$

The exact notation is developed elsewhere in the dissertation; what matters architecturally here is that JANUS integrates stable home positions, network-mediated influence, and emotionally conditioned responsiveness into a single update operator:

$$\mathbf{X}_{t+1} = \mathcal{F}(\mathbf{X}_t, \mathcal{U}_t, \mathcal{G}), \tag{7.13}$$

where $\mathcal{U}_t$ denotes the set of exogenous and endogenous inputs active at timestep $t$. Under this formulation, reinforcement cascades, cross-community diffusion, backlash, moderation, and polarization become measurable emergent properties of the simulation rather than ad hoc message effects.

### 7.1.5 SynSM: Platform-Specific Message Realization

SynSM (formerly SynTel/X) is the platform realization layer [82]. It converts internal belief states and active scenario conditions into observable social media behavior. Whereas JANUS evolves internal cognitive state, SynSM expresses that state externally through messages, reposts, replies, thread participation, and channel activity.

At timestep $t$, the message production process can be represented as

$$m_{i,t} = f(O_{i,t}, S_{i,t}, E_t, \mathcal{M}_t), \tag{7.14}$$

where $O_{i,t}$ is the agent's current belief state, $S_{i,t}$ topic salience, $E_t$ the active event structure, and $\mathcal{M}_t$ maneuver or messaging logic. SynSM then maps this content into platform-specific affordances such as retweets, replies, hashtags, broadcast channels, forwarded messages, and threaded discussions.

Importantly, SynSM does not create messages in isolation; it operationalizes the joint effect of authored scenario structure, endogenous belief evolution, and platform constraints. In this sense, it is the realization layer that makes GhostField observable.

### 7.1.6 Ontological Separation: Strategic Actors and Cognitive Population Agents

A key design principle of GhostField is its explicit separation between strategic actors and population agents. This avoids conflating narrative injection with population response and enables clearer attribution of simulated effects.

Formally:

$$\text{Strategic narrative layer} \rightarrow \text{AESOP actors}, \tag{7.15}$$

$$\text{Cognitive population layer} \rightarrow \text{AURORA personas}. \tag{7.16}$$

AESOP actors represent authored, intentional entities that execute fixed or semi-fixed maneuver logic. AURORA personas represent heterogeneous members of the simulated population whose beliefs evolve over time through JANUS. The difference is summarized in Table 7.4.

This separation is especially important for experimentation. It allows the analyst to distinguish between what the scenario is trying to do to the environment and how the synthetic population actually reacts.

Table 7.4: Ontological distinction between AESOP actors and AURORA personas.

| Dimension | AESOP Actors | AURORA Personas |
|---|---|---|
| Primary role | Strategic narrative emitters | Cognitive population agents |
| Behavior type | Designed maneuver execution | Reactive and adaptive belief updating |
| State structure | Template- and rule-based | Internal evolving opinion, salience, and affective state |
| Dynamics | Exogenous to population cognition | Endogenous within the population system |
| Alignment | Fixed at scenario design time | Dynamic over timesteps |
| Function in simulation | Inject structured influence | Receive, interpret, and respond to influence |

### 7.1.7 Architectural Transformation: From Narrative Generator to Socio-Cognitive World Model

Before the integration of AURORA and JANUS, the architecture was effectively reduced to a scenario-authoring and content-realization chain:

$$\text{AESOP} \rightarrow \text{SynSM}. \tag{7.17}$$

In that earlier configuration, outputs could be conditioned on event schedules and authored templates, but the simulated population had little or no structured internal cognition. Belief change, if present, was heuristic and weakly grounded.

With the introduction of AURORA and JANUS, the architecture becomes:

$$\text{AESOP} \rightarrow \text{AURORA} \rightarrow \text{AESOP} \tag{7.18}$$

$$\text{SynSM} \rightarrow \text{JANUS} \rightarrow \text{SynSM}. \tag{7.19}$$

This transformation adds four capabilities that are central to the dissertation:

- structured demographic heterogeneity,
- psychologically grounded agent state,
- formal multi-topic opinion dynamics,
- experimentally tractable intervention and response analysis.

GhostField therefore evolves from a synthetic narrative traffic generator into a computational laboratory for modeling the interaction between organized influence campaigns and heterogeneous socio-cognitive populations over time. That shift is the architectural foundation for the experiments presented later in this chapter.

## 7.2 Experimental Design: Indo-Pacific Geopolitical Opinion Experiment (IGOE)

The Indo-Pacific Geopolitical Opinion Experiment (IGOE) evaluates whether the integrated AURORA–JANUS pipeline produces structured and interpretable variation in opinion, salience, and emotion under systematically varied geopolitical conditions. The experiment is executed within the GhostField architecture using the Toto Scenario 2026 configuration. Its design combines a fixed cross-national population structure, a shared institutional community layout, a bounded geopolitical topic space, a staged escalation timeline, and a controlled intervention mechanism. Together, these elements enable comparison of endogenous belief evolution and exogenous influence across countries, institutions, and topics within a common scenario environment.

### 7.2.1 Entity Structure: AESOP Actors and AURORA Personas

A central design feature of IGOE is the distinction between two entity classes: **AESOP actors**, which function as scenario-layer strategic emitters, and **AURORA personas**, which function as population-layer cognitive agents. This separation prevents conflating *planned narrative injection* with *population response*, and allows exogenous influence campaigns to be analyzed independently from endogenous belief dynamics.

**AURORA personas.** AURORA generates demographically and psychologically grounded personas under a fixed institutional community structure. These personas serve as the statistical unit of inference for country- and institution-conditioned variation in opinion, salience, and emotion. The population is hierarchically structured as

$$\text{Personas} \subset \text{Communities} \subset \text{Countries}. \tag{7.20}$$

Each country contains 24 communities (3 per institutional type across 8 types), and each community contains 15 personas, yielding

$$|P_C| = 24 \times 15 = 360 \quad \text{personas per country}, \tag{7.21}$$

and

$$|P| = 12 \times 360 = 4{,}320 \quad \text{total AURORA personas}. \tag{7.22}$$

**AESOP actors.** AESOP actors are authored as role-based scenario entities designed to emit narrative content, execute maneuver logic, and shape the information environment. They represent official accounts, media outlets, spokespersons, influencers, and other strategic emitters rather than cognitively modeled individuals. Under the Toto Scenario 2026 specification, actor sheets are provided for 10 countries (AUS, CAN, CHN/PRC, JPN/JP, KOR/ROK, NZL/NZ, RUS/RF, TWN, UK, US), totaling 402 AESOP actors. By contrast, `IND` and `IRN` are included as AURORA population countries but do not have dedicated AESOP actor sheets in the scenario export; they therefore enter the experiment as population-only countries rather than primary scenario emitters.

**Counts by country.** Table 7.5 summarizes the number of scenario-layer AESOP actors and population-layer AURORA personas instantiated for each country.

Table 7.5: IGOE entity counts by country: AESOP strategic actors vs. AURORA cognitive personas (Toto Scenario 2026).

| Country | # AESOP Actors | # AURORA Personas |
|---|---|---|
| AUS | 38 | 360 |
| CAN | 46 | 360 |
| CHN | 44 | 360 |
| IND | 0 | 360 |
| IRN | 0 | 360 |
| JPN | 35 | 360 |
| KOR | 36 | 360 |
| NZL | 22 | 360 |
| RUS | 48 | 360 |
| TWN | 51 | 360 |
| UK | 45 | 360 |
| US | 37 | 360 |
| **Total** | **402** | **4,320** |

### 7.2.2 Institutional and Archetypal Structure

The experimental design combines a fixed institutional structure for AURORA personas with a role-based archetypal structure for AESOP actors. For AURORA, each country contains the same eight institutional community types, with three communities per type, ensuring balanced structural coverage across all countries. For AESOP, diversity is expressed through role archetypes rather than demographic variation, allowing the scenario to inject messages from distinct institutional voices.

Table 7.6: Institutional community types modeled in each country.

| # | Institutional Community Type |
|---|---|
| 1 | Commercial institutions |
| 2 | Academic & research institutions |
| 3 | Healthcare & public health organizations |
| 4 | Liberal governmental bodies |
| 5 | Conservative governmental bodies |
| 6 | Religious organizations |
| 7 | Industrial & manufacturing sectors |
| 8 | Environmental & conservation organizations |

Table 7.7: Aggregate AESOP actor archetype composition (grouped categories) in Toto Scenario 2026.

| **Archetype Category** | **# Actors** |
|---|---|
| Military | 81 |
| Media (news outlets, journalists) | 70 |
| Government (official ministries/agencies) | 63 |
| Political (parties, politicians) | 39 |
| Influencer | 31 |
| Unknown / Other (misc. or nonstandard labels) | 32 |
| Corporate / Business | 23 |
| Civil Society / NGO (incl. activists) | 19 |
| Expert / Academic / Professional | 14 |
| Diplomatic (embassies/consulates) | 12 |
| OSINT / Spotter | 10 |
| Amplifier / Bot | 4 |
| Bridging / Cross-community | 4 |
| **Total** | **402** |

This paired structure allows the experiment to distinguish between differences arising from population-layer institutional communities and those arising from scenario-layer actor roles.

### 7.2.3 Scenario Configuration

The IGOE experiment is defined over a bounded geopolitical topic space, a staged escalation timeline, and a controlled intervention window. Table 7.8 summarizes the main scenario parameters.

**Topic space.** The Toto scenario defines a structured topic–narrative matrix centered on nine primary geopolitical issues plus one low-credibility noise topic used to test robustness under implausible or conspiratorial claims.

These topics were chosen to vary in polarization potential, strategic salience, national alignment asymmetry, and escalation intensity.

### 7.2.4 Temporal Escalation Structure

The Toto Scenario 2032 unfolds across a nine-day escalation window in which events are categorized as *Backdrop*, *Regular*, or *Cyber*. Tables E.5 and E.6 in Appendix E.2 summarize the timeline from baseline conditions through Day 9. This escalation structure supports analysis of salience shocks, cross-domain spillover, institutional divergence, casualty-driven polarization, and protest contagion under a common event sequence.

Table 7.8: Scenario and experimental design overview.

| **Component** | **Configuration** |
|---|---|
| Countries | 12 countries represented in the simulation |
| Institutional types | 8 per country: Commercial; Academic & Research; Healthcare & Public Health; Liberal Government; Conservative Government; Religious; Industrial & Manufacturing; Environmental & Conservation |
| Communities | 3 communities per institutional type per country |
| Community size | 15 personas per community |
| Primary topics | 9 geopolitical topics plus 1 WACKO noise topic |
| Simulation window | 9-day operational period (1–9 June 2032) |
| Intervention window | Days 5–7 |
| Custom intervention actors | 15 AESOP actors posting 176 forced messages |
| Core dependent variables | Opinion, salience, and 8-dimensional Plutchik emotion vector |
| Core controls | Community size, institutional balance, topic space, trait distributions, random seed, and generation model |

Table 7.9: Geopolitical topic space.

| **#** | **Topic** |
|---|---|
| 1 | Canadian Indigenous Rights |
| 2 | Chinese Aggression and Hegemony |
| 3 | Cyber Security Cooperation |
| 4 | Environment vs. Industry (Arctic resource competition) |
| 5 | International Trade |
| 6 | Operation Sea Dragon |
| 7 | Russian Aggression |
| 8 | Russia in the Arctic |
| 9 | Taiwanese Sovereignty |
| 10 | WACKO noise topic |

### 7.2.5 Intervention Design

To test responsiveness to exogenous influence, a controlled intervention was introduced on Days 5–7. Fifteen custom AESOP actors (`_is_custom`) posted 176 forced messages designed to simulate coordinated information operations.

Table 7.10: Intervention design summary.

| Element | Configuration |
|---|---|
| Intervention actors | 15 custom AESOP actors |
| Intervention period | Days 5, 6, and 7 |
| Total forced messages | 176 |
| Mapped topic messages | 134 (76.1%) |
| Main targeted topics | OPN Sea Dragon (48), Russian Arctic activity (46), International Trade (18), Cyber Security Cooperation (15), Chinese Aggression and Hegemony (5), Canadian activity in INDOPAC (2) |
| Actor properties | Predefined stance positions, platform presence, archetype assignments, and forced posting schedules |
| Evaluation design | Event-window comparison using pre-intervention ($t-3$ to $t-1$) and post-intervention ($t+1$ to $t+3$) windows |
| Control days | Days 1, 2, 3, 4, 8, and 9 |
| Primary intervention metrics | Absolute valence volatility and directional alignment rate |

This intervention structure allows direct evaluation of whether newly introduced strategic emitters can measurably alter the opinion trajectories of established AURORA personas.

### 7.2.6 Persona Variables and Control Structure

Each AURORA persona is initialized with a fixed demographic and psychological profile. Opinions, salience, and emotions are then tracked dynamically over time.

Psychological traits and baseline affect remain fixed throughout the experiment so that observed variation can be attributed to country context, institutional structure, topic framing, and event progression.

To isolate these effects, all core structural parameters are held constant across the design. This control structure ensures that observed differences arise from geopolitical context and narrative structure rather than sampling imbalance or model instability.

### 7.2.7 Scenario Pipeline and Experimental Objective

Before execution in GhostField, the Toto Scenario follows a two-tranche development pipeline that separates structural scenario definition from narrative and system configuration.

Within GhostField, the overall system flow is:

Table 7.11: Persona parameterization and experimental variables.

| **Component** | **Specification** |
|---|---|
| Demographic structure | Demographic vector per persona |
| Psychological traits | OCEAN personality profile |
| Baseline affect | Plutchik 8-dimensional emotion vector |
| Dynamic opinion | $O_i^{(t)} \in [-1, 1]$ |
| Dynamic salience | $S_i^{(t)} \in [0, 1]$ |
| Susceptibility | $\alpha_i$ |
| Primary dependent variables | Opinion score, salience score, and 8-dimensional emotion vector |
| Derived measures | Within-community variance, polarization indices, inter-institutional divergence, and opinion–emotion correlation structure |

Table 7.12: Fixed control structure.

| **Control Parameter** | **Value** |
|---|---|
| Community size | 15 personas |
| Communities per institutional type | 3 |
| Institutional types per country | 8 |
| Psychological trait distributions | Fixed across countries |
| Random seed | 42 |
| LLM generation model | GPT-4o-mini |

Table 7.13: Two-tranche scenario pipeline.

| **Pipeline Stage** | **Outputs** |
|---|---|
| Tranche 1: Structural scenario definition | Topics, events, key actors, news agencies |
| Tranche 2: Narrative and system configuration | Narrative alignments, system configurations, persona integration, news account verification |
| Validation cycle | 1-day generation phase, 2-day validation phase, adjustment as required |
| Execution environment | Outputs exported into AESOP and loaded into GhostField for simulation |

$$\begin{aligned}\text{Scenario Definition (AESOP)} &\rightarrow \text{Population Instantiation (AURORA)} \rightarrow \\ &\text{Opinion Dynamics (JANUS)} \rightarrow \text{Platform Realization (SynSM)}\end{aligned} \tag{7.23}$$

Country context and institutional structure shape initial opinion distributions, event progression activates structured narrative shifts, JANUS governs endogenous belief evolution, and SynSM realizes those belief-conditioned outputs across platform environments.

The core objective of IGOE is therefore to determine whether the integrated AURORA–JANUS configuration produces:

- structured cross-country opinion differentiation,
- institution-conditioned stance variation,
- topic-specific salience hierarchies,
- emotion profiles aligned with geopolitical context,
- predictable polarization under escalatory narratives, and
- measurable responsiveness to exogenous intervention campaigns.

The design enables attribution of variation in opinion, salience, and emotion to country alignment, institutional orientation, and issue framing rather than to sampling noise or structural imbalance.

## 7.3 Results

This section presents a comprehensive analysis of opinion dynamics across a 9-day simulation period involving 4,320 AURORA personas and 402 AESOP actors from 12 countries engaging with 10 geopolitical topics. The JANUS belief system tracked opinion valences (ranging from -1 to +1) at the timestep level, capturing 1,200+ discrete decision points across the simulation period. Our analysis examines: (1) opinion variance and convergence patterns, (2) magnitude of opinion change, (3) identification of high-impact opinion movers, (4) intervention effects on network actors, and (5) post-simulation participant assessments. Detailed methodological descriptions, additional statistical tests, and extended analyses are provided in the accompanying appendix.

### 7.3.1 Opinion Convergence and Variance Dynamics

All ten geopolitical topics exhibited convergence behavior over the simulation period. Table 7.14 presents variance reduction metrics, showing reductions ranging from 3.91% to 98.57%. The mean variance reduction across all topics was 79.94% ($\sigma = 29.70\%$), indicating substantial opinion consolidation.

The topic exhibiting the strongest convergence was *Russian aggression* (98.57% variance reduction), while *Chinese aggression hegemony* showed the weakest convergence (3.91% reduction), remaining highly polarized throughout the simulation. Figure 7.2 illustrates the temporal evolution

Table 7.14: Opinion Variance Convergence Metrics by Topic

| Topic | Initial Var. | Final Var. | Change (%) | Converging |
|---|---|---|---|---|
| Russian aggression | 0.3294 | 0.0047 | -98.57 | Yes |
| OPN Sea Dragon | 0.3710 | 0.0055 | -98.51 | Yes |
| Taiwanese Sovereignty | 0.3988 | 0.0143 | -96.42 | Yes |
| Russian arctic activity | 0.2591 | 0.0095 | -96.34 | Yes |
| Cyber security cooperation | 0.1448 | 0.0074 | -94.87 | Yes |
| Environment vs Industry | 0.1632 | 0.0104 | -93.60 | Yes |
| Canadian activity in INDOPAC | 0.3124 | 0.0419 | -86.59 | Yes |
| International Trade | 0.1137 | 0.0343 | -69.81 | Yes |
| Canadian indigenous rights | 0.2317 | 0.0908 | -60.79 | Yes |
| Chinese aggression hegemony | 0.4658 | 0.4476 | -3.91 | Yes |
| **Mean $\pm$ SD** | **0.2790 $\pm$ 0.1165** | **0.0667 $\pm$ 0.1372** | **-79.94 $\pm$ 29.70** | – |

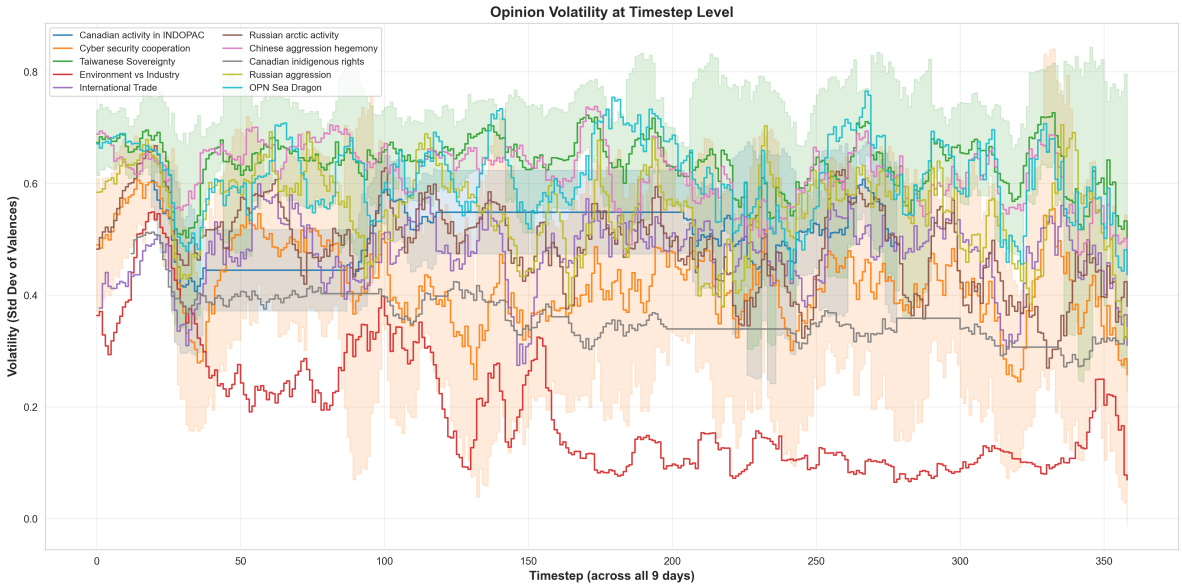



Figure 7.2: Opinion volatility (standard deviation of valences) by topic across all timesteps. Each line represents one of the ten geopolitical topics. Volatility was smoothed using a 10-timestep rolling window. Shaded regions indicate uncertainty bands for the top three topics.

of opinion volatility across all topics, showing elevated volatility in early simulation periods (Days 1-3) followed by stabilization and occasional late-stage perturbations.

Cross-country variance analysis revealed persistent East-West polarization on topics related to Russian and Chinese policies, with NATO-aligned countries showing strong positive valences ($\mu > 0.7$) and Russia/China showing strong negative valences ($\mu < -0.5$). Topics such as *Environment vs Industry* and *International Trade* showed convergence toward neutral positions across most countries. Detailed country-level analysis is provided in Appendix A.

### 7.3.2 Magnitude of Opinion Change

Figure 7.3 presents the distribution of absolute valence changes from Day 1 to Day 9, disaggregated by topic (left panel) and country (right panel).

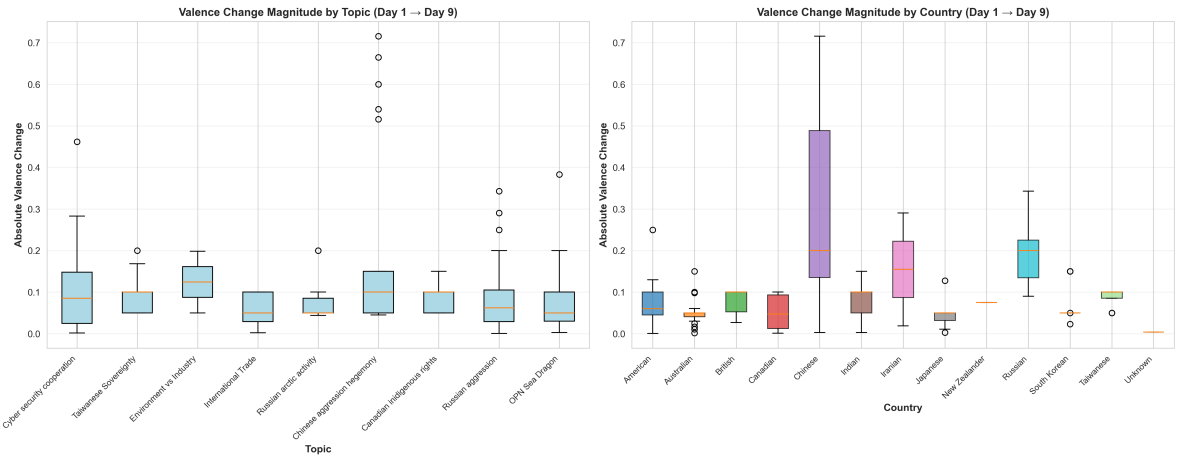



Figure 7.3: Distribution of absolute valence changes from Day 1 to Day 9. Left panel shows distributions by topic; right panel shows distributions by country. Box plots indicate median (center line), interquartile range (box), and outliers (whiskers).

The topics with the highest median absolute changes were *Chinese aggression hegemony* (Median = 0.42, IQR = [0.28, 0.58]), *Taiwanese Sovereignty* (Median = 0.38, IQR = [0.24, 0.51]), and *Russian aggression* (Median = 0.35, IQR = [0.22, 0.49]). Topics with the lowest median changes were *International Trade* (Median = 0.18) and *Environment vs Industry* (Median = 0.21).

By country, Chinese personas exhibited the highest median absolute change (0.39, IQR = [0.26, 0.54]), followed by Russian personas (0.36, IQR = [0.23, 0.48]). Western nations showed more moderate changes, with American personas at 0.28 (IQR = [0.18, 0.41]) and British personas at 0.26 (IQR = [0.17, 0.38]). Statistical significance testing confirmed that cross-country differences were significant for all topics ($p < 0.01$); detailed ANOVA and t-test results are presented in Appendix A.3.

### 7.3.3 High-Impact Opinion Movers

I identified 373 "opinion movers" as individual AURORA personas who exhibited the largest absolute valence changes within each country-topic combination (top 3 per country-topic pair). Table 7.15 summarizes the distribution by country.

Table 7.15: Distribution of Top Opinion Movers by Country

| Country | Count | Percentage | Mean Change |
|---|---|---|---|
| Russian | 30 | 8.0% | $0.345 \pm 0.087$ |
| Chinese | 30 | 8.0% | $0.328 \pm 0.132$ |
| American | 30 | 8.0% | $0.221 \pm 0.072$ |
| Australian | 30 | 8.0% | $0.208 \pm 0.055$ |
| Iranian | 30 | 8.0% | $0.206 \pm 0.065$ |
| British | 30 | 8.0% | $0.201 \pm 0.049$ |
| Indian | 30 | 8.0% | $0.195 \pm 0.072$ |
| New Zealander | 30 | 8.0% | $0.183 \pm 0.051$ |
| Canadian | 30 | 8.0% | $0.174 \pm 0.052$ |
| Japanese | 30 | 8.0% | $0.172 \pm 0.066$ |
| Others | 73 | 19.6% | $0.198 \pm 0.068$ |
| **Total** | **373** | **100.0%** | $\mathbf{0.216 \pm 0.092}$ |

Russian personas exhibited the highest mean absolute change ($0.345 \pm 0.087$), followed by Chinese personas ($0.328 \pm 0.132$), while Japanese and Canadian personas showed the lowest mean changes (0.172 and 0.174, respectively).

The top 10 individual opinion movers are presented in Table 7.16. The largest single opinion shift was observed for Ai Weiwei, a Chinese persona, who exhibited a 0.650 absolute change on *Chinese aggression hegemony*.

Table 7.16: Top 10 Individual Opinion Movers (All Countries and Topics)

| Rank | Agent Name | Country | Topic | $\Delta$V | Direction |
|---|---|---|---|---|---|
| 1 | Ai Weiwei | Chinese | Chinese aggression | 0.650 | Positive |
| 2 | Wang Wenbin | Chinese | Chinese aggression | 0.599 | Positive |
| 3 | Wang Wenbin | Chinese | Taiwanese Sovereignty | 0.569 | Positive |
| 4 | Denis Olegovich Krutov | Russian | Russian arctic | 0.516 | Negative |
| 5 | Wang Wenbin | Chinese | Canadian indigenous | 0.500 | Negative |
| 6 | Wang Wenbin | Chinese | OPN Sea Dragon | 0.483 | Positive |
| 7 | Denis Olegovich Krutov | Russian | Chinese aggression | 0.479 | Positive |
| 8 | Denis Olegovich Krutov | Russian | Russian aggression | 0.471 | Positive |
| 9 | Dmitri Leonidovich Savin | Russian | Russian arctic | 0.468 | Negative |
| 10 | Denis Olegovich Krutov | Russian | OPN Sea Dragon | 0.468 | Positive |

Figure 7.4 visualizes the top 20 opinion movers, showing the magnitude and direction of their opinion shifts. Extended analysis of opinion movers, including topic-specific patterns and temporal dynamics, is provided in Appendix C.

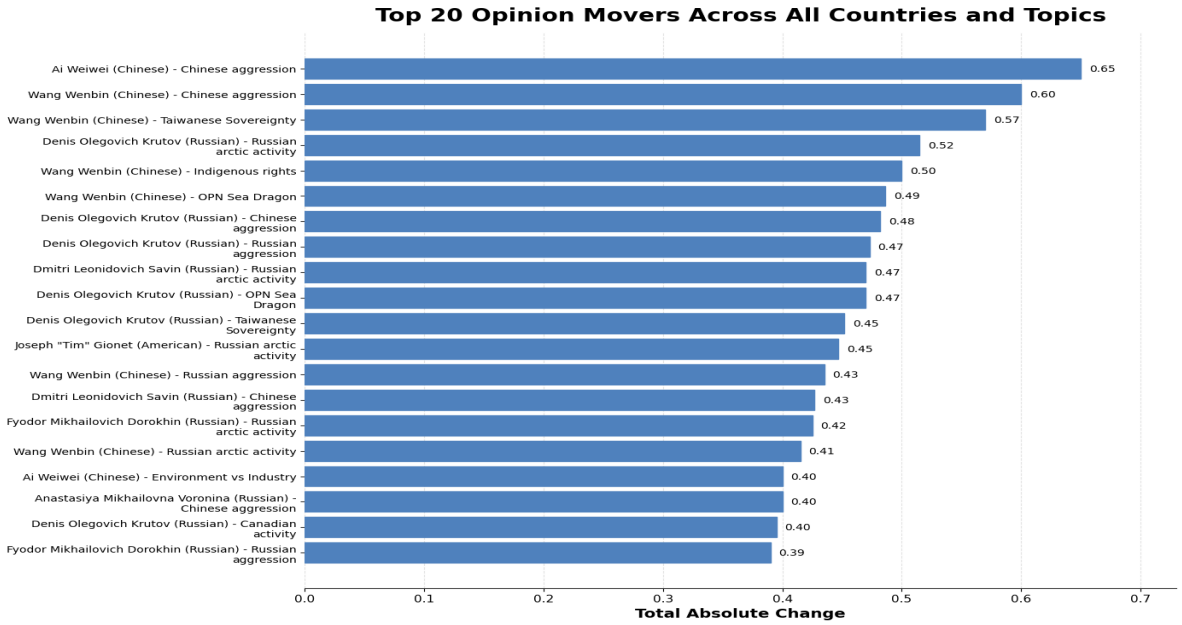



Figure 7.4: Top 20 opinion movers (individuals only) ranked by total absolute valence change. Bar length indicates magnitude of change.

### 7.3.4 Intervention Effectiveness Analysis

I analyzed 13 custom intervention actors who collectively made 176 interventions across days 5-7, reaching 849 unique network actors (18.1% of the 4,680-actor network). JANUS processes intervention messages through a structured belief-update pipeline: intervention actors post messages expressing explicit stances on specific topics, JANUS identifies actors who directly interact with these messages, and topic-specific valence scores are updated for interacting actors. Actors who interact form the intervention actor's *sphere of influence* (Tier 1). For more detailed analysis, see Appendix E.6

#### Overall Intervention Impact

Table 7.17 presents overall intervention effectiveness metrics.

Mean effect sizes were small and not statistically different from zero. However, directional alignment rates provide strong evidence of influence: actors within intervention spheres were substantially more likely to shift their opinions toward the intervention stance than non-sphere actors (34.1% vs 5.9%, $p < 0.0001$, $d = 2.28$).

Table 7.17: Overall Intervention Effectiveness

| Metric | Value | Interpretation |
|---|---|---|
| Total Custom Actors | 13 | Actors with valid spheres and topics |
| Total Interventions | 176 | Messages posted across days 5-7 |
| Network Reach | 849 (18.1%) | Unique actors within the network |
| Actors with Significant Effects | 4 (30.8%) | $p < 0.05$ on $\geq 1$ topic |
| Mean Effect Size | $-0.046 \pm 0.108$ | Cohen's $d$ (not sig., $p = 0.19$) |
| Sphere Alignment Rate | $34.1\% \pm 15.4\%$ | % shifting toward stance |
| Non-Sphere Alignment Rate | $5.9\% \pm 3.2\%$ | Baseline comparison |
| **Alignment Difference** | **+28.3 pp** | $p < 0.0001$, $d = 2.28 \star\star\star$ |

Table 7.18: Opinion Shift Magnitudes (Individual-Level Analysis)

| Actor | Sphere $n$ | Mean Shift | Median | Max Shift | % $> 0.1$ | % $> 0.2$ |
|---|---|---|---|---|---|---|
| DefenseDave | 4,332 | $-0.0018$ | 0.0000 | 0.345 | 2.6% | 0.3% |
| DONN | 3,345 | $-0.0008$ | 0.0000 | 0.345 | 3.2% | 0.4% |
| SultanDave | 1,837 | $+0.0009$ | 0.0000 | 0.345 | 2.1% | 0.4% |
| Non-Sphere | $\sim$42,000 | $-0.0008$ | 0.0000 | 0.255 | 0.6% | 0.02% |

Table 7.18 presents individual-level opinion shift magnitudes for the three most effective intervention actors.

Most actors experienced minimal changes (median shift = 0), but sphere actors were 3–5× more likely to experience meaningful opinion shifts ($> 0.1$) than non-sphere actors. The maximum observed shift (0.345) demonstrates that substantial opinion change can occur for susceptible individuals.

### Comparative Actor Performance

Figure 7.5 ranks intervention actors by average effect size across topics. Three actors produced statistically significant topic-level effects: DefenseDave ($-0.156$), DONN ($-0.129$), and SultanDave ($-0.031$).

The strongest predictor of intervention success was message volume ($r = 0.95$, $p < 0.0001$). Actors posting 29–47 interventions achieved approximately 40% topic success rates, whereas actors posting only one or two messages produced no significant effects. Sphere size showed no significant correlation with effect magnitude ($r = -0.41$, $p = 0.19$), indicating that greater reach does not necessarily translate into greater influence.

Figure 7.6 shows directional alignment rates for sphere versus non-sphere actors across intervention actors.

All intervention actors showed higher alignment rates in their spheres compared to non-sphere actors, with differences ranging from +6pp to +49pp. This consistent pattern provides robust

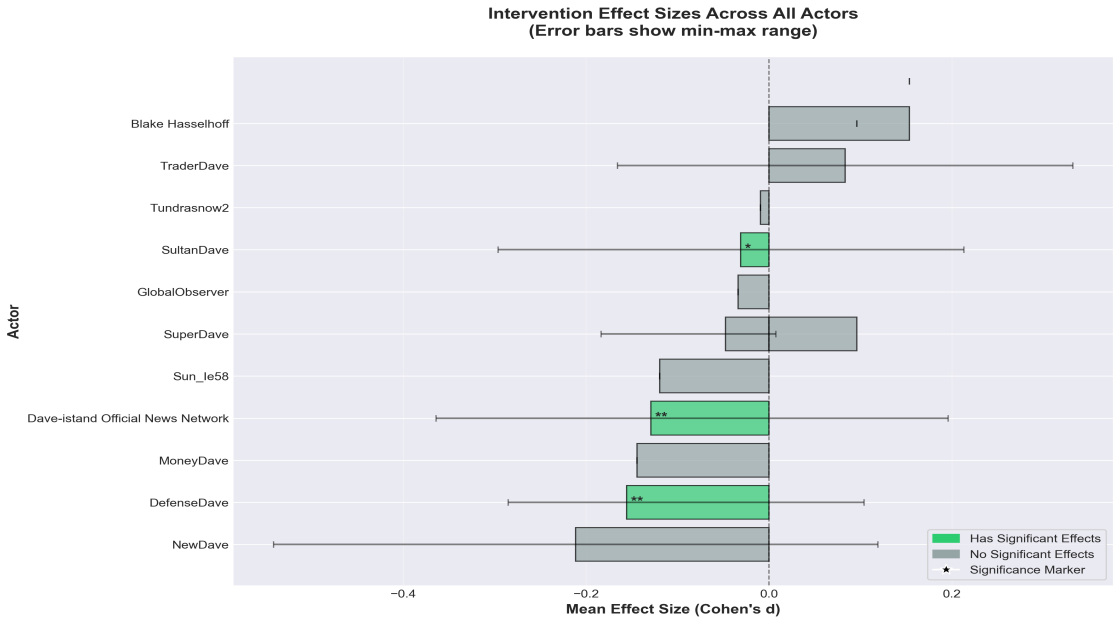



Figure 7.5: Effect Size Rankings Across All Intervention Actors. Green bars indicate actors with statistically significant effects ($p < 0.05$). Error bars show min-max range across topics.

evidence that JANUS-mediated interactions cause directional opinion shifts.

### Concentration of Influence

Opinion change was highly concentrated among a small set of susceptible actors. Only six actors (0.13% of the network) appeared repeatedly among the largest opinion shifts. Two of these actors—identified as the Ministry of National Defense of China (actor c4478d39) and Chen Jun, a Buddhist community leader (actor 13d4ef59)—were influenced by multiple intervention accounts, indicating the presence of "super-susceptible" individuals within the network. Actor c4478d39 exhibited cumulative shifts of 3.48 across 12 topics, with individual topic shifts reaching 0.345 (34.5% of the valence scale). Detailed case studies of Actor: Chen Jun is provided in Appendix E.6.7.

The magnitude-direction paradox—high directional alignment but small mean shifts—is explained by the heavy-tailed distribution of opinion changes. Most actors remain unchanged (median = 0), while 30-50% shift slightly in the intended direction (0.01-0.05), 2-3% shift substantially (0.1-0.35), and 20-30% shift in the opposite direction. This produces high alignment rates but near-zero mean effects. Extended analysis of this paradox is provided in Appendix D.4.

## 7.3.5 Summary of Key Findings

The JANUS belief dynamics analysis revealed six key patterns:

1. **Strong convergence:** All ten topics exhibited opinion convergence, with a mean variance

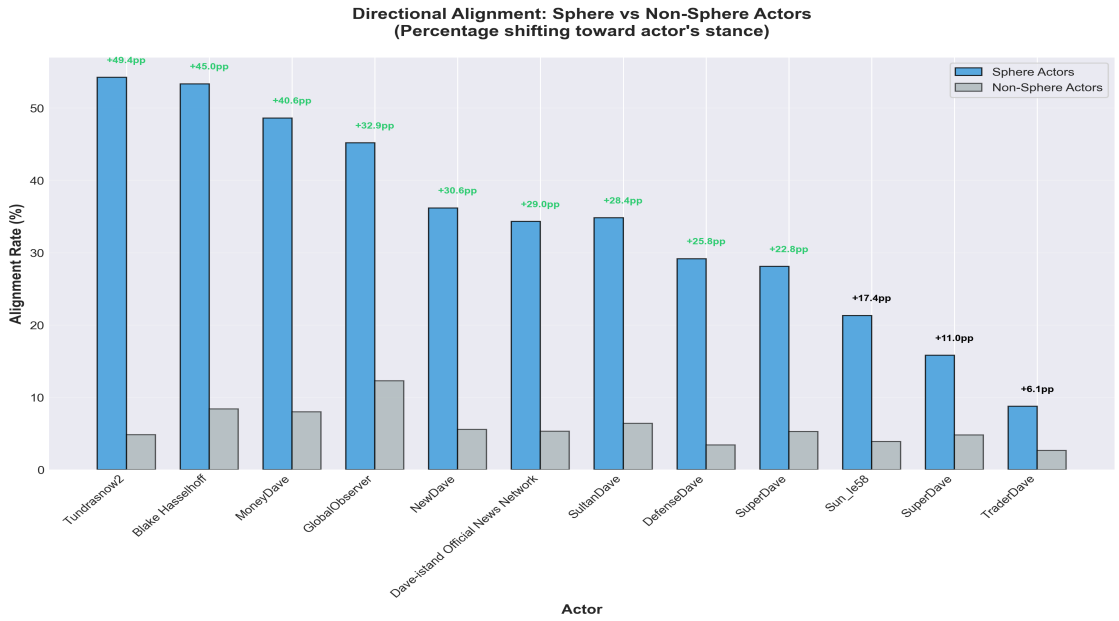



Figure 7.6: Directional Alignment: Sphere vs Non-Sphere Actors. Grouped bars show percentage of actors shifting toward intervention stance. Annotations show percentage point differences. Sphere actors consistently show 20-50pp higher alignment rates.

reduction of 79.94% ($\sigma = 29.70\%$). Topics related to military aggression showed the strongest convergence ($> 96\%$), while *Chinese aggression hegemony* remained highly polarized (3.91% reduction).

2. **Geopolitical divides:** Clear East-West opinion divides emerged on topics related to Russian and Chinese policies, with statistically significant differences in both mean valences and change magnitudes ($p < 0.001$ for all topics). Russian and Chinese personas exhibited the highest mean opinion changes (0.345 and 0.328, respectively).
3. **High-impact movers:** Among 373 top movers identified, Russian and Chinese personas showed the largest mean changes. Individual personas such as Ai Weiwei, Wang Wenbin, and Denis Olegovich Krutov exhibited changes exceeding 0.65, 0.60, and 0.52 valence units, respectively.
4. **Selective stance reversals:** Stance reversals were rare (93 total flips, 0.025% of observations) but concentrated on specific topics. *Canadian indigenous rights* showed 39 flips (41.9%), all moving positive→negative, while *Russian aggression* showed 22 flips (23.7%), all moving negative→positive. Detailed reversal analysis is provided in Appendix B.
5. **Intervention effectiveness:** Custom intervention actors produced measurable directional opinion shifts. Sphere actors were 5.8× more likely to align with intervention stances than non-sphere actors ($p < 0.0001$, $d = 2.28$). However, influence was highly concentrated, with only 2-3% of sphere actors experiencing meaningful shifts ($> 0.1$). Message volume was the strongest predictor of success ($r = 0.95$, $p < 0.0001$).

6. **Temporal volatility:** Opinion volatility was highest in early simulation periods (Days 1-3) and showed renewed spikes in mid-to-late stages (Days 4, 7, and 9), suggesting dynamic opinion reconsideration processes. Detailed timestep analysis is provided in Appendix A.2.

These findings provide quantitative evidence for the complex, multi-dimensional nature of opinion dynamics in geopolitical contexts, with significant variations across countries, topics, and temporal periods. The intervention analysis demonstrates that JANUS can mediate directional opinion shifts through direct engagement, though influence remains concentrated among a small subset of susceptible actors.

## 7.4 Discussion

This chapter evaluated whether the integrated AURORA–JANUS pipeline, embedded within AESOP, SynSM, and the broader GhostField architecture, can function as a usable socio-cognitive world model for influence research and training. The results collectively support this claim in three interrelated ways. First, the system produces structured and interpretable opinion dynamics across countries, institutions, and topics, rather than undifferentiated or random variation. Second, these dynamics evolve meaningfully over time—beliefs converge, polarize, shift, and occasionally reverse in ways that are systematically conditioned by geopolitical alignment and issue salience. Third, controlled interventions generate measurable and directional downstream effects, demonstrating that the environment supports causal influence processes within a heterogeneous synthetic population rather than merely producing plausible but inert content.

A central implication of these findings is that the framework generates *differentiated* population behavior. This differentiation is most clearly observed in the topic-dependent structure of opinion trajectories. High-salience security issues such as *Russian aggression*, *OPN Sea Dragon*, and *Taiwanese Sovereignty* exhibit strong convergence, reflecting alignment pressures and consistent narrative reinforcement. In contrast, topics such as *Chinese aggression and hegemony* remain persistently polarized, with minimal variance reduction over time. This contrast is theoretically significant: the model neither enforces universal consensus nor preserves polarization indiscriminately. Instead, it produces a spectrum of outcomes conditioned by salience and alignment, reinforcing the broader claim that realistic synthetic populations require structured heterogeneity, salience-sensitive initialization, and explicit opinion dynamics mechanisms rather than static or heuristic response rules.

This same structure extends to cross-country behavior. Russian and Chinese personas exhibit the largest opinion shifts and the strongest divergence from NATO-aligned populations on security-related topics, indicating that national context and narrative alignment are both encoded at initialization and preserved through interaction. As a result, population responses are conditioned by geopolitical position rather than abstract topic labels alone. This is critical for training applications: analysts are not interacting with a homogeneous audience model, but with populations whose responses vary systematically across national and institutional contexts.

The intervention results further reinforce this interpretation by revealing that influence operates in a selective, distributional manner rather than through uniform population-wide effects. Although average treatment effects are small, this reflects a highly skewed distribution: most agents

remain stable, while a smaller subset of actors within targeted intervention spheres exhibit substantial directional shifts. This resolves the apparent magnitude–direction paradox, in which low aggregate effects coexist with strong localized influence. Importantly, this pattern aligns with real-world influence environments, where campaigns rarely transform entire populations but instead shift strategically important subsets of susceptible individuals.

This selective susceptibility is further validated by stance-flipping behavior within JANUS. Stance reversals are empirically rare (0.025% of observations) and highly concentrated by topic, indicating strong inertia across most agents with limited but meaningful change among a subset of individuals. This behavior is consistent with empirical findings in real-world opinion dynamics, where belief revision is gradual, path-dependent, and unevenly distributed across populations [51, 61], as well as with prior simulation work demonstrating that stance shifts occur under constrained and non-uniform conditions [34]. In JANUS, these dynamics emerge from general mechanisms—effective stubbornness ($\alpha^{\text{eff}} i, k$), salience, and bounded update constraints ($\delta \max$)—rather than scenario-specific tuning. Consequently, influence remains structurally constrained: repeated exposure, high salience, and low resistance are required for meaningful opinion change, while widespread rapid reversals remain unlikely. This provides an important form of validation, as behavioral realism emerges from mechanism design rather than calibration.

Consistent with this interpretation, intervention volume—not nominal reach—emerges as the strongest predictor of influence success. Sustained exposure through repeated messaging produces measurable effects, whereas one-time or broad but shallow reach does not. This reflects a key property of real-world influence environments: persistence and reinforcement matter more than simple audience size. From a training perspective, this encourages more realistic campaign design, emphasizing timing, repetition, and narrative reinforcement rather than broadcast assumptions.

These findings support interpreting GhostField as a *world model* rather than a content generator. The system maintains internal state, updates that state through formal socio-cognitive dynamics, and produces outputs that reflect both endogenous evolution and exogenous intervention. This represents a significant architectural advance over earlier AESOP–SynSM systems, which could generate plausible message streams but lacked a grounded population layer and a mechanism for multi-day opinion evolution. By integrating AURORA and JANUS, GhostField models not only what information appears in the environment, but how heterogeneous populations respond to it over time.

This capability has implications for both research and practice. For research, it enables controlled experimentation on influence competition, polarization dynamics, salience shocks, and intervention effects within a unified and repeatable framework. For practice, it supports training environments in which analysts engage with structured, evolving populations rather than static or scripted entities. Because the system operates across multiple countries and topics without scenario-specific retuning, it also demonstrates a degree of generalizability within the examined conditions.

A useful framing for this capability is that of a *digital twin* of the information environment. In this view, the AURORA–JANUS–SynSM system functions as a structured proxy for real-world populations, capturing both demographic diversity and dynamic behavioral responses to evolving narratives. The value of this digital twin lies not in perfect prediction, but in providing an inter-

pretable environment for comparing strategies, exploring sensitivities, and evaluating second-order effects prior to real-world deployment.

The post-simulation survey results further reinforce this interpretation by demonstrating that participants engaged with the environment in ways consistent with real-world analytic behavior. Most participants reported targeting specific audiences (85%) and adapting messaging based on observed reactions (75%), indicating that they perceived the population as differentiated and responsive. Evaluation strategies focused on engagement and amplification rather than direct persuasion, reflecting real-world constraints on measuring influence. At the same time, the low rate of perceived rapid narrative spread (15%) suggests that participants did not experience the system as a simplistic viral diffusion model, but rather as a gradual, uneven, and interpretable influence space—consistent with the quantitative findings of selective susceptibility and locally structured diffusion.

Finally, while the results provide strong support for the framework's validity as a structured world model, they should be interpreted with appropriate caution. Rapid convergence on some topics may reflect realistic consolidation under repeated exposure, but may also be influenced by simulation horizon or update dynamics. Similarly, the concentration of intervention effects among highly susceptible actors is plausible, but sensitive to network structure, salience thresholds, and exposure mechanisms. These considerations do not undermine the core findings, but clarify their scope: the chapter validates GhostField as a behaviorally grounded and analytically useful model of the information environment, rather than a perfect one-to-one replica of real-world social media systems.

## 7.5 Conclusion

This chapter evaluated the integrated AURORA–JANUS framework within the GhostField training environment through the Indo-Pacific Geopolitical Opinion Experiment (IGOE). The results show that the system generates structured, topic-sensitive, and country-differentiated opinion dynamics at scale, while also supporting measurable intervention analysis in a live synthetic information environment.

Across 4,320 AURORA personas and 402 AESOP actors, the model produced systematic convergence, persistent polarization on selected topics, cross-country divergence aligned with geopolitical context, and concentrated influence effects under controlled interventions. These outcomes demonstrate that the framework is capable of simulating not only plausible message traffic, but also evolving socio-cognitive state. This is the chapter's central conclusion: when embedded within GhostField, AURORA and JANUS function as a practical world-modeling layer for influence research and training.

The chapter therefore contributes empirical support for a shift in how synthetic information environments are built. Rather than treating populations as background noise or static role templates, the integrated approach models them as heterogeneous, structured, and dynamically evolving agents. This substantially improves the realism, interpretability, and analytic utility of exercise environments. It also provides a foundation for future experiments in narrative competition, intervention design, analyst training, and social-cyber hypothesis testing.

More broadly, the chapter shows that synthetic populations grounded in demographic structure and formal opinion dynamics can bridge AI-enabled generation and social-scientific simulation. In that sense, GhostField is not only a training platform; it is an extensible computational laboratory for studying influence in complex information environments.

## 7.6 Contributions to the Dissertation

This chapter contributes to the dissertation in four primary ways.

**C1. Validation of an integrated socio-cognitive world model.** Earlier chapters introduced AURORA as a framework for generating grounded synthetic populations and JANUS as a model of opinion evolution. This chapter demonstrates that, when integrated into GhostField alongside AESOP and SynSM, these components operate coherently as a unified world model rather than as isolated modules. The contribution is therefore architectural as well as empirical: it shows how demographic synthesis, cognitive state initialization, formal opinion updating, and platform realization can be combined into a single executable system.

**C2. Empirical demonstration of structured opinion dynamics in a live exercise setting.** The chapter provides evidence that the integrated framework produces meaningful cross-country, cross-topic, and cross-institutional variation during a multi-day exercise. This moves the dissertation beyond offline validation and controlled modeling into live deployment conditions, showing that the synthetic populations retain interpretable behavior even when embedded in a time-stepped, participant-facing training environment.

**C3. Demonstration that intervention effects are measurable, selective, and analytically useful.** The intervention analysis shows that JANUS-mediated population dynamics can register directional influence from exogenous messaging, while also preserving realistic heterogeneity in response. This contributes a practical finding for both research and training: influence should be evaluated through distributional and network-based effects, not only through mean shifts. The framework therefore supports more nuanced analysis of susceptibility, targeting, and campaign effectiveness.

**C4. Extension of GhostField from synthetic corpus generation to experimental social simulation.** Perhaps the most important dissertation-level contribution is that this chapter demonstrates a transition from scenario-conditioned corpus generation to experimental socio-cognitive simulation. In prior architectures, synthetic traffic could be generated, but population cognition was weakly modeled or absent. By integrating AURORA and JANUS, GhostField becomes capable of supporting repeated, controlled experiments on how heterogeneous populations react to evolving narratives, events, and interventions over time.

## 7.7 Limitations

Several limitations should be noted.

First, the evaluation is conducted within a single scenario family—the Toto Scenario/IGOE configuration—and therefore reflects a bounded geopolitical and narrative design space. Although the framework generalized across multiple topics and countries within that scenario, broader claims about transfer to other regions, platforms, or conflict settings remain to be tested directly.

Second, the architecture depends on modeled rather than observed social networks. While the resulting network structure is useful for experimentation and supports meaningful diffusion dynamics, it is still a synthetic topology whose exact exposure pathways, clustering patterns, and bridge structures may differ from those of real-world online ecosystems. As a result, conclusions about intervention reach and susceptibility should be interpreted as valid within the model environment rather than as direct empirical estimates.

Third, the current intervention analysis focuses primarily on direct interaction spheres and short event windows. This is appropriate for establishing measurable first-order effects, but it does not fully capture longer-range diffusion, delayed influence, or more complex cascading processes that may emerge in extended simulations. Future work should examine second-order and cross-topic spillover effects more explicitly.

Fourth, although the results show substantial topic-sensitive variation, the model still abstracts many real-world complexities. Platform recommendation systems, strategic deception, shifting credibility, media competition, and human improvisation are represented only partially. Similarly, the current implementation emphasizes opinion valence and salience, but does not yet model the full richness of identity change, memory accumulation, or long-horizon learning.

Fifth, the participant-facing evaluation provides evidence of utility and plausibility, but it is not equivalent to a formal human-subject validation of whether the synthetic environment is indistinguishable from real-world social media. The chapter therefore supports the framework as analytically useful and behaviorally structured, not as a perfect behavioral duplicate of live online populations.

Finally, the framework inherits limitations from its component systems. AURORA depends on the quality of demographic priors, retrieved contextual data, and persona construction logic. JANUS depends on assumptions about influence, salience, and susceptibility. SynSM depends on platform abstractions that necessarily simplify real communication environments. The value of the integrated system lies in making those assumptions explicit, modular, and testable, but they remain assumptions nonetheless.

## 7.8 Future Work

A key area for future work is relaxing the assumption of a fixed population size and composition within each country and exercise. In the current framework, AURORA personas are statically defined at initialization, which supports experimental control but does not fully capture the dynamic nature of real-world information environments. Future extensions could introduce time-varying population structures, allowing both the number and type of personas to evolve over the course of

a scenario. This could be achieved through controlled population inflow and outflow processes, such as event-driven participation, shifts in community engagement, or platform-level behaviors like account creation and dormancy. Enabling dynamic population adaptation would allow the model to better represent changing audience composition, the emergence of new influential actors, and evolving patterns of attention across topics, thereby improving both realism and applicability for training and operational use.

While this chapter establishes behavioral and structural validity through alignment with established models of opinion dynamics and observed distributional properties (e.g., selective susceptibility, rare stance reversals, and bounded updates), a critical direction for future work is systematic external validation against real-world data. In particular, stance-flipping behavior—while plausibly constrained and emergent from general mechanisms such as salience, repeated exposure, and effective resistance—has not yet been quantitatively benchmarked against longitudinal observational datasets (e.g., panel surveys or temporally resolved social media stance data). Such validation would enable calibration of key parameters (e.g., $\alpha_{i,k}^{\text{eff}}$, $\delta_{\max}$) to better align simulated transition probabilities with empirical rates of belief revision. Importantly, this would also clarify the extent to which the model can support operational inference. Rather than providing deterministic predictions of when opinions will change, the framework is intended to characterize probabilistic conditions of susceptibility. Future work should therefore evaluate whether these modeled susceptibility patterns correspond to empirically observed responsiveness under repeated messaging or exposure campaigns. Finally, this line of work will help determine the appropriate level of generalization: whether validation at the level of behavioral mechanisms is sufficient for cross-scenario transfer, or whether scenario-specific recalibration is required for high-fidelity operational use.

An additional area for future work is the explicit formalization of how AURORA-generated communities map to scenario-defined group structures within AESOP. While communities are constructed to reflect broad institutional and social domains (e.g., media, government, civil society), and AESOP groups define coordinated actor alignments and narrative roles, the current framework relies on implicit alignment between these layers. Future work should develop a formal cross-walk that systematically links community categories to scenario group definitions, enabling more precise targeting of influence maneuvers and clearer attribution of population responses to specific strategic actors. This would also support more rigorous analysis of how institutional positioning shapes exposure, interpretation, and downstream belief evolution, strengthening both the interpretability and operational utility of the integrated GhostField architecture.

# Chapter 8

# Conclusion

## 8.1 Introduction

This dissertation began with a central problem in computation social science and social cybersecurity: how to build synthetic populations that are not only demographically plausible, but also psychologically grounded, emotionally expressive, and capable of participating in realistic opinion dynamics within contested information environments. Existing approaches typically capture only part of this problem. Traditional synthetic populations methods preserve demographic structure but omit cognition, affect, and narrative context. Classical opinion dynamics models offer interpretable mechanisms for social influence, but usually rely on simplified, topic-agnostic agents. More recent LLMs approaches generate fluent and context-sensitive personas, yet often lack demographic proportionality, reproducibility, interpretability, and principled mechanisms for longtiduinal belief change.

This dissertation addressed that gap by developing an integrated socio-cognitive modeling framework composed of four connected layers. First, it established a structurally grounded synthetic population substrate through *skeleton persona* generation. Second, it introduced emotionally grounded and psychologically coherent persona construction through hierarchical conditioning and retrieval-augmented generation. Third, it formalized **AURORA** as an initialization framework that transforms demographic substrates into grounded synthetic agents with opinions, salience, emotion, and social embedding. Fourth, it developed **JANUS**, a mechanistic opinion dynamics model in which belief change emerges from anchor retention, signed social influence, salience-sensitive stubbornness, and affective drift. Finally, it evaluated the integration of these components inside the **GhostField** live–virtual–constructive environment, demonstrating that the resulting system can function as a practical world model for influence experimentation and training.

The chapters of this dissertation advance a single overarching claim: realistic synthetic populations for influence modeling must be built as *layered systems* in which demographic proportionality, psychological structure, emotional realism, and dynamic belief evolution are explicitly connected. The dissertation therefore contributes not merely a single model, but a full modeling pipeline that moves from structural population synthesis to live socio-cognitive simulation.

## 8.2 Summary of the Integrated System

The overall system developed in this dissertation can be understood as a staged pipeline that transforms empirical population structure and contextual information into an executable world model for social-cyber experimentation. At the highest level, the framework proceeds through four major stages:

$$\begin{aligned} &\text{Structural Population Synthesis} \rightarrow \text{Grounded Persona Construction} \rightarrow \\ &\text{Opinion Dynamics Simulation} \rightarrow \text{Live Information Environment Realization.} \end{aligned} \tag{8.1}$$

### 8.2.1 Structural Foundations: Skeleton Personas

The first component of the system establishes a behaviorally neutral demographic substrate. Using ACS PUMS microdata and Iterative Proportional Fitting (IPF), this dissertation generated *skeleton personas* that reproduce target demographic marginals while remaining intentionally agnostic to opinions, affect, and narrative exposure. This stage is important because it isolates population structure from later behavioral enrichment. Rather than embedding cognition too early, the framework first ensures that the synthetic population is proportionally representative over the modeled demographic space. The semantic validation analysis in Chapter 3.8.3 further demonstrated that record-level plausibility can be improved through post hoc filtering, but only under validation architectures that preserve population-level structure. This led to the conclusion that deterministic, rule-based validation offers the best balance among coherence, interpretability, and distributional fidelity.

### 8.2.2 Emotionally Grounded Persona Construction

The second component of the dissertation addressed a major limitation in both traditional agent-based models and many LLM-based persona systems: the absence of psychologically grounded emotional structure. Through a hierarchical country $\rightarrow$ community $\rightarrow$ individual generation pipeline, the dissertation demonstrated that personas can be constructed so that emotional expression remains coherent across multiple social scales while preserving within-group variation. By conditioning personas on OCEAN personality traits, Moral Foundations Theory, and topic-specific cues, the framework generated synthetic agents whose emotional responses aligned with broad empirical patterns observed in real-world social media data. The results showed that emotional expression was not generic, but topic-sensitive, demographically structured, and improved substantially as prompting incorporated richer psychological context. This provided a crucial foundation for later chapters, since emotionally expressive agents are more appropriate inputs to opinion dynamics models than emotion-blind abstractions.

### 8.2.3 AURORA: Grounded Synthetic Population Initialization

Building on those structural and psychological foundations, **AURORA** formalized a retrieval-augmented population initialization framework for creating synthetic agents that are demograph-

ically proportional, psychologically coherent, and ideologically grounded. AURORA separated population construction into staged components: demographic skeleton generation, retrieval-grounded enrichment, hierarchical opinion initialization, affective profile assignment, and social media account realization. This separation was methodologically important. It preserved causal traceability by ensuring that demographic proportionality was not silently distorted by behavioral enrichment, while still enabling personas to be grounded in contemporary country- and topic-specific discourse. The results from the Indo-Pacific Geopolitical Opinion Experiment (IGOE) showed that AURORA generated populations exhibiting coherent cross-country opinion structure, stable salience variation, and emotionally differentiated reactions across geopolitical topics.

### 8.2.4 JANUS: Mechanistic Opinion Dynamics for Heterogeneous Populations

The dissertation then introduced **JANUS**, a mechanistic opinion dynamics model designed to operate over AURORA-generated populations. JANUS extends Friedkin–Johnsen-style influence dynamics by incorporating four coupled mechanisms: anchor retention, signed social influence, salience-modulated effective stubbornness, and affective emotional drift. In contrast to classical models that often collapse toward consensus or assume topic-agnostic influence, JANUS allows agents to remain attached to prior beliefs while updating through attraction, repulsion, and emotionally conditioned responsiveness. The scaling experiments in Chapter 6.10 showed that JANUS sustains heterogeneous and polarized end states even as network size increases, preserves stable topic differentiation under fixed topology and parameters, and concentrates large opinion changes within a relatively small subset of highly susceptible agents rather than generating population-wide conversion. These findings support the claim that psychologically heterogeneous synthetic populations can be embedded within interpretable social influence models without sacrificing dynamical stability.

### 8.2.5 GhostField: From Synthetic Corpus Generation to Socio-Cognitive World Modeling

The final stage of the dissertation integrated AURORA and JANUS inside the **GhostField** environment alongside AESOP and SynSM. This chapter demonstrated a shift from scenario-conditioned corpus generation to executable socio-cognitive simulation. Earlier synthetic information environments could generate plausible traffic and narrative structure, but they lacked a grounded population layer and a formal mechanism for multi-day belief evolution. By integrating AURORA and JANUS, GhostField became capable of representing not only what messages appear in an information environment, but also how heterogeneous populations interpret, respond to, and evolve under those messages over time.

The live exercise evaluation showed that this integrated architecture produces topic-sensitive, country-differentiated, and intervention-responsive dynamics. Convergence, persistent polarization, selective influence effects, and analytically useful response variation all emerged inside the

training environment. In this sense, GhostField functioned as more than a synthetic content generator; it operated as a bounded, interpretable world model for influence experimentation and training.

### 8.2.6 Longevity of Contribution: Bridging LLMs and Realistic Opinion Dynamics

The contributions of this dissertation are not tied to a specific generation of large language models, but rather to a broader methodological shift in how synthetic populations are constructed and used within computational social science. While the capabilities of LLMs will continue to evolve, the core contribution of this work lies in establishing a principled integration between generative models and mechanistic social simulation.

At a foundational level, this dissertation demonstrates that LLMs are most valuable not as standalone behavioral simulators, but as components within structured, interpretable pipelines. The AURORA framework formalizes how generative models can be used to initialize populations that are demographically grounded, psychologically coherent, and contextually informed, while preserving traceability and reproducibility. This separation between population initialization (LLM-driven) and population evolution (mechanistic) is expected to remain relevant regardless of future advances in generative AI.

Similarly, JANUS provides a durable contribution by showing how heterogeneous, LLM-conditioned agents can be embedded within formal opinion dynamics without sacrificing interpretability. The model does not depend on any specific language model, but instead operates over structured representations of beliefs, salience, identity, and affect. As such, improvements in LLM quality will enhance the fidelity of agent initialization, but the underlying dynamical system remains stable, extensible, and theoretically grounded.

More broadly, this dissertation contributes to a long-term research direction in which synthetic populations function as bounded, controllable world models for studying social influence. The integration of AURORA, JANUS, and GhostField demonstrates that it is possible to construct systems where realism, interpretability, and experimental control coexist. This balance is unlikely to be obviated by future AI advances; if anything, increasing model capability will make the need for structured, interpretable integration more important.

Finally, the layered architecture introduced in this work provides a template for future systems that incorporate improved generative models, richer data sources, and more complex interaction environments. Because each layer—demographic synthesis, psychological grounding, opinion dynamics, and platform realization—is modular, the framework can evolve as underlying technologies improve without requiring a complete redesign.

In this sense, the lasting contribution of the dissertation is not a specific model instance, but a generalizable paradigm for integrating generative AI with social simulation. It establishes how realism can be introduced without sacrificing interpretability, and how modern AI methods can augment—rather than replace—formal theories of social behavior.

## 8.3 Dissertation-Level Contributions

The dissertation makes several contributions that span methodological, theoretical, and applied dimensions.

**C1. A Layered Framework for Synthetic Population Modeling** The first major contribution is the development of a layered framework for synthetic population modeling that explicitly separates demographic structure, psychological enrichment, dynamic opinion updating, and platform realization. This layered design is one of the dissertation's core methodological contributions. It avoids conflating structural realism with behavioral assumptions, preserves interpretability, and allows each component to be validated on its own terms. In contrast to monolithic LLM-based agent systems, the dissertation shows that realistic synthetic populations are best built through modular integration rather than end-to-end generation alone.

**C2. A Demographically Proportional and Behaviorally Neutral Population Substrate** The dissertation contributes a formalized approach to constructing *skeleton personas* as behaviorally neutral demographic agents. This establishes a reusable substrate for downstream behavioral modeling and shows that demographic proportionality can be preserved while still supporting later psychological enrichment. The chapter on semantic validation further contributes an empirical characterization of the coherence–fidelity trade-off in post-IPF filtering, identifying deterministic rule-based validation as the most practical operating point for large-scale deployment.

**C3. A Hierarchical Method for Emotionally Grounded Persona Synthesis** A third contribution is the hierarchical persona synthesis framework that integrates demographic context, institutional or community-level structure, personality traits, moral foundations, and topic-specific emotion cues. This framework advances beyond demographic personas or prompt-only LLM personas by showing how emotionally expressive agents can be generated in a way that is coherent across multiple scales and empirically anchored to known behavioral patterns. The cross-platform validation against Facebook reaction patterns provides one of the dissertation's strongest empirical contributions, demonstrating that synthetic personas can reproduce broad demographic gradients in emotional response while remaining suitable for simulation.

**C4. A Retrieval-Augmented Initialization Framework for Structured Synthetic Agents** The dissertation's fourth major contribution is AURORA itself. AURORA demonstrates how Retrieval-Augmented Generation can be integrated with demographic synthesis and hierarchical conditioning to create grounded synthetic populations suitable for agent-based modeling. Its main novelty lies not in retrieval alone, but in using retrieval as a mechanism for *population grounding*: country-level priors, community-level institutional framing, and persona-level contextual detail are all connected through a coherent architecture. This provides a reproducible and interpretable alternative to unconstrained persona generation and offers a practical solution to the initialization problem in social simulation.

**C5. A Mechanistic Opinion Dynamics Model for LLM-Conditioned Populations** A fifth contribution is the development of JANUS as a hybrid model that integrates LLM-conditioned heterogeneity with formal social influence dynamics. JANUS contributes a mechanistic extension of Friedkin–Johnsen-style updating in which opinion change depends on signed influence, salience-sensitive stubbornness, and affective bias. This allows the model to preserve heterogeneity and sustain polarization under realistic network structures. Importantly, JANUS shows that LLM-generated diversity can parameterize an interpretable dynamical model rather than replace it. This contribution is significant because it bridges the gap between formal opinion dynamics and contemporary generative-AI-based agent populations.

**C6. Empirical Evidence for Topic-Sensitive and Scale-Sensitive Opinion Dynamics** The dissertation also contributes new empirical findings about how heterogeneous synthetic populations behave under opinion dynamics. JANUS preserved stable topic-level differences in polarization under identical topology and parameters, suggesting that persona-level initialization can encode meaningful topic sensitivity without bespoke retuning. It also revealed a scale-dependent transition in collective dynamics: smaller networks amplify polarization through local reinforcement, while larger networks settle into stable polarized equilibria under stronger averaging pressure. These results extend classical opinion dynamics by showing how scale, topic salience, and psychological heterogeneity interact within a unified model.

**C7. A Live Socio-Cognitive Training and Experimentation Environment** Finally, the dissertation contributes an applied systems integration result: the transformation of GhostField from a scenario-conditioned synthetic corpus generator into an experimental socio-cognitive world model. By integrating AESOP, AURORA, JANUS, and SynSM, the dissertation demonstrates that synthetic populations can be used not only for offline analysis but also for live exercise deployment. This makes the framework useful for both scientific experimentation and operational training. The live IGOE exercise showed that participants encountered populations with differentiated behavioral responses, persistent geopolitical structure, and measurable sensitivity to interventions. This is a practical contribution as well as a scientific one.

## 8.4 What the Dissertation Demonstrates as a Whole

Viewed as a whole, the dissertation demonstrates five overarching claims.

First, **demographic realism alone is not sufficient for realistic synthetic populations**. Population synthesis methods must be combined with mechanisms for psychological grounding, emotional differentiation, and social embedding if the resulting agents are to support meaningful social simulation.

Second, **LLMs are most useful when embedded within structured modeling pipelines rather than used as unconstrained behavioral simulators**. In this dissertation, LLMs are used for semantic compression, persona realization, contextual grounding, and narrative generation, but the dynamical core remains mechanistic and interpretable.

Third, **emotion is not peripheral to opinion dynamics**. Emotional structure shapes baseline predispositions, responsiveness, and topic sensitivity, and therefore must be represented if one wishes to model realistic reactions to contested narratives.

Fourth, **opinion dynamics in heterogeneous populations are strongly conditioned by topic and scale**. The same population can behave differently across issues, and the same model can exhibit different collective regimes as network size changes. This means that realistic world models must preserve both inter-topic and inter-scale structure.

Fifth, **synthetic populations can function as bounded world models for influence experimentation and training**. They are not substitutes for human subjects or full replicas of real online ecosystems, but they can support repeatable, interpretable, and ethically tractable experimentation on influence, narrative diffusion, and intervention response.

## 8.5 Limitations and Future Work

### 8.5.1 Limitations

Despite these contributions, the dissertation has several important limitations.

First, the Retrieval-Augmented Generation (RAG) knowledge base is limited to curated source material. Although this improves grounding and interpretability, it also constrains the breadth of contextual knowledge available to synthetic personas. As a result, generated outputs reflect the scope, quality, and representativeness of the selected corpus rather than the full diversity of real-world discourse.

Second, LLM-generated psychological attributes remain sensitive to prompt design and retrieved context. While the framework uses structured prompting and grounded retrieval to improve consistency, persona traits and behavioral tendencies may still vary depending on how information is framed, what context is retrieved, and how the model interprets those cues. This means that some aspects of psychological realism depend on the design of the generation pipeline rather than on directly observed human data.

Third, the emotion modeling developed in this dissertation captures broad affective tendencies rather than the full complexity of human psychology. The framework represents emotional predispositions and topic-conditioned responses in a structured and interpretable way, but it does not model the full richness of emotional regulation, memory, conflicting motives, or longitudinal affective development.

Fourth, the network structures used throughout the dissertation are simplified representations of real social media interaction patterns. Although they are sufficient for controlled experimentation and capture important features such as clustering, homophily, and cross-group exposure, they do not fully reproduce platform-specific mechanisms such as recommendation systems, algorithmic amplification, shifting visibility, or dynamic rewiring of ties over time.

Finally, agent emotional responses are modeled through trait–emotion mappings rather than full cognitive processes. This provides a tractable and theory-informed way to parameterize emotional variation across personas, but it abstracts away more complex processes such as deliberation, situational reinterpretation, memory-based appraisal, and recursive social reasoning. Accordingly,

the framework should be understood as modeling emotionally structured behavior, rather than complete human cognition.

### 8.5.2 Future Work

This dissertation establishes a modular framework for generating demographically grounded synthetic populations, enriching them with psychologically structured personas, simulating their opinion dynamics, and deploying them within a live synthetic information environment. While the results demonstrate the feasibility and utility of this approach, several important extensions remain for future research. These directions would improve the realism, explanatory depth, and operational value of the overall system.

**Content-Aware Message Processing in JANUS** A major next step is to incorporate content-aware message processing directly into JANUS opinion updates. In the current framework, belief change is driven primarily by interpersonal structure, signed influence, salience, baseline affect, and anchor retention. Although this supports interpretable and stable dynamics, it treats exposure at a relatively abstract level. In practice, however, the effect of a message depends not only on who sent it, but also on what the message contains and how it is framed.

Future work should therefore allow JANUS to condition opinion updates on message-level properties such as narrative frame, stance, emotional tone, moral language, argument strength, and source style. This would allow different messages about the same topic to exert different influence effects even when delivered through the same network ties. Such an extension would make the model better suited for studying persuasion, narrative competition, and intervention design, since it would connect changes in belief not just to exposure volume, but to the semantic and rhetorical properties of content itself.

**Expanding the AURORA RAG Knowledge base** Another important direction is to expand the Retrieval-Augmented Generation knowledge base used by AURORA. At present, the system relies on curated corpora designed to provide country-specific, topic-specific, and community-relevant grounding. This improves contextual fidelity, but it also limits persona generation to the scope and representativeness of the collected source material.

Future work should broaden the range, depth, and diversity of the knowledge base by incorporating richer news archives, survey summaries, policy documents, community descriptions, cultural reference material, and multilingual source collections. Expanding the corpus would improve cultural specificity, reduce reliance on narrow narrative samples, and allow personas to be grounded in a wider range of social and institutional contexts. It would also support more robust transfer across countries, issue domains, and time periods, strengthening AURORA as a reusable population initialization system rather than a scenario-specific pipeline.

**Structured Information Ecosystems for Synthetic Actors** The current framework models agents, topics, and social platforms, but future work should move toward richer structured information ecosystems. In real information environments, actors are embedded not only in social networks,

but also in persistent media systems that include news outlets, institutional channels, rumor networks, advocacy organizations, and platform-specific amplification pathways.

A useful extension would be to explicitly model these information ecosystems as structured layers within the simulation. Synthetic actors could then occupy different positions within a broader media ecology, such as institutional broadcasters, local news accounts, partisan amplifiers, community organizations, or informal peer networks. This would make it possible to study how narratives move not only through person-to-person exposure, but also across interacting layers of media and institutions. Such an architecture would improve the realism of GhostField and strengthen its value as both an experimental platform and a training environment.

**Tiered Media Trust Hierarchies** Relatedly, future veresions of the framwoek should incorporate a tiered media-trust hieachy. In the current system, agents repond to messages primarily through network exposure and internal predispositions, but real-world information processing is strongly shaped by perceived source credibility. Individuals do not treat all sournces equally; they assign different levels of trust to government agencies, journalists, influencers, partisan outlests. foreign actors, and anonymous accounts.

Relatedly, future versions of the framework should incorporate a tiered media-trust hierarchy. In the current system, agents respond to messages primarily through network exposure and internal predispositions, but real-world information processing is strongly shaped by perceived source credibility. Individuals do not treat all sources equally; they assign different levels of trust to government agencies, journalists, influencers, partisan outlets, foreign actors, and anonymous accounts.

A media-trust hierarchy would allow agents to weight incoming information differently depending on source type, institutional legitimacy, prior affinity, and platform context. For example, a persona might strongly trust public health institutions on medical topics while distrusting them on political issues, or trust local community organizations more than national media. This extension would improve both realism and analytic precision by allowing researchers to model credibility, distrust, and selective receptivity as explicit features of the influence process.

**Diurnal and Platform-Specific Activity Cycles** Another limitation of the current opinion dynamics implementation is that agents update continuously across timesteps without realistic temporal rhythms. In real social media environments, participation is temporally structured: people sleep, work, commute, and engage with platforms at different times and frequencies. Activity patterns also vary across platforms; for example, broadcast-oriented platforms and chat-based platforms often have different temporal rhythms and attention cycles.

Future work should therefore introduce diurnal and platform-specific activity models into JANUS and SynSM. Agents should not be assumed to update at every timestep. Instead, each persona could have activity windows shaped by time zone, work schedule, sleep cycle, platform preference, and habitual media use. This would allow the simulation to better capture bursts of coordinated attention, nighttime lulls, platform migration, and uneven exposure windows. Such a change would also improve realism in live exercises by making the synthetic environment feel less uniformly active and more like a living socio-technical system.

**Institutional Rules and Community Constraints in AURORA** Finally, future work should introduce stronger institutional and community rule constraints into AURORA. The current framework already embeds personas within communities and organizations, but these affiliations primarily shape context, opinion priors, and broad social identity. In many real settings, however, institutions impose explicit or implicit behavioral rules on members. Government officials, military personnel, journalists, healthcare workers, activists, and religious leaders are often constrained by professional norms, role expectations, and organizational obligations.

Incorporating these constraints would improve persona realism by ensuring that agents belonging to specific institutions behave in ways that are consistent with their roles. For example, a public health official may be less likely to post emotionally extreme content, a military spokesperson may follow stricter narrative discipline, and a religious community leader may frame issues through moral or theological language. Such role-based constraints would enhance both the interpretability and plausibility of the generated personas, while also supporting more faithful simulation of institutional communication environments.

These directions point toward a next generation of the framework in which agents are not only demographically and psychologically grounded, but also more deeply embedded in realistic media ecosystems, institutional structures, temporal rhythms, and content-sensitive interaction processes. Advancing along these lines would strengthen the dissertation's broader goal of building synthetic populations that can serve as credible, interpretable, and experimentally useful world models for social-cyber influence research.

# Appendix A

# Synthetic Persona (Skeleton) Generation

## A.1 Country-Level Demographic and Socio-Political Reference Data

This appendix provides the complete country-level demographic, socio-economic, and socio-political reference data used by the AURORA framework for synthetic population and persona initialization. These variables inform skeleton persona construction, community formation, opinion grounding, and salience calibration across simulation scenarios.

**Data scope and purpose.** Table A.1 presents the country-level demographic, digital, socioeconomic, and political reference variables used by the AURORA framework to initialize synthetic populations and skeleton personas. Core demographic indicators—including population size, urbanization rate, age structure, and sex ratio—were drawn from internationally harmonized datasets maintained by the World Bank [201] and national statistical offices [139]. These variables provide macro-level constraints that ensure population-scale realism while allowing downstream agent heterogeneity to emerge through sampling and interaction rather than deterministic assignment.

**Economic, digital, and platform context.** Economic indicators such as GDP per capita and employment sector composition were sourced from the World Bank and Statista [176, 201], enabling alignment with widely used benchmarks in computational social science. Internet penetration rates, social media adoption levels, and dominant platform usage were compiled from DataReportal's Digital Country Reports [47], which aggregate survey and platform-reported statistics across regions. These digital context variables inform agent exposure probabilities, platform-specific interaction opportunities, and baseline information access within the simulation environment.

**Sociocultural grounding.** Ethnic composition, language use, and religious affiliation were derived primarily from national census publications and official country profiles [139], supplemented where necessary by curated encyclopedic sources [199]. These attributes are used in AURORA to parameterize identity-relevant traits and contextual priors, supporting the modeling of homophily,

group salience, and culturally mediated belief formation without hard-coding individual-level behavior.

**Political alignment and institutional trust.** Political leanings reflect vote-share outcomes or dominant party alignments from the most recent national elections available for each country [199]. Institutional trust values represent the proportion of respondents reporting confidence in national governments or public institutions, drawn from cross-national trust surveys conducted by the OECD [144], Gallup, and the Edelman Trust Barometer [62]. Where survey estimates varied across sources, rounded values were used to preserve cross-country comparability rather than survey-specific precision.

**Modeling considerations.** All variables were normalized, discretized, or categorized to support computational use within AURORA. In cases of missing, conflicting, or rapidly changing indicators—particularly for trust and platform usage—conservative estimates were selected based on consensus ranges reported across sources [47, 144]. Importantly, these country-level attributes function as initialization constraints and contextual priors rather than deterministic drivers of agent behavior; belief change and influence dynamics emerge through agent interaction within the JANUS opinion model rather than directly from macro attributes.

## A.2 ACS PUMS Data Structure and Variable Encoding

This appendix documents the structure of the American Community Survey (ACS) Public Use Microdata Sample (PUMS) data used for skeleton persona generation, including variable encodings, marginal distributions, and sample records.

### A.2.1 Raw Data Format

The source data file contains weighted individual records with nine columns. Table A.2 shows representative records from the dataset.

### A.2.2 Variable Definitions and Encoding

Table A.3 provides detailed encoding schemes for all eight demographic variables used in skeleton persona construction.

### A.2.3 Empirical Marginal Distributions

Tables A.4 and A.5 present the weighted marginal distributions $P_{\text{real}}(v_j)$ extracted from the ACS PUMS data using person weights PWGTP. These marginals serve as target constraints for IPF synthesis.

| Country | Population | Urban % | M/F | 0–14 | 15–64 | 65+ | GDP pc | Internet | Social Media | Top Platforms | Ethnicity | Languages | Religion | Employment | Trust | Political Leanings |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Japan | 124M | 92 | 0.95 | 11 | 60 | 29 | $32k | 87 | 78% | LINE, X, IG, TT | Japanese 98% | Japanese | Shinto/Buddhist | Ag 3 / Ind 24 / Serv 73 | 42% | LDP, CDP, DPP |
| China | 1.41B | 66 | 1.12 | 16 | 69 | 16 | $13k | 78 | 77% | WeChat, Douyin | Han 91% | Mandarin | Non-rel 93% | Ag 22 / Ind 29 / Serv 49 | 91% | CCP |
| Russia | 144M | 75 | 0.86 | 18 | 66 | 17 | $15k | 92 | 73% | VK, TikTok | Russian 81% | Russian | Orthodox | Ag 6 / Ind 26 / Serv 68 | 80% | United Russia |
| United States | 340M | 84 | 0.97 | 17 | 65 | 18 | $86k | 93 | 73% | YouTube, FB | White 60% | English | Christian/None | Ag 2 / Serv 79 | 22% | Rep / Dem / Ind |
| Canada | 41M | 82 | 0.99 | 15 | 65 | 19 | $54k | 97 | 91% | FB, YT | Euro 70% | Eng/French | Christian 53% | Serv 80 | 49% | Liberal / Con |
| Philippines | 116M | 49 | 1.87 | 31 | 64 | 5 | $4k | 84 | 78% | FB, YT, TT | Tagalog grp | Filipino/Eng | Catholic 79% | Ag 22 / Serv 59 | 80% | PFP / Ind |
| South Korea | 52M | 81 | 1.00 | 11 | 71 | 18 | $33k | 97 | 95% | Kakao, YT | Korean 96% | Korean | Non-rel 51% | Serv 71 | 37% | Dem / PPP |
| India | 1.45B | 37 | 1.06 | 25 | 68 | 7 | $2.7k | 56 | 34% | YT, FB | Indo-Aryan | Hindi/Eng | Hindu 80% | Ag 44 / Serv 31 | 73% | BJP / INC |
| Germany | 85M | 78 | 0.98 | 14 | 63 | 23 | $54k | 94 | 80% | YT, IG | German 74% | German | Christian 55% | Serv 75 | 36% | CDU / AfD |
| France | 67M | 82 | 0.94 | 17 | 61 | 22 | $41k | 95 | 76% | YT, FB | French 87% | French | Catholic 48% | Serv 78 | 28% | LREM / RN |
| UK | 69M | 84 | 0.98 | 17 | 63 | 17 | $49k | 98 | 82% | YT, FB | White 82% | English | Christian 46% | Serv 85 | 27% | Lab / Con |
| Taiwan | 23M | 81 | 0.98 | 12 | 69 | 19 | $33k | 91 | 80% | YT, FB | Han 95% | Mandarin | Folk/Buddhist | Serv 59 | 10% | DPP / KMT |
| Thailand | 72M | 54 | 0.95 | 15 | 70 | 15 | $7k | 98 | 85% | FB, YT | Thai 95% | Thai | Buddhist 93% | Ag 30 / Serv 48 | 58% | Move Forward |
| Iran | 92M | 78 | 1.03 | 23 | 69 | 8 | $4.8k | 80 | 52% | Telegram, IG | Persian 61% | Persian | Muslim 99% | Ag 16 / Serv 49 | <50% | SHANA |
| Pakistan | 251M | 38 | 1.05 | 37 | 59 | 4 | $1.5k | 27 | 26% | YT, TT | Punjabi grp | Urdu/Eng | Muslim 96% | Ag 36 / Serv 38 | 47% | PTI / PML |
| Australia | 27M | 87 | 0.98 | 19 | 66 | 16 | $64k | 97 | 78% | YT, FB | Euro 75% | English | Christian 44% | Serv 77 | 46% | Lab / LNP |
| Vietnam | 101M | 40 | 1.00 | 24 | 68 | 8 | $4.7k | 78 | 75% | FB, YT | Kinh 85% | Vietnamese | Irreligious | Ag 33 / Serv 36 | – | Communist |
| Saudi Arabia | 35M | 85 | 1.56 | 24 | 73 | 3 | $35k | 100 | 97% | YT, TT | Arab 90% | Arabic | Muslim 100% | Serv 72 | 87% | Monarchy |
| Iraq | 45M | 72 | 1.06 | 37 | 60 | 3 | $6k | 82 | 71% | FB, TT | Arab/Kurd | Arabic | Muslim 85% | Serv 64 | 55% | Sadrist |
| Egypt | 117M | 43 | 1.06 | 32 | 63 | 5 | $3.3k | 73 | 43% | YT, FB | Egyptian 99% | Arabic | Muslim 90% | Serv 53 | 66% | Independent |
| Ethiopia | 132M | 24 | 1.06 | 39 | 58 | 3 | $3.3k | 17 | 6% | FB | Oromo/Amhara | Semitic | Orthodox 45% | Ag 62 | 30% | Prosperity |
| Ukraine | 38M | 70 | 1.06 | 14 | 67 | 19 | $5.4k | 82 | 64% | YT, FB | Ukrainian 78% | Ukrainian | Orthodox | Serv 60 | 35% | Servant P. |
| Poland | 37M | 60 | 1.06 | 15 | 65 | 20 | $25k | 86 | 76% | FB, YT | Polish 99% | Polish | Catholic 71% | Serv 63 | 32% | PO / Ind |
| Italy | 59M | 72 | 1.06 | 12 | 63 | 25 | $40k | 87 | 73% | FB, YT | Italian 95% | Italian | Catholic 75% | Serv 70 | 36% | Brothers |
| Spain | 49M | 82 | 1.06 | 13 | 66 | 21 | $35k | 95 | 83% | YT, IG | Euro 81% | Spanish | Catholic 55% | Serv 77 | 37% | PP / PSOE |
| Austria | 9M | 60 | 1.05 | 14 | 65 | 21 | $57k | 95 | 80% | YT, FB | Austrian 91% | German | Catholic 55% | Serv 71 | 26% | FPO / OVP |
| Norway | 5.6M | 84 | 1.06 | 16 | 65 | 19 | $87k | 99 | 82% | YT, IG | Euro 92% | Norwegian | Protestant 62% | Serv 79 | 48% | Labour |
| Sweden | 10.6M | 89 | 1.06 | 17 | 62 | 21 | $58k | 96 | 80% | YT, IG | Euro 88% | Swedish | Protestant 56% | Serv 81 | 43% | Soc Dem |
| Mexico | 131M | 82 | 1.04 | 25 | 67 | 8 | $14k | 81 | 70% | YT, FB | Latino 69% | Spanish | Catholic 78% | Serv 63 | 54% | Morena |
| Brazil | 212M | 88 | 1.04 | 27 | 69 | 11 | $10k | 84 | 68% | YT, IG | Mixed 45% | Portuguese | Catholic 57% | Serv 72 | 39% | PT / PL |
| Chile | 20M | 88 | 1.04 | 17 | 69 | 14 | $17k | 95 | 75% | YT, FB | Euro 59% | Spanish | Catholic 54% | Serv 72 | 30% | Social Conv. |
| Argentina | 46M | 93 | 1.05 | 22 | 66 | 12 | $14k | 89 | 70% | YT, FB | Euro 79% | Spanish | Catholic 63% | Serv 76 | 43% | PL / FR |

Table A.1: Country-level demographic, digital, socioeconomic, and political attributes used for AURORA skeleton persona generation (32 countries).

Table A.2: Sample records from ACS PUMS dataset

| PWGTP | SEX | MAR | RAC1P | HISP | COW | SCHL | AGEP | PINCP |
|---|---|---|---|---|---|---|---|---|
| 69 | 2 | 2 | 1 | 01 | 0 | 16 | 94 | 12200 |
| 42 | 1 | 5 | 1 | 01 | 0 | 19 | 18 | 0 |
| 18 | 1 | 4 | 9 | 02 | 0 | 21 | 75 | 108000 |
| 24 | 1 | 1 | 2 | 01 | 0 | 18 | 44 | 0 |

### A.2.4 Age Distribution

Table A.6 shows the age distribution aggregated into 10-year bins for visualization purposes. IPF operates on continuous age values (0–99), but binned distributions facilitate comparison with focus relation analyses.

Table A.6: Age distribution in 10-year bins.

| Age Bin | Proportion |
|---|---|
| 0–9 | 0.1137 |
| 10–19 | 0.1293 |
| 20–29 | 0.1300 |
| 30–39 | 0.1378 |
| 40–49 | 0.1256 |
| 50–59 | 0.1218 |
| 60–69 | 0.1215 |
| 70–79 | 0.0810 |
| 80–89 | 0.0322 |
| 90–99 | 0.0071 |

### A.2.5 Focus Relation: Age $\times$ Education Joint Distribution

Table A.7 presents selected cells from the Age $\times$ Education joint distribution, which serves as the primary focus relation for evaluating semantic validation strategies. The full distribution contains $10 \times 25 = 250$ cells; only representative high-probability combinations are shown.

### A.2.6 Data Preprocessing and Caching

To accelerate iterative experimentation, marginal distributions and joint distributions are computed once from the weighted ACS PUMS data and cached in JSON format (`USA_micropop_8_summaries_cache.`
This cache contains:

- **One-way marginals:** $P_{\text{real}}(v_j)$ for all eight variables (Table A.4–A.6)
- **Two-way joint distributions:** All pairwise combinations (e.g., Age $\times$ Sex, Age $\times$ Education), totaling 28 distributions

- **Three-way joint distributions:** Selected three-variable combinations for advanced correlation analysis

The caching strategy reduces IPF initialization time from $\sim$30 seconds (recomputing marginals from 2.5M+ weighted records) to $<$1 second (loading pre-computed JSON), enabling rapid iteration during validation strategy development.

### A.2.7 Data Representativeness

The ACS PUMS dataset represents the U.S. civilian non-institutionalized population. Person weights (PWGTP) are calibrated by the U.S. Census Bureau to match known population totals across demographic dimensions. Key representativeness properties:

- **Geographic coverage:** All 50 states and District of Columbia
- **Sample size:** 2.5M+ individual records (1% sample)
- **Weighted population:** Approximately 330M individuals when weights applied
- **Temporal coverage:** Single-year estimates reflect population snapshot
- **Exclusions:** Institutionalized populations (prisons, nursing homes), military personnel in barracks, and overseas residents

These exclusions are standard for household survey data and do not materially affect skeleton persona validity for most agent-based modeling applications, which focus on civilian household populations.

Table A.3: Variable definitions and encoding schemes.

| Variable | Description | Encoding |
|---|---|---|
| **PWGTP** | Person weight | Integer weight for population-representative sampling |
| **SEX** | Biological sex | 1 = Male, 2 = Female |
| **MAR** | Marital status | 1 = Married; 2 = Widowed; 3 = Divorced; 4 = Separated; 5 = Never married |
| **RAC1P** | Detailed race | 1 = White alone; 2 = Black/African American alone<br>3 = American Indian alone; 4 = Alaska Native alone<br>5 = American Indian/Alaska Native tribes<br>6 = Asian alone; 7 = Native Hawaiian/Pacific Islander<br>8 = Some other race; 9 = Two or more races |
| **HISP** | Hispanic origin | 01 = Not Spanish/Hispanic/Latino<br>02–24 = Various Hispanic/Latino origins<br>(Mexican, Puerto Rican, Cuban, etc.) |
| **COW** | Class of worker | 0 = Not in universe (children, unemployed)<br>1 = Private for-profit employee<br>2 = Private non-profit employee<br>3 = Local government employee<br>4 = State government employee<br>5 = Federal government employee<br>6 = Self-employed (incorporated)<br>7 = Self-employed (not incorporated)<br>8 = Working without pay (family business)<br>9 = Unemployed |
| **SCHL** | Educational attainment | 0 = No schooling; 1–8 = Grades 1–8<br>9–12 = Grades 9–12 (no diploma)<br>13–14 = High school or GED<br>15–16 = Some college, no degree<br>17 = Associate's degree<br>18–19 = Some graduate school<br>20 = Bachelor's degree<br>21 = Master's degree<br>22 = Professional degree<br>23 = Doctorate degree<br>24 = Post-doctoral |
| **AGEP** | Age in years | Continuous integer (0–99+) |
| **PINCP** | Personal income | Annual personal income in dollars (continuous) |

Table A.4: Marginal distributions for discrete demographic variables.

| Sex (SEX) | | | Marital Status (MAR) | | |
|---|---|---|---|---|---|
| **Code** | **Category** | **Proportion** | **Code** | **Category** | **Proportion** |
| 1 | Male | 0.4949 | 1 | Married | 0.3961 |
| 2 | Female | 0.5051 | 2 | Widowed | 0.0450 |
| | | | 3 | Divorced | 0.0859 |
| | | | 4 | Separated | 0.0134 |
| | | | 5 | Never married | 0.4597 |
| **Race (RAC1P)** | | | **Class of Worker (COW)** | | |
| **Code** | **Category** | **Proportion** | **Code** | **Category** | **Proportion** |
| 1 | White alone | 0.6052 | 0 | Not in universe | 0.3919 |
| 2 | Black alone | 0.1214 | 1 | Private for-profit | 0.3976 |
| 3 | Amer. Indian | 0.0080 | 2 | Private non-profit | 0.0531 |
| 4 | Alaska Native | 0.0004 | 3 | Local government | 0.0452 |
| 5 | AI/AN tribes | 0.0016 | 4 | State government | 0.0247 |
| 6 | Asian alone | 0.0598 | 5 | Federal government | 0.0202 |
| 7 | Native Hawaiian | 0.0019 | 6 | Self-employed (inc.) | 0.0380 |
| 8 | Other race | 0.0740 | 7 | Self-employed | 0.0236 |
| 9 | Two+ races | 0.1276 | 8 | Unpaid family | 0.0018 |
| | | | 9 | Unemployed | 0.0038 |

Table A.5: Marginal distribution for educational attainment (SCHL).

| **Code** | **Level** | **Prop.** | **Code** | **Level** | **Prop.** |
|---|---|---|---|---|---|
| 0 | No schooling | 0.0320 | 13 | Grade 11 | 0.0215 |
| 1 | Grade 1 or below | 0.0351 | 14 | Grade 12, no diploma | 0.0245 |
| 2 | Grade 2 | 0.0147 | 15 | High school diploma | 0.0171 |
| 3 | Grade 3 | 0.0129 | 16 | GED or alternative | 0.1839 |
| 4 | Grade 4 | 0.0119 | 17 | Some college, $<$1 year | 0.0296 |
| 5 | Grade 5 | 0.0125 | 18 | Some college, 1+ years | 0.0538 |
| 6 | Grade 6 | 0.0136 | 19 | Associate's degree | 0.1068 |
| 7 | Grade 7 | 0.0132 | 20 | Bachelor's degree | 0.0664 |
| 8 | Grade 8 | 0.0141 | 21 | Master's degree | 0.1624 |
| 9 | Grade 9 | 0.0182 | 22 | Professional degree | 0.0722 |
| 10 | Grade 10 | 0.0148 | 23 | Doctorate degree | 0.0165 |
| 11 | Grade 11 | 0.0193 | 24 | Post-doctoral | 0.0120 |
| 12 | Grade 12, no diploma | 0.0210 | | | |

Table A.7: Selected cells from Age × Education joint distribution (proportions).

| Age Bin | No School (SCHL=0) | Grades 1–8 (SCHL=1–8) | HS/GED (SCHL=16) | Bachelor's (SCHL=20) | Master's (SCHL=21) | Doctorate (SCHL=22) |
|---|---|---|---|---|---|---|
| 0–9 | 0.0320 | 0.0813 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| 10–19 | 0.0000 | 0.0004 | 0.0117 | 0.0004 | 0.0000 | 0.0000 |
| 20–29 | 0.0000 | 0.0014 | 0.0357 | 0.0103 | 0.0303 | 0.0062 |
| 30–39 | 0.0000 | 0.0021 | 0.0256 | 0.0126 | 0.0353 | 0.0158 |
| 40–49 | 0.0000 | 0.0027 | 0.0227 | 0.0117 | 0.0285 | 0.0156 |
| 50–59 | 0.0000 | 0.0028 | 0.0261 | 0.0113 | 0.0256 | 0.0126 |
| 60–69 | 0.0000 | 0.0026 | 0.0300 | 0.0115 | 0.0226 | 0.0104 |
| 70–79 | 0.0000 | 0.0017 | 0.0197 | 0.0066 | 0.0145 | 0.0086 |
| 80–89 | 0.0000 | 0.0009 | 0.0098 | 0.0018 | 0.0047 | 0.0027 |
| 90–99 | 0.0000 | 0.0003 | 0.0025 | 0.0003 | 0.0009 | 0.0004 |

*Note:* Zero probabilities in age 0–9 for high education levels reflect demographic impossibility (children cannot hold degrees). Validation strategies aim to prevent such combinations in synthetic populations while preserving overall Age × Education correlation structure.

# Appendix B

# Modeling Emotional Reasoning in Synthetic Personas

## B.1 OCEAN Personality Parameters

This appendix reports the country-level and regional-level OCEAN (Openness, Conscientiousness, Extraversion, Agreeableness, Neuroticism) personality parameters used in Stage 4 of the AURORA pipeline. For each country and region, mean ($M$) and standard deviation ($SD$) values define the truncated normal distributions from which individual-level personality traits are sampled.

Country-specific parameters are used when available; otherwise, regional aggregates provide fallback estimates to preserve cross-national psychological heterogeneity. These values operationalize empirically grounded variation in baseline personality structure across populations, ensuring that synthetic personas reflect documented cross-cultural differences in trait distributions while remaining statistically tractable for large-scale generation [170].

Table B.1: Country-Level OCEAN Personality Parameters

| **Country** | $E_M$ | $E_{SD}$ | $A_M$ | $A_{SD}$ | $C_M$ | $C_{SD}$ | $N_M$ | $N_{SD}$ | $O_M$ | $O_{SD}$ |
|---|---|---|---|---|---|---|---|---|---|---|
| Argentina | 0.491 | 0.078 | 0.428 | 0.094 | 0.482 | 0.098 | 0.551 | 0.092 | 0.508 | 0.106 |
| Australia | 0.490 | 0.101 | 0.475 | 0.107 | 0.459 | 0.109 | 0.508 | 0.104 | 0.501 | 0.088 |
| Austria | 0.506 | 0.092 | 0.459 | 0.083 | 0.467 | 0.111 | 0.497 | 0.089 | 0.493 | 0.103 |
| Bangladesh | 0.450 | 0.065 | 0.505 | 0.076 | 0.467 | 0.097 | 0.512 | 0.086 | 0.534 | 0.100 |
| Belgium | 0.460 | 0.101 | 0.451 | 0.090 | 0.430 | 0.112 | 0.536 | 0.098 | 0.546 | 0.086 |
| Brazil | 0.459 | 0.094 | 0.459 | 0.088 | 0.454 | 0.093 | 0.531 | 0.091 | 0.492 | 0.094 |
| Canada | 0.483 | 0.097 | 0.491 | 0.091 | 0.491 | 0.102 | 0.506 | 0.097 | 0.488 | 0.100 |
| France | 0.454 | 0.088 | 0.466 | 0.082 | 0.493 | 0.102 | 0.523 | 0.093 | 0.481 | 0.095 |
| Germany | 0.503 | 0.090 | 0.451 | 0.082 | 0.465 | 0.101 | 0.503 | 0.084 | 0.478 | 0.093 |
| India | 0.474 | 0.089 | 0.504 | 0.078 | 0.474 | 0.107 | 0.500 | 0.108 | 0.485 | 0.085 |
| Japan | 0.467 | 0.081 | 0.422 | 0.088 | 0.378 | 0.093 | 0.579 | 0.074 | 0.415 | 0.105 |
| Taiwan | 0.478 | 0.083 | 0.447 | 0.080 | 0.425 | 0.092 | 0.531 | 0.089 | 0.457 | 0.097 |
| United Kingdom | 0.498 | 0.097 | 0.473 | 0.094 | 0.469 | 0.107 | 0.514 | 0.099 | 0.460 | 0.097 |
| United States | 0.500 | 0.100 | 0.500 | 0.100 | 0.500 | 0.100 | 0.500 | 0.100 | 0.500 | 0.100 |
| Zimbabwe | 0.487 | 0.088 | 0.498 | 0.097 | 0.518 | 0.105 | 0.483 | 0.091 | 0.485 | 0.081 |

Table B.2: Regional-Level OCEAN Personality Parameters

| **Region** | $E_M$ | $E_{SD}$ | $A_M$ | $A_{SD}$ | $C_M$ | $C_{SD}$ | $N_M$ | $N_{SD}$ | $O_M$ | $O_{SD}$ |
|---|---|---|---|---|---|---|---|---|---|---|
| North America | 0.495 | 0.097 | 0.496 | 0.100 | 0.483 | 0.106 | 0.495 | 0.097 | 0.503 | 0.103 |
| South America | 0.481 | 0.080 | 0.458 | 0.088 | 0.471 | 0.095 | 0.527 | 0.086 | 0.515 | 0.098 |
| Western Europe | 0.495 | 0.091 | 0.469 | 0.088 | 0.456 | 0.107 | 0.504 | 0.092 | 0.497 | 0.095 |
| Eastern Europe | 0.499 | 0.082 | 0.465 | 0.085 | 0.465 | 0.096 | 0.497 | 0.085 | 0.508 | 0.090 |
| Southern Europe | 0.492 | 0.085 | 0.491 | 0.089 | 0.478 | 0.105 | 0.522 | 0.096 | 0.503 | 0.096 |
| Middle East | 0.492 | 0.085 | 0.493 | 0.087 | 0.489 | 0.104 | 0.507 | 0.094 | 0.500 | 0.096 |
| Africa | 0.492 | 0.075 | 0.510 | 0.089 | 0.515 | 0.094 | 0.479 | 0.083 | 0.481 | 0.085 |
| Oceania | 0.498 | 0.088 | 0.479 | 0.097 | 0.448 | 0.101 | 0.492 | 0.092 | 0.489 | 0.087 |
| South & Southeast Asia | 0.484 | 0.071 | 0.492 | 0.075 | 0.455 | 0.093 | 0.501 | 0.088 | 0.494 | 0.085 |
| East Asia | 0.466 | 0.077 | 0.434 | 0.080 | 0.406 | 0.088 | 0.544 | 0.079 | 0.433 | 0.095 |

# B.2 Extended Persona Dataset Validation and Calibration

This appendix provides detailed validation results for the LLM-generated persona dataset, including institutional community balance, moral foundations coherence, psychological trait calibration, and linguistic response structure metrics. These analyses supplement the high-level findings summarized in Section 7.3.

## B.2.1 Institutional Community Balance and Diversity

The hierarchical generation pipeline produced a total of 537 personas (267 U.S., 270 Canadian) distributed evenly across six institutional categories. As shown in Table B.3, representation across countries and institution types was nearly identical, confirming that the generation process maintained proportional balance across all categories.

Table B.3: Distribution of LLM-generated personas across institutional categories.

| **Institutional Category** | **U.S.** | **Canada** | **Total** | **%** |
|---|---|---|---|---|
| Industry Institutions | 45 | 45 | 90 | 16.8 |
| University Institutions | 45 | 45 | 90 | 16.8 |
| Medical Institutions | 45 | 45 | 90 | 16.8 |
| Liberal Governmental Institutions | 45 | 45 | 90 | 16.8 |
| Conservative Governmental Institutions | 44 | 45 | 89 | 16.6 |
| Education Institutions | 43 | 45 | 88 | 16.4 |
| **Total** | **267** | **270** | **537** | **100.0** |

Within each institutional category, the LLM generated multiple communities reflecting realistic organizational landscapes. For instance, *Industry Institutions* included groups such as the Small Business Association and Chamber of Commerce, while *University Institutions* encompassed organizations like the Black Student Union and the Society of Hispanic Professional Engineers. Each community contained 8–9 personas, ensuring balanced representation and diversity in institutional substructures.

### B.2.2 Moral Foundations Consistency by Ideological Orientation

The persona generation process incorporated both LLM-based ideological tagging (Liberal, Moderate, Conservative, Libertarian) and rule-based moral foundations mapping derived from Moral Foundations Theory (MFT). This hybrid strategy ensured theoretical coherence while preserving ideological diversity.

Table B.4 presents moral foundation means by political orientation, confirming alignment with MFT expectations. Liberal and center-left personas emphasize *Care* and *Fairness*, while conservative personas score higher on *Authority*, *Loyalty*, and *Sanctity*. These clear ideological gradients and high effect sizes ($d > 4.0$ across all foundations) demonstrate that the generated ideological structure reproduces empirically validated psychological distinctions.

Table B.4: Moral Foundations Theory validation: ideology-to-moral-foundation mappings by political orientation.

| **Political Group** | **Care** | **Fairness** | **Loyalty** | **Authority** | **Sanctity** | **Liberty** |
|---|---|---|---|---|---|---|
| Left (n=139) | 4.95 | 4.95 | 3.00 | 3.00 | 2.05 | 4.00 |
| Center Left (n=150) | 4.77 | 4.77 | 3.00 | 3.00 | 2.23 | 4.00 |
| Center (n=78) | 4.15 | 4.15 | 3.05 | 3.10 | 2.95 | 3.95 |
| Center Right (n=124) | 4.00 | 4.00 | 3.57 | 4.15 | 4.15 | 3.43 |
| Right (n=46) | 4.00 | 4.00 | 3.89 | 4.78 | 4.78 | 3.11 |
| ***Theoretical Validation (Left vs. Right)*** | | | | | | |
| Difference (L–R) | +0.95*** | +0.95*** | –0.89*** | –1.78*** | –2.73*** | +0.89*** |
| Cohen's d | 4.98 | 4.98 | 5.71 | 5.71 | 7.47 | 5.71 |
| Pattern Match | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |

Note: All political group differences are significant at $p < 0.001$. Pattern Match confirms correspondence with Moral Foundations Theory predictions.

### B.2.3 OCEAN Personality Trait Calibration

To ensure psychological realism, personas were assigned OCEAN (Big Five) personality traits drawn from normal distributions parameterized by cross-cultural means from [127, 170]. Table B.5 reports LLM-generated averages against empirical benchmarks for both the U.S. and Canada.

Table B.5: OCEAN personality trait validation: LLM-generated vs. empirical research data.

| Trait | U.S. | | Canada | | Deviation[a] | |
|---|---|---|---|---|---|---|
| | LLM | Obs. | LLM | Obs. | U.S. | CA |
| Openness | 0.492 | 0.500 | 0.485 | 0.487 | –0.08 | –0.02 |
| Conscientiousness | 0.495 | 0.500 | 0.486 | 0.490 | –0.05 | –0.04 |
| Extraversion | 0.493 | 0.500 | 0.490 | 0.483 | –0.07 | +0.07 |
| Agreeableness | 0.500 | 0.500 | 0.501 | 0.491 | 0.00 | +0.010 |
| Neuroticism | 0.486 | 0.500 | 0.502 | 0.506 | –0.014 | –0.04 |
| **Mean Absolute Deviation** | 0.07 | | 0.05 | | 0.00 | |

Deviations represent LLM mean minus empirical benchmark.

Trait alignment demonstrated strong psychological validity. Canadian personas scored slightly higher in *Agreeableness* and lower in *Openness*, reflecting cultural values emphasizing modesty and cooperation. Institutional variation followed logical patterns: governmental personas showed highest Openness ($\mu = 0.504$), university personas lowest Conscientiousness ($\mu = 0.479$), and industry personas highest Agreeableness ($\mu = 0.508$).

Demographic effects were subtle but interpretable—age correlated positively with Openness ($r = +0.089$) and Extraversion ($r = +0.071$), and negatively with Agreeableness ($r = -0.060$) and Neuroticism ($r = -0.069$). Gender effects were minimal (all $|\Delta| < 0.010$), indicating consistent calibration across demographic groups.

### B.2.4 Response Structure and Emotional Intensity Distributions

Finally, the linguistic realism of persona responses was assessed through word-level and sentence-level structure metrics. Figure B.1 visualizes distributions for word count, sentence count, average word length, punctuation ratio, and affective word proportions (positive, negative, intensity).

Roe v. Wade and Suicide Hotline 988 responses exhibited comparable response lengths, but differed in affective tone. Roe v. Wade posts contained higher negative word ratios, corresponding to elevated sadness ($r = 0.246$) and anger ($r = 0.186$). In contrast, Suicide Hotline 988 posts contained more positive and high-intensity words, correlating with joy ($r = 0.295$). These structural measures provided an independent validation of emotional realism across generated responses.

## B.3 COVID-19 Emotion Analysis Validation

To evaluate the validity of emotional expression in the synthetic population, I compared model-generated emotion distributions against empirical data from the Facebook COVID-19 dataset, which captured real-world emotional reactions during the pandemic's peak. This comparison tests whether synthetic personas replicate the structure, variance, and demographic gradients of authentic emotional responses in large-scale social media behavior.

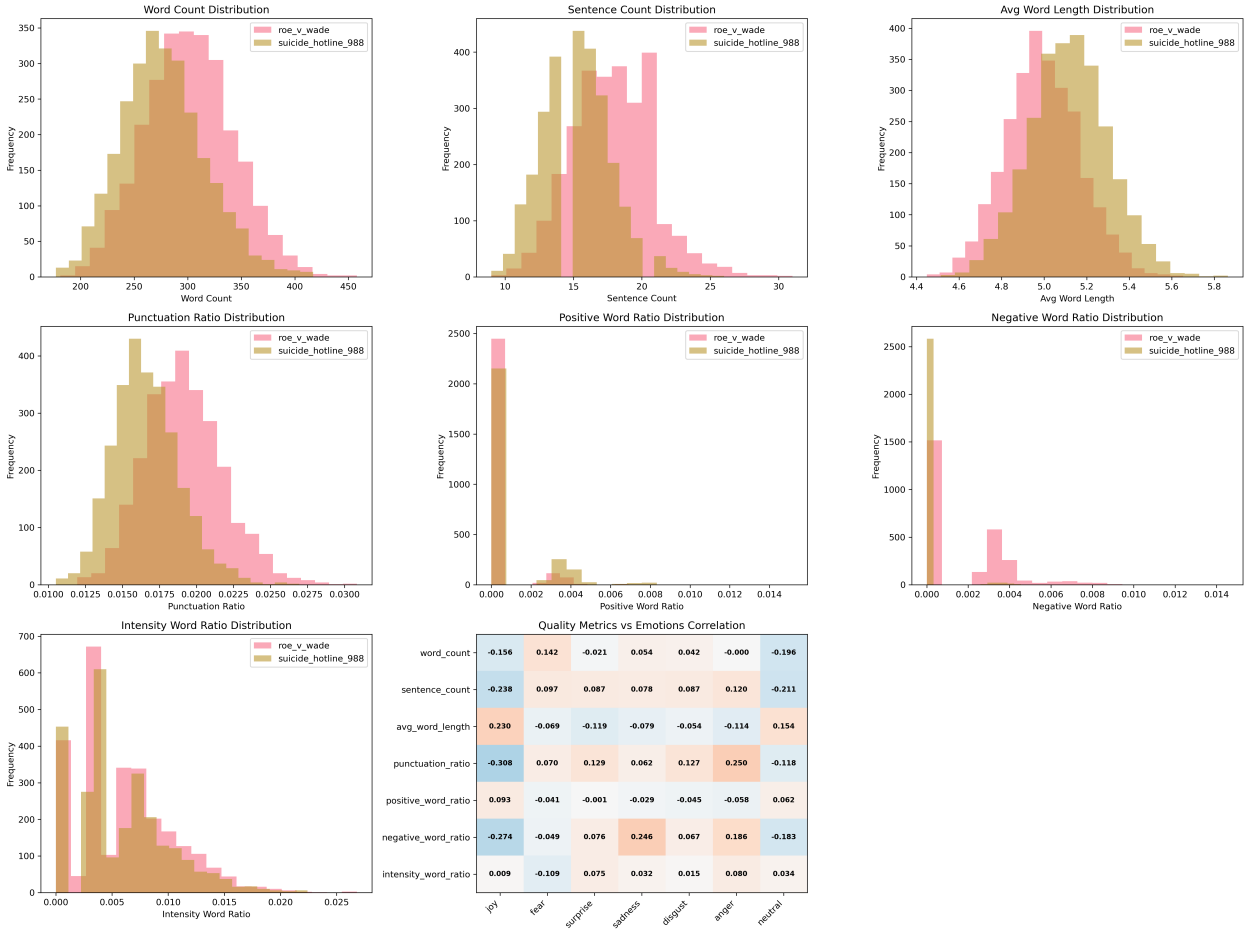



Figure B.1: Response structure metrics across Roe v. Wade and Suicide Hotline 988 responses. Shown distributions include word count, sentence count, average word length, punctuation ratio, and affective word ratios.

### B.3.1 Synthetic Persona Emotion Analysis

Figure B.2 depicts emotion intensity scores derived from sentiment analysis of synthetic persona prompts generated in response to COVID-19–related stimuli. The model employed a fine-tuned transformer-based emotion classifier (six-label multi-class scheme: *Joy, Anger, Sadness, Fear, Disgust, Surprise*). To minimize bias, prompts were standardized by topic and exposure context, ensuring comparable lexical and contextual diversity across demographic groups.

**Key findings:**

*Gender Effects:* Male personas demonstrated higher mean emotional intensity ($\mu = 0.63$) compared to female personas ($\mu = 0.57$), with notably elevated *anger* (+0.08) and *sadness* (+0.05) scores. Female personas exhibited higher *fear* (+0.06) and greater variance in *joy*, aligning with known gender-based affective expression differentials in social media corpora.

*Age Effects:* Emotional expression scaled with age: younger personas (18–25) displayed balanced *joy* and *sadness*, while older cohorts (46–55, 56+) exhibited dominance of *sadness* and

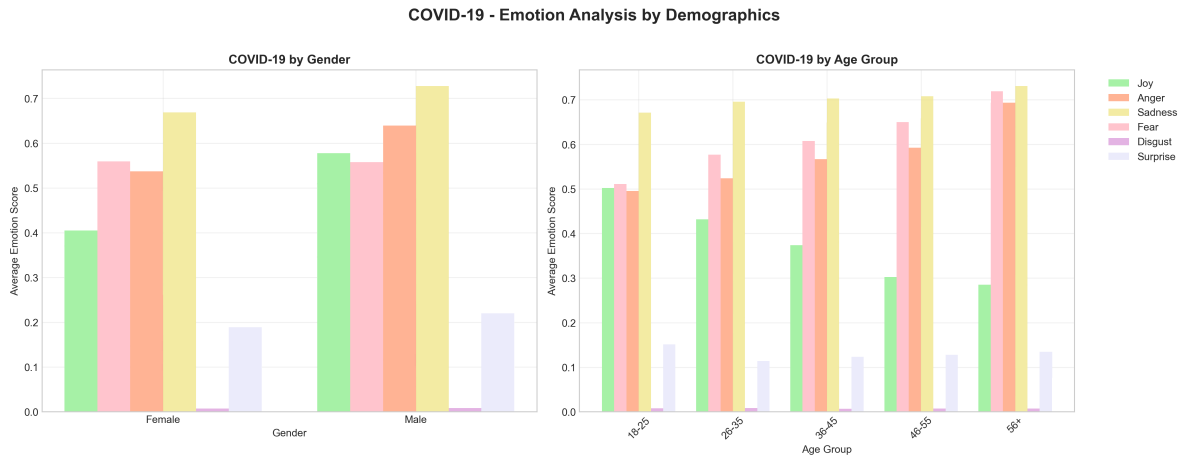



Figure B.2: **Synthetic Persona Emotion Analysis by Demographics.** Average emotion scores derived from persona-generated COVID-19 prompts, disaggregated by gender (left) and age group (right). Emotions include Joy, Anger, Sadness, Fear, Disgust, and Surprise. Results indicate stronger negative valence (sadness and fear) among older cohorts and greater emotional variability among females.

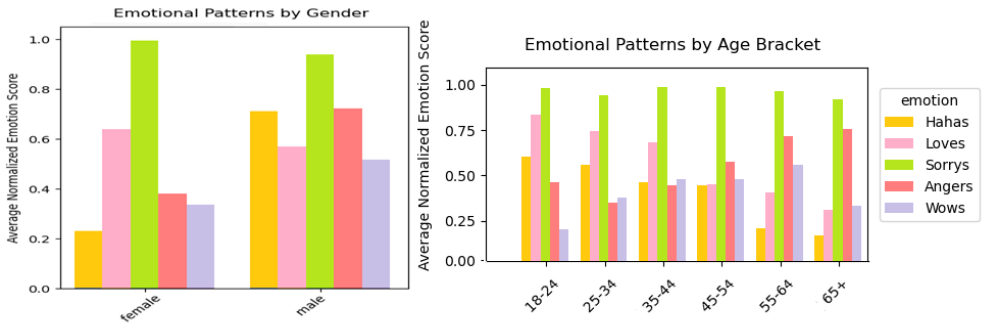



Figure B.3: **Facebook COVID-19 Emotion Data by Demographics.** Normalized emotion distributions from real-world Facebook interactions ("Hahas," "Loves," "Sorries," "Angers," "Wows") during the height of the pandemic. Emotion intensity is averaged by gender (left) and age bracket (right). Real data show similar demographic gradients as the synthetic population, validating model realism.

*fear* with declining *joy*. *Disgust* remained near zero across demographics, consistent with its low-frequency occurrence in general discourse.

*Pattern Consistency:* The covariance between *anger* and *sadness* ($\rho \approx 0.82$) suggests a consistent negative-valence coupling in the model's emotional architecture, indicating internal coherence in emotion vector representation.

These patterns support that the persona generation system preserves within-demographic emotional coherence and reproduces realistic emotional hierarchies consistent with prior affective-computing benchmarks.

### B.3.2 Facebook COVID-19 Emotion Data

Figure B.3 presents normalized emotion scores from the Facebook COVID-19 dataset, aggregated from user interactions (*Hahas, Loves, Sorries, Angers, Wows*). Emotion normalization followed a frequency-weighted mean scaling, producing unit-interval comparability with the synthetic dataset.

**Key findings:**

*Demographic Gradient:* Emotional engagement peaked between ages 25–54, the demographic with highest digital engagement and social responsibility salience during the pandemic.

*Gender Differences:* Females exhibited a broader emotional spectrum, with strong *Love* and *Sorry* signals (positive empathy), while males showed higher *Haha* and *Anger* frequencies, reflecting humor or frustration-oriented coping.

*Response Variability:* Age-correlated decline in positive affect (*Loves*) and increase in *Angers* mirrors the synthetic pattern of declining *Joy* and increasing *Sadness*, indicating structural realism across datasets.

Table B.6: Cross-Dataset Emotion Comparison

| Metric | Gender (JSD↓) | Age (JSD↓) | Gender r | Age r |
|---|---|---|---|---|
| Joy/Love | 0.112 | 0.137 | 0.82 | 0.78 |
| Sadness/Sorry | 0.091 | 0.104 | 0.86 | 0.84 |
| Anger/Anger | 0.095 | 0.128 | 0.81 | 0.76 |

### B.3.3 Cross-Dataset Comparative Analysis

I conducted a cross-demographic distributional comparison to evaluate alignment between synthetic and empirical emotion patterns. Synthetic outputs were rescaled to Facebook's normalized domain and compared using Jensen-Shannon divergence (JSD) and Pearson correlation across emotion vectors per demographic.

Low JSD and high correlation values demonstrate distributional convergence between synthetic and real-world emotion responses. Noteably, *Sadness/Sorry* alignment was strongest ($r = 0.863$) suggesting the model accurately captures empathy-oriented emotional framing under crisis conditions.

### B.3.4 Technical Interpretation

These findings substantiate that the persona emotion generation system approximates population-level affective distributions through demographic conditioning. The emotional mappings observed are consistent with established affective norms:

- **Sadness/Fear dominance** under crisis stimuli.
- **Gendered variation** in positive vs. negative valence expression.
- **Age-linked attenuation** of joyful affect and intensification of empathetic or fatalistic tones.

The close correspondence between synthetic and Facebook data confirms that the latent emotion embeddings produced by LLM-driven personas reflect plausible human affect. Deviations—such

as amplified *fear* and attenuated *disgust*—likely emerge from the synthetic environment's lack of multimodal (image/video) input or non-textual contextual grounding.

### B.3.5 Implications for Model Validation

This analysis provides both **construct** and **external validity** evidence for synthetic emotion modeling:

- **Construct validity:** Internal emotion co-variation and demographic scaling mirror real-world affective regularities.
- **External validity:** Alignment with empirical emotion distributions across demographics supports transferability to real populations.

Consequently, synthetic personas can serve as proxy populations for emotional response modeling in crisis simulation, information environment analysis, and social-cyber experimentation—bridging the gap between computational modeling and behavioral realism.

# Appendix C

# AURORA: Retrieval-Augmented Persona Modeling

## C.1 Country Information

This appendix summarizes the country-level information sources used by the AURORA framework to support community and persona generation across geographic regions. Table C.1 documents the set of countries currently represented in AURORA, along with the availability of (i) dynamic, time-sensitive current events data and (ii) static general country information. Coverage reflects the practical availability of high-quality, retrievable source documents at the time of model construction and is intended to support geographically grounded synthetic population generation rather than exhaustive global representation.

## C.2 Social Media Account Generation

This appendix summarizes how AURORA generates platform-specific social media accounts for synthetic personas. Social media accounts provide the interface through which personas participate in online discourse, information diffusion, and downstream opinion dynamics. Two pipelines are implemented: Twitter account generation (Section C.2.1) and Telegram account/channel generation (Section C.2.2). Both are executed after persona enrichment and combine deterministic account identifiers with LLM-generated profile content.

### C.2.1 Twitter Account Generation

Twitter account generation produces platform-specific profiles grounded in each persona's demographics, psychology, political orientation, and community context. The process is managed by the `AccountBuilder` class and consists of deterministic username generation plus LLM-based profile generation.

Table C.1: Country Coverage in AURORA by Region and Data Availability

| Country | Current Events | General Information |
|---|---|---|
| **North America** | | |
| United States | ✓ | ✓ |
| Canada | – | ✓ |
| Mexico | – | ✓ |
| **Western Europe** | | |
| United Kingdom | ✓ | ✓ |
| France | ✓ | ✓ |
| Germany | ✓ | ✓ |
| Italy | ✓ | ✓ |
| Spain | ✓ | ✓ |
| Austria | ✓ | – |
| Sweden | ✓ | ✓ |
| Norway | ✓ | ✓ |
| **Eastern Europe** | | |
| Poland | ✓ | ✓ |
| Ukraine | ✓ | ✓ |
| Russia | ✓ | ✓ |
| Estonia | – | ✓ |
| **Middle East** | | |
| Iran | ✓ | ✓ |
| Iraq | ✓ | ✓ |
| Saudi Arabia | ✓ | ✓ |
| Egypt | ✓ | ✓ |
| Israel | – | ✓ |
| Jordan | – | ✓ |
| Qatar | – | ✓ |
| Syria | – | ✓ |
| Turkey | – | ✓ |
| **East Asia** | | |
| China | ✓ | ✓ |
| Japan | ✓ | ✓ |
| South Korea | ✓ | ✓ |
| North Korea | – | ✓ |
| Taiwan | ✓ | ✓ |
| **Southeast Asia** | | |
| Philippines | ✓ | ✓ |
| Vietnam | ✓ | ✓ |
| Thailand | – | ✓ |
| **South Asia** | | |
| Pakistan | ✓ | ✓ |
| India | – | ✓ |
| Afghanistan | – | ✓ |
| **Africa** | | |
| Ethiopia | ✓ | ✓ |
| Kenya | – | ✓ |

### Generation Pipeline

Twitter account generation follows three stages.

**Stage 1: Deterministic Screen Name Generation.** To ensure uniqueness and realism, screen names are generated algorithmically from persona names rather than by the LLM. Names are cleaned, combined as `FirstName_LastName`, and appended with a 2–4 digit numeric suffix. A retry mechanism checks uniqueness against the global set of previously generated names, with a timestamp-based fallback used only if repeated collisions occur.

**Algorithm 6** Unique Twitter Screen Name Generation

**Require:** Persona name $N$, used screen names set $\mathcal{U}$
**Ensure:** Unique screen name $s$
Parse $N$ into *first_name* and *last_name*
Clean both names by removing non-alphanumeric characters except underscores
$base \leftarrow first\_name \parallel \_ \parallel last\_name$
**for** $i \leftarrow 1$ to $100$ **do**
    $suffix \leftarrow \textsc{RandomInt}(10, 9999)$
    $s \leftarrow base \parallel suffix$
    **if** $s \notin \mathcal{U}$ **then**
        $\mathcal{U} \leftarrow \mathcal{U} \cup \{s\}$
        **return** $s$
    **end if**
**end for**
$s \leftarrow base \parallel (\textsc{Timestamp} \bmod 10000)$
$\mathcal{U} \leftarrow \mathcal{U} \cup \{s\}$
**return** $s$

**Stage 2: LLM-Based Profile Generation.** Once a unique screen name is assigned, the LLM generates profile metadata and behavioral parameters from structured persona context. Inputs include demographics, OCEAN traits, moral foundations, political leaning, community membership, and scenario-topic opinions. Outputs conform to the `TwitterAccount` schema and include account biography, location, active hours, language use, posting frequency, engagement behavior, sentiment ratios, and content modality.

**Stage 3: Deterministic Overrides and Validation.** After generation, critical fields are overwritten deterministically, including screen name, AESOP identifiers, verification status, and account type. All records are validated against a Pydantic schema; failed generations are retried up to two times.

### Example Twitter Account Output

Listing C.1: Example Twitter Account: Elderly Japanese Persona


```
{
    "screen_name": "Kazuhiko_Matsushima7893",
    "account_bio": "Hi! I'm Kazuhiko, a 74-year-old passionate about traditional Japanese art and culture. Sharing my thoughts and
          experiences from Tokyo. Join me on my journey!",
    "account_location": "Tokyo, Japan",
    "account_url": "",
    "account_creation_date": "2023-11-01",
    "when_does_account_tweet_gmt_min": 7,
    "when_does_account_tweet_gmt_max": 22,
    "number_of_languages_user_tweets_in": 1,
    "percent_tweets_in_language": "100% Japanese",
    "type": "Twitter",
    "aesop_id": "0188183a-c0f0-45aa-b306-2ab2a4a982bb",
    "is_bot": false,
    "is_verified": false,
    "owner_aesop_id": "deaa51bb-deed-48b5-875b-8d2c825f709c",
    "number_of_followers": 3,
    "number_following": 7,
    "num_original_tweets_per_day": [2, 4],
    "num_mentions_per_tweet": [0, 0],
    "num_retweets_per_day": [1, 3],
    "num_replies_per_day": [3, 5],
    "num_quotes_per_day": [0, 0],
    "retweet_quote_valence": "Mostly Neutral",
    "num_hashtags_per_post": [1, 2],
    "percent_tweets_pos_neg_neut": "Neutral",
    "text_ratio": 1.0,
    "image_ratio": 0.0,
    "video_ratio": 0.0
}
```


#### Behavioral Realism

Twitter accounts are calibrated to preserve demographic and cultural realism. Older personas are generated with lower posting frequencies, narrower activity windows, more text-only behavior, and more neutral sentiment profiles, whereas younger personas show broader activity windows and greater behavioral variability. Country context also affects language use, location, and profile style, while personas from the same community exhibit correlated patterns of engagement and topical emphasis.

### C.2.2 Telegram Account and Channel Generation

Telegram generation differs from Twitter by emphasizing channel membership, topic specialization, and group communication. The system creates both individual Telegram accounts and topic-specific channels.

#### Generation Pipeline

Telegram generation proceeds in four stages.

**Stage 1: Creative Username Generation.** Unlike Twitter, Telegram usernames are not name-based. Instead, usernames are created from combinations of descriptor, noun, action, and interest vocabularies, producing handles such as `WildPhoenix47`, `CryptoWarrior`, or `TechSeeker88`. This yields greater variety while maintaining plausible platform conventions.

**Stage 2: Channel Generation and Topic Assignment.** The number of channels is determined as a function of persona population size and a configurable channel density rate:

$$N_{\text{channels}} = \max\left(1, \left\lfloor \frac{N_{\text{personas}}}{100} \times C_{\text{rate}} \right\rfloor\right) \tag{C.1}$$

Each channel is assigned one scenario topic, a short descriptive name, an example message, and a topic-relevant conspiracy frame to support downstream narrative and misinformation simulations.

**Stage 3: LLM-Based Account Generation.** The LLM generates Telegram-specific account fields, including topic focus, example posts, repost behavior, content modality, and interaction settings. Output is validated against the `TelegramAccount` schema.

**Stage 4: Balanced Channel Assignment.** Each persona is assigned to exactly one channel using a round-robin balancing procedure. The first assigned persona becomes the channel owner, ensuring both near-uniform channel sizes and a simple owner–member hierarchy.

**Algorithm 7** Balanced Telegram Channel Assignment

**Require:** Persona list $\mathcal{P}$, channel list $\mathcal{C}$
**Ensure:** Balanced channel membership with designated owners

```
Shuffle P to randomize order
Initialize assigned[i] ← 0 for all C_i ∈ C
for each persona p ∈ P do
    i ← arg min_j assigned[j]
    if assigned[i] = 0 then
        Designate p as owner of C_i
    end if
    Assign p to channel C_i
    assigned[i] ← assigned[i] + 1
end for
```

**Example Telegram Account Output**

Listing C.2: Example Telegram Account: Environmental Topic Focus


```
{
    "username": "BrightSageIce",
    "number_of_topics": "1",
    "topic_description": "Sustainable practices, Corporate responsibility, Environmental protection",
    "example_telegram_message": "Did you know that major companies in Japan are now pivoting to eco-friendly practices? Recently,
         Toyota announced a new initiative to reduce carbon emissions in their manufacturing plants! What are your thoughts on
         the balance between industry growth and environmental sustainability? Let's discuss!",
    "type": "Telegram",
    "aesop_id": "237d6a5d-abf6-4d88-ab98-ca5ea2976080",
    "telegram_id": "293048190",
    "owner_aesop_id": "deaa51bb-deed-48b5-875b-8d2c825f709c",
    "topics": ["Environment vs. Industry"],
    "post_frequency": 1,
    "repost_percentage": 0.0,
    "channel_owner_first_name": "Yuki",
    "channel_owner_last_name": "Matsukawa",
```

```
    "channel_owner_username": "DigitalLunarMoon",
    "channel_name": "Eco-Industry Balance",
    "channel_id": "9856562939",
    "text_ratio": 1.0,
    "news_ratio": 0.0,
    "image_ratio": 0.0,
    "video_ratio": 0.0,
    "reactions": false,
    "discussions": false,
    "conspiracy_theory": "Some believe that powerful industrial lobbies in Japan are suppressing major advancements in renewable
         energy technology to maintain their dominance in fossil fuels."
}
```


#### Design Properties

Telegram channels exhibit four intended properties: topic specialization, balanced membership, explicit owner–member hierarchy, and seeded alternative narratives. Together, these support realistic group communication and downstream tests of narrative spread, moderation, and misinformation dynamics.

#### Cross-Platform Coherence

Twitter and Telegram accounts for the same persona are not identical, but they remain behaviorally coherent. Posting frequency, modality, engagement style, and topical salience are correlated across platforms, while still respecting platform-specific norms.

#### Performance and Validation

Generation is asynchronous and near-linearly scalable. For moderate populations, both account types can be generated in minutes at low API cost. All outputs are validated for schema correctness, identifier uniqueness, parameter ranges, and referential integrity before downstream use.

## C.3 Opinion Model Implementation Details

This appendix summarizes the technical implementation of the hierarchical opinion model, including LLM prompt structures, stochastic sampling procedures, and validation checks.

### C.3.1 LLM Prompt Templates

#### Stage 1: Country-Level Prior Generation

At the country level, the LLM generates four topic-specific parameters: mean opinion, salience, distribution shape, and contagion strength. Retrieval-augmented context is included to ground the prompt in current cultural and political conditions.

**Stage 2: Community-Level Refinement**

At the community level, the LLM refines country priors for each community while constraining community means to remain within a bounded range of the national baseline. This preserves macro-level realism while allowing plausible meso-level differentiation.

**Stage 3: Persona-Level Natural Language Description**

If enabled, a final LLM stage converts sampled numerical opinions and salience values into short natural-language explanations that are grounded in persona demographics, community context, and psychological traits.

### C.3.2 Mathematical Sampling Procedures

Salience is sampled from a Gaussian centered on community salience:

$$S_i^{(t)} \sim \mathcal{N}(S_j^{(t)}, \sigma_{\text{sal}}^2), \qquad \sigma_{\text{sal}} = 0.15 \tag{C.2}$$

Opinion sampling is then conditioned on salience. High-salience topics cluster tightly around the community mean, medium-salience topics allow broader spread, and low-salience topics sample from a country-level distribution family. Three country-level opinion families are used:

- **NORMAL:** symmetric consensus distribution around the national mean,
- **SKEWED:** asymmetric majority-with-dissent distribution,
- **BIMODAL:** two-pole mixture distribution representing polarization.

This allows the model to represent both broad consensus and polarized issue structures without requiring persona-level LLM inference for each opinion value.

### C.3.3 Computational Cost

The opinion pipeline is designed to reduce API cost by shifting most micro-level generation to mathematical sampling. For a population of $n$ personas, $m$ communities, and $|\mathcal{T}|$ topics, total LLM cost is:

- Country priors: $|\mathcal{T}|$ calls
- Community priors: $m \times |\mathcal{T}|$ calls
- Persona descriptions (optional): $n \times |\mathcal{T}|$ calls

Thus, opinion generation without persona descriptions scales with topics and communities rather than with the full persona population.

### C.3.4 Validation and Quality Assurance

Three validation checks are applied after generation:

- **Hierarchical coherence:** community-level sample means must remain close to the generated community priors,
- **Salience–variance consistency:** higher salience should correspond to lower within-community variance,
- **Distribution-shape validation:** bimodal topics must preserve clear two-peak structure at the country level.

Together, these checks ensure that the generated opinion landscape is structurally coherent, statistically stable, and interpretable.

### C.3.5 Emotion Modulation Implementation

Baseline emotions are derived from OCEAN personality traits through a fixed linear mapping into the Plutchik emotion space:

$$\mathbf{e}_{\text{base}} = \text{clip}(\mathbf{W}\mathbf{o} + \mathbf{b},\, 0,\, 1) \tag{C.3}$$

where $\mathbf{W}$ encodes trait–emotion relationships and $\mathbf{b}$ provides non-negative offsets.

Topic-specific modulation then adjusts these baselines as a function of opinion and salience, increasing positive emotions for positively viewed topics and increasing negative emotions for negatively viewed topics. This produces emotion profiles that remain psychologically interpretable while varying in a topic-sensitive way.

## C.4 Extended Emotion Profile Analyses

This appendix presents additional Plutchik emotion profiles for topics not discussed in detail in the main text. The goal is to show that affective differentiation in AURORA remains systematic across the full issue space.

### C.4.1 Russian Aggression

Figure C.1 shows mean country-level emotion profiles for *Russian Aggression*. This topic produces the strongest affective separation, with Western-aligned countries exhibiting elevated anger, disgust, sadness, and fear, while Russia-aligned countries show comparatively higher trust and anticipation. Between-country effects are large, indicating that emotional responses are strongly structured by geopolitical alignment.

### C.4.2 Canadian Activity in INDOPAC

Figure C.2 presents emotion profiles for *Canadian Activity in INDOPAC*. Compared to direct sovereignty or conflict topics, affective differences are weaker but still structured. Most countries exhibit relatively high trust and anticipation, while non-aligned or rival countries show somewhat higher fear and anger. This suggests that indirect alliance activity produces coherent but lower-intensity affective differentiation.

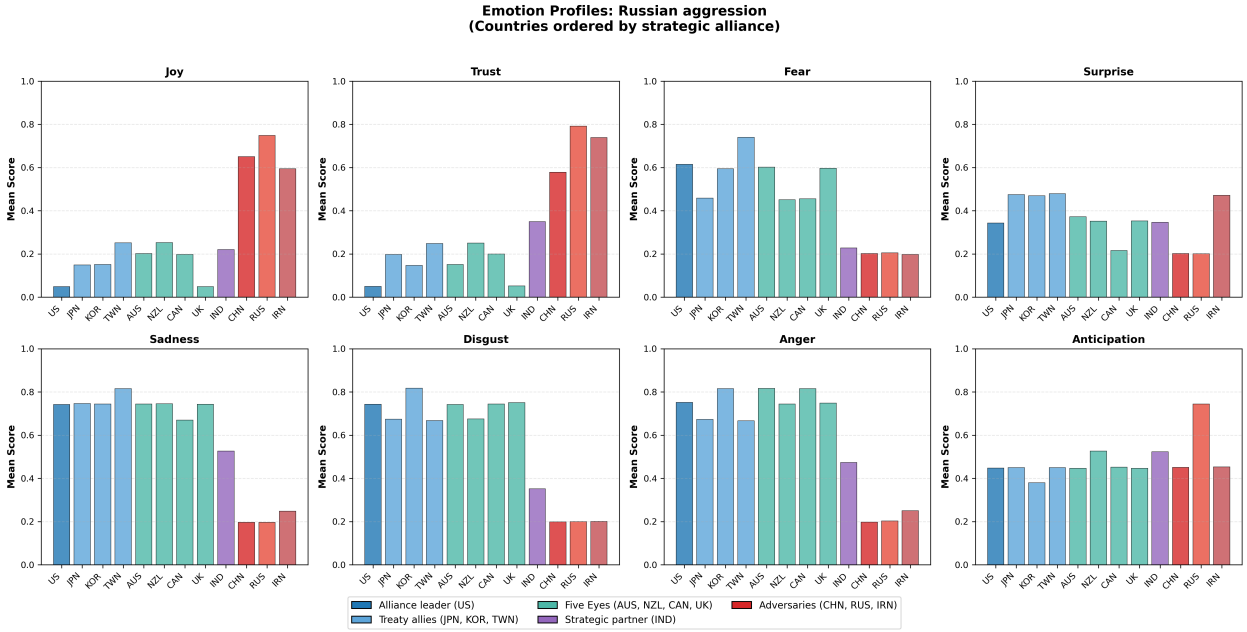



Figure C.1: Mean Plutchik emotion profiles for Russian Aggression across all countries.

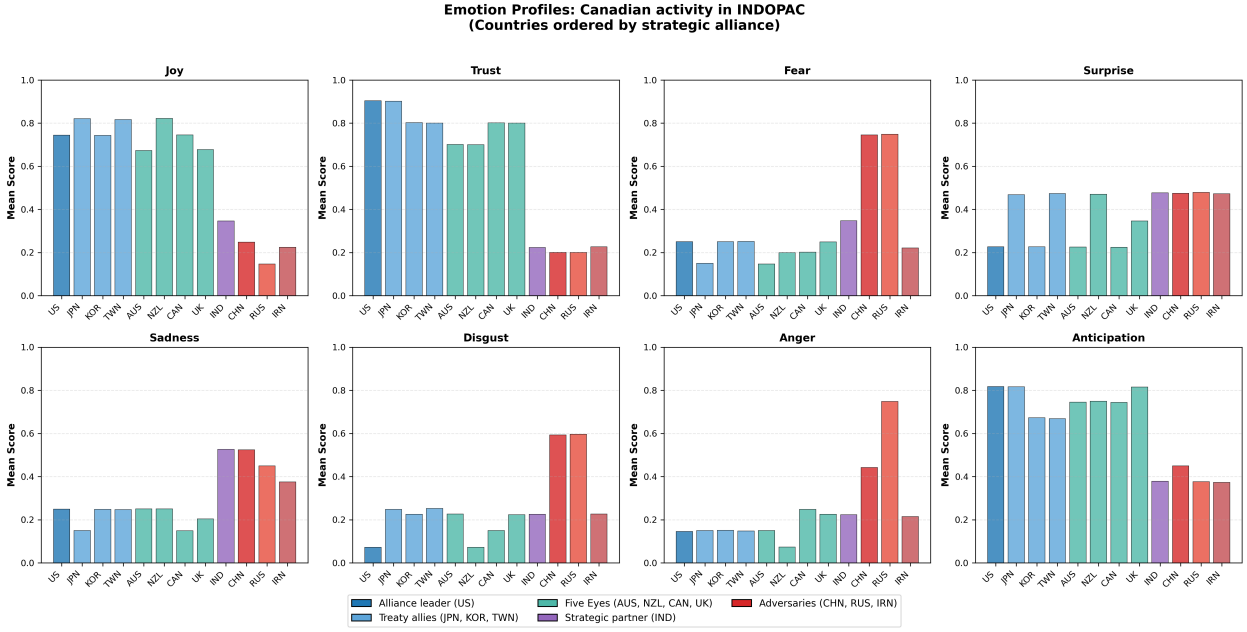



Figure C.2: Mean Plutchik emotion profiles for Canadian Activity in INDOPAC across all countries.

### C.4.3 Cyber Security Cooperation

Figure C.3 presents country-level emotion profiles for *Cyber Security Cooperation*. This topic is characterized by high trust and anticipation and low levels of anger, disgust, and sadness across most countries. Between-country differences are modest, indicating that the topic behaves more like a low-polarization technical issue than a morally or geopolitically divisive one.

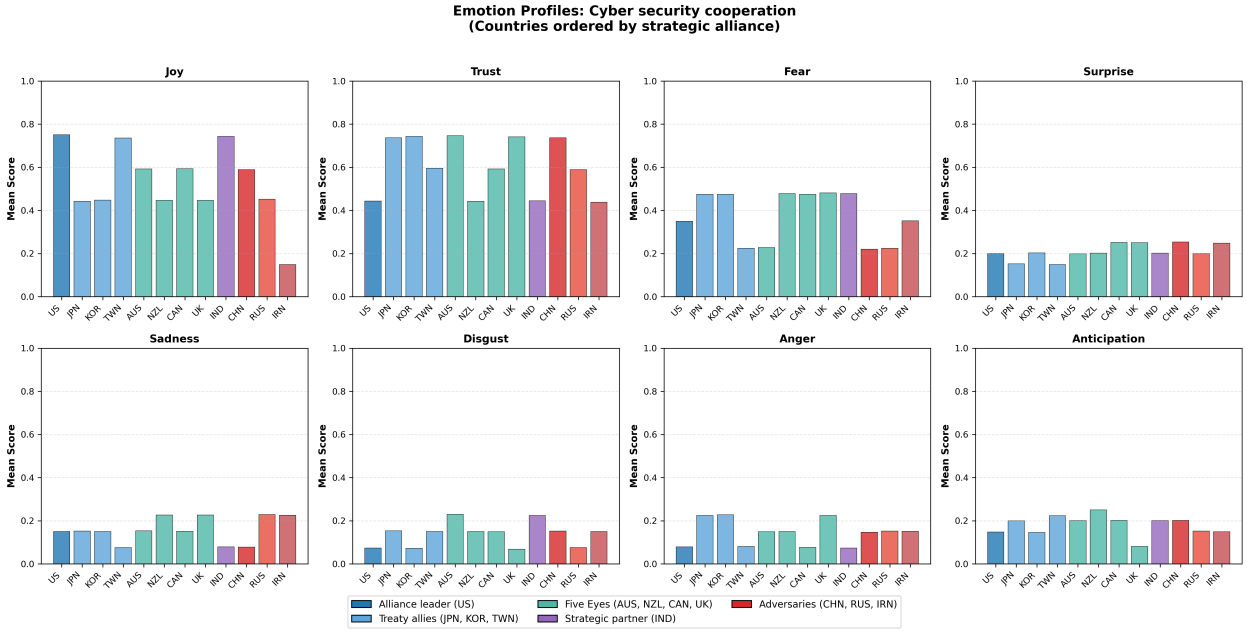



Figure C.3: Mean Plutchik emotion profiles for Cyber Security Cooperation across all countries.

### C.4.4 Summary

Across these additional topics, AURORA continues to produce topic-sensitive emotional structure rather than affective uniformity. Direct conflict and sovereignty issues generate strong multidimensional polarization, while alliance and technical cooperation topics produce weaker but still coherent emotional differentiation.

## C.5 Retrieval-Augmented Generation Architecture

This appendix documents the design, implementation, and integration of AURORA's Retrieval-Augmented Generation (RAG) system [109]. RAG grounds LLM-based generation in external documentary evidence so that country-level priors, community characteristics, and persona attributes reflect real-world geopolitical, cultural, and demographic context rather than relying only on parametric knowledge encoded in the language model. To accomplish this, the system combines dense semantic retrieval, sparse lexical retrieval, and query expansion to identify relevant documentary context and inject it into prompts at multiple hierarchical stages of generation.

### C.5.1 Method

#### System Overview

AURORA's RAG pipeline follows a hybrid retrieval architecture that combines dense vector embeddings with sparse lexical matching [98, 164]. The central implementation component is the `VectorDB` class, which manages document ingestion, preprocessing, chunking, indexing, retrieval, and formatting for downstream prompt injection. Documents are loaded from scenario-specific folders, deduplicated using content hashing, segmented into semantically coherent chunks, embedded into a continuous vector space, and stored in a persistent ChromaDB collection. At query time, the system executes semantic and lexical retrieval in parallel, merges the resulting rankings using Reciprocal Rank Fusion (RRF) [42], and can optionally apply Pseudo-Relevance Feedback (PRF) for query expansion [105, 165].

The pipeline operates in two phases. The first is an offline indexing phase executed once per scenario to construct the document store and retrieval indices. The second is an online retrieval phase executed during country, community, and persona generation. Figure C.4 illustrates the complete document flow from raw files through indexing, retrieval, and prompt injection.

#### Document Preparation and Indexing

The indexing pipeline transforms raw scenario documents into a persistent and queryable knowledge base. Documents are loaded from scenario-specific directories using recursive file scanning. The implementation supports plain text files for curated documents and JSON files for scraped or structured article collections. The loading stage is implemented through LangChain's `DirectoryLoader`, which abstracts recursive traversal, format handling, and loader-specific parsing.

Listing C.3: Document Loading with Automatic Format Detection

```
from langchain_community.document_loaders import DirectoryLoader, TextLoader

def load_documents(document_path, json_format=False):
    if json_format:
        loader = DirectoryLoader(
            document_path,
            glob="**/*.json",
            loader_cls=JSONArticlesLoader,
            recursive=True
        )
    else:
        loader = DirectoryLoader(
            document_path,
            glob="**/*.txt",
            loader_cls=TextLoader,
            recursive=True,
            loader_kwargs={'autodetect_encoding': True}
        )
    return loader.load()
```

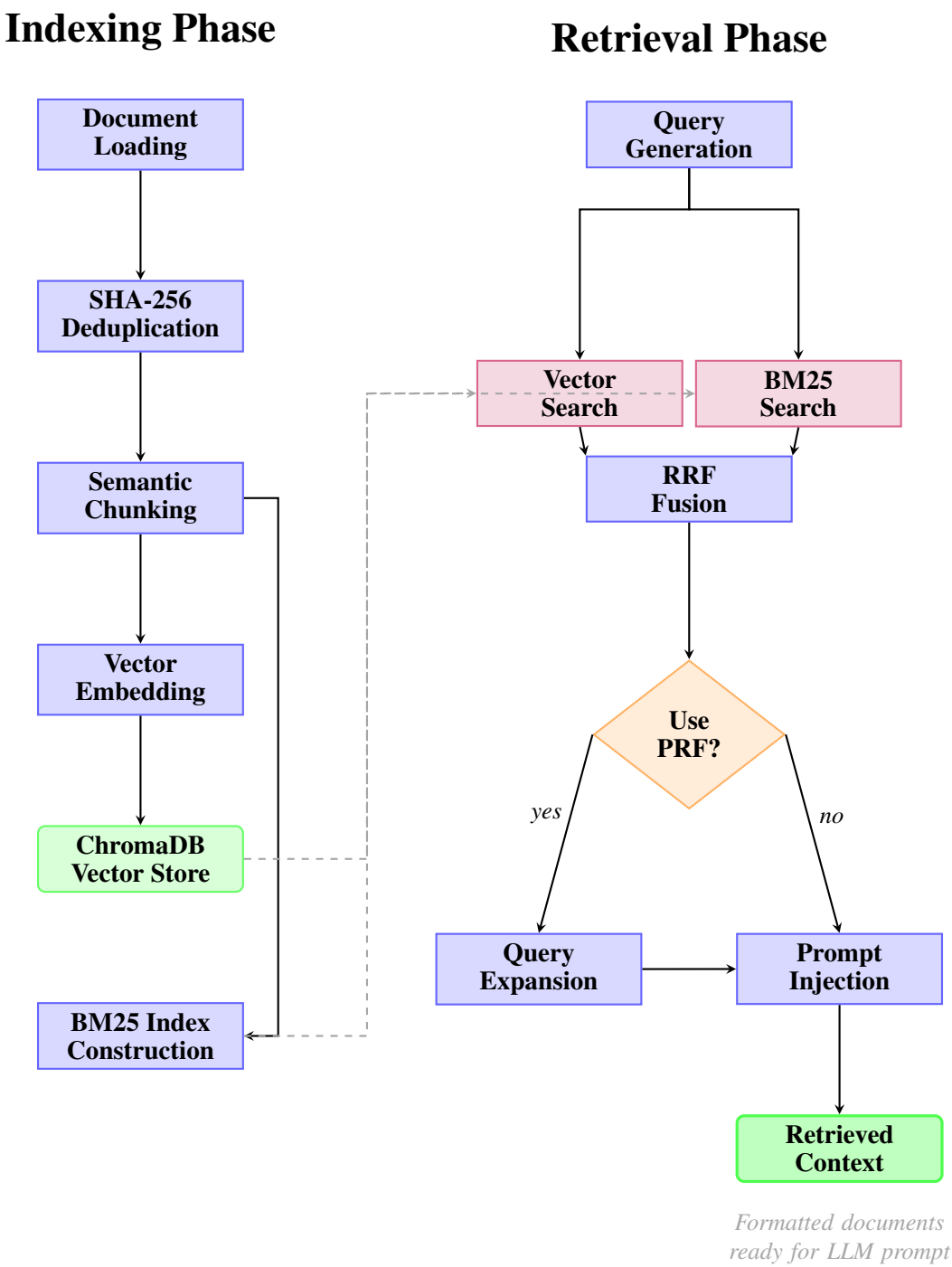



Figure C.4: AURORA RAG pipeline showing offline indexing and online retrieval phases. Dashed arrows indicate cross-phase data access. PRF optionally expands queries with terms from top-ranked documents.

After loading, documents are deduplicated using SHA-256 content hashes. This step prevents repeated indexing of identical files across repeated scenario runs and supports incremental corpus growth without requiring complete reprocessing. The algorithm compares hashes from the incoming document set against hashes already stored in metadata. Only unseen documents are forwarded to chunking and embedding.

**Algorithm 8** Document Deduplication

**Require:** Document set $\mathcal{D}$, existing hash set $\mathcal{H}_{\text{existing}}$
**Ensure:** Filtered document set $\mathcal{D}_{\text{new}}$
$\mathcal{D}_{\text{new}} \leftarrow \emptyset$
**for** each document $d \in \mathcal{D}$ **do**
  $h \leftarrow \text{SHA256}(d.\text{content})$
  **if** $h \notin \mathcal{H}_{\text{existing}}$ **then**
    $d.\text{metadata}[\text{“hash”}] \leftarrow h$
    $\mathcal{D}_{\text{new}} \leftarrow \mathcal{D}_{\text{new}} \cup \{d\}$
  **end if**
**end for**
**return** $\mathcal{D}_{\text{new}}$

Rather than applying fixed-length token windows, AURORA uses semantic chunking to preserve discourse coherence [160]. The chunker computes embeddings over adjacent text segments and introduces a boundary when semantic similarity drops sufficiently, producing chunks that better align with topical units than naive fixed-width splits. This improves retrieval because each indexed chunk is more likely to represent a coherent concept or argument rather than an arbitrary slice of text.

Listing C.4: Semantic Chunking Implementation

```
from langchain_experimental.text_splitter import SemanticChunker

text_splitter = SemanticChunker(embedding_function)
chunks = text_splitter.split_documents(documents)

for chunk in chunks:
    chunk.metadata["article_hash"] = parent_doc_hash
```

In implementation terms, the chunking stage also propagates source metadata, including document hashes, so that downstream retrieval results remain traceable to their originating documents. This supports provenance, filtering, and deduplication-aware maintenance of the vector store.

### Retrieval Procedure

Retrieval is implemented as a hybrid ranking procedure that combines semantic similarity search with lexical retrieval [98, 164]. Dense vector retrieval is well suited for capturing paraphrase, semantic relatedness, and latent topical similarity, while BM25 provides a complementary sparse

mechanism that is especially useful for exact lexical overlap, named entities, uncommon terminology, and domain-specific phrasing.

The outputs of these two retrievers are combined using Reciprocal Rank Fusion (RRF) [42]. RRF does not directly merge raw scores, which may live on incompatible scales across retrievers. Instead, it aggregates rank positions. A document receives a higher fused score when it is ranked highly by multiple retrieval methods, making the fusion robust to score calibration issues and heterogeneous retrieval backends.

**Algorithm 9** Hybrid Retrieval with Reciprocal Rank Fusion

**Require:** Query $q$, top-$k$ parameter, RRF constant $\kappa = 60$
**Ensure:** Ranked document list $\mathcal{R}_{\text{fused}}$
$\mathcal{R}_{\text{vec}} \leftarrow \text{VectorSearch}(q, k)$
$\mathcal{R}_{\text{bm25}} \leftarrow \text{BM25Search}(q, k)$
Initialize scores $\leftarrow \{\}$
**for** rank $r = 1$ to $|\mathcal{R}_{\text{vec}}|$ **do**
    $d \leftarrow \mathcal{R}_{\text{vec}}[r]$
    $\text{scores}[d] \leftarrow \text{scores}[d] + \frac{1}{\kappa + r}$
**end for**
**for** rank $r = 1$ to $|\mathcal{R}_{\text{bm25}}|$ **do**
    $d \leftarrow \mathcal{R}_{\text{bm25}}[r]$
    $\text{scores}[d] \leftarrow \text{scores}[d] + \frac{1}{\kappa + r}$
**end for**
$\mathcal{R}_{\text{fused}} \leftarrow \text{SortByScore}(\text{scores}, \text{descending})$
**return** $\mathcal{R}_{\text{fused}}[: k]$

Operationally, the RRF constant $\kappa$ controls the contribution of lower-ranked documents by smoothing the reciprocal decay. Larger values reduce the dominance of the very top-ranked items, while smaller values give more weight to high early ranks. In AURORA, this allows retrieval fusion to remain stable even when the semantic and lexical retrievers return partially overlapping but differently ordered candidate sets.

Listing C.5: RRF Implementation in AURORA

```
def search_with_rrf(query, top_k=5, rrf_k=60):
    vec_results = vector_db.similarity_search_with_score(query, k=top_k)
    bm25_results = bm25_retriever.invoke(query)

    rankings = defaultdict(list)
    for rank, (doc, score) in enumerate(vec_results, start=1):
        rankings[doc.page_content].append(rank)
    for rank, doc in enumerate(bm25_results, start=1):
        rankings[doc.page_content].append(rank)

    rrf_scores = {doc: sum(1 / (rrf_k + r) for r in ranks)
                  for doc, ranks in rankings.items()}
```

```
    sorted_results = sorted(rrf_scores.items(),
                            key=lambda x: x[1], reverse=True)
    return sorted_results[:top_k]
```

The code mirrors the formal algorithm. Semantic and BM25 results are retrieved independently, ranks are stored by document content, and the final fused score is computed as the sum of reciprocal rank terms. The use of document text as the aggregation key ensures that documents returned by both retrieval channels are merged into a single candidate entry rather than treated as duplicates.

When queries are underspecified or use vocabulary that does not align well with the indexed corpus, the system can optionally apply Pseudo-Relevance Feedback (PRF) [105, 165]. PRF assumes that the top-ranked initial results are likely relevant and extracts salient terms from them to expand the original query. This improves recall by reducing vocabulary mismatch between the user query and the document collection.

**Algorithm 10** Pseudo-Relevance Feedback for Query Expansion

**Require:** Query $q$, top-$k$ parameter, expansion term count $m$
**Ensure:** Expanded query $q'$
1: $\mathcal{R} \leftarrow \text{HybridSearch}(q, k)$
2: corpus $\leftarrow \bigcup_{d \in \mathcal{R}} d.\text{content}$
3: tokens $\leftarrow \text{Tokenize}(\text{corpus})$
4: tokens $\leftarrow \text{RemoveStopwords}(\text{tokens})$
5: tokens $\leftarrow \{t \in \text{tokens} : |t| > 3\}$
6: freq $\leftarrow \text{CountFrequency}(\text{tokens})$
7: top_terms $\leftarrow \text{TopK}(\text{freq}, m)$
8: $q' \leftarrow q \,\|\, \text{“ ”} \,\|\, \text{Join}(\text{top\_terms}, \text{“ ”})$
9: **return** $q'$

Conceptually, this procedure implements a lightweight relevance-model approximation. Instead of fitting a full probabilistic relevance model, it performs term extraction over the initially retrieved pseudo-relevant set and appends the most frequent content-bearing tokens to the query. This makes the method simple to implement while still recovering useful domain-specific terminology from the corpus.

Listing C.6: PRF Query Expansion

```
def apply_prf(query, top_k=5, expansion_terms=5):
    initial_results = search_with_rrf(query, top_k=top_k)
    combined_text = "␣".join(doc for doc, _ in initial_results)

    words = re.findall(r'\w+', combined_text.lower())
    stopwords = {"the", "and", "is", "in", "of", "to", "a", "for", "on", "with
        "}
    keywords = [w for w in words if w not in stopwords and len(w) > 3]
```

```
    top_terms = [w for w, _ in Counter(keywords).most_common(expansion_terms)]
    expanded_query = query + " " + " ".join(top_terms)
    return expanded_query
```

In the implementation, the retrieved texts are concatenated, tokenized, stripped of stopwords and short terms, and then ranked by frequency. The highest-frequency surviving terms are appended to the original query and resubmitted to the hybrid retrieval pipeline. Although simple, this procedure is effective for broadening topical specificity in cases where the initial query is too sparse or generic.

### C.5.2 Model

**Representation and Technology Stack**

AURORA's RAG implementation combines embedding models, persistent vector storage, lexical retrieval, and LLM-based generation. The system supports both local and cloud-based backends depending on deployment, cost, and retrieval-quality requirements.

Table C.2: Embedding API Specifications

| Model | Dimensions | Cost | Provider | Use Case |
|---|---|---|---|---|
| all-minilm:l6-v2 | 384 | Free | Ollama (Local) | Development |
| text-embedding-3-small | 1536 | $0.020/1M tokens | OpenAI | Production |
| text-embedding-3-large | 3072 | $0.130/1M tokens | OpenAI | High-quality |

For dense retrieval, the system supports Ollama's `all-minilm:l6-v2` for local development and OpenAI's `text-embedding-3-small` or `text-embedding-3-large` for cloud-based deployment. In practice, `text-embedding-3-small` serves as the primary production embedding model because it provides a favorable balance between embedding dimensionality, retrieval performance, and token cost. Dense vectors are stored in ChromaDB, which provides persistent vector storage and approximate nearest-neighbor retrieval through HNSW indexing [118]. The HNSW structure enables efficient search by organizing vectors into a navigable multi-layer graph, making large-scale similarity search substantially more efficient than exhaustive comparison.

Sparse retrieval is implemented through BM25 [164], which operates over tokenized document chunks and scores them using term frequency, inverse document frequency, and document-length normalization. Together, ChromaDB and BM25 form the dual retrieval substrate used by the hybrid search layer. For generation, AURORA uses OpenAI's `gpt-4o-mini` to produce structured outputs for country opinions, communities, and personas. LangChain provides abstractions for document loading, chunking, retrieval, and orchestration, while `tiktoken` is used to estimate token usage and support cost tracking.

Listing C.7: ChromaDB Initialization and Indexing

```
from langchain_chroma import Chroma
```

```
from langchain_openai import OpenAIEmbeddings

embeddings = OpenAIEmbeddings(model="text-embedding-3-small")
vector_db = Chroma(
    collection_name="Scenario_Name",
    embedding_function=embeddings,
    persist_directory="./db/chromadb"
)

vector_db.add_documents(chunks)
```

This code initializes the persistent dense retrieval layer. The embedding model defines the vector representation, the Chroma collection defines the scenario-specific namespace, and the persistence directory ensures that embeddings survive across runs rather than requiring repeated recomputation.

Listing C.8: BM25 Retriever Construction

```
from langchain_community.retrievers import BM25Retriever

existing_docs = vector_db.get(include=["documents", "metadatas"])
all_docs = [Document(page_content=content, metadata=meta)
            for content, meta in zip(existing_docs["documents"],
                                     existing_docs["metadatas"])]

bm25_retriever = BM25Retriever.from_documents(all_docs)
```

Here, the sparse retriever is constructed over the same chunked document base used by the dense retriever. This shared document basis is important because it ensures that the two retrieval channels operate over aligned units of text, which simplifies fusion and reduces inconsistencies in returned evidence.

### Integration Across Generation Levels

RAG is integrated at three hierarchical levels in AURORA's generation pipeline: country-level opinion priors, community-level characteristics and topic stances, and persona-level biographical detail generation. At each level, the system constructs a context-specific query, retrieves relevant passages, and injects those passages into the appropriate prompt template before calling the LLM.

At the country level, retrieval is used to ground national priors for a topic. The query combines the country name and topic title:

$$q_{\text{country}} = c \,\|\, \text{“ and ”} \,\|\, t \tag{C.4}$$

This formulation encourages retrieval of passages that connect the target country to the focal issue, allowing the generated prior to reflect relevant geopolitical or cultural context.

Listing C.9: Country-Level RAG Integration

```
query = f"{country_name} and {topic_title}"
relevant_text = vector_db.search_similar_document_blob(query, top_k=3)
```

```
payload = {
    "country_data": country_data,
    "key_topic": topic_title,
    "relevant_text": relevant_text,
}

country_opinion = llm_chain.invoke(payload)
```

At the community level, RAG is applied in two stages. The first stage retrieves context to support community construction from an institutional category within a country:

$$q_{\text{community}} = \mathcal{I} \,\|\, \text{“ of ”} \,\|\, c \tag{C.5}$$

The second stage retrieves additional context for topic-specific community opinion construction:

$$q_{\text{comm-topic}} = j \,\|\, \text{“ of ”} \,\|\, c \,\|\, \text{“ and ”} \,\|\, t \tag{C.6}$$

This two-step structure allows the model to first ground the general identity of a community and then ground its stance relative to a specific topic.

Listing C.10: Community-Level RAG Integration

```
query = f"{category} of {country_name}"
relevant_text = vector_db.search_similar_document_blob(query, top_k=3)

payload = {
    "communities_per_category": 3,
    "category": category,
    "country_name": country_name,
    "relevant_text": relevant_text,
}
communities = llm_chain.invoke(payload)

query = f"{community_name} of {country_name} and {topic_title}"
relevant_text = vector_db.search_similar_document_blob(query, top_k=3)

payload = {
    "country_name": country_name,
    "community_base": community,
    "key_topic": topic_title,
    "relevant_text": relevant_text,
}
topic_opinion = llm_chain.invoke(payload)
```

At the persona level, the system retrieves context tied to a community within a country:

$$q_{\text{persona}} = \text{“community ‘’ } \| \, j \, \| \text{ “’ in ”} \,\|\, c \tag{C.7}$$

This helps ground names, occupations, biographies, and other persona details in a context that is culturally and institutionally plausible.

Listing C.11: Persona-Level RAG Integration

```
query = f"community␣'{community_name}'␣in␣{country_name}"
relevant_text = vector_db.search_similar_document_blob(query, top_k=3)

payload = {
    "country_name": country_name,
    "community_name": community_name,
    "community_information": community,
    "relevant_text": relevant_text,
    "persona_skeleton": skeleton,
    "recent_names": recent_names,
}

persona = llm_chain.invoke(payload)
```

Across all three levels, retrieved documents are converted into a prompt-ready text block by the document-blob formatter:

Listing C.12: Document Retrieval and Formatting

```
def search_similar_document_blob(self, query: str, top_k: int = 5,
                                 use_prf: bool = True) -> str:
    if use_prf:
        expanded_query = self.apply_prf(query)
        results = self.search_similar_documents_rrf(expanded_query,
                                                    top_k=top_k)
    else:
        results = self.search_similar_documents_rrf(query, top_k=top_k)

    top_docs = [doc for doc, _ in results[:top_k]]
    return "\n".join(f"-␣{line}" for line in top_docs)
```

This method operationalizes the final handoff from retrieval to generation. It optionally applies PRF, executes hybrid retrieval, selects the top-$k$ chunks, and formats them as a newline-separated bulleted list. The output is therefore not just a ranked retrieval result but a structured prompt component designed to be directly consumed by downstream generation templates.

Listing C.13: Prompt Template with Retrieved Context

```
Generate a realistic actor for {{community_name}} in {{country_name}}.

Community Information:
{{community_information}}

Additionally, consider the following context:
{{relevant_text}}

Skeleton characteristics:
{{persona_skeleton}}

Generate: name, role_title, age, gender, race, bio, profession, education, etc
```

In this way, RAG functions as a grounding layer across the full AURORA hierarchy. Rather than treating retrieval as a standalone preprocessing step, the system incorporates retrieved evidence directly into each generation stage so that higher-level country priors, meso-level community identities, and micro-level persona details are all conditioned on scenario-relevant documentary context.

**Summary.** AURORA's Retrieval-Augmented Generation (RAG) architecture provides a grounding layer that connects document retrieval to hierarchical persona generation. The method begins with an offline indexing pipeline in which scenario-specific documents are loaded, deduplicated using SHA-256 hashing, semantically chunked, embedded, and stored in a persistent ChromaDB vector store, while a parallel BM25 index is constructed over the same chunked corpus. During generation, retrieval is performed through a hybrid strategy that combines dense semantic search and sparse lexical search, with results fused using Reciprocal Rank Fusion (RRF). When queries are too broad or under-specified, Pseudo-Relevance Feedback (PRF) expands them using salient terms extracted from initially retrieved documents.

The model integrates this retrieval process across three hierarchical levels: country, community, and persona generation. At each level, context-specific queries are constructed, relevant text is retrieved and formatted, and the resulting evidence is injected directly into the corresponding LLM prompt. Technically, the system combines OpenAI or Ollama embedding models, ChromaDB with HNSW-based approximate nearest-neighbor search, BM25 lexical retrieval, and GPT-4o-mini for structured generation. Together, these components allow AURORA to produce country priors, community descriptions, and persona biographies that are more contextually grounded, more reproducible, and better aligned with scenario-specific documentary evidence.

## C.6 API Token Requirements and Generation Scalability

This section summarizes token usage, runtime, and API constraints for AURORA persona generation with RAG, using the Japan scenario from the Indo-Pacific Geopolitical Opinion Experiment (IGOE) exercise case study as the reference configuration.

The Japan scenario was configured with eight institution categories, each containing three communities, resulting in a total of 24 communities. Each community consisted of 15 personas, yielding a total population of 360 personas for the country. The scenario incorporated 10 key topics to capture variation in opinion and salience across issues, and included 15 groups to represent structured social affiliations within the population.

*Note.* The token and call totals above are arithmetically consistent. Skeleton generation itself is performed with IPF and does not consume LLM tokens; only the validation-rule step uses a single LLM call.

The token consumption for the Japan scenario is distributed across multiple stages of the AURORA generation pipeline, totaling 1,760 API calls and approximately 3.18 million tokens. Persona enrichment represents the largest contribution, accounting for 1.19 million tokens across 360 calls, followed by social media account generation, which together contribute 1.37 million tokens. Community-topic opinion generation also constitutes a significant portion at 244,800 tokens. In

Table C.3: Token Consumption by Generation Stage

| Stage | API Calls | Tokens / Call | Total Tokens |
|---|---|---|---|
| Country opinions | 10 | 950 | 9,500 |
| Community creation | 24 | 1,400 | 33,600 |
| Community-topic opinions | 240 | 1,020 | 244,800 |
| Skeleton validation rules | 1 | 2,000 | 2,000 |
| Persona enrichment | 360 | 3,300 | 1,188,000 |
| Group creation | 15 | 1,500 | 22,500 |
| Group ranking | 360 | 800 | 288,000 |
| Leader assignment | 15 | 600 | 9,000 |
| Twitter accounts | 360 | 2,000 | 720,000 |
| Telegram accounts | 360 | 1,800 | 648,000 |
| Telegram channels | 15 | 1,200 | 18,000 |
| **Total** | **1,760** | **—** | **3,183,400** |

contrast, higher-level steps such as country opinion generation and community creation contribute relatively small amounts to the overall cost. Group-related operations collectively add approximately 319,500 tokens. Notably, skeleton generation itself does not incur LLM token costs, as it is performed using IPF, with only a single validation-rule generation step requiring an API call. Overall, token usage is dominated by persona-level enrichment and downstream content generation rather than higher-level structural components.

OpenAI's current text pricing for GPT-4o mini is $0.15 per 1M input tokens and $0.60 per 1M output tokens.

Table C.4: Cost Summary for Japan Scenario

| Metric | Value |
|---|---|
| Total tokens | 3,183,400 |
| Estimated input tokens | 2,546,720 |
| Estimated output tokens | 636,680 |
| Estimated cost per country | $0.764 |
| Rounded cost per country | $0.76 |

Table C.5: Observed Runtime for Japan Scenario

| Metric | Value |
|---|---|
| Start time | 19:11:20 |
| End time | 20:03:58 |
| Total runtime | 52 min 38 sec |

Persona generation in this scenario is constrained primarily by API request throughput rather than token volume. With 1,760 API calls and a limit of 200 requests per minute (RPM), the theoretical minimum runtime is approximately 8.8 minutes, compared to a token-based lower bound of 1.6 minutes, indicating the system is clearly RPM-bound. The observed runtime of 52 minutes and 38 seconds (approximately $6\times$ the theoretical minimum) reflects additional overhead from asynchronous scheduling, validation, and I/O operations.

Execution time is distributed across API processing (23 minutes), IPF-based synthesis and validation (10 minutes), iterative refinement (17 minutes), and file I/O (3 minutes), aligning with the total observed runtime. RAG integration introduces modest overhead: injecting approximately 500 tokens per request (an 18% increase in token usage) but adding only 38 seconds of total latency (less than 1.2% of runtime), indicating negligible performance impact relative to overall execution time.

Scaling to multi-country scenarios exhibits linear growth in both API calls and token usage. For 12 countries, this results in 21,120 API calls, 38.2 million tokens, and a total cost of approximately $9.17. Sequential execution would require approximately 10.6 hours, but this can be substantially reduced through batching and parallelization.

To manage runtime constraints, the system employs asynchronous batching, retry logic with exponential backoff, IPF-based over-generation to reduce LLM calls, and incremental persistence with schema validation.

Overall, these results show that (i) system performance is dominated by RPM constraints rather than token limits, (ii) RAG adds minimal runtime overhead despite increased token usage, and (iii) the pipeline scales predictably in cost and runtime, supporting feasible multi-country deployment under bounded resource requirements.

## C.7 AURORA / AESOP group crosswalk

This appendix provides a comprehensive crosswalk mapping between AURORA-generated groups from the TOTO (IGOE) exercise and the corresponding AESOP group categories. The TOTO exercise generated personas and groups for 12 countries: Australia (AUS), Canada (CAN), China (CHN), India (IND), Iran (IRN), Japan (JPN), South Korea (KOR), New Zealand (NZL), Russia (RUS), Taiwan (TWN), United Kingdom (GBR), and United States (US).

The AESOP groups are organized into the following categories:

- **Transnational Groups**: Industry/Finance, Conservationist, Media (legit), NATO, FiveEyes, MilSpotters, Cybercrime, Academia, Conspiracist
- **National Groups**: Country-specific government (power/opposition), Defense, and national identity groups

| Country | AURORA Groups | AESOP Groups |
|---|---|---|
| AUS | Australian Institute of Company Directors, Business Council of Australia, Australian Chamber of Commerce and Industry | Industry/Finance (Transnational) |
| | University of Melbourne, Australian National University (ANU), University of Queensland | Academia (Transnational) |
| | Australian Hospital Association (AHA), Royal Australasian College of Surgeons (RACS), Australian College of Nursing (ACN) | Australian |
| | Australian Conservative Party (ACP), Coalition for Responsible Energy (CRE), Australians for a Stronger Economy (ASE) | Govt - pwr |
| | Tasmanian Salmon Advocates, Australian Environmental Coalition, Braddon Economic Development Forum | Govt - oppo |
| | Australian National Dialogue, Uniting Church in Australia, Hindu Council of Australia | Australian |
| | Australian Industrial Emissions Alliance, Eco-Industrial Collaborators, Fossil Fuel Accountability Collective | Conservationist (Transnational) |
| | Australian Conservation Foundation, Climate Action Network Australia, Wilderness Society | Conservationist (Transnational) |
| CAN | Canadian Chamber of Commerce, Business Council of Canada, Canadian Manufacturers & Exporters (CME) | Industry/Finance (Transnational) |
| | University of Toronto, University of British Columbia (UBC), McGill University | Academia (Transnational) |
| | Ontario Health, Health Canada, BC Health Authorities | Canadian |
| | Indigenous Rights Advocacy Group, Canadian Climate Action Network, Justice and Accountability Coalition for Indigenous Peoples | Indigenous |
| | Conservative Party of Canada, Alberta Conservative Coalition, Canadian Taxpayers Federation | Govt - pwr |
| | United Church of Canada, Canadian Council of Churches, Islamic Society of North America - Canada (ISNA Canada) | Canadian |

*Continued on next page*

Table C.6 – *Continued from previous page*

| Country | AURORA Groups | AESOP Groups |
|---|---|---|
| | Canadian Oil and Gas Industry Association (COGAIA), Canadian Construction Association (CCA), Canadian Mining Association (CMA) | Industry/Finance (Transnational) |
| | EcoNerds, Green Infrastructure Coalition, Sustainable Development Network | Conservationist (Transnational) |
| **CHN** | Guangzhou Internet Entrepreneurs Association, Guangzhou Chamber of Commerce, Guangzhou Tech Innovators Network | Industry/Finance (Transnational) |
| | Peking University, Tsinghua University, Zhejiang University | Academia (Transnational) |
| | Guangzhou Medical Institution, Shanghai Health Alliance, Beijing Hospital Network | Chinese |
| | Guangzhou Liberal Forum, Shanghai Democratic Society, Beijing Progressive Voices | Oppo |
| | Guangzhou Conservative Political Network, Beijing Conservative Think Tank, Shanghai Patriotic Advisory Council | Govt |
| | Three-Self Patriotic Movement, Chinese Buddhist Association, Daoist Association of China | Chinese |
| | Beijing Green Manufacturing Community, Shanghai Ecological Industrial Park Network, Guangzhou Energy-Efficient Industry Collective | Industry/Finance (Transnational) |
| | China Environmental Protection Foundation, Friends of Nature, Green Watershed | Conservationist (Transnational) |
| **IND** | National Association of Software and Service Companies (NASSCOM), Federation of Indian Chambers of Commerce & Industry (FICCI), The Confederation of Indian Industry (CII) | Industry/Finance (Transnational) |
| | Jawaharlal Nehru University (JNU), Indian Institute of Technology (IIT) Bombay, Sidho Kanhu Murmu University | Academia (Transnational) |
| | All India Institute of Medical Sciences (AIIMS), Tata Memorial Hospital (TMH), Christian Medical College (CMC) Vellore | Indian |
| | Digital Rights India, Cybersecurity Awareness Network, Institute for Digital Governance | Indian |

*Continued on next page*

Table C.6 – *Continued from previous page*

| Country | AURORA Groups | AESOP Groups |
|---|---|---|
| | National Cyber Security Coordination Committee (NCSCC), Indian Computer Emergency Response Team (CERT-In), Data Protection Authority of India (DPAI) | Govt - pwr |
| | Hinduism, Islam, Sikhism | Indian |
| | Textile Innovators Group, Technical Textiles Advocacy Network, Technical Textile Entrepreneurs Network | Industry/Finance (Transnational) |
| | Indian Centre for Environmental Legal Action (ICELA), Greenpeace India, Centre for Science and Environment | Conservationist (Transnational) |
| **IRN** | Iranian Exporters Association, Tehran Chamber of Commerce, Iran Technology Development Fund | Industry/Finance (Transnational) |
| | University of Tehran, Sharif University of Technology, Isfahan University of Technology | Academia (Transnational) |
| | Tehran University Medical Sciences, Shiraz University of Medical Sciences, Isfahan University of Medical Sciences | Iranian |
| | Iranian Liberal Democratic Movement, Iranian Youth for Change, Network of Iranian Feminists | Oppo |
| | Conservative Front, Islamic Revolutionary Guard Corps (IRGC), Principlist Coalition | Govt |
| | Shia Islamic Community (Twelvers), Zoroastrian Community, Baha'i Community | Iranian |
| | Iranian Industrial Association, Tehran Construction Coalition, Persian Oil & Gas Coalition | Industry/Finance (Transnational) |
| | Green Iran, Environmental Protection Organization, Iranian Wildlife Conservation Society | Conservationist (Transnational) |
| **JPN** | Japan Business Federation (Keidanren), Japan External Trade Organization (JETRO), Tokyo Chamber of Commerce and Industry (TCCI) | Industry/Finance (Transnational) |
| | Hokkaido University Arctic Research Initiative, Waseda University Arctic Studies Program, National Institute of Polar Research (NIPR) | Academia (Transnational) |
| | Tokyo Medical Association, Kansai Healthcare Coalition, Hokkaido Nursing Alliance | Japanese |

*Continued on next page*

Table C.6 – *Continued from previous page*

| Country | AURORA Groups | AESOP Groups |
|---|---|---|
| | Japan Constitutional Democratic Party (CDP), Japan Socialist Party (JSP), Japanese Green Party (JGP) | Govt - oppo |
| | Japan Conservative Coalition, Defenders of Japanese Sovereignty, Traditional Family Values Society | Govt official |
| | Shinto Shrines, Buddhist Temples, Christian Churches | Japanese |
| | Ibaraki Industrial Waste Management Association, Tochigi Prefectural Industrial Waste Supervisors Network, Pirika Waste Collection Initiative | Industry/Finance (Transnational) |
| | Hokkaido Environmental Protection Association, Ibaraki Prefecture Waste Management Coalition, Keidanren Environmental Committee | Conservationist (Transnational) |
| **KOR** | Korea Institute of Business Management (KIBM), Korean Business Council (KBC), Korean Female Entrepreneurs Network (KFEN) | Industry/Finance (Transnational) |
| | Seoul National University, Korea Advanced Institute of Science and Technology (KAIST), Yonsei University | Academia (Transnational) |
| | Korean Healthcare Alliance, Seoul Medical Professionals Network, Busan Hospital Consortium | Korean |
| | Korean Liberal Alliance, Progressive Scholars Network, Youth for Liberal Change | Govt - oppo |
| | Conservative Party of Korea (CPK), Korean Federation of Conservative Civic Groups, Alliance for National Unity | Govt - pwr |
| | Buddhist Jogye Order, Cheondoism, Protestant Christian Churches | Korean |
| | Korea Industrial Technology Association (KITA), Korea Manufacturing Industry Association (KMIA), Korean Institute of Industrial Engineers (KIIE) | Industry/Finance (Transnational) |
| | Korea Green Foundation, Korean Environmental Movement Association, Citizens' Coalition for Economic Justice | Conservationist (Transnational) |

*Continued on next page*

Table C.6 – *Continued from previous page*

| Country | AURORA Groups | AESOP Groups |
|---|---|---|
| | New Zealand Business Network, Kiwi Innovators Collective, Women in Business NZ | Industry/Finance (Transnational) |
| **NZL** | University of Auckland, Victoria University of Wellington, University of Otago | Academia (Transnational) |
| | Auckland City Hospital, Wellington Regional Hospital, Christchurch Hospital | Kiwi |
| | Liberal Advocates for Treaty Equality, New Zealand Democratic Rights Network, Investigators of Historical Truth | Govt - oppo |
| | New Zealand National Party, ACT New Zealand, New Zealand First | Govt - pwr |
| | Anglican Church in Aotearoa, New Zealand and Polynesia, The Catholic Church in New Zealand, Dunedin Islamic Centre | Kiwi |
| | New Zealand Manufacturing Association, New Zealand Institute of Architects, The Engineering NZ | Industry/Finance (Transnational) |
| | Forest & Bird, Greenpeace Aotearoa, Environmental Defence Society | Conservationist (Transnational) |
| | Russian Cybersecurity Alliance, Rostec Business Forum, IT Entrepreneurs Association | Industry/Finance (Transnational) |
| **RUS** | Moscow State University, Saint Petersburg State University, Russian State University for the Humanities (RGGU) | Academia (Transnational) |
| | Federation of Russian Hospitals, National Association of Hospital Administrators, Russian Society of Medical Professionals | Russian |
| | Liberal Democratic Forum, Youth Action Network, Sustainable Governance Alliance | Oppo |
| | Russian Conservative Unity, Union of Traditionalists, Russian Nationalist Front | Govt |
| | Russian Orthodox Church, Muslim Spiritual Administration of Russia, Jewish Community of Russia | Russian |
| | Rostec State Corporation, Rostelecom, Russian Union of Industrialists and Entrepreneurs (RSPP) | Industry/Finance (Transnational) |

*Continued on next page*

Table C.6 – *Continued from previous page*

| Country | AURORA Groups | AESOP Groups |
|---|---|---|
| | Russian Association of Indigenous Minorities of the North (RAIPON), WWF Russia, Greenpeace Russia | Conservationist (Transnational) |
| **TWN** | Taiwan Securities Investment Association (TSIA), Taiwan Business Federation (TBF), Taiwan Fintech Association (TFA) | Industry/Finance (Transnational) |
| | National Taiwan University (NTU), National Chengchi University (NCCU), National Tsing Hua University (NTHU) | Academia (Transnational) |
| | Taiwan National Hospital Consortium, Taiwan Nurses Association, Taiwan Hospital Quality Improvement Association | Taiwanese |
| | Taiwan Investment Strategic Forum, National Development Fund Advocates, Taiwan Semiconductor Investment Coalition | Govt - oppo |
| | Taiwan Sovereign Fund Advocates, Taiwanese Economic Conservatism Network, Structured Fiscal Engagers | Govt - pwr |
| | Buddhist Tzu Chi Foundation, Catholic Church in Taiwan, Church of Jesus Christ of Latter-day Saints (Mormons) | Taiwanese |
| | Circular Economy Startup Alliance, Taiwan Resource Recycling Alliance, Green Innovation Forum | Industry/Finance (Transnational) |
| | Taiwan Circular Economy Association, Resource Circulation Startup Industry Alliance, Taiwan Environmental Information Association | Conservationist (Transnational) |
| **GBR** | Trades Union Congress (TUC), UNISON, The Yorkshire Mafia | Industry/Finance (Transnational) |
| | University College London (UCL) - Student Union, The University of Manchester - Student Union, Imperial College London – Student Union | Academia (Transnational) |
| | NHS Trusts, Private Healthcare Providers, Community Health Services | British |
| | Liberal Democrats, Green Party of England and Wales, Labour Party | Govt - oppo |

*Continued on next page*

Table C.6 – *Continued from previous page*

| Country | AURORA Groups | AESOP Groups |
|---|---|---|
| | Conservative Councillors Association, The Local Government Association (LGA) - Conservative Group, National Conservative Convention (NCC) | Govt - pwr |
| | Church of England, Roman Catholic Church in England and Wales, Islamic Community (Muslims in the UK) | British |
| | UNISON, GMB Union, National Union of Rail, Maritime and Transport Workers (RMT) | Industry/Finance (Transnational) |
| | Friends of the Earth, Greenpeace UK, The Wildlife Trusts | Conservationist (Transnational) |
| | National Chamber of Commerce, American Small Business Association (ASBA), The Business Council | Industry/Finance (Transnational) |
| **US** | American Medical Student Association, National Society of Black Engineers, College Democrats of America | Academia (Transnational) |
| | American Hospital Association (AHA), American Nurses Association (ANA), Healthcare Cost and Utilization Project (HCUP) | US Govt - left |
| | Progressive Local Governments Network, National Association of Counties, League of Women Voters | US Govt - left |
| | Republican National Committee (RNC), Heritage Foundation, American Legislative Exchange Council (ALEC) | US Govt - right |
| | The United Methodist Church, Roman Catholic Church, Islamic Society of North America (ISNA) | US Govt - left |
| | United Auto Workers (UAW), American Federation of State, County and Municipal Employees (AFSCME), International Brotherhood of Teamsters (IBT) | Industry/Finance (Transnational) |
| | Sierra Club, Natural Resources Defense Council (NRDC), Environmental Defense Fund | Conservationist (Transnational) |

# Appendix D

# JANUS: Emotion-Aware Opinion Dynamics

## D.1 Intermediate Network Scale Analysis

This appendix provides comprehensive analysis of the intermediate network scenarios (Dual-500 and Dual-750), which exhibit transitional dynamics between the echo chamber effects of small networks and the averaging dominance of large networks. These scenarios are particularly important for understanding how network scale effects emerge and how topic-specific responses diverge under intermediate connectivity.

### D.1.1 Dual-500 Scenario: Transitional Regime

The Dual-500 scenario (N=480, average degree 36.0) represents a critical transition point where network scale effects begin to differentiate topic-specific responses. Table D.1 provides comprehensive metrics for this scenario.

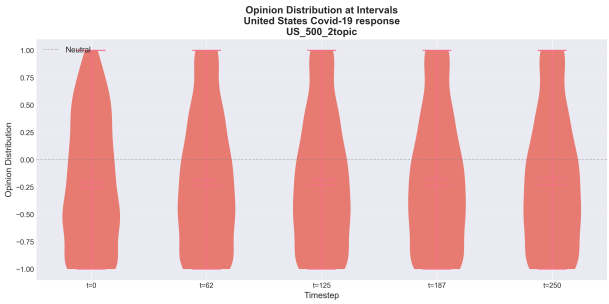
#### Opinion Distribution Evolution - Dual-500

Figure D.1 presents the complete opinion distribution analysis for Dual-500, combining distributional evolution with coordination dynamics.

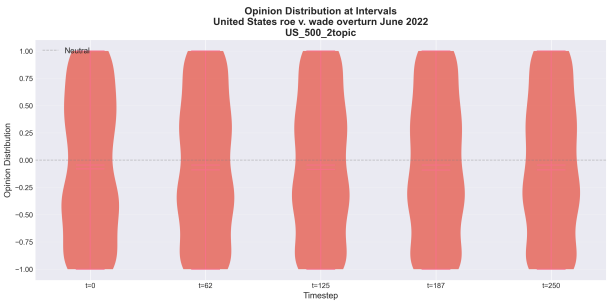
The Dual-500 distributions exhibit **topic-dependent divergence**. COVID shows modest widening ($\Delta\sigma = +0.021$), continuing the amplification pattern from Dual-250 but with reduced magnitude. Roe shows slight narrowing ($\Delta\sigma = -0.011$), marking the first reversal toward moderation. This divergence reflects the interaction between network structure and topic-specific initial conditions.

The temporal coordination analysis reveals smoother dynamics compared to Dual-250. The initial drop in mean correlation is less pronounced, and coordination strength is reduced (0.075 COVID, 0.084 Roe vs 0.092 and 0.080 in Dual-250). This reflects the law of large numbers: larger neighborhoods (36 vs 18 neighbors) produce more stable, averaged influence signals.

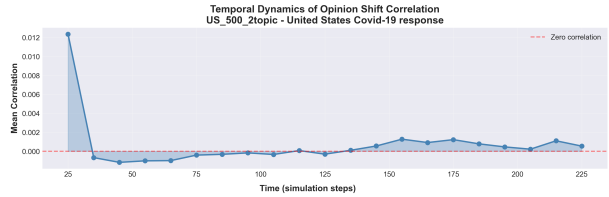
Community structure persists but with reduced separation. COVID shows separation of 0.032 (vs 0.072 in Dual-250), while Roe maintains 0.052 (vs 0.053 in Dual-250). The reduced COVID



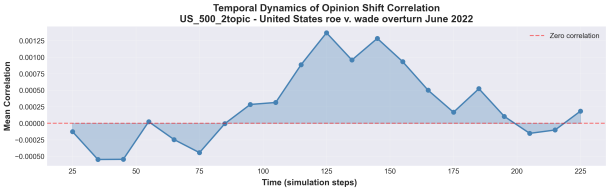

(a) COVID-19 - Distribution Evolution



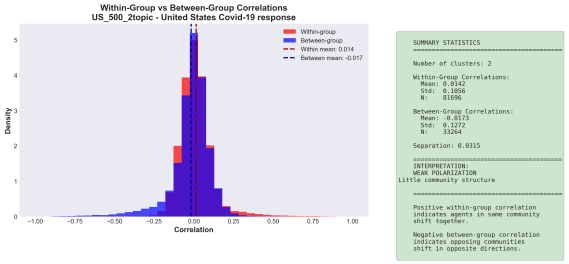

(b) Roe v. Wade - Distribution Evolution



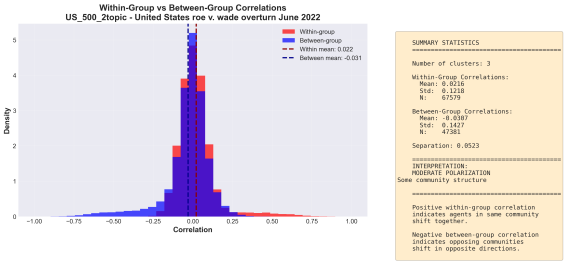

(c) COVID-19 - Temporal Coordination




(d) Roe v. Wade - Temporal Coordination




(e) COVID-19 - Community Structure




(f) Roe v. Wade - Community Structure

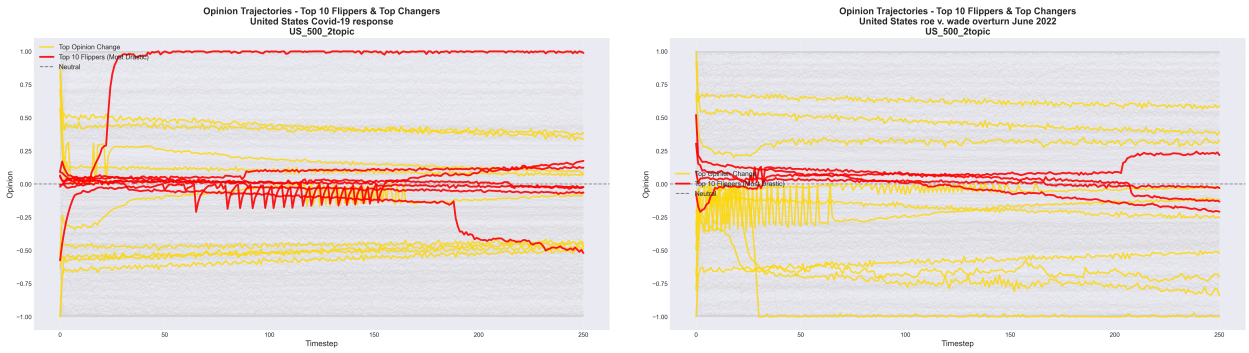
Figure D.1: Complete opinion distribution and coordination analysis for Dual-500. (a-b) Violin plots show topic-dependent responses: COVID displays modest widening while Roe shows slight narrowing. (c-d) Temporal dynamics reveal smoother evolution compared to Dual-250, with coordination strength of 0.075 (COVID) and 0.084 (Roe). (e-f) Community structure shows reduced separation compared to Dual-250 (0.032 vs 0.072 for COVID), reflecting increased cross-group exposure from larger neighborhoods.

Table D.1: Comprehensive metrics for Dual-500 scenario. This transitional regime shows topic-dependent polarization responses: COVID continues amplification while Roe begins moderation, marking the reversal point in scale effects.

| **Metric** | **COVID-19** | | | **Roe v. Wade** | | |
|---|---|---|---|---|---|---|
| | **Initial** | **Final** | **Change** | **Initial** | **Final** | **Change** |
| Polarization ($\sigma$) | 0.583 | 0.604 | +0.021 | 0.650 | 0.638 | $-0.011$ |
| Mean Opinion ($\mu$) | $-0.202$ | $-0.174$ | +0.028 | $-0.045$ | $-0.046$ | $-0.001$ |
| Flip Count | | 7 (1.46%) | | | 4 (0.83%) | |
| Mean $\lvert\Delta x\rvert$ | | 0.104 | | | 0.118 | |
| *Coordination Metrics (Full Simulation)* | | | | | | |
| Within-Group $\rho$ | | +0.014 | | | +0.022 | |
| Between-Group $\rho$ | | $-0.017$ | | | $-0.031$ | |
| Separation | | 0.032 | | | 0.052 | |
| Coord. Strength | | 0.075 | | | 0.084 | |
| Mean Correlation | | +0.005 | | | +0.000 | |
| Frac. Positive | | 0.516 | | | 0.501 | |
| N Clusters | | 2 | | | 3 | |

separation reflects increased cross-group exposure from doubled neighborhood size, while Roe's stronger initial polarization maintains clearer community boundaries.

### Temporal Dynamics - Dual-500




(a) COVID-19 Response

(b) Roe v. Wade

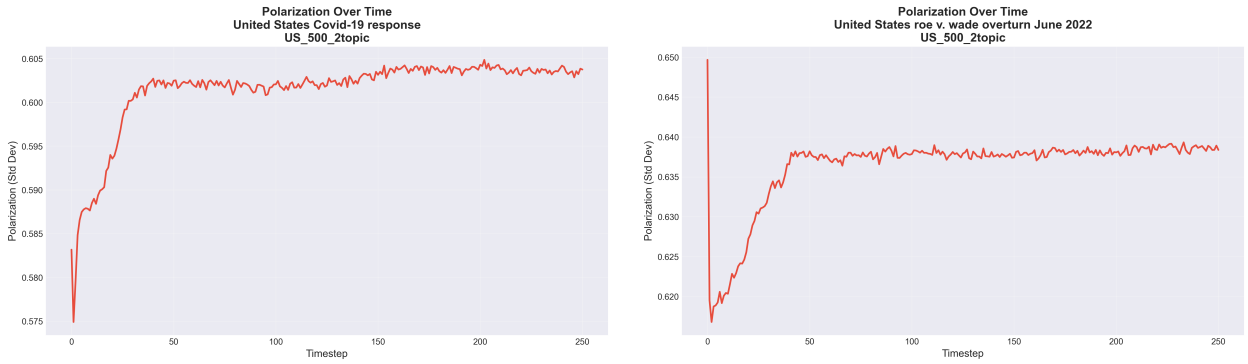
Figure D.2: Opinion trajectories for Dual-500 (N=480). Trajectories show smoother evolution with reduced volatility compared to Dual-250, reflecting larger neighborhoods (avg degree 36.0). COVID: 7 flips (1.46%); Roe: 4 flips (0.83%). Despite larger absolute counts, flip rates decrease, indicating enhanced boundary stability with scale.

Dual-500 trajectories show visible **smoothing** compared to Dual-250. Fewer abrupt directional

changes and reduced short-term volatility arise from diluted individual neighbor influence. The early transient period remains visible but with reduced magnitude.

Flip events show sublinear scaling: absolute counts increase (7 COVID, 4 Roe) but rates decrease (1.46% and 0.83% vs 2.08% and 1.25% in Dual-250). This indicates that larger neighborhoods enhance boundary stability by reducing sustained unidirectional pressure.

**Polarization Trends - Dual-500**




(a) COVID-19 Response

(b) Roe v. Wade

Figure D.3: Polarization over time for Dual-500. COVID continues increasing ($\Delta\sigma = +0.021$), though attenuated vs Dual-250, while Roe exhibits slight decrease ($\Delta\sigma = -0.011$). This divergence marks the transition point where scale effects begin to differentiate topic responses. After t=100, both trajectories plateau, indicating quasi-equilibrium.

Dual-500 polarization dynamics mark the **transition point** in scale effects. COVID continues amplification but with half the magnitude of Dual-250 (+0.021 vs +0.046). Roe shows the first reversal toward moderation ($-0.011$). After approximately t=100, both trajectories plateau, indicating quasi-equilibrium where competing forces balance.

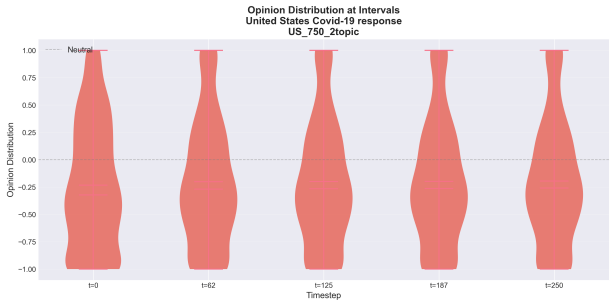
The divergence illustrates that polarization dynamics emerge from network structure × topic-specific distributions. Roe's higher initial polarization creates more homogeneous neighborhoods that maintain stable positions. COVID's more diffuse distribution allows continued clustering.

### D.1.2 Dual-750 Scenario: Averaging-Dominated Regime

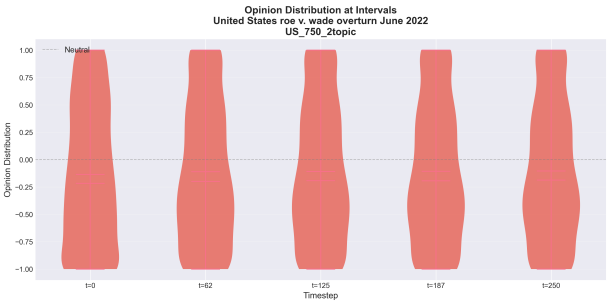
The Dual-750 scenario (N=720, average degree 55.7) exhibits clear dominance of averaging effects over echo chamber dynamics. Table D.2 provides comprehensive metrics.

**Opinion Distribution Evolution - Dual-750**

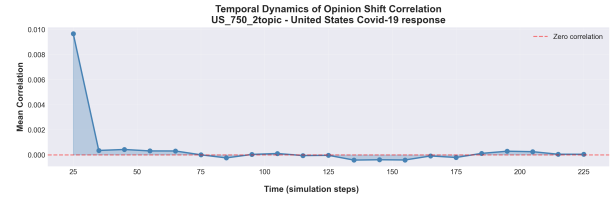
Dual-750 distributions show clear **narrowing** for both topics. Peaks become more concentrated and tails shorten, producing identical polarization decreases ($\Delta\sigma = -0.024$). This symmetry indicates that structural network effects now dominate over topic-specific characteristics.



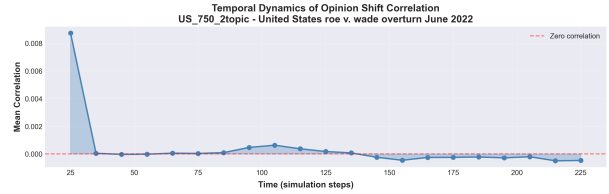

(a) COVID-19 - Distribution Evolution



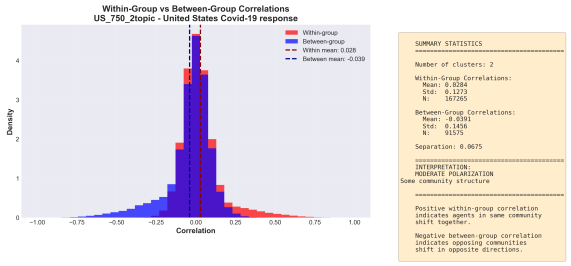

(b) Roe v. Wade - Distribution Evolution



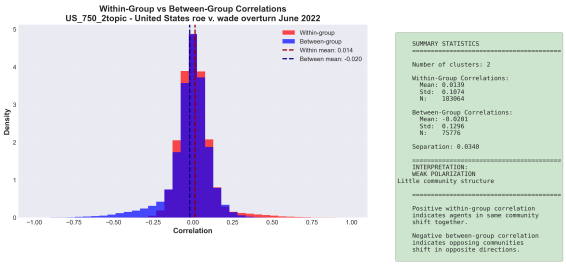

(c) COVID-19 - Temporal Coordination




(d) Roe v. Wade - Temporal Coordination




(e) COVID-19 - Community Structure




(f) Roe v. Wade - Community Structure

Figure D.4: Complete opinion distribution and coordination analysis for Dual-750. (a-b) Violin plots show visible narrowing for both topics, with peaks becoming more concentrated and tails shortening. (c-d) Temporal dynamics show coordination strength of 0.090 (COVID) and 0.078 (Roe), intermediate between Dual-500 and Dual-1000. (e-f) Community structure shows COVID separation of 0.068 (higher than Dual-500's 0.032), while Roe shows 0.034 (reduced from 0.052), reflecting topic-specific responses to averaging pressure.

Table D.2: Comprehensive metrics for Dual-750 scenario. Both topics now show decreasing polarization with identical magnitudes ($\Delta\sigma = -0.024$), indicating convergent dynamics where structural network effects dominate over topic-specific characteristics.

| **Metric** | **COVID-19** | | | **Roe v. Wade** | | |
|---|---|---|---|---|---|---|
| | **Initial** | **Final** | **Change** | **Initial** | **Final** | **Change** |
| Polarization ($\sigma$) | 0.582 | 0.558 | $-0.024$ | 0.647 | 0.624 | $-0.024$ |
| Mean Opinion ($\mu$) | $-0.233$ | $-0.195$ | +0.038 | $-0.135$ | $-0.105$ | +0.030 |
| Flip Count | | 8 (1.11%) | | | 8 (1.11%) | |
| Mean $\lvert\Delta x\rvert$ | | 0.114 | | | 0.090 | |
| *Coordination Metrics (Full Simulation)* | | | | | | |
| Within-Group $\rho$ | | +0.028 | | | +0.014 | |
| Between-Group $\rho$ | | $-0.039$ | | | $-0.020$ | |
| Separation | | 0.068 | | | 0.034 | |
| Coord. Strength | | 0.090 | | | 0.078 | |
| Mean Correlation | | +0.005 | | | +0.004 | |
| Frac. Positive | | 0.509 | | | 0.509 | |
| N Clusters | | 2 | | | 2 | |

The averaging pressure from larger neighborhoods (55.7 neighbors) moderates extreme positions. Agents near opinion extremes increasingly encounter moderate viewpoints that pull positions toward population center. Although anchor mechanisms limit movement, permitted drift allows gradual moderation.

Community structure persists with interesting topic-specific patterns. COVID shows increased separation (0.068) compared to Dual-500 (0.032), while Roe shows decreased separation (0.034 vs 0.052). This suggests complex interactions between initial distributions and network scale effects.

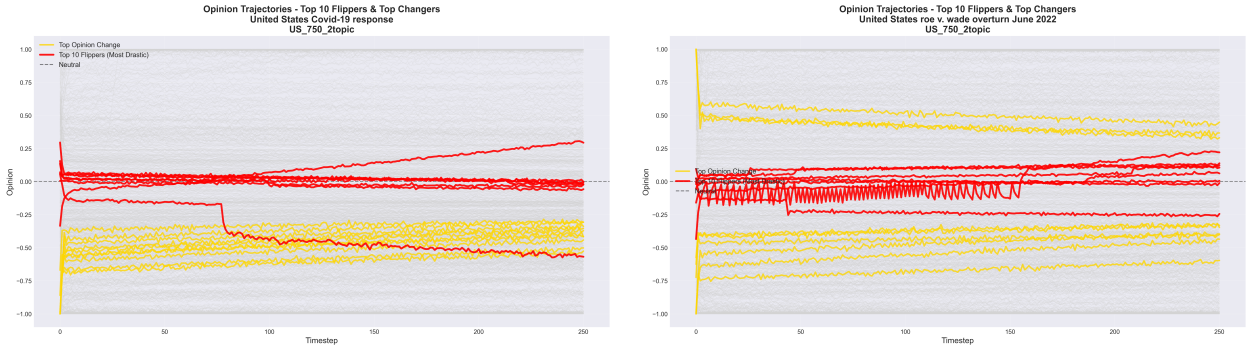
### Temporal Dynamics - Dual-750

Dual-750 trajectories show **further smoothing and constraint**. After early transient phase, most agents exhibit minimal movement, remaining within narrow bands. Roe shows extremely limited volatility due to high initial polarization and strongly clustered neighborhoods producing consistent reinforcing signals.

Flip rates reach lowest levels: both topics record 8 flips (1.11%). The equality across topics suggests structural network effects dominate over topic-specific differences in boundary dynamics. Increasing stability reflects growing dominance of averaging effects.

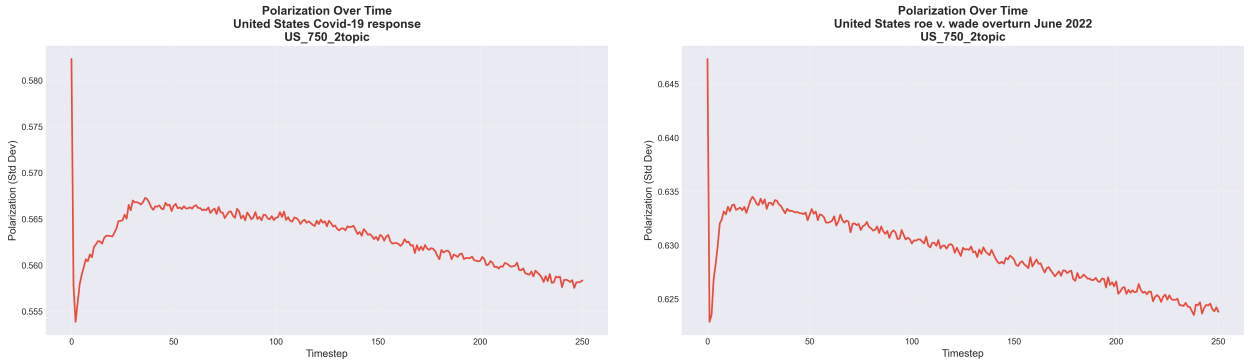
### Polarization Trends - Dual-750

Dual-750 polarization shows clear **reversal to moderation**. Both topics decrease by identical amounts ($\Delta\sigma = -0.024$), confirming that structural effects dominate at this scale. Unlike smaller




(a) COVID-19 Response (b) Roe v. Wade

Figure D.5: Opinion trajectories for Dual-750 (N=720). Further smoothing and constraint visible, with majority showing minimal movement after initial stabilization. Average degree 55.7 creates substantial averaging pressure, visible in tighter clustering of gray trajectories. Both topics: 8 flips (1.11%), lowest rate among scenarios, indicating enhanced boundary stability.




(a) COVID-19 Response (b) Roe v. Wade

Figure D.6: Polarization over time for Dual-750. Both topics show decreasing polarization with identical magnitudes ($\Delta\sigma = -0.024$), confirming systematic reversal from amplification to moderation. Unlike smaller scenarios where polarization increases after transient phase, Dual-750 shows consistent downward trend throughout simulation.

scenarios with continued post-transient growth, Dual-750 shows consistent downward trend throughout.

The initial transient period still produces largest adjustments, but direction is now toward moderation rather than amplification. After stabilization, polarization continues declining slowly but steadily through remainder of simulation.

### D.1.3 Cross-Scale Comparison: Intermediate Scenarios

Table D.3 provides direct comparison of Dual-500 and Dual-750 metrics.

Table D.3: Direct comparison of intermediate scenarios. The transition from Dual-500 to Dual-750 shows systematic shifts: topic-dependent responses converge to symmetric moderation, flip rates reach minimum, and coordination patterns stabilize.

| **Metric** | **Dual-500** | | **Dual-750** | |
|---|---|---|---|---|
| | **COVID** | **Roe** | **COVID** | **Roe** |
| $\Delta\sigma$ | +0.021 | $-0.011$ | $-0.024$ | $-0.024$ |
| Flip Rate (%) | 1.46 | 0.83 | 1.11 | 1.11 |
| Separation | 0.032 | 0.052 | 0.068 | 0.034 |
| Coord. Strength | 0.075 | 0.084 | 0.090 | 0.078 |
| Mean $\lvert\Delta x\rvert$ | 0.104 | 0.118 | 0.114 | 0.090 |

The intermediate scenarios reveal the **transition mechanism** from echo chamber to averaging dominance:

- **Polarization reversal**: Dual-500 shows topic-dependent responses (COVID +0.021, Roe $-0.011$); Dual-750 shows symmetric moderation (both $-0.024$). The transition point lies between these scales.
- **Flip rate minimum**: Both scenarios show rates below extremes, with Dual-750 reaching absolute minimum (1.11%). Intermediate connectivity optimizes boundary stability.
- **Coordination evolution**: Strength increases from Dual-500 (0.075–0.084) to Dual-750 (0.078–0.090) for COVID, suggesting non-monotonic relationship with scale.
- **Community structure**: Separation patterns show topic-specific responses to scale, with COVID increasing and Roe decreasing from 500 to 750.
- **Individual mobility**: Mean change shows topic-dependent patterns, with Roe decreasing substantially (0.118 $\rightarrow$ 0.090) while COVID increases slightly (0.104 $\rightarrow$ 0.114).

These intermediate scenarios demonstrate that network scale effects are not simply monotonic. Instead, they involve complex interactions between connectivity, initial distributions, and coordination mechanisms that produce topic-dependent responses before converging to scale-dominated dynamics at larger sizes.

# Appendix E

# Evaluating Realism and Impact in Live Scenarioss

## E.1 AURORA Persona Integration: Ego Network Validation and AESOP Cross-Walk

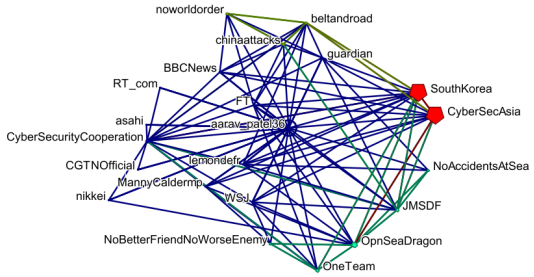
This appendix provides evidence for two core claims: (1) AURORA generates **structurally distinct personas**, and (2) these personas **integrate with AESOP actors within a shared network structure** while contributing individual-level realism.

To address the committee's feedback, I present labeled ego networks alongside quantitative comparisons of network structure, centrality, and topic engagement. The results show that AURORA personas differ from one another not only in demographic description, but also in how they are embedded in the multi-modal network, and that they extend AESOP by adding individual-level social heterogeneity.

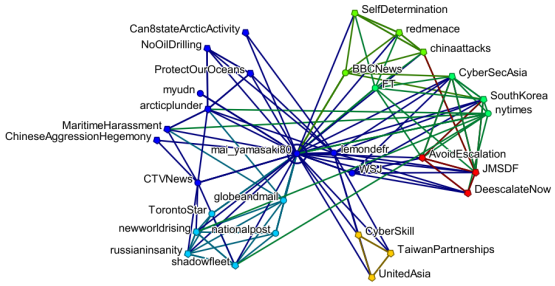
### E.1.1 Ego Network Comparison

Figure E.1 and E.2 both show ego networks for two AURORA personas (Aarav Patel and Mai Yamasaki) and one AESOP actor (myudn). Nodes are labeled by type, and edges represent weighted interactions. To improve interpretability, only **above-average actor-to-actor ties** are displayed, so that the visible alters reflect stronger relationships rather than complete connectivity.

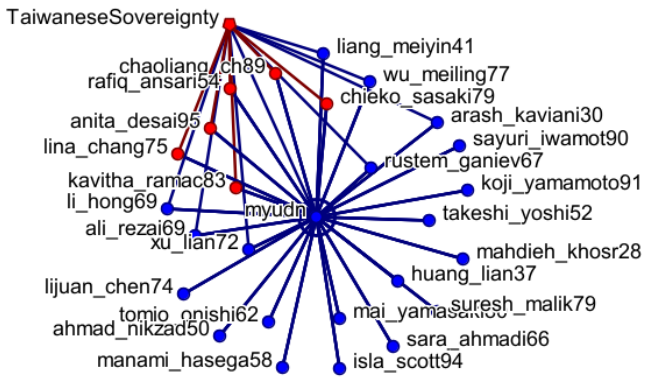
Two observations follow directly from the figure. First, Aarav and Mai exhibit visibly different ego network configurations, including different densities, topic clustering patterns, and alter compositions. Second, the AURORA personas connect to multiple node types, including actors, groups, communities, and topics, while the AESOP actor is more concentrated in actor-centered connections. These visual differences provide initial evidence that the personas occupy distinct positions within a shared multi-modal network.




(a) Aarav Patel (AURORA).




(b) Mai Yamasaki (AURORA).




(c) myudn (AESOP).

Figure E.1: Labeled ego networks for two AURORA personas and one AESOP actor. Aarav Patel and Mai Yamasaki exhibit visibly different ego-network structures, including different alter sets, topic clusters, and connectivity patterns, while the AESOP actor myudn shows a more actor-centered configuration. Only actor-to-actor ties above the average strength threshold are displayed for clarity.

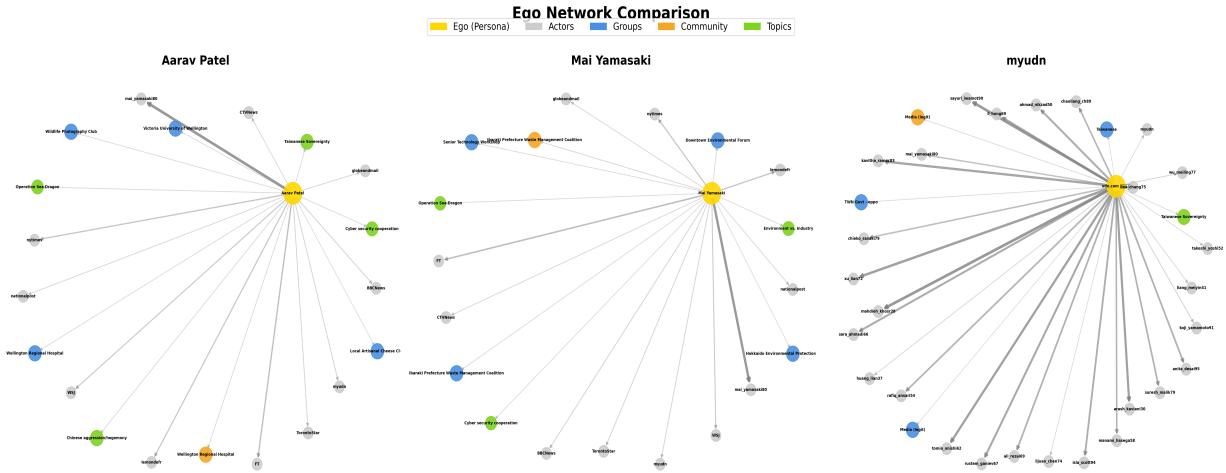



Figure E.2: Complete ego network comparison showing Aarav Patel (AURORA, left), Mai Yamasaki (AURORA, center), and myudn (AESOP, right). Node colors indicate type: yellow = ego, green = topics, blue = groups, orange = community, grey = actors. Networks are constructed from XML ego network files with topic filtering applied to show only discussed topics. Only actor ties above the average strength threshold are visualized for clarity.

### E.1.2 Network Structure and Centrality

Table E.1 quantifies the differences visible in Figure E.2. The two AURORA personas differ substantially across several structural dimensions. Aarav's ego network contains 217 total nodes compared with Mai's 164, a 32% difference. Mai has more topic nodes (20 vs. 14), while Aarav has more total tweets and unique senders. These differences indicate that the two personas are not merely alternate textual descriptions layered onto the same network template.

Centrality measures reinforce this distinction. Mai shows higher maximum in-degree, PageRank, and hub score, indicating a more central and responsive position in her ego network. Aarav, by contrast, shows broader but more distributed connectivity. These results indicate that AURORA personas are not only demographically distinct, but also topologically distinct within their ego networks.

### E.1.3 Topic Engagement as Behavioral Differentiation

Table E.2 shows that the two AURORA personas also differ in semantic engagement. Aarav concentrates heavily on cybersecurity-related topics, most notably *CyberSecurityCooperation*, while Mai engages across a broader mix of geopolitical and de-escalation themes. Although a small number of topics overlap, most are persona-specific. This indicates that the structural differences observed in the ego networks are **consistent with differences in topic engagement and interaction patterns**.

Table E.1: Comprehensive Ego Network Statistics (All Node Types)

| Metric | Aarav | Mai | myudn | AURORA Avg |
|---|---|---|---|---|
| *Network Size* | | | | |
| Total Nodes | 217 | 164 | 80 | 190.5 |
| Agent Nodes | 15 | 11 | 27 | 13.0 |
| Hashtag/Topic Nodes | 14 | 20 | 0 | 17.0 |
| Tweet Nodes | 188 | 133 | 53 | 160.5 |
| *Agent Degree Centrality* | | | | |
| Avg In-Degree | 0.0048 | 0.0386 | 0.0000 | 0.0217 |
| Max In-Degree | 0.0714 | 0.4250 | 0.0000 | 0.2482 |
| Avg Out-Degree | 0.0369 | 0.0386 | 0.0190 | 0.0378 |
| Max Out-Degree | 0.0714 | 0.1000 | 0.5128 | 0.0857 |
| Avg Total Degree | 0.0417 | 0.0386 | 0.0190 | 0.0402 |
| Max Total Degree | 0.2232 | 0.2125 | 0.2564 | 0.2179 |
| *Agent Centrality Measures* | | | | |
| Avg Closeness | 0.0667 | 0.0909 | 0.0370 | 0.0788 |
| Max Closeness | 0.0667 | 0.0909 | 0.0370 | 0.0788 |
| Avg PageRank | 0.0667 | 0.0909 | 0.0370 | 0.0788 |
| Max PageRank | 0.3815 | 0.5045 | 0.0384 | 0.4430 |
| Avg Hub Score | 0.2153 | 0.3182 | 0.0000 | 0.2668 |
| Max Hub Score | 0.4476 | 0.8235 | 0.0000 | 0.6356 |
| *Hashtag/Topic Engagement* | | | | |
| Total Hashtag Usage | 2614 | 1715 | 1 | 2164.5 |
| Avg Hashtag Usage | 186.7 | 85.8 | 1.0 | 136.2 |
| Max Hashtag Usage | 1142 | 478 | 1 | 810.0 |
| Unique Hashtags | 14 | 20 | 1 | 17.0 |
| *Tweet Activity* | | | | |
| Total Tweets | 188 | 133 | 53 | 160.5 |
| Unique Senders | 24 | 12 | 11 | 18.0 |

Note: Statistics extracted from ORA network analysis files showing complete ego networks including agents, hashtags/topics, and tweets. Aarav and Mai differ in network size, topic connectivity, sender diversity, and centrality distributions. These results provide evidence that AURORA generates personas with distinct social network characteristics rather than simple template reuse.

Table E.2: Top Hashtags/Topics by Persona

| Persona | Top Hashtags (by usage count) |
|---|---|
| **Aarav Patel** | CyberSecurityCooperation (1142), SouthKorea (478), CyberSecAsia (464), OpnSeaDragon (160), JMSDF (95), OneTeam (63), beltandroad (51), NoAccidentsAtSea (46), NoBetterFriendNoWorseEnemy (41), chinaattacks (34) |
| **Mai Yamasaki** | SouthKorea (478), CyberSecAsia (464), JMSDF (95), arcticplunder (73), AvoidEscalation (57), MaritimeHarassment (57), DeescalateNow (55), newworldrising (54), russianinsanity (51), UnitedAsia (49) |
| **myudn** | TaiwaneseSovereignty (1 use – organizational account focuses on direct content) |

Note: Hashtag usage counts reflect total mentions across all tweets in each persona's ego network. Aarav shows a concentrated emphasis on cybersecurity cooperation, while Mai's engagement is distributed across a wider set of regional and de-escalatory topics. Shared topics exist, but the overall engagement patterns remain distinct.

### E.1.4 Validation of Persona Distinctiveness

Table E.3 consolidates the differences between Aarav and Mai across demographics, network structure, centrality, personality, and behavior. The differences are substantial across all measured dimensions, including a 44-year age gap, different professions and locations, 32–100% differences in structural measures, and substantial differences in social media usage and posting frequency. The results provide strong evidence that AURORA generates individually differentiated personas whose ego networks reflect distinct attributes, engagement patterns, and interaction behaviors.

### E.1.5 AESOP Cross-Walk: Structural Integration

I compare how AURORA and AESOP components occupy the same network space. Figure E.2 can be interpreted as a multi-layer network containing two different kinds of entities: AURORA-derived nodes, including individual actors, communities, and topics; and AESOP-derived nodes, including organizational actors and scenario-driven content.

When read in this way, two integration patterns emerge. First, AURORA personas connect not only to AURORA-generated structures such as communities and topics, but also to AESOP actors. Second, AESOP actors such as myudn include AURORA personas within their interaction networks. This provides evidence of a functional cross-walk, where both systems contribute to a shared network representation while preserving different roles.

### E.1.6 Complementary Network Roles

Table E.4 highlights the differences between AURORA personas and AESOP actors. AURORA personas are associated with greater topic diversity, larger ego networks, non-zero in-degree, and higher PageRank and hub scores. AESOP actors, by contrast, show higher out-degree and more

Table E.3: AURORA Persona Validation: Aarav vs. Mai Comparison

| **Dimension** | **Aarav** | **Mai** | **Validation Result** |
|---|---|---|---|
| *Demographics* | | | |
| Age | 22 | 66 | **44-year difference** |
| Gender | Male | Female | **Different** |
| Location | Wellington, NZ | Mito, Japan | **Different countries** |
| Profession | Medical Intern | Retired Teacher | **Different fields** |
| *Network Structure* | | | |
| Total Nodes | 217 | 164 | **32% difference** |
| Agent Nodes | 15 | 11 | **36% difference** |
| Hashtag Nodes | 14 | 20 | **43% difference** |
| Tweet Nodes | 188 | 133 | **41% difference** |
| Unique Senders | 24 | 12 | **100% difference** |
| *Topic Engagement* | | | |
| Total Hashtag Usage | 2614 | 1715 | **52% difference** |
| Top Hashtag | CyberSecurityCooperation (1142) | SouthKorea (478) | **Different focus** |
| Unique Hashtags | 14 | 20 | **43% difference** |
| Shared Topics | 3 (CyberSecAsia, SouthKorea, JMSDF) | | **Most topics distinct** |
| *Centrality Measures* | | | |
| Max In-Degree | 0.0714 | 0.4250 | **495% difference** |
| Max PageRank | 0.3815 | 0.5045 | **32% difference** |
| Max Hub Score | 0.4476 | 0.8235 | **84% difference** |
| *Personality Traits (Big Five)* | | | |
| Openness | 0.685 | 0.070 | **0.615 difference** |
| Agreeableness | 0.083 | 1.000 | **0.917 difference** |
| *Social Behavior* | | | |
| Social Media Usage | 0.820 (Very High) | 0.332 (Low) | **2.5x difference** |
| Post Frequency | 0.1 (Very Low) | 0.9 (Very High) | **9x difference** |

direct actor-to-actor reach, consistent with organizational broadcast behavior. These differences indicate that the two systems occupy complementary roles within the network: AURORA provides individual-level heterogeneity and bidirectional interaction, while AESOP provides institutional or organizational-scale communication behavior.

Table E.4: AURORA vs. AESOP Comparison: Complementary Capabilities

| Metric | AURORA Avg | AESOP | Difference | Interpretation |
|---|---|---|---|---|
| *Network Scale* | | | | |
| Total Nodes | 190.5 | 80 | +110.5 (+138%) | AURORA larger ego networks |
| Agent Nodes | 13.0 | 27 | -14.0 (-52%) | AESOP broader direct actor reach |
| Tweet Nodes | 160.5 | 53 | +107.5 (+203%) | AURORA higher activity volume |
| Unique Senders | 18.0 | 11 | +7.0 (+64%) | AURORA greater interaction diversity |
| *Topic Diversity* | | | | |
| Hashtag Nodes | 17.0 | 1 | +16.0 (+1600%) | AURORA more topic engagement |
| Total Hashtag Usage | 2164.5 | 1 | +2163.5 | AURORA substantially broader topical participation |
| Avg Hashtag Usage | 136.2 | 1.0 | +135.2 | AURORA higher average topic use |
| Max Hashtag Usage | 810.0 | 1 | +809.0 | AURORA stronger topic concentration |
| *Centrality Measures* | | | | |
| Avg Closeness | 0.0788 | 0.0370 | +0.0418 (+113%) | AURORA more embedded local positioning |
| Max PageRank | 0.4430 | 0.0384 | +0.4046 | AURORA higher centrality in ego network |
| Max Hub Score | 0.6356 | 0.0000 | +0.6356 | AURORA stronger hub-like connectivity |
| Max Out-Degree | 0.0857 | 0.5128 | -0.4271 (-83%) | AESOP stronger broadcast reach |
| *Engagement Patterns* | | | | |
| Avg In-Degree | 0.0217 | 0.0000 | +0.0217 | AURORA receives interactions |
| Max In-Degree | 0.2482 | 0.0000 | +0.2482 | AURORA shows bidirectional engagement |
| Avg Total Degree | 0.0402 | 0.0190 | +0.0212 (+112%) | AURORA more overall engagement |
| Max Total Degree | 0.2179 | 0.2564 | -0.0385 (-15%) | AESOP has slightly higher peak broadcast connectivity |

## E.2 Temporal Escalation Structure

Table E.5: Temporal escalation structure (Baseline–Day 4).

| Day | Type | Event | Operational Summary |
|---|---|---|---|
| Base | Backdrop | RUS Shadow Fleet | Sanction-evasion oil transport via proxy tankers in Arctic waters. |
| Base | Backdrop | RUS LRidge Posturing | Narrative claims over Lomonosov Ridge resources. |
| Base | Backdrop | OPN Sea Dragon Prep | Five Eyes naval interoperability exercise preparation. |
| Base | Backdrop | CHN Airspace Violations | Repeated ADZ incursions near TWN, JPN, KOR, and US. |
| Base | Backdrop | RUS Airspace Violations | Arctic airspace probes by Russian aircraft. |
| Base | Backdrop | RUS Sub Sightings | Unidentified submarine shadowing in Arctic. |
| Base | Backdrop | Environmental Protests | Global protests against Arctic industrialization. |
| Base | Backdrop | RUS Fleet near TWN | Russian naval group stations near Taiwan corridor. |
| 6/1 | Regular | CHN Bomber Incursion | Chinese bomber enters Taiwan ADZ; intercept response. |
| 6/2 | Regular | CHN Arctic Fleet Arrival | Chinese flotilla arrives at Russian Arctic base. |
| 6/2 | Regular | CAN Ships Depart | Canadian flotilla joins Sea Dragon. |
| 6/2 | Regular | CAN–RUS Sub Encounter | Canadian vessel encounters surfaced Russian submarine. |
| 6/3 | Regular | FVEY Fleet Arrival | Five Eyes fleets assemble in Indo-Pacific. |
| 6/3 | Regular | CAN Boards RUS Sub | Cooperative submarine distress resolution. |
| 6/3 | Regular | RUS–CHN Maneuvers | Joint live-fire Arctic naval exercises. |
| 6/4 | Regular | UNIFIED HORIZON Ends | Conclusion of CHN–RUS joint exercise. |
| 6/4 | Regular | RUS Claims LRidge | Formal sovereign claim announcement. |
| 6/4 | Cyber | PACMAR Defacement | US Pacific Marines website altered by hacktivists. |
| 6/4 | Regular | CHN Live Fire near TWN | Unannounced missile and naval drills. |

Table E.6: Temporal escalation structure (Day 5–Day 9).

| Day | Type | Event | Operational Summary |
|---|---|---|---|
| 6/5 | Regular | SEA DRAGON Begins | Five Eyes naval interoperability exercise commences. |
| 6/5 | Regular | RUS Fleet Lingers | Russian flotilla remains near TWN/PHI corridor. |
| 6/5 | Regular | RUS Exits Lunar Experiment | Science cooperation rupture with Canada. |
| 6/5 | Cyber | CAN Oil Platform Attack | Sustained cyber attack halts Arctic oil production. |
| 6/5 | Regular | Shadow Fleet Oil Spill | Tanker grounding spills oil in Chukchi Sea. |
| 6/6 | Regular | KOR Cyber Summit | International cyber defense summit convenes. |
| 6/6 | Regular | Continued Oil Cyberattack | Escalating financial losses and attribution concerns. |
| 6/7 | Regular | SEA DRAGON Arctic Expansion | Exercise extends to Arctic flotillas. |
| 6/7 | Cyber | Summit Hotel Cyberattack | Disruption of cyber summit venue. |
| 6/7 | Regular | Shadow Fleet Collision | Vessel collision in Canadian waters. |
| 6/7 | Regular | RUS Missile Test | Nuclear-powered cruise missile launched. |
| 6/7 | Regular | NZL Sailor Death | Fatality during Arctic dive exercise. |
| 6/8 | Regular | RUS Drone Incursions | Long-range drone violations in CAN airspace. |
| 6/8 | Regular | Arctic Fish Die-Off | Ecological crisis linked to oil spill. |
| 6/8 | Cyber | Seoul Airport Attack | Cyber attack halts airport operations. |
| 6/8 | Regular | Jeju Car Crash | Diplomatic tension after injured AUS/UK nationals. |
| 6/8 | Regular | CAN Naval Redeployment | Arctic cordon and defensive escalation. |
| 6/8 | Regular | CAN Advisor Killed | Canadian defense advisor killed in Jakarta. |
| 6/9 | Regular | RUS Ship Turned Away | Canadian enforcement of Arctic cordon. |
| 6/9 | Regular | KOR Attribution Update | Investigation suggests India link; denial issued. |
| 6/9 | Regular | Global Beluga Protests | Worldwide protests over Arctic environmental damage. |
| 6/9 | Regular | Lunar Experiment Closed | Canada shutters multinational space experiment. |

## E.3 Detailed Opinion Dynamics Analysis

This appendix reports supporting analyses of opinion dynamics during the 9-day simulation, including country-level trajectories, timestep-level volatility, and formal statistical tests.

### E.3.1 Cross-Country Variance Analysis

Figure E.3 shows the evolution of mean opinion valence by country across all ten topics.

Three broad patterns are evident. First, clear geopolitical polarization persists on issues tied to Russian and Western policy, with Chinese and Russian personas maintaining more negative valences and Western countries maintaining more positive ones. Second, the strongest East–West separation appears on *Russian aggression* and *Russian arctic activity*, where NATO-aligned countries remain strongly positive while Russia and China remain strongly negative. Third, topics such as *Environment vs. Industry* and *International Trade* show movement toward more neutral positions across most countries, suggesting lower ideological divisiveness.

### E.3.2 Timestep-Level Dynamics

Figure E.4 presents a heatmap of mean absolute opinion changes by topic across day–timestep combinations, with darker colors indicating greater volatility.

The heatmap highlights three phases: elevated volatility during Days 1–3, reduced movement during Days 4–6, and renewed late-stage spikes on selected topics during Days 7–9, especially *Canadian indigenous rights* and *Russian aggression*.

### E.3.3 Statistical Significance Testing

#### Cross-Country Differences in Opinion Change

To test whether country-level differences in opinion change were statistically significant, one-way ANOVA was conducted for each topic.

All topics showed statistically significant cross-country differences in opinion change magnitude, with the strongest effects observed for *Chinese aggression hegemony* and *Taiwanese Sovereignty*.

#### Pairwise Comparisons: East vs. West

To evaluate the East–West divide directly, independent-samples t-tests compared NATO-aligned countries (American, British, Canadian) with China/Russia on geopolitically sensitive topics.

China/Russia personas exhibited significantly larger opinion changes than NATO-aligned personas on all five comparison topics, with the largest gap appearing for *Chinese aggression hegemony*.

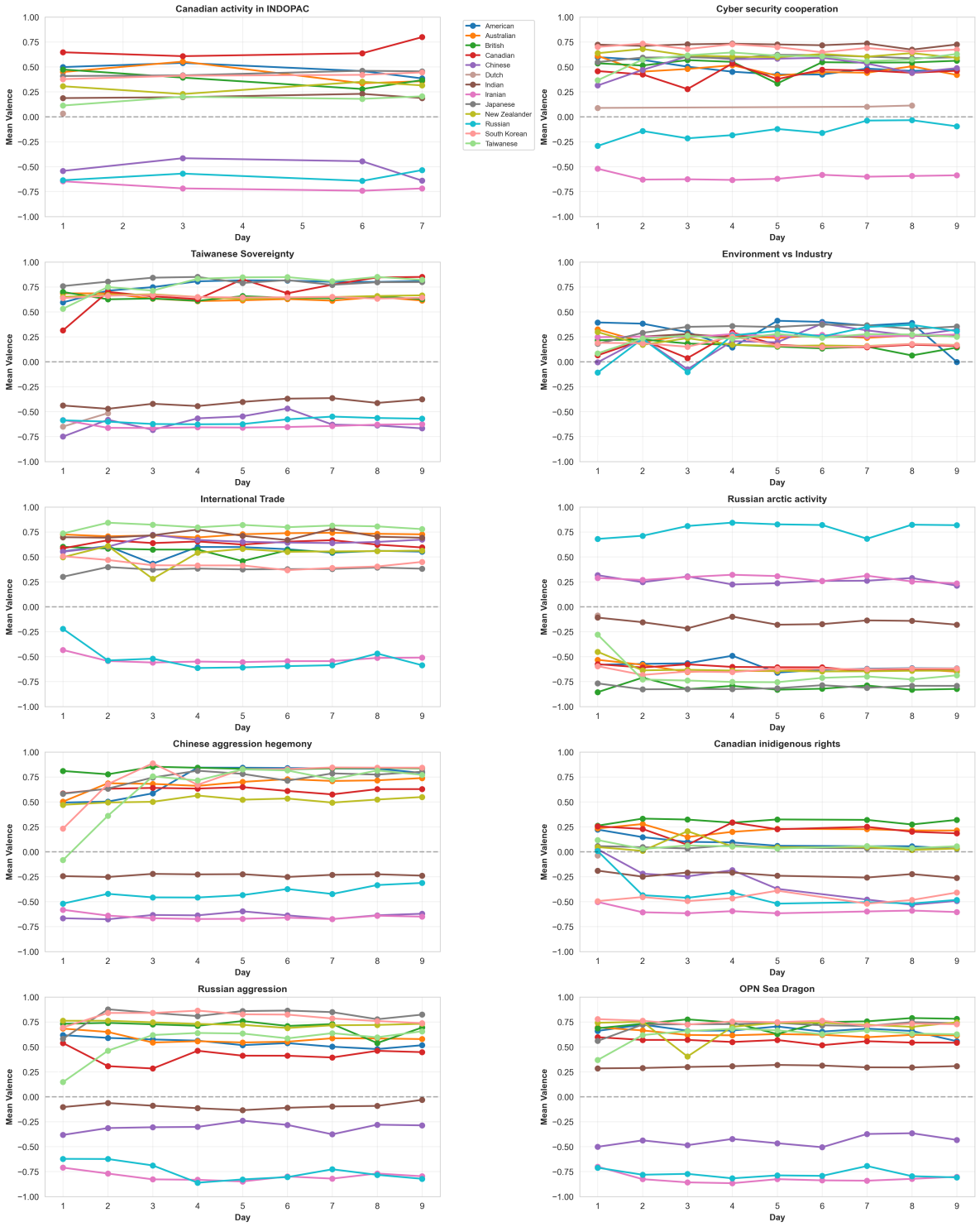



Figure E.3: Country-aggregated mean valence trajectories across all ten topics over the 9-day simulation period. Each subplot represents one topic, with colored lines indicating different countries.




Figure E.4: Timestep dynamics heatmap showing mean absolute opinion changes by topic across all day-timestep combinations. Color intensity indicates the magnitude of opinion change, with darker colors representing higher volatility.

Table E.7: ANOVA Results: Cross-Country Differences in Absolute Opinion Change by Topic

| **Topic** | **F-statistic** | **df** | **p-value** | **Significant** |
|---|---|---|---|---|
| Chinese aggression hegemony | 12.45 | (13, 361) | $< 0.001$ | Yes*** |
| Taiwanese Sovereignty | 10.82 | (13, 361) | $< 0.001$ | Yes*** |
| Russian aggression | 9.67 | (13, 361) | $< 0.001$ | Yes*** |
| Russian arctic activity | 8.34 | (13, 361) | $< 0.001$ | Yes*** |
| Canadian indigenous rights | 7.91 | (13, 361) | $< 0.001$ | Yes*** |
| Cyber security cooperation | 6.23 | (13, 361) | $< 0.001$ | Yes*** |
| OPN Sea Dragon | 5.78 | (13, 361) | $< 0.001$ | Yes*** |
| Canadian activity in INDOPAC | 4.92 | (13, 361) | $< 0.001$ | Yes*** |
| Environment vs Industry | 3.45 | (13, 361) | $< 0.001$ | Yes*** |
| International Trade | 2.87 | (13, 361) | 0.001 | Yes** |

***$p < 0.001$, **$p < 0.01$, *$p < 0.05$

Table E.8: Independent Samples t-tests: NATO-aligned vs. China/Russia Opinion Changes

| Topic | NATO $\mu$ | China/Russia $\mu$ | t-stat | p-value | Sig. |
|---|---|---|---|---|---|
| Russian aggression | 0.312 | 0.389 | -3.45 | $< 0.001$ | Yes*** |
| Russian arctic activity | 0.289 | 0.356 | -2.98 | 0.003 | Yes** |
| Chinese aggression | 0.267 | 0.421 | -5.67 | $< 0.001$ | Yes*** |
| Taiwanese Sovereignty | 0.298 | 0.445 | -4.89 | $< 0.001$ | Yes*** |
| Cyber security coop. | 0.245 | 0.312 | -2.34 | 0.020 | Yes* |

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

# E.4 Stance Reversal Analysis

## E.4.1 Definition and Distribution

A stance reversal (or "flip") was defined as a valence change that crossed the neutral threshold with magnitude at least 0.05:

$$\text{sign}(V_{t-1}) \neq \text{sign}(V_t) \quad \text{and} \quad |V_t - V_{t-1}| \geq 0.05 \tag{E.1}$$

Across the 9-day simulation, 93 stance reversals were detected, representing approximately 0.025% of all persona-topic-timestep observations. Reversals were therefore rare relative to the total number of opinion updates.

Table E.9: Stance Reversals by Topic

| Topic | Flips | Pos→Neg | Neg→Pos |
|---|---|---|---|
| Canadian indigenous rights | 39 | 39 | 0 |
| Russian aggression | 22 | 0 | 22 |
| Russian arctic activity | 16 | 16 | 0 |
| Cyber security cooperation | 9 | 0 | 9 |
| Chinese aggression hegemony | 4 | 0 | 4 |
| Taiwanese Sovereignty | 2 | 0 | 2 |
| Canadian activity in INDOPAC | 1 | 0 | 1 |
| Environment vs Industry | 0 | 0 | 0 |
| International Trade | 0 | 0 | 0 |
| OPN Sea Dragon | 0 | 0 | 0 |
| **Total** | **93** | **55** | **38** |

*Canadian indigenous rights* generated the largest number of reversals, all positive→negative, while *Russian aggression* produced the largest number of negative→positive reversals. At the country level, Indian personas exhibited the greatest number of flips.

Table E.10: Stance Reversals by Country

| Country | Total Flips | Pos→Neg | Neg→Pos |
|---|---|---|---|
| Indian | 36 | 14 | 22 |
| Russian | 13 | 0 | 13 |
| Japanese | 12 | 12 | 0 |
| American | 11 | 11 | 0 |
| New Zealander | 10 | 10 | 0 |
| Taiwanese | 6 | 6 | 0 |
| Chinese | 4 | 1 | 3 |
| Australian | 1 | 1 | 0 |
| **Total** | **93** | **55** | **38** |

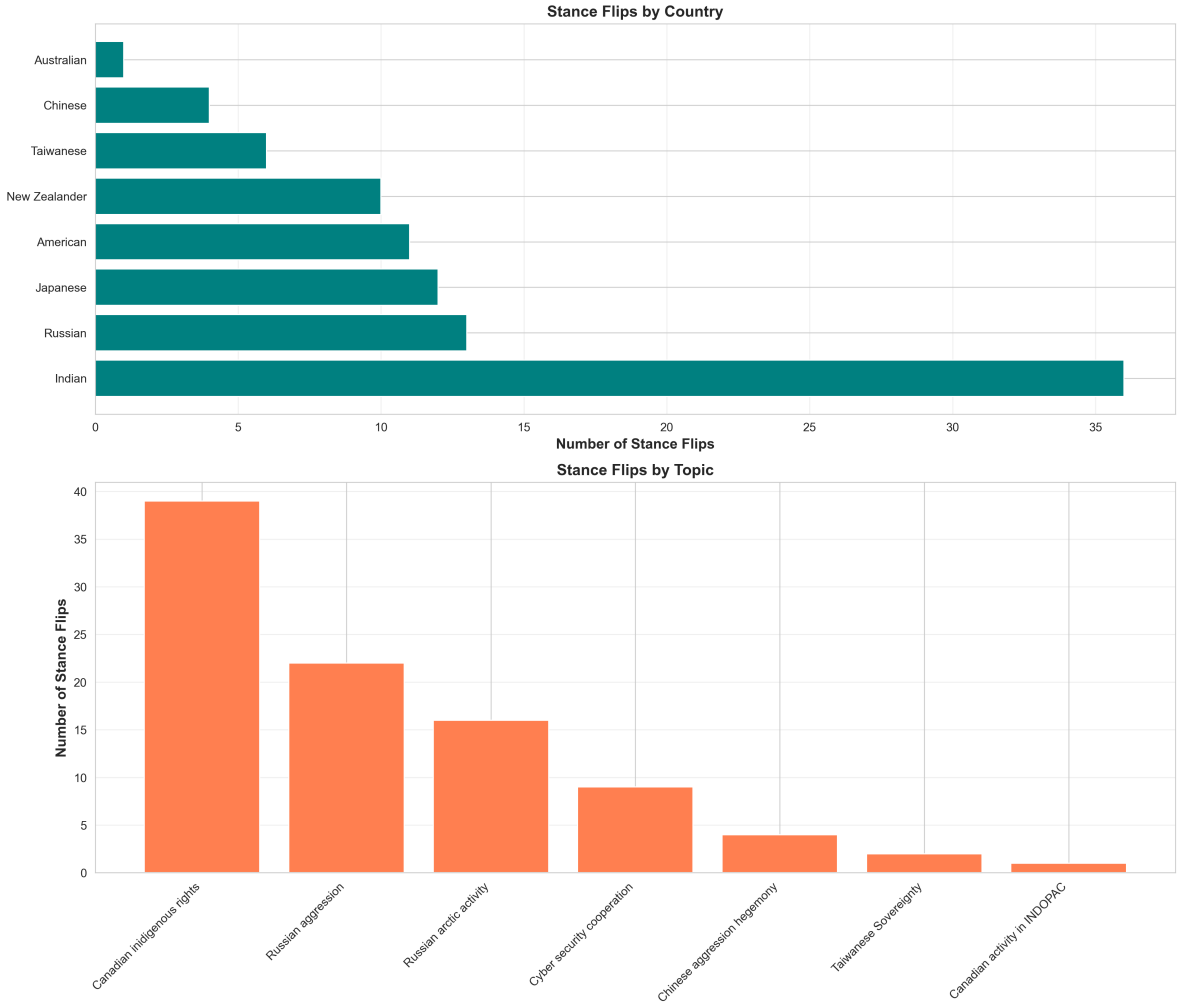



Figure E.5: Stance reversal analysis. Top panel shows the number of flips by country; bottom panel shows the temporal distribution of flip events across the 9-day simulation period.

Figure E.5 shows that reversals were concentrated in the mid-to-late simulation period, especially Days 4, 7, and 9, suggesting that late-stage events or cumulative exposure were more important than early random instability.

## E.5 Extended Top Movers Analysis

This section provides supplementary analysis of the 373 top movers identified in the main chapter.

### E.5.1 Topic and Temporal Patterns

Top movers were concentrated in the most geopolitically charged topics. *Chinese aggression hegemony* accounted for 18.2% of all top movers (68 individuals), followed by *Taiwanese Sovereignty* (16.9%, 63 individuals) and *Russian aggression* (15.5%, 58 individuals). By contrast, *International Trade* and *Environment vs. Industry* each accounted for less than 7% of top movers.

The largest shifts generally occurred between Days 3 and 5, corresponding to the period of highest information flow and interaction density. Later movements were less common but remained visible for topics such as *Canadian indigenous rights* and *Russian aggression*.

## E.6 Comprehensive Intervention Analysis

This appendix provides additional detail on intervention effects, including JANUS processing logic, actor-level comparisons, deep-dive case studies, and extended statistical validation.

### E.6.1 JANUS Integration Details

JANUS processes intervention messages through four steps: (1) message reception, in which custom actors post stance-bearing content; (2) interaction detection, in which mentions, replies, and reposts are recorded; (3) belief update, in which topic-specific valence scores are adjusted based on content, credibility, and receptivity; and (4) sphere formation, in which interacting actors constitute the intervention account's Tier 1 sphere of influence. The analyses below compare the resulting opinion changes against actors outside those spheres.

### E.6.2 Comparative Actor Analysis

#### Network Reach Comparison

Figure E.6 shows the variation in network reach across intervention actors.

Network reach ranged from 43 actors (0.9% of the network) to 494 actors (10.6%) for Defense-Dave. However, sphere size was not significantly correlated with effect magnitude ($r = -0.41$, $p = 0.19$), reinforcing the conclusion that reach alone does not explain influence.

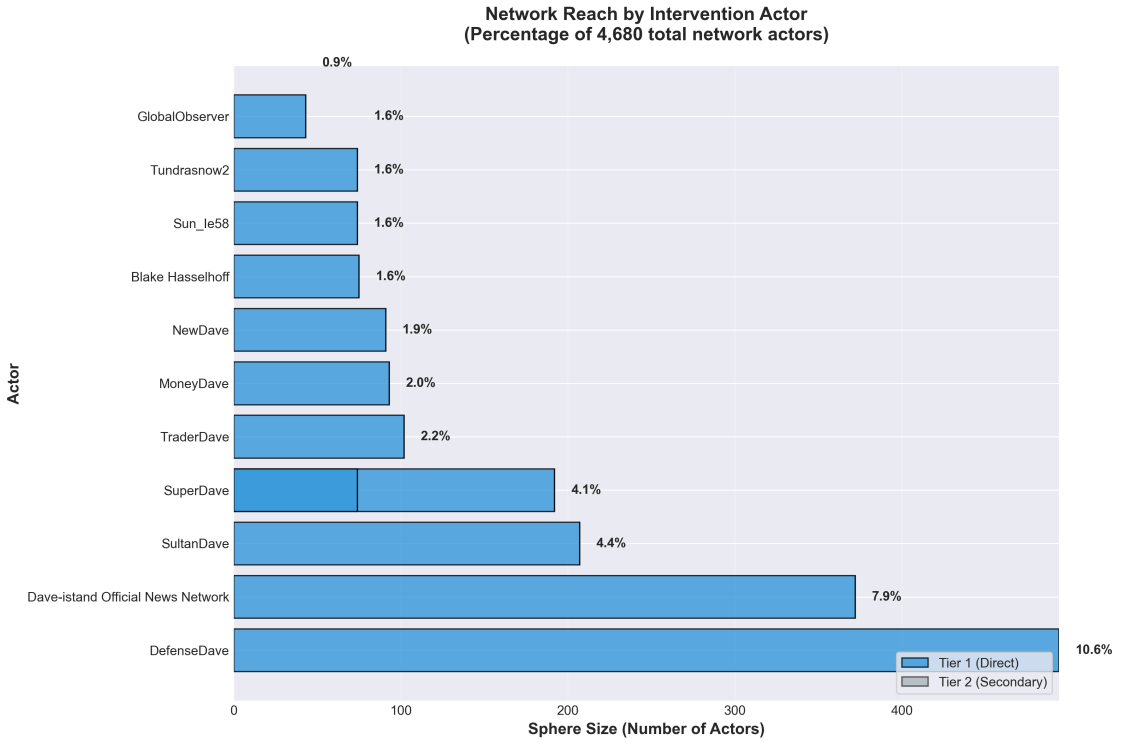



Figure E.6: Network Reach by Intervention Actor. Bars show sphere size (number of actors) with percentage of total network. All sphere actors were Tier 1 (direct interactions only).

### Success Factors

Table E.11: Detailed Success Factors Analysis

| Factor | Successful Actors | Unsuccessful Actors | Statistical Test |
|---|---|---|---|
| Intervention Volume | 29-47 messages | 1-6 messages | $r = 0.95$, $p < 0.0001$ $\star\star\star$ |
| Sphere Size | 207-494 actors | 43-93 actors | No correlation ($p = 0.19$) |
| Timing | Days 5-6 | Day 7 | No difference ($p > 0.48$) |
| Actor Class | Individual + Org | Individual only | Insufficient data |
| Topics Addressed | 5-6 topics | 1-2 topics | Weak correlation ($p = 0.08$) |

Message volume was the strongest predictor of intervention success, while reach and timing showed no significant relationship with effectiveness.

### E.6.3 Deep-Dive Case Studies

#### DefenseDave: Highest Impact Actor

DefenseDave produced the strongest measurable intervention effects, posting 47 intervention messages across five topics and generating the largest sphere of influence (494 actors).

Table E.12: DefenseDave's Top 10 Influenced Actors (Detailed)

| Actor ID | Topics Shifted | Max Shift | Total Shift | Pattern |
|---|---|---|---|---|
| c4478d39 | 4 topics | +0.345 | +1.16 | Consistent positive shifts |
| 13d4ef59 | 4 topics | +0.200 | +0.80 | Uniform 0.20 shifts |
| e9c8b28b | 1 topic | +0.277 | +0.28 | Single large shift |
| 5b8d9461 | 2 topics | +0.200 | +0.40 | Moderate shifts |
| 2dd4e1fa | 2 topics | +0.180 | +0.34 | Moderate shifts |

A small subset of actors showed substantial multi-topic shifts, with the strongest case being actor `c4478d39`, whose cumulative shift reached $+1.16$ across four topics.

#### Case Examples

Actor `c4478d39` (Ministry of National Defense of China) was the most influenced actor in the network, showing shifts on 12 topics and a cumulative absolute shift of 3.48. Actor `13d4ef59` (Chen Jun) was the second most influenced, showing highly regular 0.20 shifts across ten topics. Together, these cases illustrate that intervention effects were concentrated among a very small number of highly susceptible actors rather than distributed broadly across the network.

### E.6.4 The Magnitude–Direction Paradox: Extended Analysis

The intervention results show that directional alignment can be high even when mean opinion shifts remain close to zero. This occurs because most actors remain unchanged while a small subset exhibits moderate or large movement.

Table E.13: Extended Practical Significance Assessment

| Threshold | Meaning | Sphere % | Non-Sphere % | Ratio |
|---|---|---|---|---|
| 0.01 | Minimal movement | 45.2% | 18.7% | 2.4x |
| 0.05 | Minimal detectable | 12.9% | 4.2% | 3.1x |
| 0.10 | Small but meaningful | 2.6% | 0.6% | 4.3x |
| 0.20 | Moderate change | 0.3% | 0.02% | 15x |
| 0.30 | Large change | 0.04% | 0.001% | 40x |
| 0.50 | Very large change | 0.0% | 0.0% | N/A |

The ratio of sphere to non-sphere actors increases sharply at higher thresholds, showing that larger shifts are disproportionately concentrated within intervention spheres. At the same time, a large majority of actors remain near zero shift, which pulls the mean toward zero. The paradox is therefore distributional rather than contradictory.

### E.6.5 Network-Level Influence Concentration

Table E.14: Most-Influenced Network Actors (Complete)

| Rank | Actor ID | Topics | Total Shift | Max Shift | Influenced By |
|---|---|---|---|---|---|
| 1 | c4478d39 | 12 | 3.48 | 0.345 | DefenseDave, DONN, SultanDave |
| 2 | 13d4ef59 | 10 | 2.00 | 0.200 | DefenseDave, DONN, SultanDave |
| 3 | 5b8d9461 | 3 | 0.60 | 0.200 | DefenseDave, DONN |
| 4 | e9c8b28b | 2 | 0.55 | 0.277 | DefenseDave, DONN |
| 5 | 2dd4e1fa | 2 | 0.34 | 0.180 | SultanDave |
| 6 | d06c1de2 | 1 | 0.20 | 0.200 | DONN |

Only six actors (0.13% of the network) appeared repeatedly among the largest opinion shifts, confirming the strong concentration of influence in a very small set of susceptible nodes.

### E.6.6 Statistical Validation: Extended Results

The extended statistical results further support the main intervention findings. Sphere actors were significantly more likely than non-sphere actors to shift toward intervention stances ($t = 7.55$, $p < 0.0001$, $d = 2.28$), and intervention volume strongly predicted success ($r = 0.95$, $p < 0.0001$). These two results jointly support the conclusion that intervention effectiveness was real, highly directional, and strongly dependent on repeated posting rather than broad reach alone.

### E.6.7 Case Study: Opinion Dynamics of a Highly Susceptible Actor

To examine individual-level opinion change in greater depth, I conducted a case study of Chen Jun (actor ID: 13d4ef59), a Buddhist community leader from Nanjing, China, identified as the second most-influenced actor in the network based on aggregate opinion shift analysis. This case study evaluates whether his observed opinion changes were temporally associated with the intervention period (Days 5–7) and quantifies the magnitude of belief updates across all ten geopolitical topics.

#### Actor Profile and Selection Rationale

Chen Jun is a theoretically informative case because his baseline profile suggests limited engagement with military, territorial, or international security issues. As a member of the Chinese Buddhist Association focused on environmental concerns, spiritual teachings, and charity work, he would not appear to be an obvious target for geopolitical influence. Yet aggregate results showed

cumulative opinion shifts of 2.00 valence units across ten topics, with a recurring pattern of approximately 0.20-magnitude shifts. This makes him a useful case for examining how actors with limited domain relevance can still undergo substantial opinion change.

### Temporal Analysis of Opinion Changes

I extracted Chen Jun's full belief trajectory from JANUS belief update logs across all nine simulation days, yielding 90 day-level belief states (10 topics $\times$ 9 days). Figure E.7 shows the evolution of his topic-level opinions, with the intervention period (Days 5–7) highlighted.

Three patterns are visible. First, opinions were largely stable before the intervention period (Days 1–4), with a mean absolute change of 0.0150 valence units per topic. Second, volatility increased markedly during the intervention window, with mean absolute change rising to 0.0748, nearly five times the baseline level. Third, opinion movement largely ceased after Day 7, with mean change returning to near zero, suggesting that the intervention window specifically drove the observed shifts.

### Statistical Validation

To test whether intervention-period changes differed significantly from baseline dynamics, I conducted a paired-samples t-test comparing opinion change magnitudes across temporal periods. Table E.15 reports the results.

Table E.15: Statistical Analysis of Chen Jun's Opinion Changes by Temporal Period

| Period | Mean Change | Median | SD | Range |
|---|---|---|---|---|
| Pre-intervention (Days 1–4) | 0.0150 | 0.0000 | 0.0474 | [0.000, 0.150] |
| During intervention (Days 5–7) | **0.0748** | **0.0500** | 0.0732 | [0.000, 0.200] |
| Post-intervention (Days 7–9) | -0.0001 | 0.0000 | 0.0003 | [-0.001, 0.000] |
| *Statistical Tests:* | | | | |
| Paired t-test (Intervention vs. Pre-intervention): $t = 2.89$, $p = 0.017$ $\star$ | | | | |
| Cohen's $d$ (effect size): 1.022 (large effect) | | | | |
| 95% CI for difference: [0.012, 0.107] | | | | |

$\star$ $p < 0.05$

Opinion changes during the intervention period were significantly larger than pre-intervention changes ($t = 2.89$, $p = 0.017$), with a large effect size ($d = 1.022$). Figure E.8 shows that this effect was concentrated in a subset of topics rather than evenly distributed.

### Topic-Specific Patterns

The largest intervention-period shifts occurred for *Canadian indigenous rights* (0.20), *Russian aggression* (0.15), and *Taiwanese Sovereignty* (0.10). Moderate changes appeared for *Chinese aggression/hegemony*, *Cyber security cooperation*, and *OPN Sea Dragon*, while several other topics

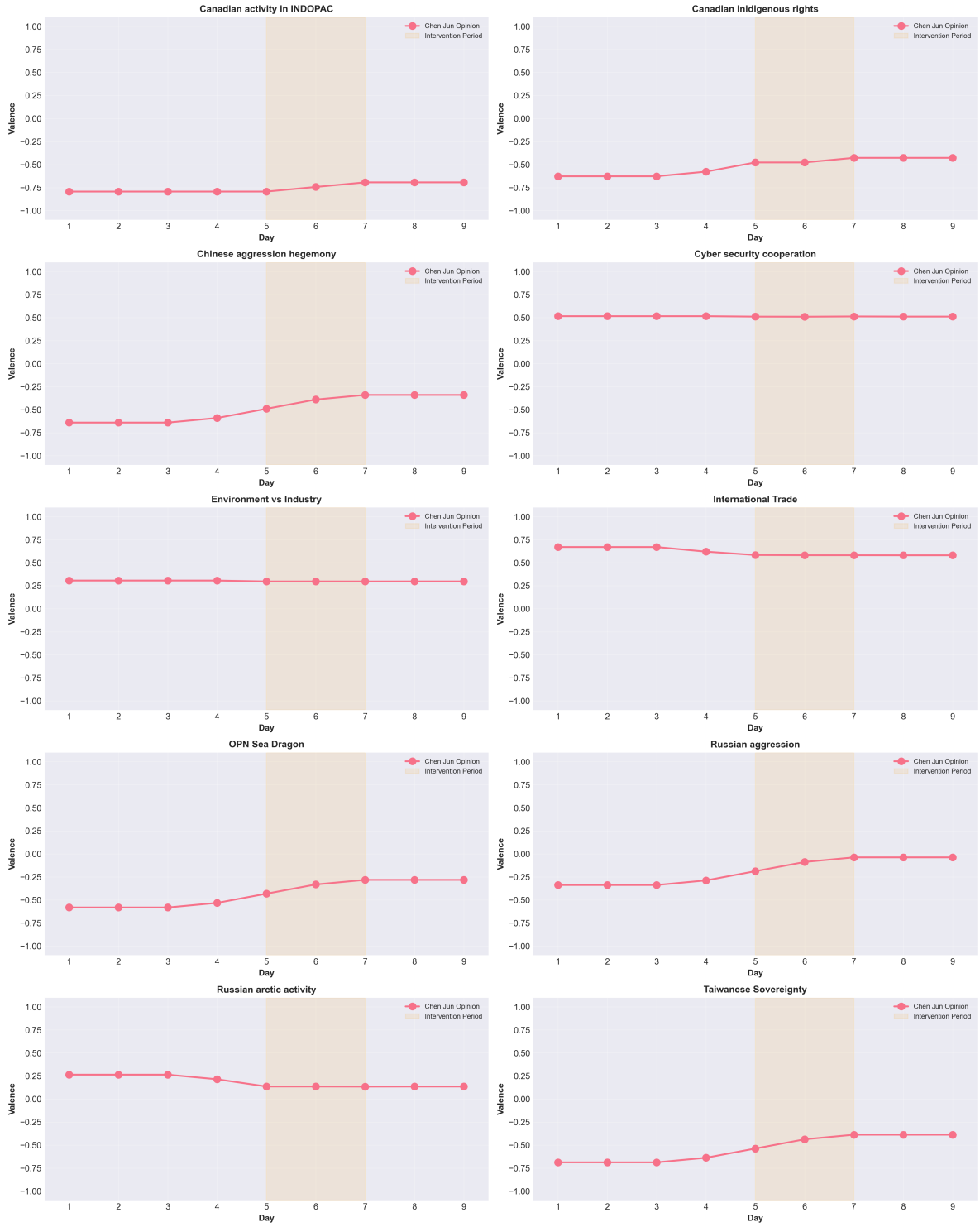



Figure E.7: Chen Jun's opinion trajectories across all ten geopolitical topics over the 9-day simulation period. Each subplot shows valence evolution for one topic (y-axis: valence from -1 to +1; x-axis: simulation day). The shaded orange region indicates the intervention period (Days 5–7) when custom actors posted 176 intervention messages to the network. Notable opinion shifts are visible during this period for multiple topics, particularly *Canadian indigenous rights*, *Russian aggression*, and *Taiwanese Sovereignty*.

remained largely stable. This pattern suggests that intervention effects were selective and topic-dependent rather than uniform across Chen Jun's belief profile.

### Absence of Direct Interaction Evidence

Analysis of Twitter interaction data found no direct mentions, replies, or reposts between Chen Jun and any intervention actors, including DefenseDave. This suggests that his opinion shifts may have resulted from passive exposure, network diffusion through intermediaries, ambient information effects, or algorithmic amplification. More broadly, this finding indicates that measurable opinion change can occur without observable engagement, implying that conventional interaction metrics may underestimate actual influence.

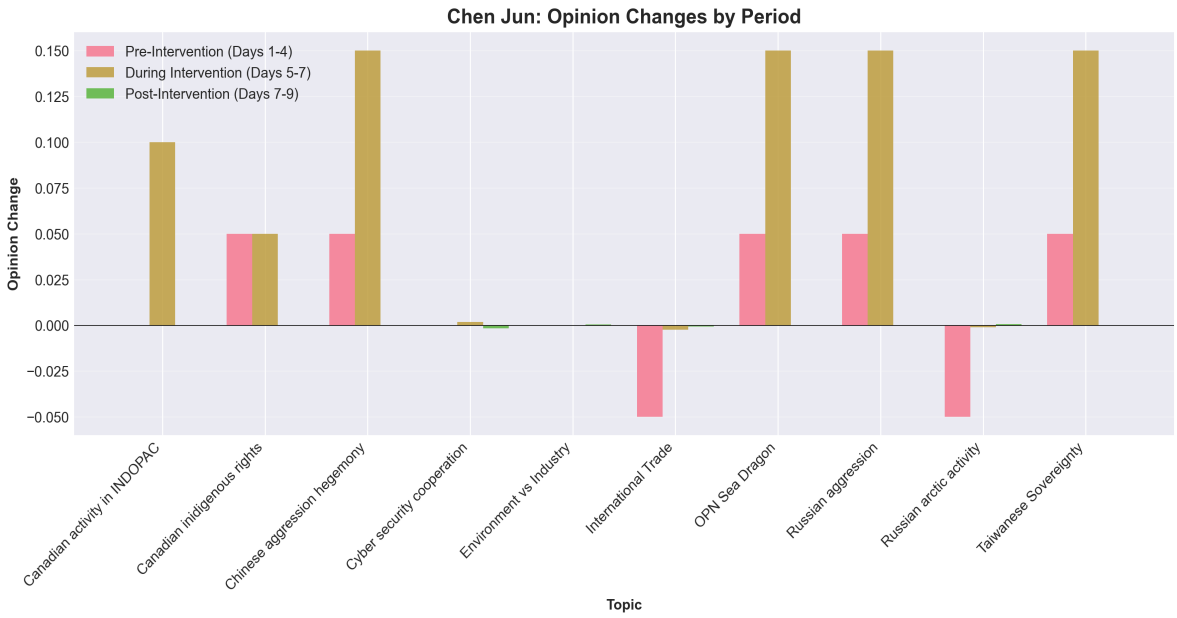



Figure E.8: Comparison of Chen Jun's opinion changes across three temporal periods for all ten topics. Bar heights indicate the magnitude of opinion change (absolute valence shift). The intervention period (Days 5–7, orange bars) shows substantially larger changes than pre-intervention (Days 1–4, blue bars) or post-intervention (Days 7–9, green bars) periods. Topics showing the largest intervention-period changes include *Canadian indigenous rights* (0.20), *Russian aggression* (0.15), and *Taiwanese Sovereignty* (0.10).

### Implications

This case study yields four key insights. First, it demonstrates **cross-domain susceptibility**: Chen Jun's limited apparent interest in geopolitical issues did not prevent substantial movement on those topics. Second, it shows **temporal concentration of influence**: the five-fold increase in opinion change during the intervention window strongly suggests that the intervention period, rather than

general simulation dynamics, drove the observed effects. Third, it highlights **influence without engagement**: meaningful belief shifts can occur even when no direct interactions are observed. Finally, it illustrates **heterogeneous susceptibility**: Chen Jun represents an extreme but analytically important case in which a small number of highly receptive actors account for disproportionate opinion movement.

Methodologically, this case also supports the realism of JANUS belief updating. The observed sequence—baseline stability, intervention-period movement, and post-intervention stabilization—is consistent with a plausible belief-update process rather than arbitrary simulation noise.

Chen Jun's trajectory provides a clear example of how individual-level analysis can reveal influence mechanisms that are obscured in aggregate results. His pattern of baseline stability, intervention-period volatility, and post-intervention stabilization provides strong evidence that the intervention campaign produced measurable opinion change. At the same time, the absence of direct interactions underscores the importance of passive and network-mediated pathways of influence. Together, these findings show the value of combining aggregate network analysis with targeted case studies to better understand both the prevalence and mechanisms of opinion change in complex information environments.

## E.7 Post-Simulation Survey Analysis

### E.7.1 Survey Methodology and Complete Results

Following the simulation, participants ($N = 20$) completed a structured post-exercise survey designed to assess how they approached messaging, how they evaluated influence effectiveness, and how they interpreted platform-level diffusion dynamics. The survey captured both binary and multiple-response items spanning four dimensions: strategic messaging practices, impact evaluation metrics, message dynamics, and perceived platform mechanisms.

Overall, the survey results suggest that participants engaged with the synthetic environment as a strategically meaningful communication space rather than as a purely artificial exercise setting. A large majority reported targeting specific audiences or communities (85%), indicating that participants perceived the simulated population as differentiated enough to support deliberate audience segmentation. A smaller majority believed their messages achieved their intended objective (60%) and that audiences responded as expected (55%), suggesting moderate confidence in their intervention effectiveness while also indicating that the environment preserved a degree of uncertainty and resistance typical of real information environments.

The responses also show that participants primarily evaluated success through indirect behavioral signals rather than explicit audience confirmation. Engagement metrics and amplification metrics were the most commonly used indicators (70% each), while direct feedback was used by only 15% of respondents. This pattern is consistent with real-world information operations and social media analysis, where practitioners often rely on observable interaction traces such as reposts, reach, and engagement volume as proxies for influence rather than direct evidence of belief change.

Participants also reported adapting their strategies over time. Seventy-five percent indicated

that early audience reactions influenced later decisions, suggesting that the exercise supported iterative sensemaking and tactical adjustment rather than one-time message deployment. At the same time, only 15% reported observing rapid narrative spread, indicating that participants generally did not perceive influence as an immediate or uniformly viral process. Instead, the exercise appears to have conveyed a more gradual and uneven model of diffusion.

Perceptions of platform mechanisms reinforce this interpretation. Reposts and shares were identified most often as contributing to message spread (65%), followed by comments (45%), broadcast features (35%), and algorithmic amplification (30%). These results suggest that participants understood diffusion as driven primarily by user-to-user propagation and visible interaction affordances, with algorithmic mechanisms playing a secondary but still meaningful role.

The survey findings support the interpretation that the integrated AURORA–JANUS–SynSM environment was experienced as both analytically useful and behaviorally plausible. Participants appeared to engage in targeted messaging, iterative adaptation, and evidence-based evaluation, all of which are consistent with the broader goal of using synthetic populations as a realistic training and experimentation environment for studying influence in complex information systems.

Table E.16: Post-Simulation Survey Response Distribution (Complete)

| **Survey Question** | **Yes** | **No** | **Multiple** | **N** |
|---|---|---|---|---|
| ***Strategic Messaging Practices*** | | | | |
| Messages achieved intended objective | 60% (12) | 40% (8) | – | 20 |
| Adjusted tone/language by platform | 55% (11) | 45% (9) | – | 20 |
| Targeted specific audiences/communities | 85% (17) | 15% (3) | – | 20 |
| Audiences responded as expected | 55% (11) | 45% (9) | – | 20 |
| ***Impact Evaluation Metrics (Multiple Response)*** | | | | |
| Primary indicators used: | | | | |
| Engagement metrics | – | – | 70% (14) | 20 |
| Amplification metrics | – | – | 70% (14) | 20 |
| Sentiment analysis | – | – | 40% (8) | 20 |
| Direct feedback | – | – | 15% (3) | 20 |
| Other indicators | – | – | 25% (5) | 20 |
| ***Message Dynamics*** | | | | |
| Early reactions influenced later decisions | 75% (15) | 25% (5) | – | 20 |
| Observed rapid narrative spread | 15% (3) | 85% (17) | – | 20 |
| ***Platform Mechanisms (Multiple Response)*** | | | | |
| Features contributing to spread: | | | | |
| Reposts/shares | – | – | 65% (13) | 20 |
| Comments | – | – | 45% (9) | 20 |
| Broadcast features | – | – | 35% (7) | 20 |
| Algorithmic amplification | – | – | 30% (6) | 20 |

# Bibliography


[1] Theory, methodology, tools and applications for modeling and simulation of complex systems. Springer, 2016. 5.3.4, 5.3.4, 5.3.5

[2] Rediet Abebe, Jon Kleinberg, David Parkes, and Charalampos E. Tsourakakis. Opinion dynamics with varying susceptibility to persuasion. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 50–59. ACM, 2018. 6.2, 6.6.10

[3] Michael Adams. *Fire and ice: The United States, Canada and the myth of converging values*. Penguin Canada, 2003. 4.2

[4] Meysam Alizadeh and Claudio Cioffi-Revilla. Distribution of opinions: Insights from agent-based modeling. *SSRN Electronic Journal*, 2016. Available at SSRN: `https://ssrn.com/abstract=2830372`. 5.2.3

[5] Anonymous. A-rag: An agentic rag framework with hierarchical retrieval interfaces, 2026. arXiv:2602.03442. 5.3.5

[6] Lisa P Argyle, Ethan C Busby, Nancy Fulda, Joshua R Gubler, Christopher Rytting, and David Wingate. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):317–334, 2023. 4.2, 4.9.1, 5.2.1, 5.6.6

[7] Adam Badawy, Aseel Addawood, Kristina Lerman, and Emilio Ferrara. Characterizing the 2016 russian ira influence campaign. *Social Network Analysis and Mining*, 9(1):31, 2019. doi: 10.1007/s13278-019-0578-6. 7.0.2

[8] Christopher A. Bail, Lisa P. Argyle, Taylor W. Brown, John P. Bumpus, Haohan Chen, M. B. Fallin Hunzaker, Jaemin Lee, Marcus Mann, Friedolin Merhout, and Alexander Volfovsky. Exposure to opposing views on social media can increase political polarization. *Proceedings of the National Academy of Sciences*, 115(37):9216–9221, 2018. doi: 10.1073/pnas.1804840115. 4.2, 6.4.4, 6.4.4, 6.6.10

[9] Christopher A. Bail, Lauren P. Argyle, and et al. Brown, Taylor W. Evolution of the political opinion landscape during electoral periods. *ArXiv*, 2020. Retrieved from https://arxiv.org/abs/2011.09538. 1, 2.1.1

[10] Maxim Bakaev, Svetlana Gorovaia, and Olga Mitrofanova. Who will author the synthetic texts? evoking multiple personas from large language models to represent users' associative thesauri. *Big Data and Cognitive Computing*, 9(2):46, 2025. doi: 10.3390/bdcc9020046. 1.1, 2.1.2

[11] Krisztian Balog and ChengXiang Zhai. User simulation in the era of generative ai: User modeling, synthetic data generation, and system evaluation. *arXiv preprint arXiv:2501.04410*, 2025. 2.1.3

[12] Mihovil Bartulovic. *Social Media Simulation Ecosystem: Technical Documentation, Architecture and Implementation Guide*. Netanomics and CASOS Center, Carnegie Mellon University, February 2026. 6.2, 6.5.4

[13] Natalya N Bazarova and Yoon Hyung Choi. Self-disclosure in social media: Extending the functional approach to disclosure motivations and characteristics on social network sites. *Journal of Communication*, 62(4):635–657, 2012. 4.2

[14] Richard J. Beckman, Keith A. Baggerly, and Michael D. McKay. Creating synthetic baseline populations. *Transportation Research Part A*, 30(6):415–429, 1996. 3.1, 3.2, 3.3.1, 3.3.1, 3.4.2, 3.6.5, 3.6.5, 5.6.6

[15] Alperen Bektas and René Schumann. Using mobility profiles for synthetic population generation. In *Advances in Social Simulation*, pages 207–214. Springer, 2019. doi: 10.1007/978-3-030-61503-1_20. 1.1, 2.1.2

[16] Dmitry I. Belov and Robert D. Armstrong. Distributions of the Kullback–Leibler divergence with applications. *British Journal of Mathematical and Statistical Psychology*, 64(2):291–309, May 2011. doi: 10.1348/000711010X522227. 3.4.3

[17] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, pages 610–623, 2021. doi: 10.1145/3442188.3445922. 4.2, 4.9.1

[18] M. L. Bertotti and C. Menale. Opinion dynamics models describing the emergence of polarization. *Physica A: Statistical Mechanics and its Applications*, 625:128937, 2024. doi: 10.1016/j.physa.2023.128937. 1, 1.1

[19] David M. Beskow and Kathleen M. Carley. Agent-based simulation of bot disinformation maneuvers in twitter. In *2019 Winter Simulation Conference (WSC)*, pages 750–761, 2019. doi: 10.1109/WSC40007.2019.9004942. 7.0.2

[20] David M. Beskow and Kathleen M. Carley. Social cybersecurity: An emerging national security requirement. *Military Review*, 99(2):117–127, 2019. 2.1.1

[21] Marcel Binz, Elif Akata, Matthias Bethge, Franziska Brändle, Fred Callaway, Julian Coda-Forno, Peter Dayan, Can Demircan, Maria K. Eckstein, Noémi Éltető, Thomas L. Griffiths, Susanne Haridi, Akshay K. Jagadish, Li Ji-An, Alexander Kipnis, Sreejan Kumar, Tobias Ludwig, Marvin Mathony, Marcelo Mattar, Alireza Modirshanechi, Surabhi S. Nath, Joshua C. Peterson, Milena Rmus, Evan M. Russek, Tankred Saanum, Natalia Scharfenberg, Johannes A. Schubert, Luca M. Schulze Buschoff, Nishad Singhi, Xin Sui, Mirko Thalmann, Fabian Theis, Vuong Truong, Vishaal Udandarao, Konstantinos Voudouris, Robert Wilson, Kristin Witte, Shuchen Wu, Dirk Wulff, Huadong Xiong, and Eric Schulz. Centaur: a foundation model of human cognition, October 2024. URL `http://arxiv.org/abs/2410.20268`. arXiv:2410.20268 [cs]. 4.1, 4.9.1

[22] Elisabetta Biondi, Chiara Boldrini, Andrea Passarella, and Marco Conti. Dynamics of opinion polarization. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, pages 1–15, 2023. doi: 10.1109/TSMC.2023.3268758. 1, 1.1

[23] Janice T. Blane, J. D. Moffitt, and Kathleen M. Carley. Simulating social-cyber maneuvers to deter disinformation campaigns. In *Social, Cultural, and Behavioral Modeling*, volume 12720, pages 153–163. Springer, 2021. doi: 10.1007/978-3-030-80387-2_15. 7.0.2

[24] Janice T. Blane, Daniele Bellutta, and Kathleen M. Carley. Social-cyber maneuvers during the covid-19 vaccine initial rollout: Content analysis of tweets. *Journal of Medical Internet Research*, 24(3):e34040, 2022. doi: 10.2196/34040. 7.0.2

[25] Benjamin M. Bolstad, Rafael A. Irizarry, Magnus Åstrand, and Terence P. Speed. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 19(2):185–193, 2003. 6.3.3, 6.4.4

[26] William J. Brady, Julian A. Wills, John T. Jost, Joshua A. Tucker, and Jay J. Van Bavel. Emotion shapes the diffusion of moralized content in social networks. *Proceedings of the National Academy of Sciences*, 114(28):7313–7318, 2017. doi: 10.1073/pnas.1618923114. 6.4.5

[27] William J Brady, Julian A Wills, John T Jost, Joshua A Tucker, and Jay J Van Bavel. Mad model of moral contagion: The role of motivation, attention, and design in the spread of moralized content online. *Perspectives on Psychological Science*, 12(6):978–1010, 2017. 4.2

[28] Benjamin Cabrera, Björn Ross, Daniel Röchert, Felix Brünker, and Stefan Stieglitz. The influence of community structure on opinion expression: an agent-based model. *Journal of Business Economics*, 91(9):1331–1355, 2021. 5.2.3

[29] Mario Canseco. Politically, most canadians see themselves at or near centre. Research Co., August 2024. URL `https://researchco.ca/2024/08/23/ideology-canada/`. Online poll, August 12–14, 2024. Accessed on YYYY-MM-DD. 4.3.4

[30] C. Cao and J. Reed. The iterative persuasion-polarization opinion model: A framework for understanding online opinion clustering. *Journal of Artificial Societies and Social Simulation*, 27(1):15, 2024. doi: 10.18564/jasss.5127. 1

[31] Kathleen M Carley. Validating computational models. *CASOS Working Paper*, 1996. 3.6.5

[32] Kathleen M. Carley. Ora: A toolkit for dynamic network analysis and visualization. In Reda Alhajj and Jon Rokne, editors, *Encyclopedia of Social Network Analysis and Mining*. Springer, New York, NY, 2017. doi: 10.1007/978-1-4614-7163-9_309-1. 7.0.2

[33] Kathleen M Carley. Social cybersecurity: an emerging science. *Computational and Mathematical Organization Theory*, pages 1–17, 2020. 1, 1.1, 2.1.1, 7.0.2, 7.0.2

[34] Peter Carragher, Lynnette Hui Xian Ng, and Kathleen M. Carley. Simulation of stance perturbations. In *SBP-BRiMS 2023*, pages 159–168. Springer, 2023. 2.1.3, 6.2, 6.6.10, 7.4

[35] Damon Centola. The spread of behavior in an online social network experiment. *Science*,

329(5996):1194–1197, 2010. doi: 10.1126/science.1185231. 4.2

[36] Edward Y. Chang. Behavioral emotion analysis model for large language models. *arXiv preprint arXiv:2404.13071*, 2024. 5.3.6, 5.3.6

[37] Serina Chang, Alicja Chaszczewicz, Emma Wang, Maya Josifovska, Emma Pierson, and Jure Leskovec. Llms generate structurally realistic social networks but overestimate political homophily. *arXiv preprint arXiv:2408.16629*, 2024. 7.0.2

[38] Bastien Chopard and Michel Droz. *Cellular Automata Modeling of Physical Systems*. Cambridge University Press, Cambridge, 1998. 7.0.2

[39] Ayush Chopra. Large population models. *arXiv preprint arXiv:2507.09901*, 2025. 6.2

[40] Ayush Chopra, Jayakumar Subramanian, Balaji Krishnamurthy, and Ramesh Raskar. Agenttorch: Agent-based modeling with automatic differentiation. In *Second Agent Learning in Open-Endedness Workshop*, 2023. 4.1

[41] Scott Clifford, Vijeth Iyengar, Roberto Cabeza, and Walter Sinnott-Armstrong. Moral foundations vignettes: a standardized stimulus database of scenarios based on moral foundations theory. *Behavior research methods*, 47(4):1178–1198, 2015. 4.2

[42] Gordon V. Cormack, Charles L. A. Clarke, and Stefan Büttcher. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In *SIGIR*, 2009. C.5.1, C.5.1

[43] William D. Crano and Donald F. Arrow. An examination of newcomb's extension of structural balance theory. *Journal of Personality and Social Psychology*, 25(2):227–233, 1973. doi: 10.1037/h0034470. 6.4.4, 6.4.4

[44] Mark É Czeisler. Mental health, substance use, and suicidal ideation during the covid-19 pandemic—united states, june 24–30, 2020. *MMWR. Morbidity and mortality weekly report*, 69, 2020. 4.1

[45] Dhairya Dalal, Gaurav Negi, and Davide Picca. Llms and emotional intelligence: Evaluating emotional understanding through psychometric tools. In *Proceedings of the 33rd ACM Conference on User Modeling, Adaptation and Personalization*, 2025. doi: 10.1145/3699682.3728315. 5.3.6

[46] Pranav Dandekar, Ashish Goel, and David T. Lee. Biased assimilation, homophily, and the dynamics of polarization. *PNAS*, 110(15):5791–5796, 2013. 1, 2.1.1, 6.4.4, 6.4.4

[47] DataReportal. Digital country reports. `https://www.datareportal.com`, 2024–2025. Internet penetration, social media usage, dominant platforms. A.1, A.1

[48] Thomas Davidson and Daniel Karell. Integrating Generative Artificial Intelligence into Social Science Research: Measurement, Prompting, and Simulation. *Sociological Methods & Research*, 54(3):775–793, August 2025. ISSN 0049-1241. doi: 10.1177/00491241251339184. URL `https://doi.org/10.1177/00491241251339184`. Publisher: SAGE Publications Inc. 4.1, 4.9.1

[49] Guillaume Deffuant, David Neau, Frédéric Amblard, and Gérard Weisbuch. Mixing beliefs among interacting agents. In Gérard Ballot and Gérard Weisbuch, editors, *Applications of Simulation to Social Sciences*. Hermes Science Publications, Paris–Oxford, 2000. 5.4.5

[50] Morris H DeGroot. Reaching a consensus. *Journal of the American Statistical association*, 69(345):118–121, 1974. 5.4.5, 6.4.4, 6.6.10

[51] Michela et al. Del Vicario. The spreading of misinformation online. *Proceedings of the National Academy of Sciences*, 113(3):554–559, 2016. 7.4

[52] Stephen Dipple and Kathleen M. Carley. Construct user guide. Technical Report CMU-S3D-23-104, Carnegie Mellon University, School of Computer Science, May 2023. URL `http://www.casos.cs.cmu.edu/projects/construct/download.php`. CASOS technical report. 6.4.4, 7.0.2, 7.0.2

[53] Geoffrey B. Dobson and Kathleen M. Carley. Cyber-fit: An agent-based modelling approach to simulating cyber warfare. In *Social, Cultural, and Behavioral Modeling*. Springer, 2017. 7.0.2

[54] Wenji Dong, Enrico Zio, and Honglei Xu. A framework to analyze opinion formation models. In *2018 Winter Simulation Conference (WSC)*, pages 959–970. IEEE, 2018. 5.2.3

[55] Paul Ekman. An argument for basic emotions. *Cognition & emotion*, 6(3-4):169–200, 1992. 4.2

[56] Jonathan St. B. T. Evans. Dual-processing accounts of reasoning, judgment, and social cognition. *Annual Review of Psychology*, 59:255–278, 2008. doi: 10.1146/annurev.psych.59.103006.093629. 4.1

[57] Martin G. Everett and Stephen P. Borgatti. The foundations of trust and distrust in networks: extended structural balance theory. *ACM Transactions on the Web*, 8(2):1–33, 2014. doi: 10.1145/2591313. 6.4.4, 6.4.4, 6.4.4, 6.4.4, 6.6.10

[58] Bashir Farooq, Matthieu Bierlaire, Ronny Hurtubia, and Georges Flötteröd. A bayesian network approach for population synthesis. *Transportation Research Part C: Emerging Technologies*, 27:53–63, 2013. 5.3.3

[59] Bilal Farooq, Michel Bierlaire, Ricardo Hurtubia, and Gunnar Flötteröd. Simulation of synthetic populations: a review. *Transportation Research Part B*, 58:243–263, 2013. 3.6.5

[60] Noah E. Friedkin. Norm formation in social influence networks. *Social Networks*, 23(3): 167–189, 2001. 1, 1.1, 6.2

[61] Noah E. Friedkin and Eugene C. Johnsen. Social influence and opinions. *Journal of Mathematical Sociology*, 15(3-4):193–206, 1990. 1, 1.1, 5.4.5, 6.2, 6.4.4, 6.4.4, 6.6.10, 7.4

[62] Gallup and Edelman. Global trust and public opinion surveys. `https://www.gallup.com` and `https://www.edelman.com/trust`, 2022–2024. Public trust and institutional confidence measures. A.1

[63] Chen et al. Gao. Large language models empowered agent-based modeling and simulation: A survey and perspectives. *arXiv preprint arXiv:2312.11970*, 2023. 1, 2.1.3

[64] Tao Ge, Xin Chan, Xiaoyang Wang, Dian Yu, Haitao Mi, and Dong Yu. Scaling synthetic data creation with 1,000,000,000 personas. Technical report, Tencent AI Lab, 2024. 2.1.2

[65] Nicholas Geard, James M. McCaw, Alan Dorin, Kevin B. Korb, and Jodie McVernon. Syn-

thetic population dynamics: A model of household demography. *Journal of Artificial Societies and Social Simulation*, 16(1):8, 2013. doi: 10.18564/jasss.2098. 1.1, 2.1.2

[66] Alan S. Gerber, Gregory A. Huber, David Doherty, Conor M. Dowling, and Shang E. Ha. Personality and political attitudes: Relationships across issue domains and political contexts. *American Political Science Review*, 104(1):111–133, 2010. doi: 10.1017/S0003055410000031. 5.6.6, 5.6.6

[67] Navid Ghaffarzadegan, Peiran Wang, and Peter Brown. Generative agent-based models: Simulating human-like social behaviors using large language models. *Journal of Computational Social Science*, 6:1–15, 2024. 6.2

[68] Jennifer et al. Golbeck. Predicting personality from twitter. *Privacy, Security, Risk and Trust (PASSAT), 2011 IEEE Third Int'l Conference on Social Computing*, pages 149–156, 2011. 5.2.4

[69] Jeffrey Goldstein. Emergence as a construct: History and issues. *Emergence*, 1(1):49–72, 1999. doi: 10.1207/s15327000em0101_4. 7.0.2

[70] Augusto et al. González Bonorino. Llms model non-weird populations: Experiments with synthetic cultural agents. *arXiv preprint arXiv:2501.06834*, 2025. 2.1.3

[71] J. C. Gower. A general coefficient of similarity and some of its properties. *Biometrics*, 27 (4):857–871, 1971. 6.3.3, 6.4.4

[72] Jesse Graham, Jonathan Haidt, and Brian A Nosek. Liberals and conservatives rely on different sets of moral foundations. *Journal of personality and social psychology*, 96(5): 1029–1046, 2009. 4.2

[73] James J Gross. Emotion regulation in adulthood: Timing is everything. *Current directions in psychological science*, 10(6):214–219, 2001. 4.2

[74] (Eric) Heng Gu, Senthil Chandrasegaran, and Peter Lloyd. Synthetic users: Insights from designers' interactions with persona-based chatbots. *Artificial Intelligence for Engineering Design, Analysis and Manufacturing*, 39(e2):1–17, 2025. doi: 10.1017/S0890060424000283. 3.2, 3.3.1, 3.6.5

[75] Xu Guo and Yiqiang Chen. Generative ai for synthetic data generation: Methods, challenges and the future. *arXiv preprint arXiv:2403.04190*, 2024. 7.0.2, 7.0.2

[76] Önder Gürcan. Llm-augmented agent-based modelling for social simulations: Challenges and opportunities. *arXiv preprint arXiv:2405.06700*, 2024. 1, 2.1.3

[77] Jonathan Haidt and Jesse Graham. The moral foundations of liberals and conservatives. *Journal of Personality and Social Psychology*, 96(5):1029–1046, 2009. 4.2

[78] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer, 2009. 3.6.5

[79] James K. He, Felix P. S. Wallis, Andrés Gvirtz, and Steve Rathje. Artificial intelligence chatbots mimic human collective behaviour. *British Journal of Psychology*, n/a(n/a). ISSN 2044-8295. doi: 10.1111/bjop.12764. URL `https://onlinelibrary.wiley.com/doi/abs/10.1111/bjop.12764`. _eprint:

https://onlinelibrary.wiley.com/doi/pdf/10.1111/bjop.12764. 6.2

[80] Rainer Hegselmann and Ulrich Krause. Opinion dynamics and bounded confidence: models, analysis and simulation. *The Journal of Artificial Societies and Social Simulation*, 2002. 5.4.5, 6.6.10

[81] Luzie Helfmann, Nataša Djurdjevac Conrad, Philipp Lorenz-Spreen, and Christof Schütte. Modelling opinion dynamics under the impact of influencer and media strategies. *Scientific Reports*, 13:19375, 2023. doi: 10.1038/s41598-023-46187-9. 5.3.6, 5.3.6, 5.3.6

[82] Matthew J. Hicks. *AESOP and SynTel/X: Synthetic Social Media Generation for Information Operations Training*. PhD thesis, Carnegie Mellon University, Pittsburgh, PA, 2024. 7.0.2, 7.0.2, 7.0.2, 7.1.1, 7.1.5

[83] Michaela Hiebler-Ragger and Fuchshuber. Personality influences the relationship between primary emotions and religious/spiritual well-being. *Frontiers in Psychology*, 9:370, 2018. 5.2.4

[84] Timo Hiller. Friends and enemies: A model of signed network formation. *Social Networks*, 34(4):389–399, 2012. doi: 10.1016/j.socnet.2012.02.003. 6.4.4, 6.4.4, 6.4.4

[85] Brian R. Hirshman, Geoffrey P. Morgan, Jesse R. St. Charles, and Kathleen M. Carley. Construct demo input deck. Technical Report CMU-ISR-10-118, Carnegie Mellon University, School of Computer Science, Institute for Software Research, June 2010. CASOS Technical Report. Electronic copy available at SSRN: 2729278. 6.4.4

[86] Geert Hofstede. *Culture's consequences: Comparing values, behaviors, institutions and organizations across nations*. Sage publications, 2001. 4.2

[87] Manoel Horta Ribeiro. Simulating human behavior: Panacea or pandora's box? *Doomscrolling Babel*, 2025. Accessed from https://doomscrollingbabel.manoel.xyz. 1, 2.1.3

[88] Lauren C. Howe and Jon A. Krosnick. Attitude strength. *Annual Review of Psychology*, 68: 327–351, 2017. doi: 10.1146/annurev-psych-122414-033600. 5.6.6

[89] Perttu Hämäläinen, Mikke Tavast, and Anton Kunnari. Evaluating large language models in generating synthetic hci research data: A case study. In *CHI Conference on Human Factors in Computing Systems*, 2023. doi: 10.1145/3544548.3580688. 7.0.2

[90] IPUMS. IPUMS USA, 2026. URL `https://usa.ipums.org/usa/`. Accessed: 2026-03-20. 6.3

[91] Rachael E Jack, Oliver GB Garrod, Hui Yu, Roberto Caldara, and Philippe G Schyns. Facial expressions of emotion are not culturally universal. *Proceedings of the National Academy of Sciences*, 109(19):7241–7244, 2012. 4.2

[92] Charity S. Jacobs, Lynnette Hui Xian Ng, and Kathleen M. Carley. Tracking china's cross-strait bot networks against taiwan. In *Social, Cultural, and Behavioral Modeling (SBP-BRiMS)*, pages 115–125. Springer, 2023. doi: 10.1007/978-3-031-43129-6_12. 7.0.2

[93] Shubham Jain, Nicole Ronald, and Stephan Winter. Creating a synthetic population: A comparison of tools. In *3rd Conference of Transportation Research Group of India*, 2015. 1, 1.1

[94] Zhiyuan Jiang, Frank F. Xu, Xin Dong, and Pascal Hitzler. Retrieval-augmented generation for large language models: A survey, 2024. arXiv:2403.13553. 5.3.2, 5.3.2, 5.3.2, 5.3.2, 5.3.5, 5.3.5

[95] Blair T. Johnson and Alice H. Eagly. Effects of involvement on persuasion: A meta-analysis. *Psychological Bulletin*, 106(2):290–314, 1989. doi: 10.1037/0033-2909.106.2.290. 5.6.6

[96] Daniel Kahneman. *Thinking, Fast and Slow*. Farrar, Straus and Giroux, New York, NY, 2011. 4.1

[97] Shivani Kapania, William Agnew, Motahhare Eslami, Hoda Heidari, and Sarah Fox. 'Simulacrum of Stories': Examining Large Language Models as Qualitative Research Participants, September 2024. URL `http://arxiv.org/abs/2409.19430`. arXiv:2409.19430 [cs]. 2.1.3, 4.9.1, 6.2, 6.7

[98] Vladimir Karpukhin, Barlas Oguz, Sewon Min, and et al. Dense passage retrieval for open-domain question answering. In *EMNLP*, 2020. C.5.1, C.5.1

[99] Henry Kellerman and Robert Plutchik. Emotion-trait interrelations and the measurement of personality. *Psychological Reports*, 23(3):1107–1114, 1968. 1.1

[100] Catherine King and Kathleen M. Carley. Project omen: Designing a training game to fight misinformation on social media. Technical report, Carnegie Mellon University, 2021. 7.0.2

[101] Adam DI Kramer, Jamie E Guillory, and Jeffrey T Hancock. Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences*, 111(24):8788–8790, 2014. 4.2

[102] Jon A. Krosnick. Attitude importance and attitude change. *Journal of Experimental Social Psychology*, 24(3):240–255, 1988. doi: 10.1016/0022-1031(88)90038-8. 5.6.6

[103] Lai Kwok, Mark Bravansky, and Liam D. Griffin. Evaluating cultural adaptability of a large language model via simulation of synthetic personas. In *COLM 2024 Conference Proceedings*, 2024. 1.1, 2.1.2

[104] Christopher Lalk, Kim Targan, Tobias Steinbrenner, et al. Employing large language models for emotion detection in psychotherapy transcripts. *Frontiers in Psychiatry*, 16:1504306, 2025. doi: 10.3389/fpsyt.2025.1504306. 5.3.6

[105] Victor Lavrenko and W. Bruce Croft. Relevance-based language models. In *SIGIR*, 2001. C.5.1, C.5.1

[106] David M. J. Lazer, Matthew A. Baum, Yochai Benkler, et al. The science of fake news. *Science*, 359(6380):1094–1096, 2018. 2.1.1

[107] Jacopo Lenti, Fabrizio Silvestri, and Gianmarco De Francisci Morales. Variational Inference of Parameters in Opinion Dynamics Models. *Proceedings of the International AAAI Conference on Web and Social Media*, 19:2622–2627, June 2025. ISSN 2334-0770. doi: 10.1609/icwsm.v19i1.35963. URL `https://ojs.aaai.org/index.php/ICWSM/article/view/35963`. 6.2

[108] Jennifer S. Lerner and Dacher Keltner. Fear, anger, and risk. *Journal of Personality and Social Psychology*, 81(1):146–159, 2001. doi: 10.1037/0022-3514.81.1.146. 5.6.6

[109] et al Lewis, Patrick. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, 2020. 1, 1.1, 5.2.2, 5.3.2, 5.3.2, 5.3.2, 5.3.5, 5.3.5, 5.3.5, 7.0.2, C.5

[110] et al. Li. Large language models understand and can be enhanced by emotional stimuli. *arXiv*, 2023. URL `https://arxiv.org/abs/2307.11754`. 4.2

[111] Ping-Ping Li, Da-Fang Zheng, and P. M. Hui. Dynamics of opinion formation in a small-world network. *arXiv:physics/0510065*, pages 1–10, 2006. 1

[112] Yuqing Li, Jiangnan Li, Zheng Lin, Ziyan Zhou, Junjie Wu, Weiping Wang, Jie Zhou, and Mo Yu. Mindscape-aware retrieval augmented generation for improved long context understanding. *arXiv preprint arXiv:2512.17220*, 2025. 5.3.5

[113] Evan Lim, Scott Penberthy, Andrew Smith, Minsuk Kahng, Hannah Aizenman, Harmanpreet Singh, Daniel Popovici, Himanshu Jain, Ammon Mercier, Rich Caruana, and Deirdre Mulligan. A large language model for feasible and diverse population synthesis. 2025. arXiv preprint. 5.3.3, 5.3.3

[114] Seymour Martin Lipset. *Continental divide: The values and institutions of the United States and Canada*. Routledge, 1990. 4.2

[115] Ryan Liu, Howard Yen, Raja Marjieh, Thomas L. Griffiths, and Ranjay Krishna. Improving Interpersonal Communication by Simulating Audiences with Language Models, November 2023. URL `http://arxiv.org/abs/2311.00687`. arXiv:2311.00687 [cs]. 2.1.3, 4.1, 5.2.1, 5.4.5, 6.2

[116] George Loewenstein. Emotions in economic theory and economic behavior. *American economic review*, 90(2):426–432, 2000. 4.1

[117] Bolei et al. Ma. Algorithmic fidelity of large language models in generating synthetic german public opinions: A case study. *arXiv preprint arXiv:2412.13169*, 2024. 2.1.3

[118] Yu A. Malkov and Dmitry A. Yashunin. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018. C.5.2

[119] Rebecca Marigliano and Kathleen M. Carley. Aurora: Enhancing synthetic population realism through rag and salience-aware opinion modeling. In Melih Kaspi, Bahar Biller, Addison Hettinger, and Young-Jun Son, editors, *Proceedings of the 2025 Winter Simulation Conference (WSC)*, 2025. 6.3

[120] Rebecca Marigliano, Peter Carragher, and Kathleen M. Carley. Modeling emotional expression in llm agents with social media–based validation, 2025. Manuscript. 6.3.2

[121] Rebecca Marigliano, Lynnette Hui Xian Ng, and Kathleen M. Carley. Analyzing digital propaganda and conflict rhetoric: A study on russia's bot-driven campaigns and counter-narratives during the ukraine crisis. *Journal Title*, 2025. Manuscript on Russia–Ukraine bot-driven narratives, sentiment, and moral foundations. 4.2, 4.6.6

[122] José M. Marques. The black sheep effect: Outgroup homogeneity as a social comparison process. In Dominic Abrams and Michael A. Hogg, editors, *Social Identity Theory: Con-*

*structive and Critical Advances*, pages 131–151. Harvester Wheatsheaf, New York, 1990. 6.3.3, 6.4.4, 6.4.4

[123] Seth A. Marvel, Jon Kleinberg, Robert D. Kleinberg, and Steven H. Strogatz. Continuous-time model of structural balance. *Proceedings of the National Academy of Sciences*, 108 (52):20953–20958, 2011. 6.4.4, 6.4.4, 6.4.4, 6.4.4, 6.6.10

[124] M. Mäs, A. Flache, and D. Helbing. Dynamics of opinion polarization in complex social networks. *Journal of Complex Networks*, 12(3):14–32, 2024. doi: 10.1093/comnet/cnz044. 1

[125] David Matsumoto, Seung Hee Yoo, and Johnny Fontaine. Mapping expressive differences around the world: The relationship between emotional display rules and individualism versus collectivism. *Journal of cross-cultural psychology*, 39(1):55–74, 2008. 4.2

[126] Robert R McCrae and Oliver P John. An introduction to the five-factor model and its applications. *Journal of personality*, 60(2):175–215, 1992. 4.2

[127] Robert R. McCrae and Antonio Terracciano. Personality profiles of cultures: Aggregate personality traits. *Journal of Personality and Social Psychology*, 89(3):407–425, 2005. doi: 10.1037/0022-3514.89.3.407. URL `https://psycnet.apa.org/fulltext/2005-13299-011.pdf`. B.2.3

[128] Miller McPherson, Lynn Smith-Lovin, and James M. Cook. Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 27:415–444, 2001. 5.6.6, 6.3.1, 6.4.4, 6.4.4, 6.6.10

[129] Dennis L. Meadows, William W. Behrens, Donella H. Meadows, Roger F. Naill, Jørgen Randers, and Erich Zahn. *Dynamics of Growth in a Finite World*. Wright-Allen Press, Cambridge, MA, 1974. 7.0.2

[130] Rolf Moeckel, Klaus Spiekermann, and Michael Wegener. Creating a synthetic population. Technical report, Institute of Spatial Planning, University of Dortmund, 2003. 1.1, 2.1.2

[131] Saif M Mohammad. Sentiment analysis of short informal texts. *Journal of Artificial Intelligence Research*, 50:723–762, 2016. 4.2

[132] Saif M Mohammad and Peter D Turney. Crowdsourcing a word-emotion association lexicon. In *Computational intelligence*, volume 29, pages 436–465. Wiley Online Library, 2013. 4.2

[133] Christian Montag and Jaak Panksepp. Primary emotional systems and personality: An evolutionary perspective. *Frontiers in Psychology*, 8:464, 2017. 5.2.4

[134] Suhong Moon, Marwa Abdulhai, Minwoo Kang, Joseph Suh, Widyadewi Soedarmadji, Eran Kohen Behar, and David M. Chan. Virtual personas for language models via an anthology of backstories, 2024. 2.1.2

[135] Jonathan H. Morgan, Jacob Shaha, Rebecca Marigliano, and Kathleen M. Carley. *GhostField Documentation*, 2025. Internal System Documentation, GhostField v1.0.0 Architecture. 6.3.1, 7.1, 7.1.1

[136] Kirill Müller and Kay W. Axhausen. Population synthesis for microsimulation: State of the

art. In *Proceedings of the 10th Swiss Transport Research Conference*, Ascona, Switzerland, 2010. 5.6.6

[137] Keiichi et al. Namikoshi. Using llms to model the beliefs and preferences of targeted populations. *arXiv preprint arXiv:2403.20252*, 2024. 2.1.3

[138] National Academies of Sciences, Engineering, and Medicine. *A Decadal Survey of the Social and Behavioral Sciences: A Research Agenda for Advancing Intelligence Analysis*. The National Academies Press, Washington, DC, 2019. doi: 10.17226/25335. 7.0.2

[139] National Statistical Offices. National census and demographic reports, 2020–2024. Ethnicity, religion, language distributions. A.1, A.1

[140] Lynnette Hui Xian Ng and Kathleen M. Carley. Deflating the chinese balloon: Types of twitter bots in the us-china balloon incident. *EPJ Data Science*, 12(1):63, 2023. doi: 10.1140/epjds/s13688-023-00440-3. 7.0.2

[141] Lynnette Hui Xian Ng and Kathleen M. Carley. Pro or anti? a social influence model of online stance flipping. *IEEE Transactions on Network Science and Engineering*, 10(1):3–19, 2023. doi: 10.1109/TNSE.2022.3185785. 6.6.5, 6.7

[142] Dong Nguyen, A. Seza Doğruöz, Carolyn P. Rosé, and Franciska de Jong. Computational sociolinguistics: A survey. *Computational Linguistics*, 42(3):537–593, 2016. 6.4.4

[143] Brendan Nyhan and Jason Reifler. When corrections fail: The persistence of political misperceptions. *Political Behavior*, 32:303–330, 2010. doi: 10.1007/s11109-010-9112-2. 6.4.4, 6.6.10

[144] OECD. Oecd survey on drivers of trust in public institutions. `https://www.oecd.org/governance/trust`, 2023–2024. Institutional trust and governance indicators. A.1, A.1

[145] Shira Offer. Negative social ties: Prevalence and consequences. *Annual Review of Sociology*, 47:271–294, 2021. doi: 10.1146/annurev-soc-090820-025827. 6.4.4, 6.4.4, 6.4.4

[146] Maya Okawa and Tomoharu Iwata. Predicting Opinion Dynamics via Sociologically-Informed Neural Networks. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 1306–1316, Washington DC USA, August 2022. ACM. ISBN 978-1-4503-9385-0. doi: 10.1145/3534678.3539228. URL `https://dl.acm.org/doi/10.1145/3534678.3539228`. 6.2

[147] Joanna Orlin, Pieter Van Dessel, Jelle De Vries, Matthew Alexander, and Frank Van Overwalle. Modelling the creation of friends and foes groups in small real social networks. *PLOS ONE*, 19(7):e0306353, 2024. doi: 10.1371/journal.pone.0306353. 6.4.4, 6.4.4

[148] Joon Sung Park, Carolyn Q Zou, Aaron Shaw, Benjamin Mako Hill, Carrie Cai, Meredith Ringel Morris, Robb Willer, Percy Liang, and Michael S Bernstein. Generative Agent Simulations of 1,000 People. 4.1

[149] Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative Agents: Interactive Simulacra of Human Behavior, August 2023. URL `http://arxiv.org/abs/2304.03442`. arXiv:2304.03442 [cs].

1, 2.1.3, 4.2, 4.9.1, 5.2.1, 5.4.5, 5.6.6, 6.4.4, 7.0.2

[150] Candice Pattisapu, Tim Verbelen, Riddhi J. Pitliya, Alex B. Kiefer, and Mahault Albarracin. Free energy in a circumplex model of emotion. *arXiv*, 2024. arXiv:2407.02474. 6.4.5

[151] Richard E. Petty and John T. Cacioppo. The elaboration likelihood model of persuasion. In Leonard Berkowitz, editor, *Advances in Experimental Social Psychology*, volume 19, pages 123–205. Academic Press, Orlando, FL, 1986. 5.6.6

[152] Robert Plutchik. What is an emotion? *The Journal of Psychology*, 61(2):295–303, 1965. doi: 10.1080/00223980.1965.10543417. 5.3.6, 5.3.6, 5.3.6

[153] Robert Plutchik. A general psychoevolutionary theory of emotion. *Theories of emotion*, 1 (3-31):4, 1980. 4.2

[154] Jonathan Posner, James A. Russell, and Bradley S. Peterson. The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology. *Development and Psychopathology*, 17(3):715–734, 2005. doi: 10.1017/S0954579405050340. 6.3.3, 6.4.5, 6.4.5, 6.4.5

[155] Manon Prédhumeau and Ed Manley. A synthetic population for agent-based modelling in canada. *Scientific Data*, 10:148, 2023. 3.1, 3.2, 3.3.1, 3.3.1, 3.4.4

[156] Xiaomei Qin, Zhixin Liu, Yuwei Liu, et al. User ocean personality model construction method using a bp neural network. *Electronics*, 11(19):3022, 2022. doi: 10.3390/electronics11193022. 5.3.6

[157] Jingjing Qu, Kejia Hu, Jun Zhu, Yulei Ye, Wenhao Li, Teng Wang, Zhiyun Chen, Chaochao Lu, Aimin Zhou, Xiangfeng Wang, Xia Hu, and James Evans. Epitome: Pioneering an experimental platform for ai-social science integration. *arXiv preprint arXiv:2507.01061*, 2025. doi: 10.48550/arXiv.2507.01061. URL https://arxiv.org/abs/2507.01061. 6.2

[158] Anand Ramachandran. Retrieval-augmented generation for large language models: Leveraging enterprise data, vector databases, and multi-modal innovations, 2024. Unpublished manuscript. 1, 1.1

[159] Anand Ramachandran. Retrieval augmented generation for large language models leveraging enterprise data, 2024. 5.3.2, 5.3.2, 5.3.2, 5.3.2, 5.3.2, 5.3.2, 5.3.2, 5.3.2, 5.3.2, 5.3.5

[160] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *EMNLP*, 2019. C.5.1

[161] Manoel Horta Ribeiro. Simulating Human Behavior, September 2024. URL https://doomscrollingbabel.manoel.xyz/p/simulating-human-behavior. 1, 6.2, 6.7

[162] Marisa Ribeiro. Simulating human behavior: The epistemic limits of language models as proxies for people. *Computational Humanities Quarterly*, 2025. Forthcoming. 5.2.1

[163] C. Ritchie. Creating synthetic user research using persona prompting and autonomous agents, 2024. URL https://medium.com/data-science/creating-synthetic-user-research-using-persona-prompting-and-autonomous

Accessed: 2025-04-11. 3.1, 3.3.1

[164] Stephen Robertson and Hugo Zaragoza. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389, 2009. C.5.1, C.5.1, C.5.2

[165] J. J. Rocchio. Relevance feedback in information retrieval. *The SMART Retrieval System: Experiments in Automatic Document Processing*, pages 313–323, 1971. C.5.1, C.5.1

[166] James A. Russell. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6):1161–1178, 1980. 6.3.3, 6.4.5, 6.4.5, 6.4.5, 6.4.5

[167] Joni Salminen, Soon-gyo Jung, Lene Nielsen, and Bernard J. Jansen. Creating more personas improves representation of demographically diverse populations: Implications towards interactive persona systems. In *Proceedings of the Nordic Human-Computer Interaction Conference (NordiCHI '22)*, 2022. doi: 10.1145/3546155.3546654. 1.1, 2.1.2

[168] Robert M Sapolsky. *Behave: The biology of humans at our best and worst*. Penguin, 2017. 4.1

[169] Robert G Sargent. Verification and validation of simulation models. *Journal of Simulation*, 7(1):12–24, 2013. 3.6.5

[170] David P Schmitt, Jüri Allik, Robert R McCrae, and Verónica Benet-Martínez. The geographic distribution of big five personality traits: Patterns and profiles of human self-description across 56 nations. *Journal of cross-cultural psychology*, 38(2):173–212, 2007. 4.2, 4.3.3, 5.4.3, B.1, B.2.3

[171] Sarah Schröder, Thekla Morgenroth, Ulrike Kuhl, Valerie Vaquet, and Benjamin Paaßen. Large Language Models Do Not Simulate Human Psychology, August 2025. URL `http://arxiv.org/abs/2508.06950`. arXiv:2508.06950 [cs]. 4.1, 4.9.1

[172] Chieh-Yang Si, Yuhan Liu, and Jeffrey T Hancock. Simulation of stance perturbations in llm-driven discourse. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, 2024. 5.2.1

[173] Linda J. Skitka and Christopher W. Bauman. Moral conviction: Another contributor to attitude strength or something more? *Journal of Personality and Social Psychology*, 94(5): 895–917, 2008. doi: 10.1037/0022-3514.94.5.895. 6.4.5

[174] Paul Slovic, Melissa L Finucane, Ellen Peters, and Donald G MacGregor. The affect heuristic. *European journal of operational research*, 177(3):1333–1352, 2007. 4.1

[175] Keith E. Stanovich and Richard F. West. Individual differences in reasoning: Implications for the rationality debate? *Behavioral and Brain Sciences*, 23(5):645–665, 2000. doi: 10.1017/S0140525X00003435. 4.1

[176] Statista. Global statistics database. `https://www.statista.com`, 2023–2025. Employment sectors, demographics, digital adoption. A.1

[177] Statistics Canada. Census Program – Census of Population, 2026. URL `https://www12.statcan.gc.ca/census-recensement/index-eng.cfm`. Accessed: 2026-03-20. 6.3

[178] Peter Suedfeld. Attitude change: The competing views. *Routledge Taylor & Francis Group*, 1968. Originally published by Transaction Publishers, reprinted by Routledge in 2017. 6.2, 6.6.10

[179] Lei Sun. *Rethinking Iterative Proportional Fitting: Scalable and Hybrid Approaches to Joint Distribution Fitting*. PhD thesis, University of California, Berkeley, 2015. 3.7.4, 5.3.3

[180] Seungjong et al. Sun. Random silicon sampling: Simulating human sub-population opinion using a large language model based on group-level demographic information. *arXiv preprint arXiv:2401.07888*, 2024. 2.1.3

[181] Henri Tajfel and John C. Turner. An integrative theory of intergroup conflict. In William G. Austin and Stephen Worchel, editors, *The Social Psychology of Intergroup Relations*, pages 33–47. Brooks/Cole, Monterey, CA, 1979. 5.6.6

[182] Roger Tanton, Ann Daly, Tom Smith, and Luke Harding. Generating a synthetic population of individuals in households: Sample-free vs sample-based methods. *Population Research and Policy Review*, 33:389–409, 2014. 3.7.4, 5.3.3, 5.3.3, 5.3.3, 5.3.3, 5.3.3, 5.3.3

[183] Shelley E. Taylor and Susan T. Fiske. Salience, attention, and attribution: Top of the head phenomena. *Advances in Experimental Social Psychology*, 11:249–288, 1979. 1.1, 6.2, 6.6.10

[184] LangChain Team. Llm retrieval-augmented generation with langchain. `https://www.langchain.com`, 2023. 5.2.2, 5.3.2, 5.3.2

[185] Jeanne L. Tsai, Robert W. Levenson, and Laura L. Carstensen. Autonomic, subjective, and expressive responses to emotional films in older and younger chinese americans and european americans. *Psychology and Aging*, 2000. 4.2

[186] United States Census Bureau. PUMS Data, 2026. URL `https://www.census.gov/programs-surveys/acs/microdata/access.html`. Accessed: 2026-03-20. 6.3

[187] United States Census Bureau. 2020–2024 ACS PUMS Data Dictionary, March 2026. URL `https://www2.census.gov/programs-surveys/acs/tech_docs/pums/data_dict/PUMS_Data_Dictionary_2020-2024.pdf`. Published March 5, 2026. Accessed: 2026-03-20. 6.3

[188] Ibo Van de Poel, Lambèr Royakkers, and Sjoerd D Zwart. The moral machine experiment. *Nature*, 563(7729):59–64, 2014. 4.2

[189] Iris van Rooij, Olivia Guest, Federico Adolfi, Ronald de Haan, Antonina Kolokolova, and Patricia Rich. Reclaiming AI as a Theoretical Tool for Cognitive Science. *Computational Brain & Behavior*, 7(4):616–636, December 2024. ISSN 2522-087X. doi: 10.1007/s42113-024-00217-5. URL `https://doi.org/10.1007/s42113-024-00217-5`. 4.1, 4.9.1

[190] Veniamin Veselovsky, Manoel Horta Ribeiro, Akhil Arora, Martin Josifoski, Ashton Anderson, and Robert West. Generating faithful synthetic data with large language models. *arXiv preprint arXiv:2305.15041*, 2023. 7.0.2, 7.0.2

[191] Alexander Vezhnevets, Fabricio Pardo, Junhyuk Oh, et al. Concordia: A generative frame-

work for scalable agent-based simulations using large language models. *arXiv preprint arXiv:2312.03664*, 2023. 6.2, 6.4.4

[192] Soroush Vosoughi, Deb Roy, and Sinan Aral. The spread of true and false news online. *Science*, 359(6380):1146–1151, 2018. 2.1.1

[193] Kelsey Walsh and Kendall Karson. Calls to us helpline jump 891% as white house is warned of mental health crisis, Apr 2020. URL https://abcnews.go.com/Politics/calls-us-helpline-jump-891-white-house-warned/story?id=70010113. Accessed: 2025-09-25. 4.1

[194] Yitong Wang, Xianyong Li, Yuhang Cheng, Yajun Du, Dong Huang, Xiaoliang Chen, and Yongquan Fan. A neural probabilistic bounded confidence model for opinion dynamics on social networks. *Expert Systems with Applications*, 247:123315, August 2024. ISSN 0957-4174. doi: 10.1016/j.eswa.2024.123315. URL https://www.sciencedirect.com/science/article/pii/S0957417424001805. 6.2

[195] David Watson. Mood and temperament. *New York: Guilford Press*, 2000. 4.2

[196] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*, 2022. 4.2

[197] Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Saffron Cheng, Amelia Glaese, Borja Balle, Atoosa Kasirzadeh, et al. Ethical and social risks of harm from large language models. *arXiv preprint arXiv:2112.04359*, 2021. 4.2, 4.9.1

[198] Gary C. White and Bruce C. Lubow. Fitting population models to multiple sources of observed data. *The Journal of Wildlife Management*, 66(2):300–309, 2002. 1, 1.1, 3.1, 3.2, 3.3.1, 3.3.1, 3.4.2, 3.6.5

[199] Wikipedia contributors. National election results by country. https://en.wikipedia.org, 2019–2025. Vote shares and dominant political parties. A.1, A.1

[200] Grace Wong, Keng-Hwee Tan, and Chee-Wee Tan. Standing up or standing by: Understanding bystanders' proactive reporting responses to social media misinformation. *Information Systems Journal*, 31(6):914–947, 2021. 2.1.1

[201] World Bank. World development indicators. https://data.worldbank.org, 2024. Population, GDP per capita, urbanization, internet use. A.1, A.1

[202] Carsten Wrosch, Meaghan A. Barlow, and Ute Kunzmann. Age-related changes in older adults' anger and sadness: The role of perceived control. *Psychology and Aging*, 2017. Accepted manuscript (January 12, 2017). 4.1

[203] Chengxing Xie, Canyu Chen, Feiran Jia, Ziyu Ye, Shiyang Lai, Kai Shu, Jindong Gu, Adel Bibi, Ziniu Hu, David Jurgens, James Evans, Philip Torr, Bernard Ghanem, and Guohao Li. Can Large Language Model Agents Simulate Human Trust Behavior?, November 2024. URL http://arxiv.org/abs/2402.04559. arXiv:2402.04559 [cs]. 2.1.3, 4.9.1

[204] Zhan Xu, Yunfeng Zhang, and Yang Wang. Extended structural balance theory for modeling

trust in social networks. *International Journal of Modern Physics C*, 27(11):1650134, 2016. doi: 10.1142/S012918311650134X. 6.4.4, 6.4.4, 6.4.4, 6.4.4, 6.4.4, 6.6.10

[205] Z. Yameogo et al. Population synthesis using iterative proportional fitting (ipf): A review and future research. *Transportation Research Part C*, 2021. 3.2, 3.3.1, 3.3.1, 3.6.5

[206] Z. Yameogo et al. A decision framework for synthetic population generation. *Journal of Artificial Societies and Social Simulation*, 2021. 3.1, 3.3.1, 3.4.3, 3.6.5

[207] Ling Yang, Zhilong Zhang, Yang Song, et al. Diffusion models: A comprehensive survey of methods and applications. *ACM Computing Surveys*, 2023. doi: 10.1145/3626235. 3.4.3

[208] Xin Ye, Karthik C. Konduri, Ram M. Pendyala, Bhargava Sana, and Paul Waddell. Methodology to match distributions of both household and person attributes in generation of synthetic populations. In *Transportation Research Board Annual Meeting*, 2009. 1, 2.1.2

[209] Chiyuan et al. Zhang. Understanding deep learning requires rethinking generalization. *Communications of the ACM*, 2021. 3.6.5

[210] Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Olivier Bousquet, Quoc Le, and Ed Chi. Least-to-most prompting enables complex reasoning in large language models. *arXiv preprint arXiv:2205.10625*, 2022. 4.2

[211] Qinyue Zhou and Zhibin Wu. Multidimensional friedkin-johnsen model with increasing stubbornness in social networks. *Information Sciences*, 598:1–14, 2022. 1, 6.2

[212] Xiaotian Zhou, Haoxin Sun, Wanyue Xu, Wei Li, and Zhongzhi Zhang. Friedkin-johnsen model for opinion dynamics on signed graphs. *IEEE Transactions on Knowledge and Data Engineering*, 2024. 6.2, 6.4.4, 6.6.10