Emergent Cultural Dynamics in Agent-Based Systems through Evolutionary Language Models: A Conceptual Framework for Modeling Sociocultural Evolution via LLM-Guided Simulation

Divyansh Agrawal
B.Tech Computer Engineering
Mukesh Patel School of Technology, Management & Engineering, NMIMS
Mumbai, 400056, India
divyansh.agrawal20@gmail.com

Diviyaj Bellare
B.Tech Computer Engineering
Mukesh Patel School of Technology, Management & Engineering, NMIMS
Mumbai, 400056, India
diviyajbellare@gmail.com

Vedansh Paliwal
B.Tech Computer Engineering
Mukesh Patel School of Technology, Management & Engineering, NMIMS
Mumbai, 400056, India
paliwal.vedansh@gmail.com

Abhay Kolhe
Department of Computer Engineering
Mukesh Patel School of Technology, Management & Engineering, NMIMS
Mumbai, 400056, India
abhay.kolhe@nmims.edu

Abstract

How do the different ideologies and methods of the world's cultures come to fruition from people of the same seed? Where the traditional agent-based models (ABMs) separate communication from individual rules of behavior, we present a coevolutionary framework for the agents. We enact agent behavior as performable policy code and argumentation as natural language text. Both of them are developed through a constrained Large Language Model (LLM) that acts as a mutation engine, with the selection pressures rewarding both the performance of the task and the cohesion of the tribe's language. We assigned agents to four distinct councils of tribes located around the world, each with a different philosophical starting point, conflict through the midst of their resource allocation, and made group decisions that influenced the metrics of the shared colony. The paper makes three contributions: (1) a reproducible pipeline for the coevolution of policy code and rhetoric; (2) a set of embedding-based indices to quantify cultural dynamics; and (3) an experimental demonstration showing the way different philosophical starting points lead to different performance outcomes and rhetorical styles.

Index Terms

Agent-Based Modeling (ABM), Large Language Models (LLMs), Computational Social Science, Cultural Evolution, Multi-Agent Systems, AI Governance

I. Introduction

How have different political cultures and common rhetorical frames been developed from groups of initially alike decision makers? Classical agent-based models (ABM) portray the evolution of complex macro-levels through simple rules of behavior, but these models typically treat behavior (rules) and the communicative layer (language or arguments) separately. However, in real social systems, decision-making and persuasive traits accompany each other: the players not only modify their moves but also the way they justify these moves. The present paper proposes and validates a computational framework in which the action of an agent is represented as a small executable policy code and the argumentation as a brief natural language text, both of which are subjected to selection pressures for task performance and linguistic cohesion.

We propose the large language model (LLM) as a mutation engine that provides microscopic yet verifiable changes to policy code, where the alterations are constrained to be safe and runnable. The agents are placed in tribal councils whose objective is to resolve the narrative Dilemma Cards-these are concise scenarios that necessitate making compromises among four colony metrics (Prosperity, Security, Health, Future Outlook). Evolution within tribes happens via selection based on fitness (which is a combination of contribution to colony metrics and intra-tribe linguistic cohesion). This co-evolutionary framework enables us to ask: What are the conditions for recognizing distinct policy archetypes and rhetorical conventions? How will different philosophical foundations at the very beginning influence the cultures that emerge? The operationalization of these questions is done through embedding-based cohesion indices, semantic drift calculations, frequency analysis, and sentiment analysis in controlled experiments among four tribes with varying philosophical foundations.

II. Literature Review

Our research is based on recent advances in LLM-driven ABM. In the following, we expand the core literature into four focused subsections that highlight theoretical foundations, methodological innovations, evaluation constraints, and how our work addresses remaining empirical and reproducibility gaps.

A. Foundations & Emergence

Recent frameworks demonstrate that LLMs substantially expand agent expressivity and decision making beyond classical rule-based ABMs, while requiring careful experimental controls (prompt/seed provenance, model selection, validation) [1], [2]. LLM populations exhibit emergent social phenomena that resemble human dynamics, including iterated games, [3], naming-game experiments, [4], and opinion dynamic [14], [15]. However, many value-embedded ABMs fix agent priors top-down [7], [8], limiting insight into how values and heuristics emerge inherently.

B. Evolutionary Methods and Communication Emergence

Automated evolutionary pipelines treat agent policies as evolvable artifacts with LLMs orchestrated in iterative propose, evaluate, select loops [10], [11], [19]. Complementary emergent language research shows that natural language grounding and structured communication bottlenecks produce interpretable, transferable shared representations [6], [16], [17], [22].

C. Methodological Critiques and Open Gaps

Key critiques emphasize the need for sensitivity to design choices [20], as well as concerns about identity flattening, homophily amplification, and inconsistent reporting [11]- [13]. The literature reveals gaps, such as insufficient provenance documentation, the tendency of LLM towards average behaviors, and limited exploration of long-term cultural evolution. [5], [18], [23].

D. Our Contribution

To overcome the challenges mentioned above, we propose a constrained evolutionary pipeline in which the compact policy code is mutated by LLM-guided proposals with strict provenance logging, hybrid fitness based on governance, performance and embedding-based intra-tribe linguistic cohesion is used for selection, and systematic comparison of how initial cultural framing influences long-term evolution is enabled by four different philosophical starting conditions. While value-embedded ABMs restrict priors, we permit endogenous co-evolution from different conditions, and keep track of divergence through the use of innovative semantic metrics (drift calculation, frequency analysis, sentiment patterns, cohesion indices) over 100-round longitudinal simulations.

III. Methodology and Simulation Design

A. Simulation Overview

The simulation involves four independent tribes of 15 agents each, evolving in parallel over 100 rounds under identical environmental challenges. Each tribe has distinct philosophical orientations (Table I), initialized with two persona archetypes embodying different decision making principles.

TABLE I
Tribal Philosophical Foundations

Tribe	**Core Philosophy**
The Collective	Communal well-being, equity, health focus
The Forge	Data-driven optimization, security emphasis
The Vanguard	Bold innovation, paradigm-shifting risk-taking
The Frontiers	Territorial expansion, growth through adversity

B. Core Components

Colony State: Tracked by four metrics (Health, Prosperity, Security, Future Outlook, range 0-1000) and Energy Units (EU) treasury. Simulation ends if any metric reaches zero. **Dilemma Cards:** Narrative scenarios presenting resource allocation trade-offs, e.g., "invest 300 EU in immediate pesticide (boosts Prosperity, 5 rounds) or 250 EU in resistant crop strain (boosts Future Outlook permanently after 10-round delay)." **Agent Policy:** Co-evolving components: (1) Policy Code—Python function mapping ColonyState + DilemmaCard to EU allocation; (2) Rhetorical Frame—natural language justification string.

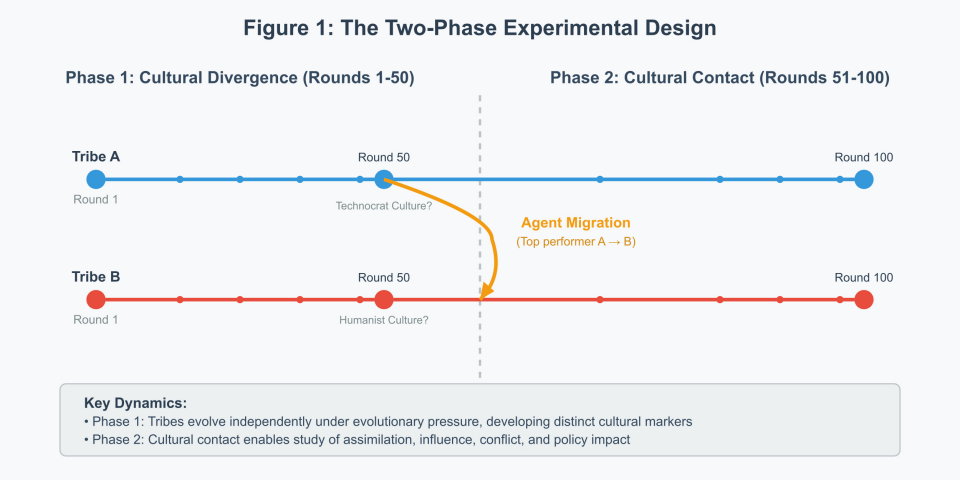

Fig. 1. The Four-Tribe Experimental Design. Parallel timelines show the evolution process of the four tribes. Over 100 rounds under identical environmental challenges with migrations. Each timeline visualizes internal selection events, mutation bursts, major regime shifts, and final metric trajectories.

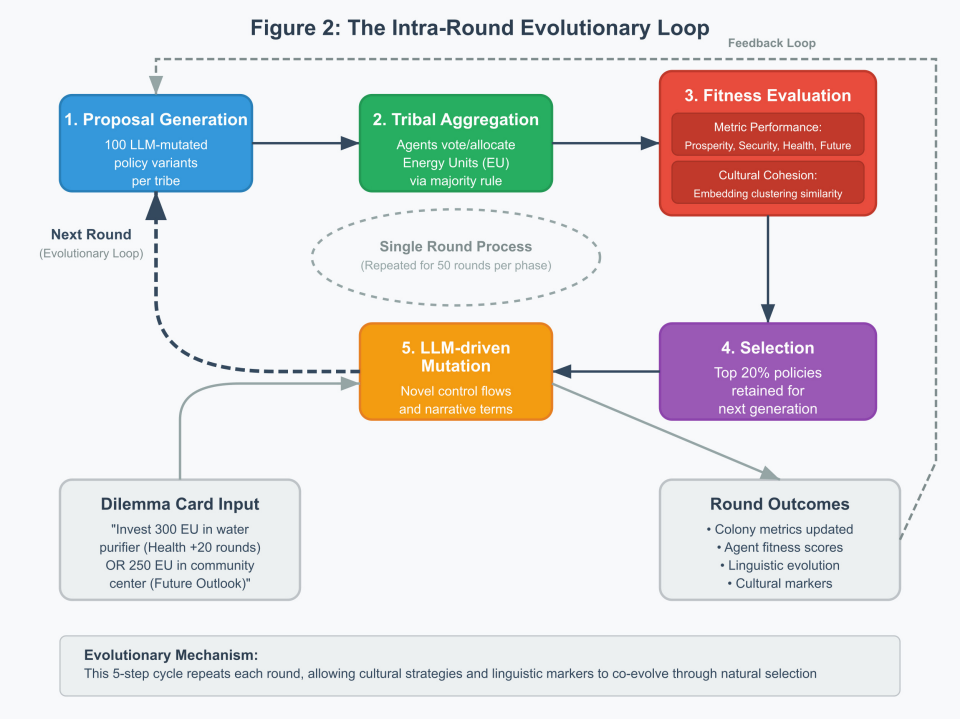

Fig. 2. The Intra-Round Evolutionary Loop. Cyclical flowchart showing: (1) Proposal Generation → (2) Tribal Aggregation → (3) Fitness Evaluation (Metric Performance + Linguistic Cohesion) → (4) Selection (Top 20%) → (5) LLM-driven Mutation, looping to the next round.

C. Evolutionary Loop

Each round followed a five-step cycle: (1) **Proposal:** agents produced resource-allocation plans and supporting arguments; (2) **Aggregation:** a majority vote selected the collective decision; (3) **Fitness evaluation:** performance was assessed using a hybrid metric combining colony-health and balance indicators with linguistic cohesion (cosine similarity in embedding space); (4) **Selection:** the top 20% of agents advanced; and (5) **Mutation:** each survivor spawned four variants generated by the LLM using targeted prompts (e.g., ”make this more risk-averse” or “emphasize security”), with outputs checked in a sandboxed environment.

D. Metrics

Linguistic: Frequency analysis, semantic drift (embedding cosine similarity over time), cohesion index (intra-tribe similarity), sentiment (polarity, subjectivity), philosophical profile mapping. **Performance:** It was evaluated using three components: a balance score (computed as 100 minus the coefficient of variation across metrics), sustainability (minimum metric value), and an overall success likelihood calculated as a weighted composite: 30% sustainability, 25% total performance, 25% balance, and 20% projected future growth.

IV. Results

A. Emergence of Distinct Cultural Archetypes

The patterns of word frequency showed a clear divergence between groups (Fig. 3). The Collective repeatedly highlighted terms such as ”health“ (550+ mentions), “well being” (150), and ”community“ (150). The Forge favored “prosperity“ (500+), “security“ (400), “metric“ (150), and “analysis“ (120). The Frontiers emphasized “expansion“ (150), “growth“ (120), and “strength“ (100), while The Vanguard consistently used “innovation“ (150), “transformative“ (120), and “bold“ (100). Philosophical orientation analysis (Fig. 4) reinforced these distinctions: The Collective leaned toward Harmony (35%) and Community (30%); The Forge toward Efficiency (30%) and Power (20%); The Frontiers toward Expansion (70%) and Progress (25%); and The Vanguard toward Innovation (65%) with a relatively balanced spread across other dimensions. Word-cloud visualizations (Fig. 5) further illustrated each group’s semantic identity, which remained stable over 100 rounds despite uniform external conditions.

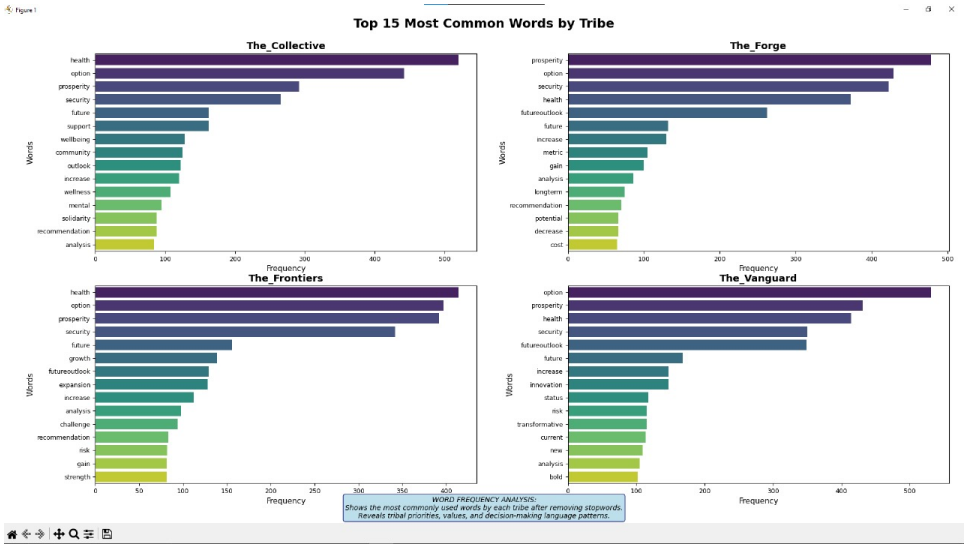

Fig. 3. Top 15 most frequent words by tribe after 100 rounds. Four grouped bar charts present the dominant terms used by The Collective, The Forge, The Frontiers, and The Vanguard, illustrating the distinct vocabulary patterns that developed over the course of the simulation.

B. Performance Outcomes

Table II shows final rankings. The Vanguard achieved dominant performance (total 595.0, balance 48.5, sustainability 90.0, success 106.2), with final metrics H:105, P:200, S:90, F:200. The Collective finished last (435.0 total, O sustainability) despite health focus. Middle tribes (Frontiers 535.0, Forge 495.0) showed over-specialization (balance 21.9, 16.7).

The Vanguard’s achievements came from taking risks that were oriented to the future, thus receiving the environmental rewards, measuring with a balanced metric distribution that avoided over-specialization, and adaptive exploration that allowed the discovery of better strategies. The Collective’s downfall was due to the emphasis on health, the aversion to risks that hindered the making of necessary trade-offs, and the tolerance of inefficiency that worsened the deficits.

C. Rhetorical Patterns

Sentiment analysis (see Fig. 6) showed near-neutral polarity across tribes (Vanguard 0.136, Frontiers 0.112, Forge 0.071, Collective 0.088) but substantial subjectivity differentiation (Vanguard 0.523 most opinion-driven, Collective 0.409 most fact-based). The most subjective tribe achieved best performance while most objective tribe performed worst, suggesting vision-driven rhetoric may enhance persuasiveness in this environment. Scatter plots (see Fig. 7) revealed tribes separated primarily along subjectivity rather than polarity axis, with Vanguard showing wider dispersion and Collective tighter clustering.

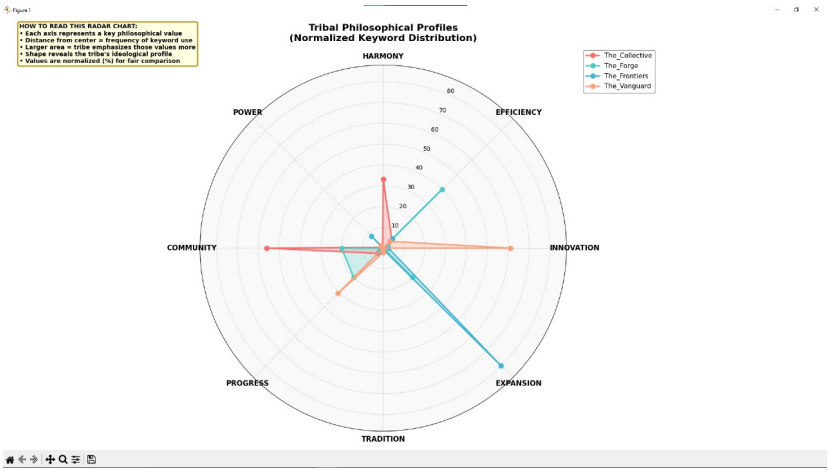

Fig. 4. Tribal Philosophical Profiles (Normalized Keyword Distribution). Radar chart showing each tribe's emphasis across eight philosophical dimensions: Harmony, Efficiency, Power, Community, Innovation, Expansion, Progress, and Tradition. Values are normalized to the same scale to facilitate visual comparison.

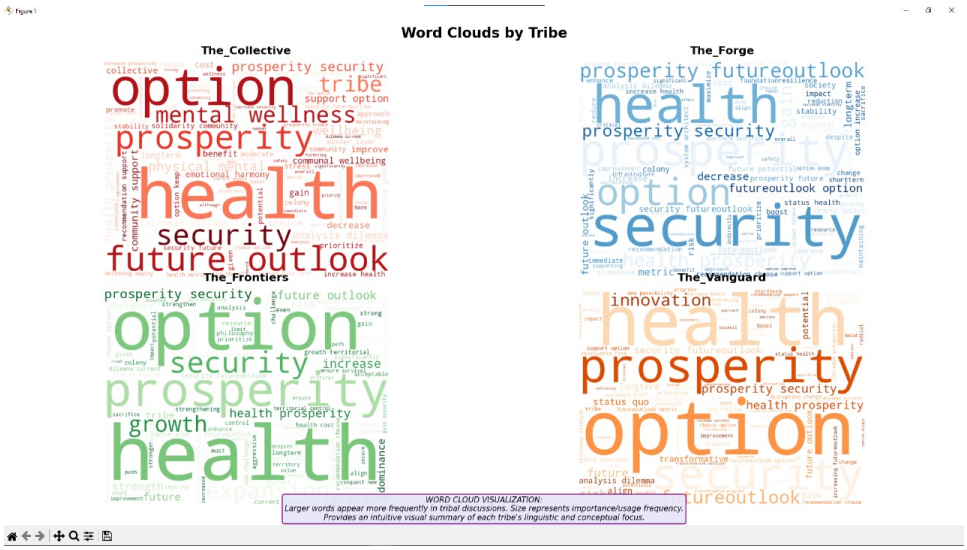

Fig. 5. Word Clouds by Tribe. Four word clouds (one per tribe) where word size is proportional to frequency. The visual highlights salient priorities, values, and rhetorical tokens used within each tribe's discourse over 100 rounds.

TABLE II
FINAL PERFORMANCE RANKINGS

Rank	Tribe	Total	Balance	Success
1	The Vanguard	595.0	48.5	106.2
2	The Frontiers	535.0	21.9	64.3
3	The Forge	495.0	16.7	54.6
4	The Collective	435.0	28.6	54.6

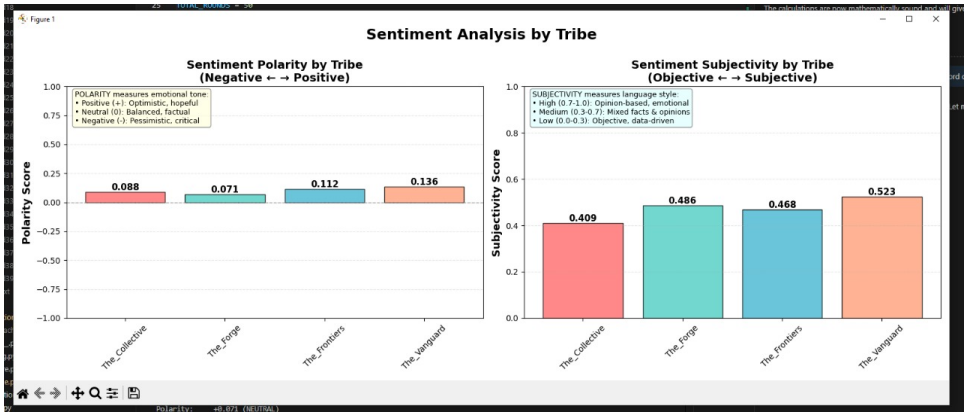

Fig. 6. Sentiment Analysis by Tribe. Left: Polarity (emotional tone) showing near-neutral values across all tribes. Right: Subjectivity showing The Vanguard as most opinion-driven and The Collective as most fact-based.

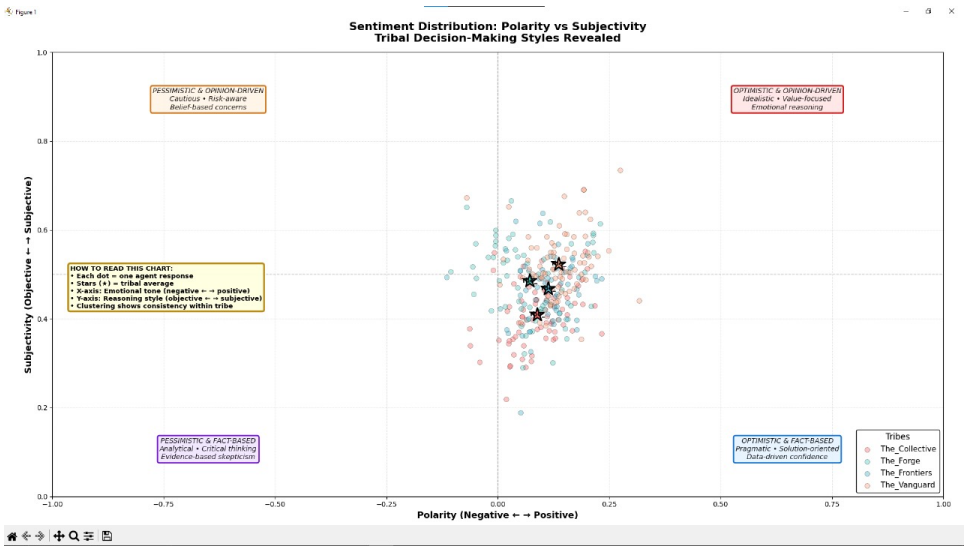

Fig. 7. Sentiment Distribution - Polarity vs. Subjectivity. Each point represents an individual agent response, with stars indicating tribe-level averages. Clustering patterns reveal consistent decision-making styles within tribes and clear separation between tribes along the subjectivity dimension.

D. Linguistic Dynamics

All tribes developed strong internal coherence by round 50 (cohesion indices 0.65-0.80) with semantic drift after 30 rounds. The Collective and Forge showed conservative evolution while the Vanguard and Frontiers exhibited faster exploration. This confirms stable cultural equilibria rather than chaotic drift.

V. Discussion

A. Key Findings

Our findings reveal five key patterns: (1) **Path dependence:** early philosophical framing led to enduring differences in discourse and performance; (2) **Innovation advantage:** bold, experimental strategies consistently outperformed stability-focused ones by balancing exploration and exploitation and compounding gains over time; (3) **Cost of risk aversion:** harmony-oriented groups under-explored high-reward strategies; (4) **Rhetoric-performance link:** greater subjectivity correlated with stronger outcomes, possibly due to persuasive effects or broader exploratory dynamics; and (5) **Persistent divergence:** policy-language co-evolution produced stable trajectories that remained distinct despite shared external conditions.

B. Limitations

Key limitations include the designer-defined (rather than fully emergent) initial philosophies, potential bias in dilemma-card environments that may advantage particular orientations, a 100-round horizon that may not capture very long-run dynamics, reliance on a single aggregation rule (majority voting), absence of inter-tribal migration or contact that could illuminate

assimilation or conflict dynamics, and a semantic analysis scope focused on cohesion and drift rather than full topological structure.

C. Implications

For AI governance: (1) **Cultural initialization** can serve as indirect alignment, shaping policies without hard-coding; (2) **diversity as robustness**—maintaining multiple cultural sub-populations improves resilience; (3) **communication co-design**—rhetorical form materially affects outcomes and requires explicit design attention; (4) **lock-in risks**—cultural stability necessitates periodic interventions (migration, crisis simulations) to prevent persistent underperformance.

VI. Conclusion and Future Work

This paper presented a computational framework for studying co-evolution of policy heuristics and rhetorical frames in LLM-based agent populations. Through controlled experiments with four philosophically distinct tribes, we demonstrated that: distinct cultures emerge and persist despite identical pressures; cultural orientation predicts performance (innovation outperformed stability); rhetoric and policy co-evolve meaningfully (subjectivity correlated with governance success); and embedding-based metrics successfully quantify cultural dynamics.

By coupling executable code with natural language argumentation under hybrid evolutionary pressure, we created a generalizable platform for studying sociocultural dynamics in computational systems. As multi-agent AI systems proliferate, understanding cultural emergence, persistence, and performance effects becomes practically essential for governance and safety.

Future directions include: (1) cultural contact experiments via inter-tribal migration to study assimilation/resistance/polarization; (2) alternative governance mechanisms (consensus, weighted voting, markets) to test institutional-cultural interactions; (3) extended temporal horizons (500-1000 rounds) to observe cyclical patterns and regime shifts; (4) real-world applications to organizational design, policy communication, and AI safety.

References

[1] T.-Y. Gao, J. Smith, A. Kumar, et al., "Large language models empowered agent-based modeling and simulation: a survey and perspectives," *Nature Computational Science*, 2024.

[2] J. Haase and M. Pokutta, 'Beyond Static Responses: Multi-Agent LLM Systems as a New Paradigm for Social Science Research," *arXiv:2506.01839*, 2025.

[3] A. Vallinder and E. Hughes, 'Cultural Evolution of Cooperation among LLM Agents," *arXiv:2412.10270*, 2024.

[4] O. Ashery, et al., 'Emergent social conventions and collective bias in LLM populations," *Science Advances*, 2025.

[5] J. Perez, C. Léger, et al., 'Cultural Evolution in Populations of Large Language Models," *arXiv:2403.08882*, 2024.

[6] T. Taniguchi, R. Ueda, et al., 'Generative Emergent Communication: Large Language Model is a Collective World Model," *arXiv:2501.00226*, 2025.

[7] W. J. Wildman, et al., 'The Role of Values in Pandemic Management: An Agent-Based Model," *JASSS*, vol. 27, no. 1, p. 19, 2024.

[8] C. G. Boshuijzen-van Burken, R. van der Veen, et al., 'Agent-Based Modelling of Values: The Case of Value Sensitive Design for Refugee Logistics," *JASSS*, vol. 23, no. 4, p. 6, 2020.

[9] R. Li, et al., 'Schema-Guided Culture-Aware Complex Event Simulation with Multi-Agent Role-Play (MIRIAM)," in *Proc. ACL*, 2025, pp. 421.

[10] S. Yuan, K. Song, et al., 'EvoAgent: Towards Automatic Multi-Agent Generation via Evolutionary Algorithms," *arXiv:2406.14228*, 2024.

[11] L. Fan, H. Zhang, et al., 'AlphaEvolve: Evolutionary Code Generation with LLMs," *Technical report/preprint*, 2023.

[12] A. Wang, J. Morgenstern, et al., 'Large language models that replace human participants can harmfully misportray and flatten identity groups," *arXiv:2402.01908*, 2024.

[13] F. Hashemi and M. Macy, 'Collective Social Behaviors in LLMs," *OpenReview/platform analysis (Chirper.ai study)*, 2024/2025.

[14] I. Horiguchi, T. Yoshida, and T. Ikegami, 'Evolution of Social Norms in LLM Agents using Natural Language," *arXiv:2409.00993*, 2024.

[15] Y.-S. Chuang, A. Goyal, et al., 'Simulating Opinion Dynamics with Networks of LLM-based Agents," in *Proc. NAACL Findings*, 2024, pp. 211.

[16] J. Park, et al., 'A framework for the emergence and analysis of language in social learning agents," *Nature Communications*, 2024.

[17] H. Li, H. Nourkhiz Mahjoub, et al., 'Language Grounded Multi-agent Reinforcement Learning with Human-interpretable Communication," in *Proc. NeurIPS*, 2024.

[18] A. Nisioti, et al., 'Collective Innovation in Groups of Large Language Models," *arXiv:2505.06904*, 2024.

[19] Y.-S. Chuang, S. Suresh, et al., 'The Wisdom of Partisan Crowds: Comparing Collective Intelligence in Humans and LLM-based Agents," *arXiv:2311.09665*, 2023/2024.

[20] D. Ju, A. Williams, et al., 'Sense and Sensitivity: Evaluating the simulation of social dynamics via Large Language Models," *arXiv:2412.05093*, 2024.

[21] Z. Wu, R. Peng, et al., 'LLM-Based Social Simulations Require a Boundary," *arXiv:2506.19806*, 2025.

[22] R. Peters, et al., 'Emergent language: a survey and taxonomy," *Artificial Intelligence Review*, 2025.

[23] A. Nisioti, et al., 'Collective Innovation in Groups of Large Language Models," *arXiv:2505.06904*, 2024.