

Agent-based Modeling: Equilibrium, Echo Chambers, and Efficiency in Hybrid Coevolutionary Opinion Games

Ming-Zhi Jiang* An-Tzu Teng† Jun-En Liu‡ Po-An Chen§ Yung-Ming Li¶

Abstract

Opinions that move closer together in an online network are often considered a sign that people have genuinely reached consensus; whether this narrowing carries a cost is a question existing methods cannot answer. Coevolutionary opinion formation games measure the social cost of equilibria through the Price of Anarchy (PoA), but their agents follow numerical update rules, while opinion simulations with large language model (LLM) agents capture language-based reasoning and report only descriptive indices.

We introduce the Hybrid Coevolutionary Opinion Game (H-COG), in which analytical and LLM-driven agents share one network that is rewired by opinion similarity and initialized from real Reddit data. To our knowledge, no prior framework lets both kinds of agents coevolve with the network inside one formal game, and none has measured the social cost or PoA of LLM-driven populations in opinion formation. We prove that on any fixed network the opinion stage has a unique equilibrium and a closed-form social optimum, and we show that the hybrid game satisfies the assumptions of a convergence theorem for coevolutionary opinion games, so its dynamics reach an approximate Nash equilibrium.

On this foundation, all simulation runs converge. Echo chambers form under every composition, and the initial topology has no detectable effect on them, which points to the rewiring rule as their source. LLM-driven populations show lower polarization, that is, a narrower spread of opinions, with about five times the PoA of analytical populations. Decomposing social cost shows why: about half of the gap is disagreement with neighbors, and the other half comes from agents being pulled away from their own prior positions. In these populations, opinions draw closer largely because agents give up their own positions, at a cost in efficiency.

Keywords: Opinion dynamics, Echo chambers, Price of Anarchy, Language models, Agent-based modeling

1 Introduction

Online platforms have turned the public square into a coevolving system: what a user reads shapes what they think, and what they think reshapes whom they follow. Echo chambers are the visible result. A population that mostly meets what already agrees with it polarizes further, which is why the phenomenon has become a societal concern. Explaining it means tracking opinions and ties together, not either alone.

Two lines of work approach this, and each gives up something the other keeps. Analytical opinion models, among them DeGroot, Friedkin–Johnsen (FJ), bounded confidence and the coevolutionary opinion games [DeGroot, 1974, Friedkin and Johnsen, 1990, Fotakis et al., 2016, Bhawalkar et al., 2013], supply

*Department of Information Management & Finance, National Yang Ming Chiao Tung University, Taiwan. s112701018.mg12@nycu.edu.tw

†Institute of Information Management, National Yang Ming Chiao Tung University, Taiwan. a0783403.c@nycu.edu.tw

‡Department of Information Management & Finance, National Yang Ming Chiao Tung University, Taiwan. neil.mg12@nycu.edu.tw

§Institute of Information Management, National Yang Ming Chiao Tung University, Taiwan. poanchen@nycu.edu.tw

¶Institute of Information Management, National Yang Ming Chiao Tung University, Taiwan. yml@nycu.edu.tw

explicit update rules and formal tools for equilibrium, social cost and the Price of Anarchy (PoA). What they give up is language: each exchange becomes a numerical update, and the semantic content through which people actually persuade one another drops out. LLM-based agent simulations keep that content, letting agents read, reason and remember, and they reproduce echo chambers and polarization [Wang et al., 2025, Chuang et al., 2024, Gu et al., 2025, Piao et al., 2025]. What they give up is the yardstick. These studies aim at behavioral realism, and the simulators that mix LLM agents with rule-based ones do so without a game or a rewiring network [Mou et al., 2024, Yao et al., 2025, Chen, 2026]. No framework yet measures social cost and PoA where language-driven and cost-driven agents coevolve with one network.

H-COG closes that gap by putting both kinds of agent in one network that rewires as opinions change. A Type-C agent minimizes a numerical disagreement cost, while a Type-L agent reads what its neighbors wrote and answers in language. The parameter α fixes how many of each the population holds, both are read on the same continuous scale, and every agent re-chooses its neighbors by closeness of position. Sweeping α therefore compares purely numerical, purely LLM-driven and mixed populations under one set of conditions. Efficiency is measured by social cost and a topology-conditioned PoA, with the observed cost and the planner optimum computed on the same realized network after each rewiring.

The simulations rest on 5,199 `r/politics` comments about gun control and abortion. A fine-tuned RoBERTa regressor maps each comment to a continuous stance in $[-1, +1]$, and stratified samples from the resulting distributions set the agents’ intrinsic opinions. The two topics differ in how their stances are spread, which lets us ask whether the effect of composition survives a change in the starting landscape. We sweep nine mixing levels of α from 0 to 1 across three initial topologies.

1.1 Research Questions

- RQ1.** How large is the efficiency gap between LLM-driven and cost-minimizing populations, and which component of social cost accounts for it?
- RQ2.** How does agent composition affect polarization and echo-chamber structure?
- RQ3.** What role does the rewiring mechanism play, relative to the initial topology, in the formation of echo chambers?

1.2 Contributions

This study makes five contributions:

1. **Hybrid coevolutionary framework.** We introduce H-COG, which integrates cost-minimizing Type-C agents and LLM-driven Type-L agents in a common network whose opinions and connections coevolve.
2. **Equilibrium analysis.** We formalize H-COG as a multi-player general-sum Markov game, prove that its opinion stage has a unique best response, an invariant opinion interval, a unique Nash equilibrium on every fixed graph and a closed-form planner optimum, show that K-NN rewiring is the high-intensity limit of a stochastic kernel, and combine these results with a convergence theorem for Markov coevolutionary opinion games.
3. **Empirically grounded initialization.** We construct a two-stage pipeline that extracts real-world stance distributions from Reddit using semantic clustering and a fine-tuned RoBERTa regressor, then initializes agents through stratified sampling.

4. **Common efficiency evaluation.** We formulate social cost and topology-conditioned PoA on a shared opinion scale, and decompose social cost into disagreement and conformity components to identify the source of efficiency differences.
5. **Systematic population comparison.** We compare nine Type-C/Type-L mixing levels across two topics and three initial topologies, evaluating efficiency, polarization, echo-chamber structure, and convergence behavior.

2 Related Work

2.1 Opinion Dynamics and Coevolutionary Games

Opinion dynamics begins with agents that hold a number and update it by averaging their neighbors. DeGroot formalized that process and showed that under mild conditions repeated averaging drives the whole network to consensus [DeGroot, 1974]. Friedkin and Johnsen refined it by separating the belief an agent holds privately from the one it expresses, which lets the system settle where disagreement survives instead of vanishing [Friedkin and Johnsen, 1990].

Bhawalkar et al. took the further step of letting the network move as well [Bhawalkar et al., 2013]. In their K -NN game each agent links to the K agents whose expressed opinions lie closest to its intrinsic opinion and picks an expressed opinion that trades social pressure against private belief. That setup admits a Price of Anarchy (PoA) analysis. The PoA of the K -NN game is bounded by a constant when agents weight their own opinion above their neighbors’ ($\rho > 1$) and is at least $1/\rho^2$ below it, and a pure equilibrium need not exist; H-COG with $\alpha = 1$ is this game with heterogeneous stubbornness. On fixed undirected graphs the Friedkin–Johnsen game loses at most a factor of $9/8$, and on directed graphs the loss is unbounded [Bindel et al., 2015]. Chen et al. bound the PoA for directed graphs in which each node influences the others no more than it is influenced by them [Chen et al., 2016]. Chen et al. [2026b] make K -NN rewiring stochastic, cast the game as a general-sum Markov game, and show that optimistic gradient ascent reaches an approximate Nash equilibrium; Section 4 builds on this result. Related work extends the analysis to competitive opinion optimization [Chen et al., 2024], designs networks that minimize polarization and disagreement at the Friedkin–Johnsen equilibrium [Musco et al., 2018], and proves convergence of Hegselmann–Krause dynamics with restricted neighborhoods [Fotakis et al., 2016]. All of these models use numerical update rules and leave out the reasoning, memory and meaning through which people form and state an opinion.

2.2 LLM-based Agent Simulation for Echo Chambers

Large language models let an agent read text, reason about it and answer in kind. Park et al. showed that such agents produce believable social behavior in interactive settings [Park et al., 2023].

Wang et al. placed LLM agents on fixed network structures, with a recommendation algorithm selecting whom each agent hears, and found that the agents track observed polarization more closely than Friedkin–Johnsen or bounded-confidence models do [Wang et al., 2025]. Chuang et al. showed that networks of LLM agents drift toward consensus unless confirmation bias is induced [Chuang et al., 2024]. Gu et al. let the language model decide rewiring as well, first on real Twitter data [Gu et al., 2025] and then with paired rewiring validated on three real datasets [Gu et al., 2026], and Piao et al. observed polarization emerging in networks that thousands of LLM agents form themselves [Piao et al., 2025].

A second line mixes LLM agents with rule-based ones. HiSim pairs LLM core users with ordinary users governed by bounded-confidence and related models, initialized from real tweets [Mou et al., 2024]; FDE-LLM couples LLM opinion leaders with dynamics equations on real Weibo data [Yao et al., 2025]; and AgoraSim runs populations that mix LLM and classical agents at controlled ratios [Chen, 2026]. None

of these simulators rewires the network or places the two types in a game. Qasmi et al. place LLM agents in a non-cooperative game of opinion polarization without an equilibrium or efficiency analysis [Qasmi et al., 2025], and welfare measures for LLM agent societies exist outside opinion formation [Piatti et al., 2024].

These studies report descriptive indices such as polarization and modularity. None measures what the resulting equilibria cost, and none places LLM agents and Friedkin–Johnsen best-response agents in one coevolving game.

3 Preliminaries: Formalization of the Hybrid Markov Coevolutionary Opinion Formation Game

3.1 Agent Population and Opinion Variables

Let $\mathcal{N} = \{1, \dots, N\}$ be a set of N agents. Each agent i has an **intrinsic opinion** $s_i \in [-1, +1]$, representing a fixed prior stance, and an **expressed opinion** $z_i^{(t)} \in [-1, +1]$, which evolves as the agent responds to its neighbors.

The mixing parameter $\alpha \in [0, 1]$ partitions the population into two types. **Type-C** agents (fraction α) update their expressed opinions by minimizing a local disagreement cost derived from the Friedkin–Johnsen model [Friedkin and Johnsen, 1990]. **Type-L** agents (fraction $1 - \alpha$) generate a natural-language opinion $o_i^{(t)} \in \mathcal{V}^*$, where $\mathcal{V}$ is a finite token vocabulary. To align text-based interaction with numerical game-theoretic optimization, we define a continuous semantic mapping

$$e : \mathcal{V}^* \rightarrow [-1, +1]$$

that places this opinion on the common numerical scale, $z_i^{(t)} = e(o_i^{(t)})$. The cases $\alpha = 1$ and $\alpha = 0$ give the purely numerical and purely LLM-driven baselines, respectively.

We write $\mathcal{N}_C$ and $\mathcal{N}_L$ for the sets of Type-C and Type-L agents. Every agent carries a stubbornness coefficient $\rho_i > 0$. For a Type-C agent it enters the update rule, and for a Type-L agent it enters only the social cost. Opinions are represented at finite precision: we write $\mathcal{Z} \subset [-1, +1]$ for the finite set of representable opinions and $\mathcal{V}^{\leq M}$ for texts of at most M tokens.

3.2 State Space and Opinion Updates

The global state at step t is

$$S^{(t)} = \left(G^{(t)}, \mathbf{z}^{(t)}, \mathbf{X}^{(t)} \right),$$

where $G^{(t)} = (S_1^{(t)}, \dots, S_N^{(t)})$ is the directed network, $S_i^{(t)} \subset \mathcal{N} \setminus \{i\}$ is agent i ’s out-neighborhood with $|S_i^{(t)}| = K$ for $t \geq 1$ ($S_i^{(0)}$ is i ’s neighbor set in the initial graph, on which Eq. (1) is applied with K replaced by $|S_i^{(0)}|$, and with $z_i^{(1)} = s_i$ if $S_i^{(0)} = \emptyset$), $\mathbf{z}^{(t)}$ is the opinion profile from which $G^{(t)}$ was built, and $\mathbf{X}^{(t)} = (\mathbf{x}_i^{(t)})_{i \in \mathcal{N}_L}$ collects the structured text records of the Type-L agents that produced $\mathbf{z}^{(t)}$.

Type-C analytical agents

A Type-C agent chooses an expressed opinion from $\mathcal{A}_i = \mathcal{Z}$. Holding its neighbors’ latest opinions fixed, its myopic stage-cost minimizer is

$$z_i^{(t+1)} = \arg \min_{z \in [-1, +1]} C_i \left(G^{(t)}, (z, \mathbf{z}_{-i}^{(t)}) \right) = \frac{\sum_{j \in S_i^{(t)}} z_j^{(t)} + \rho_i K s_i}{K(1 + \rho_i)}. \quad (1)$$

This is the unique minimizer of the stage cost, and it stays in $[-1, +1]$ (Lemmas 1 and 2). In the Markov-game representation, a more general Type-C policy $\pi_i^C(\cdot | S^{(t)})$ may instead optimize expected discounted costs; the simulations use the myopic update in Eq. (1).

Type-L generative agents

A Type-L action is a structured sequence $\mathbf{x}_i^{(t+1)} \in \mathcal{V}^{\leq M}$ containing reasoning, opinion, and memory in that order. Its prompt combines the previous record, current neighbor messages, and intrinsic stance:

$$P_i^{(t)} = \text{Prompt}\left(\mathbf{x}_i^{(t)}, \{\mathbf{x}_j^{(t)}\}_{j \in S_i^{(t)}}, s_i\right).$$

An autoregressive LLM with parameters θ generates $\mathbf{x}_i^{(t+1)} = (y_1, \dots, y_m)$ according to

$$\mathbb{P}_\theta\left(\mathbf{x}_i^{(t+1)} | P_i^{(t)}\right) = \prod_{k=1}^m \mathbb{P}_\theta\left(y_k | y_{<k}, P_i^{(t)}\right).$$

Given a temperature $\tau > 0$, each next-token probability is the softmax

$$\mathbb{P}_\theta\left(y_k | y_{<k}, P_i^{(t)}\right) = \frac{\exp\left(g_\theta\left(y_k | y_{<k}, P_i^{(t)}\right) / \tau\right)}{\sum_{y \in \mathcal{V}} \exp\left(g_\theta\left(y | y_{<k}, P_i^{(t)}\right) / \tau\right)},$$

where $g_\theta(\cdot)$ denotes the unnormalized logit output by the LLM. This temperature- τ softmax sampling induces a stochastic behavior policy over $\mathcal{Z}$,

$$\pi_i^L\left(z | S^{(t)}\right) = \sum_{\mathbf{x} \in \mathcal{V}^{\leq M}: e(o(\mathbf{x}))=z} \mathbb{P}_\theta\left(\mathbf{x} | P_i^{(t)}\right),$$

where $o(\mathbf{x})$ is the opinion field of $\mathbf{x}$. The simulations use deterministic decoding (the limit $\tau \rightarrow 0$). Each Type-L agent makes one LLM call per step, and each call rewrites the memory field. Only the free-form opinion field $o_i^{(t+1)}$ is passed to the semantic mapping, giving $z_i^{(t+1)} = e(o_i^{(t+1)}) \in [-1, +1]$.

3.3 Coevolutionary Rewiring and Transition Kernel

After all opinions are updated, the network changes from $G^{(t)}$ to $G^{(t+1)}$. Let $d_{ij}^{(t+1)} = |z_j^{(t+1)} - s_i|$ be the distance between peer j 's expressed opinion and agent i 's intrinsic opinion. A general stochastic rewiring kernel assigns attachment weight

$$w_{ij}^{(t+1)} = \exp\left(-\beta d_{ij}^{(t+1)}\right), \quad \beta > 0,$$

and draws each K -subset S with probability

$$\mathbb{P}\left(S_i^{(t+1)} = S | \mathbf{z}^{(t+1)}, s_i\right) = \frac{\prod_{j \in S} w_{ij}^{(t+1)}}{\sum_{\substack{S' \subset \mathcal{N} \setminus \{i\} \\ |S'|=K}} \prod_{k \in S'} w_{ik}^{(t+1)}}.$$

The experiments use the $\beta \rightarrow \infty$ limit:

$$S_i^{(t+1)} = \arg \operatorname{top-K}_{j \neq i} \left(-|s_i - z_j^{(t+1)}|\right), \quad K = 5. \quad (2)$$

Each agent therefore connects to the five peers whose expressed opinions lie nearest its own intrinsic stance, with ties broken uniformly at random (seeded). Lemma 4 characterizes this limit. Rewiring is a kernel of the environment, not an action any agent chooses.

Under conditional independence across agents, the transition kernel of the full state combines the rewiring draws with the Type-L texts:

$$\mathcal{P}\left(S^{(t+1)} \mid S^{(t)}, \mathbf{a}^{(t)}\right) = \left(\prod_{i=1}^N \mathbb{P}\left(S_i^{(t+1)} \mid \mathbf{z}^{(t+1)}, s_i\right)\right) \times \mathbf{1}\left[\mathbf{X}^{(t+1)} = \left(\mathbf{a}_i^{(t)}\right)_{i \in \mathcal{N}_L}\right] \times \mathbf{1}\left[\mathbf{z}^{(t+1)} = \mathbf{z}(\mathbf{a}^{(t)})\right],$$

where $\mathbf{z}(\mathbf{a}^{(t)})$ is the opinion profile the joint action is scored to. Section 4.2 shows that this state space is finite.

3.4 Social Cost and Price of Anarchy

Regardless of agent type, agent i incurs the stage cost

$$C_i\left(G^{(t)}, \mathbf{z}^{(t)}\right) = \sum_{j \in S_i^{(t)}} (z_i^{(t)} - z_j^{(t)})^2 + \rho_i K(z_i^{(t)} - s_i)^2.$$

The global social cost is the sum of individual costs:

$$SC(G, \mathbf{z}) = \underbrace{\sum_i \sum_{j \in S_i} (z_i - z_j)^2}_{\text{disagreement}} + \underbrace{\sum_i \rho_i K(z_i - s_i)^2}_{\text{conformity}}. \quad (3)$$

The first component measures disagreement with current neighbors, while the second measures displacement from intrinsic opinions.

Definition 1 (Nash equilibrium). *A joint strategy profile $\boldsymbol{\pi}^* = (\pi_1^*, \dots, \pi_N^*)$ is a Nash equilibrium if:*

1. *for every Type-C agent $i \in \mathcal{N}_C$, π_i^* minimizes its expected discounted cumulative stage cost given $\boldsymbol{\pi}_{-i}^*$ from every initial state $S^{(0)}$,*

$$\pi_i^* \in \arg \min_{\pi_i} \mathbb{E}_{(\pi_i, \boldsymbol{\pi}_{-i}^*)} \left[\sum_{t=0}^{\infty} \gamma^t C_i(G^{(t)}, \mathbf{z}^{(t+1)}) \mid S^{(0)} \right];$$

2. *for every Type-L agent $j \in \mathcal{N}_L$, π_j^* is induced by the autoregressive LLM kernel $\mathbb{P}_\theta(\mathbf{x} \mid P_j)$.*

The stage cost is charged on the graph the action faces. For $\gamma = 0$ the Type-C condition is the stage best response of Lemma 1 on $G^{(t)}$; Eq. (1) iterates this best response against the previous opinions, and the two coincide on a stationary graph (Lemma 3). Under assumption (i) of Section 4.3 the LLM kernel is a point of the Type-L policy simplex, so the Nash gap of Theorem 1 is taken over all players.

Definition 2 (Generalized Price of Anarchy). *Let $\mathcal{E}_{\text{NE}}(\alpha)$ denote the stationary state-action distributions resulting from an equilibrium at mixture level α . The generalized hybrid PoA is*

$$\text{PoA}(\alpha) = \frac{\max_{\nu \in \mathcal{E}_{\text{NE}}(\alpha)} \mathbb{E}_{(G, \mathbf{z}) \sim \nu} [SC(G, \mathbf{z})]}{\min_{G^*, \mathbf{z}^*} SC(G^*, \mathbf{z}^*)}. \quad (4)$$

The denominator ranges over all K -out graphs, a relaxation of the K -NN optimum of Bhawalkar et al. [2013].

This generalized definition allows $\text{PoA}(\alpha)$ to be analyzed and computed over the whole range $\alpha \in [0, 1]$, and it separates the effect of generative LLM reasoning on social welfare and network stability.

The experiments report a topology-conditioned operational measure. At step t , the planner optimizes opinions on the same realized graph:

$$\mathbf{z}_t^{\text{opt}} = \arg \min_{\mathbf{z} \in [-1, +1]^N} SC(G^{(t)}, \mathbf{z}),$$

and

$$\text{PoA}^{(t)} = \frac{SC(G^{(t)}, \mathbf{z}^{(t)})}{SC(G^{(t)}, \mathbf{z}_t^{\text{opt}})} \geq 1. \quad (5)$$

Here $G^{(t)}$ is the graph rebuilt from $\mathbf{z}^{(t)}$, so $\text{PoA}^{(t)}$ is evaluated after rewiring. Lemma 5 gives the minimizer in closed form and shows that this ratio is well defined. Since $SC(G, \mathbf{z}^{\text{opt}}(G)) \geq \min_{G^*, \mathbf{z}^*} SC(G^*, \mathbf{z}^*)$ for every G , every $\nu \in \mathcal{E}_{\text{NE}}(\alpha)$ satisfies $\mathbb{E}_\nu [SC(G, \mathbf{z})/SC(G, \mathbf{z}^{\text{opt}}(G))] \leq \text{PoA}(\alpha)$, so the expected topology-conditioned ratio is a lower bound on Eq. (4). Fixing the graph isolates the welfare loss associated with decentralized opinion updating from that associated with the topology produced by rewiring.

4 Equilibrium Analysis

In stochastic coevolutionary opinion formation games, network topologies evolve through randomized rewiring, so small shifts in expressed opinions may drastically alter global transition probabilities [Chen et al., 2026a,b]. By framing the hybrid population of analytical cost-minimizing agents ($\mathcal{N}_C$) and LLM generative agents ($\mathcal{N}_L$) as a multi-player general-sum Markov game and combining this formulation with Theorem 2 of Chen et al. [2026b], we establish convergence to an approximate Nash equilibrium under Optimistic Gradient Ascent (OGA) learning dynamics.

4.1 Properties of the Opinion Stage

Throughout this subsection we fix a step and suppress the superscript (t) where no confusion arises. We write z^{BR} for a best response, $\mathbf{z}^{\text{NE}}$ for a Nash fixed point and $\mathbf{z}^{\text{opt}}$ for the planner's optimum.

Lemma 1 (Type-C best response). *Fix a Type-C agent i with out-neighborhood S_i , $|S_i| = K$, and the opinions $\mathbf{z}_{-i}$ of the other agents. The function*

$$c_i(z) = \sum_{j \in S_i} (z - z_j)^2 + \rho_i K (z - s_i)^2,$$

which is the stage cost of Section 3.4 with z_i replaced by the free variable z , is strictly convex and has the unique minimizer

$$z_i^{\text{BR}} = \frac{\sum_{j \in S_i} z_j + \rho_i K s_i}{K(1 + \rho_i)}.$$

Proof. The function c_i is a sum of $K + 1$ squares, so $c_i''(z) = 2K + 2\rho_i K = 2K(1 + \rho_i)$, which is positive because $\rho_i > 0$. Hence c_i is strictly convex. Since $c_i(z) \rightarrow \infty$ as $|z| \rightarrow \infty$, a minimizer exists, and strict convexity makes it unique. Setting $c_i'(z) = 2 \sum_{j \in S_i} (z - z_j) + 2\rho_i K (z - s_i)$ to zero gives $K(1 + \rho_i) z = \sum_{j \in S_i} z_j + \rho_i K s_i$, which is the stated expression. $\square$

Lemma 2 (Invariance of the opinion interval). *If $s_i \in [-1, 1]$ for all i , then $z_i^{(t)} \in [-1, 1]$ for all $i \in \mathcal{N}$ and all $t \geq 0$.*

Proof. For a Type-L agent and every opinion text o , the score $e(o) = \tanh(\mathbf{W} \mathbf{h}_{[\text{CLS}]}(o))$ of Section 5.2 lies in $(-1, 1)$, independently of t . For a Type-C agent we argue by induction on t . At $t = 0$, $z_i^{(0)} = s_i \in [-1, 1]$. Suppose $z_j^{(t)} \in [-1, 1]$ for all j . By Lemma 1, $z_i^{(t+1)}$ is a weighted combination of $\{z_j^{(t)}\}_{j \in S_i^{(t)}}$ and s_i with weight $\frac{1}{|S_i^{(t)}|(1+\rho_i)}$ on each of the $|S_i^{(t)}|$ neighbors and $\frac{\rho_i}{1+\rho_i}$ on the intrinsic opinion (and $z_i^{(t+1)} = s_i$ if $S_i^{(t)} = \emptyset$). These weights are nonnegative and sum to $\frac{1}{1+\rho_i} + \frac{\rho_i}{1+\rho_i} = 1$, so $z_i^{(t+1)}$ is a convex combination of points of $[-1, 1]$ and lies in $[-1, 1]$. $\square$

Consequently the action space of a Type-C agent may be taken as $\mathcal{Z} \subset [-1, 1]$ without loss of generality.

Lemma 3 (Equilibrium of the opinion stage on a fixed graph). *Fix G and the Type-L opinions $\mathbf{z}_L \in [-1, 1]^{|\mathcal{N}_L|}$. Define $\Phi : [-1, 1]^{|\mathcal{N}_C|} \rightarrow [-1, 1]^{|\mathcal{N}_C|}$ by*

$$\Phi_i(\mathbf{z}_C) = \frac{\sum_{j \in S_i} z_j + \rho_i K s_i}{K(1 + \rho_i)}, \quad i \in \mathcal{N}_C,$$

where z_j is taken from $\mathbf{z}_L$ for $j \in \mathcal{N}_L$ (we write $\Phi(\mathbf{z}_C; \mathbf{z}_L)$ when the dependence on $\mathbf{z}_L$ matters). Then

1. Φ is a contraction in $\|\cdot\|_\infty$ with modulus $q = \max_{i \in \mathcal{N}_C} \frac{1}{1+\rho_i} < 1$;
2. Φ has a unique fixed point $\mathbf{z}_C^{\text{NE}}$, which is the unique Nash equilibrium of the opinion stage among Type-C agents given $(G, \mathbf{z}_L)$;
3. the synchronous iterates $\mathbf{z}_C^{(t+1)} = \Phi(\mathbf{z}_C^{(t)})$ satisfy $\|\mathbf{z}_C^{(t)} - \mathbf{z}_C^{\text{NE}}\|_\infty \leq 2q^t$, with t counted from the first step on G ;
4. if the Type-L opinions vary, $\mathbf{z}_C^{(t+1)} = \Phi(\mathbf{z}_C^{(t)}; \mathbf{z}_L^{(t)})$, and $\mathbf{z}_C^{\text{NE}}$ is taken at $\mathbf{z}_L$, then $\|\mathbf{z}_C^{(t)} - \mathbf{z}_C^{\text{NE}}\|_\infty \leq \max\{2q^t, q \max_{u < t} \|\mathbf{z}_L^{(u)} - \mathbf{z}_L\|_\infty\}$.

For $\alpha = 1$ ($\mathcal{N}_L = \emptyset$), $\mathbf{z}^{\text{NE}}$ is the Friedkin–Johnsen equilibrium on G .

Proof. Φ maps $[-1, 1]^{|\mathcal{N}_C|}$ into itself by the convex-combination argument of Lemma 2. For $\mathbf{z}_C, \mathbf{z}'_C$ and $i \in \mathcal{N}_C$, the intrinsic term and the Type-L neighbors cancel, so only Type-C neighbors contribute:

$$\begin{aligned} |\Phi_i(\mathbf{z}_C) - \Phi_i(\mathbf{z}'_C)| &\leq \frac{1}{K(1 + \rho_i)} \sum_{j \in S_i \cap \mathcal{N}_C} |z_j - z'_j| \\ &\leq \frac{|S_i \cap \mathcal{N}_C|}{K(1 + \rho_i)} \|\mathbf{z}_C - \mathbf{z}'_C\|_\infty \leq \frac{1}{1 + \rho_i} \|\mathbf{z}_C - \mathbf{z}'_C\|_\infty, \end{aligned}$$

which proves (1). The same estimate holds on all of $\mathbb{R}^{|\mathcal{N}_C|}$. Since $[-1, 1]^{|\mathcal{N}_C|}$ is complete, Banach's fixed-point theorem gives a unique fixed point, which by the same estimate is also unique in $\mathbb{R}^{|\mathcal{N}_C|}$, and $\|\mathbf{z}_C^{(t)} - \mathbf{z}_C^{\text{NE}}\|_\infty \leq q^t \|\mathbf{z}_C^{(0)} - \mathbf{z}_C^{\text{NE}}\|_\infty \leq 2q^t$, which proves (3). A profile is a fixed point of Φ if and only if every Type-C agent plays its best response to the others (Lemma 1), that is, if and only if it is a Nash equilibrium of the opinion stage; this proves (2). For (4), the estimate of (1) extends to $|\Phi_i(\mathbf{z}_C; \mathbf{z}_L) - \Phi_i(\mathbf{z}'_C; \mathbf{z}'_L)| \leq \frac{1}{1+\rho_i} \max\{\|\mathbf{z}_C - \mathbf{z}'_C\|_\infty, \|\mathbf{z}_L - \mathbf{z}'_L\|_\infty\}$, so the error $e_t = \|\mathbf{z}_C^{(t)} - \mathbf{z}_C^{\text{NE}}\|_\infty$ satisfies $e_{t+1} \leq q \max\{e_t, \|\mathbf{z}_L^{(t)} - \mathbf{z}_L\|_\infty\}$, and induction from $e_0 \leq 2$ gives $e_t \leq \max\{2q^t, q \max_{u < t} \|\mathbf{z}_L^{(u)} - \mathbf{z}_L\|_\infty\}$. $\square$

Lemma 4 (K-nearest-neighbor rewiring as the high-intensity limit). *For agent i and a K -subset $S' \subset \mathcal{N} \setminus \{i\}$, let $D_i(S') = \sum_{j \in S'} d_{ij}$ with $d_{ij} = |z_j - s_i|$, let $D_i^* = \min_{S'} D_i(S')$ and $\mathcal{M}_i = \arg \min_{S'} D_i(S')$, and order the distances from i as $d_{(1)} \leq d_{(2)} \leq \dots \leq d_{(N-1)}$. As $\beta \rightarrow \infty$, the rewiring probability of*

Section 3.3 tends to $1/|\mathcal{M}_i|$ for $S' \in \mathcal{M}_i$ and to 0 otherwise. A K -subset belongs to $\mathcal{M}_i$ if and only if it contains every peer with $d_{ij} < d_{(K)}$ and fills its remaining places with peers at distance exactly $d_{(K)}$. If $d_{(K)} < d_{(K+1)}$, $\mathcal{M}_i$ is the single K -nearest-neighbor set; otherwise the limit fills the remaining places uniformly at random from the tied peers.

Proof. Multiplying the numerator and the denominator of the rewiring probability by $\exp(\beta D_i^*)$ gives

$$P_\beta(S') = \frac{\exp(-\beta(D_i(S') - D_i^*))}{\sum_{|S''|=K} \exp(-\beta(D_i(S'') - D_i^*))}.$$

Every exponent is now nonpositive. A subset in $\mathcal{M}_i$ contributes exactly 1, and a subset outside it has $D_i(S'') - D_i^* > 0$ and contributes a term that falls to 0 as $\beta \rightarrow \infty$. The denominator therefore tends to $|\mathcal{M}_i|$, while the numerator tends to 1 when $S' \in \mathcal{M}_i$ and to 0 otherwise.

It remains to identify $\mathcal{M}_i$. Since $D_i(S')$ sums K of the ordered distances, $D_i(S') \geq \sum_{r=1}^K d_{(r)}$ for every K -subset, and a subset of the stated form attains that bound, so $D_i^* = \sum_{r=1}^K d_{(r)}$ and every such subset lies in $\mathcal{M}_i$. Conversely, let S' attain D_i^* . If S' contains a peer u with $d_{iu} > d_{(K)}$, at most $K-1$ of its members lie within $d_{(K)}$, so some peer $v \notin S'$ has $d_{iv} \leq d_{(K)} < d_{iu}$, and exchanging u for v lowers D_i . If S' omits a peer v with $d_{iv} < d_{(K)}$, then, because at most $K-1$ peers lie strictly closer than $d_{(K)}$, S' contains a peer u with $d_{iu} = d_{(K)}$, and exchanging u for v again lowers D_i . Both cases contradict minimality, so every minimizer has the stated form. When $d_{(K)} < d_{(K+1)}$, exactly K peers lie within $d_{(K)}$ and $|\mathcal{M}_i| = 1$; otherwise the limit spreads its mass evenly over the subsets that differ only in which of the tied peers they take. $\square$

The kernel factorizes across agents (Eq. (6)), so the joint limit is the product of the per-agent limits.

For a fixed G , write $m_{ij} = \mathbf{1}[j \in S_i] + \mathbf{1}[i \in S_j]$ for the weights of the undirected multigraph induced by G , L for its Laplacian, and $R = \text{diag}(\rho_1, \dots, \rho_N)$. Each unordered pair $\{i, j\}$ appears m_{ij} times in the disagreement term of Eq. (3), so

$$SC(G, \mathbf{z}) = \mathbf{z}^\top L \mathbf{z} + (\mathbf{z} - \mathbf{s})^\top K R (\mathbf{z} - \mathbf{s}).$$

Lemma 5 (The planner optimum and well-posedness of $\text{PoA}^{(t)}$). *For every G :*

1. $SC(G, \cdot)$ is strictly convex and has the unique minimizer $\mathbf{z}^{\text{opt}} = (L + KR)^{-1} K R \mathbf{s}$;
2. each z_i^{opt} is a convex combination of $s_1, \dots, s_N$; in particular $\mathbf{z}^{\text{opt}} \in [-1, 1]^N$;
3. $SC^{\text{opt}} = \min_{\mathbf{z}} SC(G, \mathbf{z}) > 0$ whenever some edge $j \in S_i$ joins agents with $s_i \neq s_j$. In that case $\text{PoA}^{(t)} = SC(G^{(t)}, \mathbf{z}^{(t)}) / SC^{\text{opt}}$ is well defined and $\text{PoA}^{(t)} \geq 1$.

Proof. (1) The Hessian is $2(L + KR)$. Since $L \succeq 0$ and, because every $\rho_i > 0$, $KR \succ 0$, we have $L + KR \succ 0$, so SC is strictly convex. Setting the gradient $2L\mathbf{z} + 2KR(\mathbf{z} - \mathbf{s})$ to zero gives $(L + KR)\mathbf{z} = KR\mathbf{s}$.

(2) Let $M = L + KR$. Its off-diagonal entries $-m_{ij}$ are nonpositive and each row is strictly diagonally dominant, since $M_{ii} = \sum_j m_{ij} + K\rho_i > \sum_{j \neq i} |M_{ij}|$. Hence M is a nonsingular M-matrix and $M^{-1} \geq 0$ entrywise, so $W = M^{-1}KR \geq 0$. Since $L\mathbf{1} = 0$, $M\mathbf{1} = KR\mathbf{1}$, and therefore $W\mathbf{1} = \mathbf{1}$. Each row of W is thus a probability vector, and $\mathbf{z}^{\text{opt}} = W\mathbf{s}$ is a row-wise convex combination of $\mathbf{s}$.

(3) Both terms of SC are nonnegative, and $SC(G, \mathbf{z}) = 0$ requires $z_i = s_i$ for all i and $z_i = z_j$ for every edge, hence $s_i = s_j$ for every edge. If some edge has $s_i \neq s_j$, then $SC(G, \mathbf{z}) > 0$ for all $\mathbf{z}$, and since the minimum is attained by (1), $SC^{\text{opt}} > 0$. Finally $SC(G^{(t)}, \mathbf{z}^{(t)}) \geq SC^{\text{opt}}$ by definition of the minimum. $\square$

In the sampled populations no three agents share an intrinsic opinion and $K = 5$, so every realized graph contains an edge joining different intrinsic opinions and the stepwise PoA is well defined.

Remark 1 (Cost decomposition). *Writing Eq. (3) as $SC = \mathcal{D} + \mathcal{C}$ (disagreement and conformity) and dividing by SC^{opt} gives $\text{PoA}^{(t)} = \mathcal{D}^{(t)}/SC^{\text{opt}} + \mathcal{C}^{(t)}/SC^{\text{opt}}$.*

Remark 2 (The 9/8 bound). *At $\alpha = 1$ H-COG is the K -NN game of Bhawalkar et al. [2013] with heterogeneous stubbornness; on a stationary graph Lemmas 3 and 5 identify $\text{PoA}^{(t)}$ with the equilibrium-to-optimum ratio of the Friedkin–Johnsen opinion game on $G^{(t)}$. Bindel et al. [2015] and Bhawalkar et al. [2013] bound this ratio by 9/8 on undirected graphs, and about 85% of the edges in H-COG graphs are reciprocal.*

4.2 The Hybrid Game as a Markov Game

We write the Hybrid Coevolutionary Opinion Game as a multi-player general-sum Markov game

$$\mathcal{M}_{\text{hybrid}} = \langle \mathcal{N}, \mathcal{S}, \mathcal{A}, \mathcal{P}, \{r^i\}_{i \in \mathcal{N}}, \gamma \rangle,$$

whose components are the following.

Players. $\mathcal{N} = \mathcal{N}_C \cup \mathcal{N}_L$, with $|\mathcal{N}_L| = \text{round}((1 - \alpha)N)$ Type-L agents and $|\mathcal{N}_C| = N - |\mathcal{N}_L|$ Type-C agents.

States. A global state at step t is the triple $S^{(t)} = (G^{(t)}, \mathbf{z}^{(t)}, \mathbf{X}^{(t)})$ of Section 3.2: the active out-edges, the opinions from which they were built, and the Type-L texts that produced those opinions. Generated texts have bounded length and opinions are represented at finite precision, so $\mathcal{S}$ is finite, the property the cited result requires.

Actions. $\mathcal{A} = \prod_{i=1}^N \mathcal{A}_i$. A Type-C agent plays $a_i = z_i \in \mathcal{Z}$ directly; a Type-L agent plays $a_i = \mathbf{x}_i \in \mathcal{V}^{\leq M}$, which the semantic mapping carries to $z_i = e(o(\mathbf{x}_i)) \in [-1, +1]$.

Transition kernel. The opinion and text components of the next state are determined by the joint action, so the kernel of Section 3.3 is governed by its graph factor. Write $\mathbf{z}^{(t+1)}$ for the opinions the joint action $\mathbf{a}^{(t)}$ is scored to, and $\mathbf{s}$ for the fixed intrinsic opinions. Rewiring snapshots those opinions and then lets every agent rank the same snapshot on its own, so the N draws are conditionally independent and

$$\mathcal{P}(G^{(t+1)} \mid S^{(t)}, \mathbf{a}^{(t)}) = \prod_{i=1}^N \mathbb{P}(S_i^{(t+1)} \mid \mathbf{z}^{(t+1)}, s_i), \quad (6)$$

with each factor the Gibbs form of Section 3.3. Two properties of Eq. (6) matter for what follows. It does not depend on $G^{(t)}$, because the rule rebuilds every out-neighborhood from the current opinions rather than editing the previous edge set, so the chain is driven by the opinions alone. And in the $\beta \rightarrow \infty$ limit used throughout the experiments each factor collapses to the uniform distribution over the minimizers $\mathcal{M}_i$ of the total distance (Lemma 4).

Rewards. The reward is the negative stage cost of the opinions the joint action is scored to, on the graph that action faces, $r^i(S^{(t)}, \mathbf{a}^{(t)}) = -C_i(G^{(t)}, \mathbf{z}^{(t+1)})$, and since $z_i, s_i \in [-1, +1]$ it is bounded by $|r^i| \leq 4(1 + \rho_i)K$.

4.3 Structural Assumptions

The learning dynamics of the hybrid Markov coevolutionary opinion game incorporate the structural assumptions of Chen et al. [2026b]. We write $\sigma \in \mathcal{S}$ for states and k for OGA iterations.

- (i) **Norm Pairs and Diameter Bounds:** Generated texts have bounded length and opinions are represented at finite precision, so each action set $\mathcal{A}_i$ is finite and each state-conditioned policy space $\mathcal{X}_i^\sigma$ is the probability simplex over $\mathcal{A}_i$. Under the Euclidean norm, which is its own dual, the diameter of the simplex gives

$$\Omega_i^\sigma \triangleq \sup_{x, x' \in \mathcal{X}_i^\sigma} \|x - x'\| = \sqrt{2} < \infty,$$

and the dual norm pair satisfies $\|\cdot\| \geq C\|\cdot\|_1$ and $\|\cdot\|_* \leq C^*\|\cdot\|_\infty$ with $C = \min_i |\mathcal{A}_i|^{-1/2}$ and $C^* = \max_i |\mathcal{A}_i|^{1/2}$ [Chen et al., 2026a].

- (ii) **Smooth Regularizer and OGA Updates:** Each player updates its state-conditioned strategy $\pi_{i,k}^\sigma$ using Optimistic Gradient Ascent with ℓ_i -smooth regularizers $\mathcal{R}_i$ [Chen et al., 2026a]:

$$\begin{aligned} \hat{x}_{i,k+1}^\sigma &= \Pi_{\mathcal{X}_i^\sigma} \{ \hat{x}_{i,k}^\sigma + \eta u_{i,k}^\sigma \} \\ x_{i,k+1}^\sigma &= \Pi_{\mathcal{X}_i^\sigma} \{ \hat{x}_{i,k+1}^\sigma + \eta u_{i,k}^\sigma \} \end{aligned}$$

where $\Pi_{\mathcal{X}_i^\sigma}$ is the Euclidean projection onto the simplex and $u_{i,k}^\sigma$ is the optimistic gradient estimator, with $\|u_{i,k}^\sigma - Q_{i,k}^\sigma(\otimes_{j \neq i} \pi_{j,k}^\sigma)\|_\infty \leq \epsilon$ [Chen et al., 2026a]. Here $Q_{i,k}^\sigma(\cdot) \in \mathbb{R}^{|\mathcal{A}_i|}$ is player i 's state-action value vector at σ against the other players' current policies, and the Nash gap of a profile at σ is the largest gain that any single player can obtain at σ by a unilateral change of its policy. With the Euclidean regularizer $\mathcal{R}_i(x) = \frac{1}{2}\|x\|_2^2$, $\ell_i = 1$.

- (iii) **Learning Rate Regime (η):** The learning rate satisfies $\eta \leq O(1/\sqrt{T})$, which offsets the positive regret terms generated by time-varying Q -value functions in general-sum games [Chen et al., 2026a].
- (iv) **Cumulative Regret Condition:** Under quadratic disagreement costs the state value functions $V_\pi^i(\sigma)$ are continuous and bounded, and the cumulative regret $\sum_i R_i^T$ is nonnegative [Chen et al., 2026a].

Condition (iv) excludes the alternative in the proof of Theorem 2 of Chen et al. [2026b] in which $\sum_i R_i^T \leq -\epsilon^2 T / (8\eta)$, so under (i)–(iv) the dynamics reach an approximate Nash equilibrium.

4.4 Formal Equilibrium Statement

Theorem 1 (Equilibrium Guarantee in Hybrid Opinion Games). *Consider the hybrid Markov coevolutionary opinion game played by Type-C and Type-L agents running OGA dynamics with learning rate $\eta \leq O(1/\sqrt{T})$ under assumptions (i)–(iv). For any $\epsilon > 0$, after*

$$T \geq \frac{4}{\epsilon^2 |\mathcal{S}|} \sum_{\sigma \in \mathcal{S}} \sum_{i \in \mathcal{N}} (\Omega_i^\sigma)^2 = \frac{8N}{\epsilon^2}$$

iterations (the threshold required in the proof of Theorem 2 of Chen et al. [2026b]), there exists an iteration $k \in [T]$ such that the joint policy profile $\pi_k^{\sigma_k} = (\pi_{1,k}^{\sigma_k}, \dots, \pi_{N,k}^{\sigma_k})$ is an approximate Nash equilibrium for the state σ_k : its Nash gap is at most a multiple of ϵ [Chen et al., 2026a,b],

$$\text{Nash-gap}(\pi_k^{\sigma_k}) \leq \epsilon \left(C^* \max_i \left\| Q_{i,k}^{\sigma_k}(\otimes_{j \neq i} \pi_{j,k}^{\sigma_k}) \right\|_\infty + \frac{2 \max_i \{\ell_i \Omega_i^{\sigma_k}\}}{\eta} \right).$$

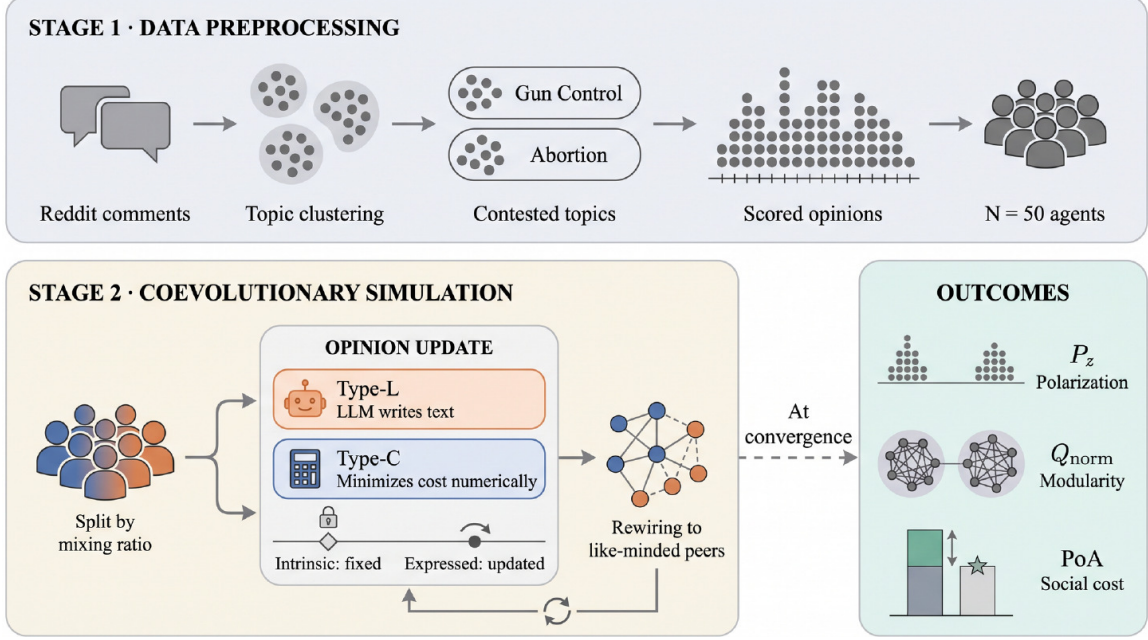


Figure 1: Two-stage pipeline: Stage 1 extracts stance distributions from Reddit; Stage 2 runs H-COG with heterogeneous agents on a coevolving network.

The result extends to multiple players [Chen et al., 2026b, Remark 1].

Theorem 1 guarantees an approximate equilibrium at some iteration; the run classes of Section 5.4.3 describe what the coevolving graph does from then on. A `fixed_point` run stays on a stationary graph, on which Lemma 3(3)–(4) drives the Type-C opinions geometrically to the stage Nash equilibrium, up to q times the residual variation of the Type-L opinions. In `limit_cycle` and `plateau` runs rewiring continues without pointwise convergence, and the statistics of the dynamics are invariant, not the state itself.

4.5 Interpretation for Hybrid Networks

Theorem 1 has a direct consequence for hybrid opinion games: the coevolutionary feedback loop between LLM prompt generation and continual K-NN rewiring reaches an approximate Nash equilibrium at some iteration, at which no agent can lower its cost by more than the Nash gap through a unilateral deviation.

5 Methodology

5.1 Simulation Framework Overview

Figure 1 summarizes the two-stage pipeline. **Stage 1** converts Reddit comments into stance distributions and samples $N = 50$ agents from them. **Stage 2** assigns the agents to Type-C and Type-L according to α and runs H-COG until the stopping rule of Section 5.4.3 is met.

5.2 Stage 1: Reddit Data Pipeline

Corpus Processing and Topic Clustering

We source top-level `r/politics` comments from the Pushshift dumps and keep only the highest-scoring comment per author and topic. Comments are encoded with the Sentence-BERT model `all-MiniLM-`

L6-v2 [Reimers and Gurevych, 2019], reduced from 384 to 50 dimensions by UMAP [McInnes et al., 2018], and clustered with HDBSCAN [Campello et al., 2013, McInnes et al., 2017] (`min_cluster_size= 50`, `min_samples= 5`) into 1,184 topic clusters.

Continuous Stance Measurement

Architecture. We replace the classification head of `RobertaForSequenceClassification` [Liu et al., 2019] with a regression head,

$$\hat{y}_i = \tanh(\mathbf{W}\mathbf{h}_{[\text{CLS}]}) \in [-1, +1], \quad (7)$$

where $\mathbf{h}_{[\text{CLS}]} \in \mathbb{R}^{768}$ is the pooled token representation and $\mathbf{W} \in \mathbb{R}^{1 \times 768}$. Negative and positive endpoints represent strong opposition and support, respectively.

Training. The regressor is trained on the IBM Argument Quality 30K corpus [Gretz et al., 2020], which contains crowd-sourced annotations for contested topics including gun control and abortion. Graded supervision is recovered through Best–Worst Scaling [Kiritchenko and Mohammad, 2017]. On 120 held-out texts, the fitted scorer is monotone in group means and correlates with the graded target at $r = 0.82$. This regression head is the semantic mapping e of Section 3.1. The same instrument scores the Reddit corpus, initializes intrinsic opinions, and measures Type-L expressed opinions, placing all three on one common scale.

Stratified Agent Sampling

For each target topic, we select $N = 50$ agents as follows:

1. Score all comments to obtain the empirical stance distribution $\hat{P}(\text{stance})$.
2. Partition $[-1, +1]$ into five strata using thresholds $\{-0.75, -0.25, +0.25, +0.75\}$.
3. For each stratum b , compute its empirical proportion p_b and sample $n_b = \lfloor 50p_b \rfloor$ comments, adjusting rounding so that $\sum_b n_b = 50$.

Agent i keeps the continuous score of its source comment as s_i .

5.3 Stage 2: H-COG Simulation

5.3.1 Agent Initialization

Each agent receives an intrinsic opinion s_i and a stubbornness $\rho_i \sim \text{Uniform}(0.3, 0.7)$ drawn independently of stance. A Type-C agent starts at $z_i^{(0)} = s_i$. A Type-L agent also carries a fixed persona written by Phi-4 [Abdin et al., 2024], giving it a name, an age, a gender, an education level and five personality traits, following the persona conditioning of Park et al. [2023]. The same model writes its opening text from that persona and the stance keyword of its stratum, and that text is scored as $z_i^{(0)} = e(o_i^{(0)})$, as at every later step.

5.3.2 Network Initialization

We manipulate the initial topology using three graph families: Barabási–Albert with preferential attachment $m = 2$ [Barabási and Albert, 1999], Erdős–Rényi [Erdős and Rényi, 1959], and Watts–Strogatz [Watts and Strogatz, 1998]. The three families are matched on mean degree while preserving differences in clustering (Section 6.2). Ten independently seeded graphs from each family are crossed with nine α levels and two topics, giving $9 \times 3 \times 10 \times 2 = 540$ runs.

K-NN rewiring (Eq. (2)) rebuilds all edges after every update, so the initial topology affects the first update directly and later dynamics only through that trajectory. Across seeds, we vary the topology and Type-C/Type-L assignment while holding the sampled agents, intrinsic opinions, personas, and stubbornness values fixed within each topic, so confidence intervals reflect topology and assignment variability, not resampling of users.

5.3.3 Opinion Updates

Type-C agents use the best response of Eq. (1). For Type-L agents, whose record is (reasoning, opinion, memory) (Section 3.2), placing reasoning first conditions the opinion on the preceding deliberation [Wei et al., 2022], and a fixed field order keeps the response format stable across steps [Turpin et al., 2023]. The opinion field is the agent’s only position output and is scored by e . We do not request a numerical self-rating because language models regress toward the center of an offered scale [Maier et al., 2025] and anchor on round values [Sato et al., 2026]. Rewriting memory at each step bounds prompt growth, and current neighbor opinions are supplied verbatim.

5.4 Structural Convergence and Stopping Rule

5.4.1 Convergence Criterion

We define convergence operationally as entry into an attractor or a stable structural regime: the rate of graph change ceases to decrease, although individual edges need not stop changing. K-NN reconstructs each out-neighborhood from a continuous ranking, so a small perturbation near the cutoff can replace an edge. Adaptive networks may consequently settle into dynamic regimes rather than frozen configurations [Gross and Blasius, 2008]. Our criterion detects either an exactly recurring edge set or a plateau in the rate of structural change.

5.4.2 Structural Convergence Metrics

Let $E(t)$ be the directed edge set and $S_i^{(t)}$ agent i ’s out-neighborhood.

Temporal correlation coefficient. Following Tang et al. [2010] and its directed extension [Büttner et al., 2016],

$$C_i^{\text{out}}(t) = \frac{|S_i^{(t-1)} \cap S_i^{(t)}|}{\sqrt{d_i^{\text{out}}(t-1)d_i^{\text{out}}(t)}}, \quad C^{\text{out}}(t) = \frac{1}{N} \sum_i C_i^{\text{out}}(t).$$

Because every out-degree equals K from $t = 1$ on, for $t \geq 2$ this is the mean fraction of neighbors retained between consecutive steps.

Secondary structural measures. We also log the normalized Hamming distance $dS(t) = |E(t) \triangle E(t-1)| / (|E(t)| + |E(t-1)|)$, the ℓ_2 distance $d_L(t)$ between the ascending eigenvalue vectors of consecutive normalized Laplacians of the undirected projection [Chung, 1997], and the DeltaCon similarity between consecutive graphs [Koutra et al., 2013].

5.4.3 Stopping Rule and Attractor Classification

A run stops at

$$t_{\text{conv}} = \min\{t : \text{exact recurrence or } C^{\text{out}} \text{ plateau}\}.$$

Table 1: Stopping-rule parameters.

Parameter	Value	Role
W	10	averaging-window length
ε_C	0.01	maximum relative change between windows
patience	5	consecutive steps satisfying the condition
post_window	10	steps recorded after detection
$P_{\max}$	20	longest tested recurrence period

Exact recurrence requires $E(t) = E(t - p)$ to hold persistently for some $p \in [1, P_{\max}]$. The plateau condition requires, for `patience` consecutive steps (parameters in Table 1),

$$\frac{|\text{mean}(C^{\text{out}}[t - W : t]) - \text{mean}(C^{\text{out}}[t - 2W : t - W])|}{\text{mean}(C^{\text{out}}[t - 2W : t - W])} < \varepsilon_C.$$

Runs are classified as `fixed_point` when the edge set becomes stationary, `limit_cycle:p` when it recurs with $2 \leq p \leq P_{\max}$, `plateau` when the retention rate stabilizes without a detected period, and `none` when no condition is met within the simulation horizon. Section 4.4 relates these classes to Theorem 1.

5.5 Evaluation Metrics

Price of Anarchy. The stepwise PoA follows Eq. (5), with the optimum of Lemma 5 and the decomposition of Remark 1.

Polarization (P_z). Following Chitra and Musco [2020], polarization is the variance of expressed opinions:

$$P_z^{(t)} = \frac{1}{N} \sum_{i=1}^N \left(z_i^{(t)} - \bar{z}^{(t)} \right)^2, \quad \bar{z}^{(t)} = \frac{1}{N} \sum_i z_i^{(t)}. \quad (8)$$

Modularity (Q and Q_{norm}). Louvain community detection [Blondel et al., 2008] is applied to the undirected projection of $G^{(t)}$, yielding modularity $Q_{\text{obs}}(t)$ [Newman and Girvan, 2004]. Because random graphs can exhibit nonzero modularity from degree fluctuations alone [Guimerà et al., 2004], we compare the observed graph with a degree-preserving null model:

$$Q_{\text{norm}}(t) = Q_{\text{obs}}(t) - \bar{Q}_{\text{rand}}(t), \quad z_Q(t) = \frac{Q_{\text{obs}}(t) - \bar{Q}_{\text{rand}}(t)}{\sigma_{\text{rand}}(t)}. \quad (9)$$

The null mean $\bar{Q}_{\text{rand}}$ and standard deviation σ_{rand} are taken over 20 degree-preserving double-edge-swap randomizations of the undirected projection and are recomputed at every step. Q_{norm} is the modularity attributable to opinion-based sorting beyond the degree sequence.

5.6 Cross-Run Comparability

Because run lengths differ, outcome metrics are compared at $t_{\text{conv}} + \text{post_window}$, each run’s own evaluation point.

6 Experiments and Numerical Results

6.1 Dataset Overview

We source data from Pushshift’s April 2019 comment dump [Baumgartner et al., 2020] (RC_2019-04, 138.5 million comments), of which 1.76 million belong to the `r/politics` subreddit. The first 500,000 valid `r/politics` comments in stream order, excluding deleted or removed comments, bodies of at most 20 characters and comments by deleted authors, enter the clustering stage. Semantic clustering (Section 5.2) identifies 1,184 topic clusters across politics; we focus on two politically contested topics relevant to our simulation: *gun control* ($n = 3,581$ comments) and *abortion* ($n = 1,618$ comments).

Figure 2 shows the stance distributions for both topics under the regressor of Section 5.2. Gun control exhibits a near-neutral mean with high variance, consistent with a bimodal and divided public; abortion displays a more concentrated opposing consensus. The resulting initial-stance split among the $N = 50$ sampled agents is 54%/46% (oppose/support) for gun control and 68%/32% for abortion, an asymmetry that Section 6.5.2 shows to be consequential.

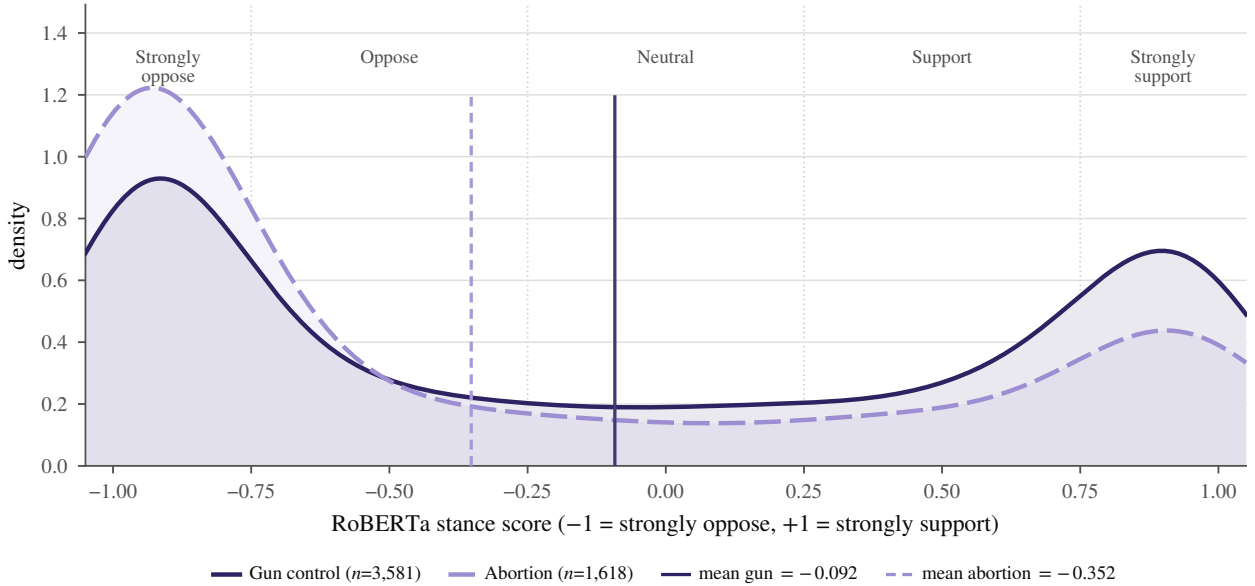


Figure 2: Stance score distributions for gun control ($n = 3,581$) and abortion ($n = 1,618$) comments in `r/politics` (April 2019).

The two topics are not independent replicates. The 50 sampled agents share identical personas and stubbornness coefficients ρ_i across the two topics; only the initial stance differs, so the two topics are compared with paired tests (Section 6.3).

6.2 Simulation Configuration

All simulations use $N = 50$ agents and $K = 5$ neighbors per agent. The mixing parameter is varied across $\alpha \in \{0, 0.125, 0.25, 0.375, 0.5, 0.625, 0.75, 0.875, 1\}$, covering the full spectrum from a purely LLM-driven population ($\alpha = 0$) to a purely cost-minimizing population ($\alpha = 1$).

Topology parameters. The three initial-topology families of Section 5.3.2 are matched on mean degree (3.84 for BA, 3.84 for ER, 4.00 for WS), so that any difference between them is attributable to structure rather than to edge count. Their clustering coefficients remain clearly distinct at 0.179, 0.079, and 0.390 respectively, preserving the contrast of interest.

Grid completeness. The design yields $n = 60$ runs at every α level and $n = 30$ at every (α, topic) cell, with no missing cells: all 540/540 runs complete and enter the analysis. The run is the unit of analysis, so statistical power comes from the number of conditions and seeds.

6.3 Statistical Analysis

Variances are markedly unequal across α , with the standard deviation of PoA falling by a factor of about 60 between $\alpha = 0$ and $\alpha = 1$, so comparisons between α levels use Welch’s t -test [Welch, 1947] with Benjamini–Hochberg FDR control [Benjamini and Hochberg, 1995] over the family of adjacent- α comparisons, and we report q -values for this family; the two per-topic contrasts in Section 6.5.2 are unadjusted p -values.

Where the claim of interest is equivalence rather than difference, specifically for the comparison among initial topologies, we test for it directly with two one-sided tests (TOST) [Schuirmann, 1987] against a margin fixed in advance at $\delta = 0.2$ PoA units.

Comparisons between topics use paired tests, for the reason given in Section 6.1. All confidence intervals are reported at the 95% level.

6.4 Implementation Details

Type-L agents are powered by Phi-4 (14.7B parameters) [Abdin et al., 2024] under 4-bit AWQ quantization [Lin et al., 2024], served via vLLM [Kwon et al., 2023]; Type-C agents are closed-form FJ updates. The full simulation codebase is available at <https://github.com/Evan-Jiang/Echo-Chamber-Simulation>.

6.5 Numerical Results

Three findings organize this section.

1. (RQ1) The efficiency gap is about fivefold. PoA falls from 5.558 in the purely language population to 1.139 in the purely numerical one (Table 3).
2. (RQ1) Language agents are displaced from their own positions as well as disagreeing with their neighbors. From the purely numerical to the purely language population the conformity term grows by a factor of 13.2 and the disagreement term by 3.3, and the conformity share of social cost rises from 16.1% to 43.6% (Table 4).
3. (RQ2) Polarization is lowest in the low- α region, where efficiency is worst. P_z and Q_{norm} rise with α , while PoA holds a statistical plateau up to $\alpha = 0.375$ and then falls (Tables 2 and 3).

Table 2 is the primary result, and Figure 3 shows the same quantities over time.

Table 2: Metrics at each run’s converged state ($t_{\text{conv}} + 10$), by topic and mixing ratio. $n = 30$ per row (3 topologies $\times$ 10 seeds), mean $\pm$ 95% CI. $\downarrow$ denotes lower-is-better; Q_{norm} and z_Q as in Section 5.5.

Topic	α	P_z	Q_{norm}	z_Q	PoA $\downarrow$
Gun Control	0	0.261 ± 0.013	0.241 ± 0.009	22.8 ± 2.1	5.681 ± 0.340
	0.125	0.312 ± 0.016	0.250 ± 0.009	23.0 ± 1.5	5.480 ± 0.312
	0.25	0.364 ± 0.013	0.259 ± 0.009	23.8 ± 1.6	5.642 ± 0.396
	0.375	0.409 ± 0.014	0.263 ± 0.009	22.5 ± 1.3	5.477 ± 0.325
	0.5	0.451 ± 0.012	0.277 ± 0.007	23.3 ± 1.3	5.374 ± 0.395
	0.625	0.494 ± 0.011	0.286 ± 0.010	23.1 ± 1.3	4.424 ± 0.344
	0.75	0.538 ± 0.008	0.314 ± 0.008	24.9 ± 1.9	3.821 ± 0.274
	0.875	0.582 ± 0.006	0.350 ± 0.009	24.5 ± 1.3	3.220 ± 0.222
	1	0.619 ± 0.000	0.427 ± 0.002	27.4 ± 1.9	1.149 ± 0.008
Abortion	0	0.347 ± 0.017	0.194 ± 0.006	18.4 ± 1.4	5.436 ± 0.535
	0.125	0.364 ± 0.011	0.199 ± 0.007	20.7 ± 1.6	6.322 ± 0.389
	0.25	0.426 ± 0.015	0.191 ± 0.007	18.8 ± 1.6	6.077 ± 0.367
	0.375	0.463 ± 0.016	0.241 ± 0.013	23.2 ± 1.9	5.815 ± 0.385
	0.5	0.480 ± 0.015	0.283 ± 0.010	24.4 ± 2.3	5.117 ± 0.298
	0.625	0.510 ± 0.013	0.297 ± 0.008	24.0 ± 1.9	4.670 ± 0.296
	0.75	0.534 ± 0.010	0.317 ± 0.008	24.4 ± 1.5	3.853 ± 0.288
	0.875	0.554 ± 0.008	0.343 ± 0.006	25.0 ± 1.4	2.758 ± 0.306
	1	0.575 ± 0.000	0.419 ± 0.002	25.8 ± 1.4	1.129 ± 0.001

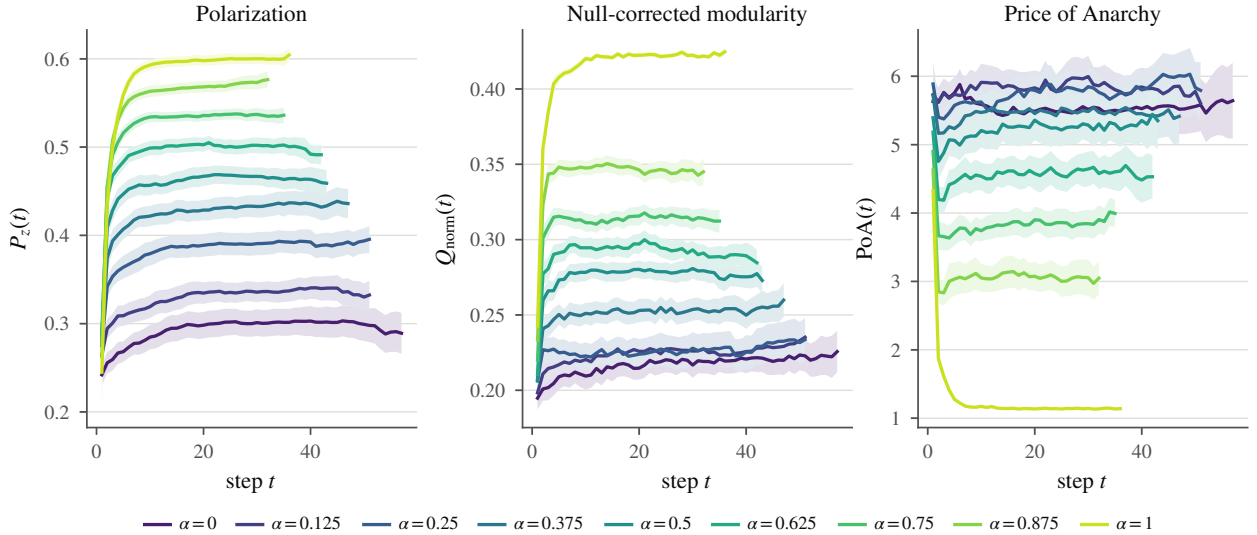
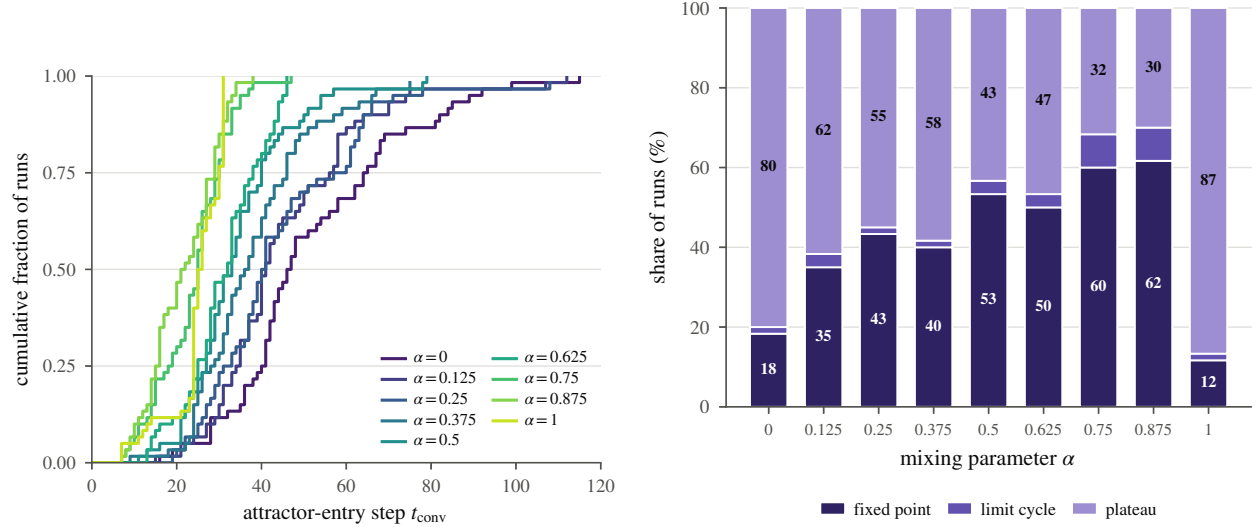


Figure 3: Per-step trajectories of P_z , Q_{norm} and PoA by mixing ratio (mean with 95% CI band). All three are flat well before the graph enters its attractor, so their values in Table 2 do not depend on the precise step at which they are read.

6.5.1 Convergence as a Precondition

Theorem 1 guarantees an approximate Nash equilibrium at some iteration; the form of the attractor the coevolving graph enters is an empirical question for both agent types, because K-NN rewiring is a discontinuous map applied on top of the Friedkin–Johnsen update and Type-L positions have no closed form. That both modes of updating converge, as shown below, licenses comparing converged states across α .



(a) Attractor-entry step t_{conv} , as empirical cumulative distribution functions.

(b) Attractor composition. Exactly stationary edge sets are rare at both pure compositions and most common in the mixed regime.

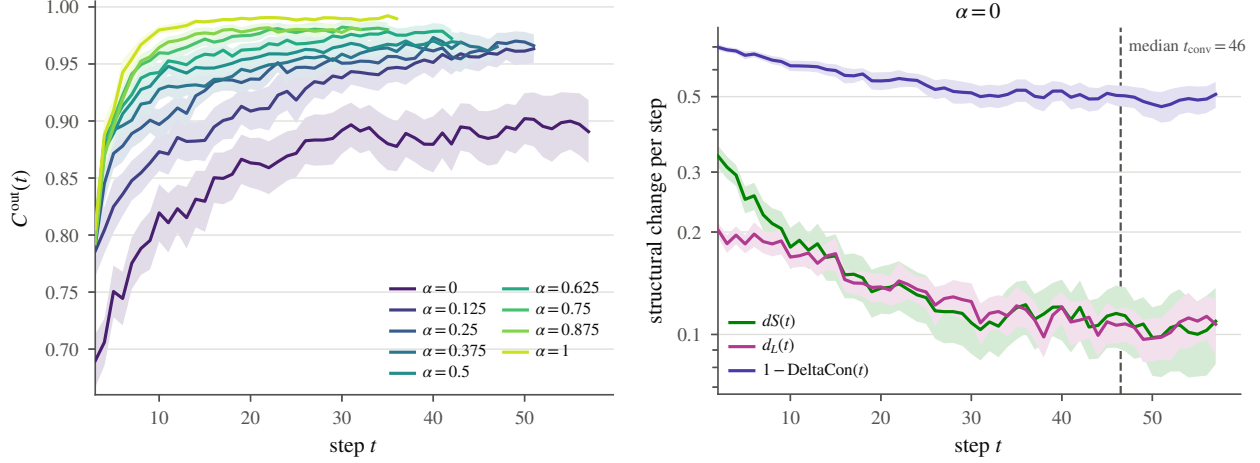
Figure 4: Settling behavior by mixing ratio ($n = 60$ per curve and per bar, pooled over three topologies and two topics).

Both pure populations converge, and mostly as a plateau. Figure 4 summarizes settling behavior across the grid. Neither pure population settles predominantly to an exact fixed point: both are classified plateau in the large majority of runs, and the purely Type-L population reaches an exactly stationary edge set slightly more often than the purely Type-C one, whose opinion dynamics converge analytically. Under K-NN rewiring, convergence takes the form of a plateau in the rate of structural change rather than a frozen graph, and Type-L agents attain that state at least as reliably as the closed-form Type-C dynamics.

Settling time is a nonmonotone function of composition. Figure 4a shows t_{conv} falling monotonically in α from 52.2 steps at $\alpha = 0$ to 21.7 at $\alpha = 0.875$, and then rising to 25.1 at $\alpha = 1.0$. Run lengths are right-skewed; at $\alpha = 0$ the median is 46 steps and the mean is 52.2.

Attractor composition shows the same reversal. Figure 4b resolves the composition by class. Exactly stationary edge sets account for 18% of runs at $\alpha = 0$ and 12% at $\alpha = 1$, whereas mixtures reach 62% at $\alpha = 0.875$: a population containing six LLM agents attains an exact fixed point five times as often as one containing none.

Rewiring churn separates the purely Type-L population. Figure 5 traces the per-step structural change, and Figure 5a the stopping statistic itself. The plateau height of C^{out} is 0.913 (SD 0.069) at $\alpha = 0$ against



(a) Temporal correlation coefficient $C^{\text{out}}(t)$, the fraction of out-neighbors retained between consecutive steps; markers give the median t_{conv} . (b) Three measures at $\alpha = 0$ on a logarithmic vertical axis, all flattening ahead of the median t_{conv} (dashed).

Figure 5: Structural change per step, defined from $t = 2$ (mean with 95% CI band). Panel (a) is the statistic the stopping rule reads; panel (b) shows three measures that take no part in it.

0.990 (SD 0.004) at $\alpha = 1$, and elevated churn is confined to $\alpha = 0$. Uniformly random rewiring gives $C^{\text{out}} \approx K/(N-1) \approx 0.10$, so both plateau heights reflect persistent neighborhoods. The Type-C population comes to rest close to a frozen structure, whereas the Type-L population reaches a state in which the rate of turnover has stabilized while a small fraction of ties continues to change.

Measures outside the stopping rule corroborate it. Convergence is declared on C^{out} alone. Figure 5b shows three measures that take no part in that decision (the normalized Hamming distance between consecutive edge sets, the spectral distance between their Laplacians, and $1 - \text{DeltaCon}$) at $\alpha = 0$, the slowest-settling composition: each flattens at around $t = 25$ to 30 , ahead of the median t_{conv} of 46, so the measure the rule reads is the last of four to register the settled regime. The criterion is satisfied within the simulation horizon in all runs (540/540), and no run is excluded.

6.5.2 RQ1: The Efficiency Gap and Its Decomposition

The purely Type-C baseline. At $\alpha = 1$ the measured PoA is 1.139 ± 0.005 , the inefficiency of the opinion formation game itself, against which every other composition is compared. On the $\alpha = 1$ runs with a stationary graph the computed opinions coincide with the fixed point of Lemma 3. The value is consistent with the $9/8$ bound (Remark 2), and it lies well below the worst case of the K -NN game for $\rho < 1$, which is at least $1/\rho^2 \geq 2$ over the range of ρ_i used here [Bhawalkar et al., 2013].

The fall in PoA is not a smooth interpolation (Table 3): no comparison up to and including the one reaching $\alpha = 0.375$ is significant ($q > 0.12$), and every comparison from $\alpha = 0.5$ onward is ($q < 0.04$). The low- α region is a statistical plateau, and the decline begins only once cost-minimizing agents approach parity. PoA is minimized at $\alpha = 1$ and no interior mixture approaches it; since polarization is lowest in the low- α plateau, where efficiency is worst, no composition optimizes both.

Table 4 and Figure 6 separate the two components. Between $\alpha = 1$ and $\alpha = 0$ the disagreement term grows by a factor of 3.3 ($0.956 \rightarrow 3.136$) while the conformity term grows by a factor of 13.2 ($0.183 \rightarrow 2.422$), and the conformity share falls as α rises. The two components contribute about equally to the gap of 4.42 between the pure populations, 2.18 from disagreement and 2.24 from conformity.

Table 3: Price of Anarchy by mixing ratio, pooled ($n = 60$ per row). The q column gives the Welch test against the preceding level under Benjamini–Hochberg control.

α	PoA	q vs. prev.
0.0	5.558 ± 0.309	—
0.125	5.901 ± 0.265	0.127
0.25	5.860 ± 0.268	0.826
0.375	5.646 ± 0.249	0.280
0.5	5.246 ± 0.242	0.036
0.625	4.547 ± 0.222	8.5×10^{-5}
0.75	3.837 ± 0.193	1.1×10^{-5}
0.875	2.989 ± 0.193	3.0×10^{-8}
1.0	1.139 ± 0.005	3.9×10^{-26}

Table 4: The same social cost separated into its two components, pooled ($n = 60$ per row). Their sum is the Price of Anarchy at left.

α	Disagreement	Conformity	Conf. share
0.0	3.136	2.422	43.6%
0.125	3.690	2.211	37.5%
0.25	3.882	1.978	33.7%
0.375	3.822	1.824	32.3%
0.5	3.610	1.636	31.2%
0.625	3.127	1.420	31.2%
0.75	2.678	1.159	30.2%
0.875	2.175	0.814	27.2%
1.0	0.956	0.183	16.1%

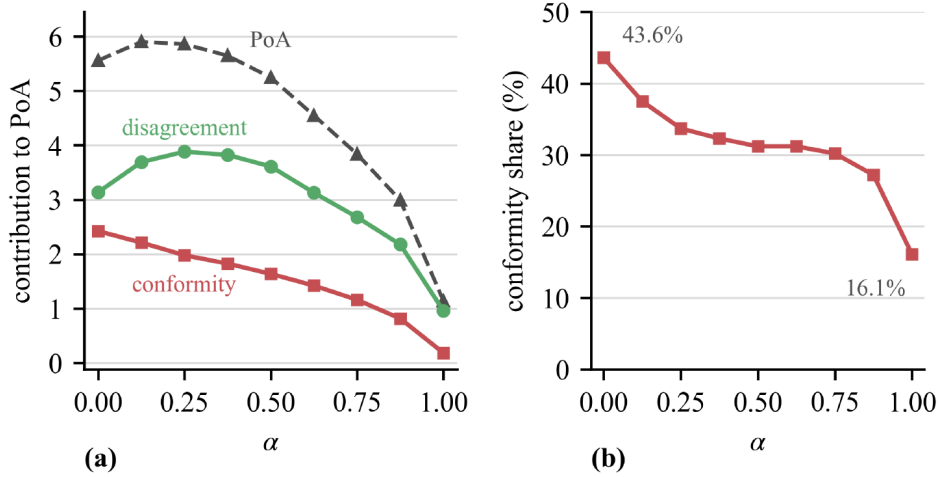


Figure 6: Social-cost decomposition against the mixing ratio, on the values of Table 4. (a) The disagreement term rises to a peak at $\alpha = 0.25$ before falling, while the conformity term declines throughout; each component is divided by the planner optimum of the same step, so the two sum to the Price of Anarchy, shown dashed. (b) The conformity share falls from 43.6% to 16.1%.

Within the plateau the two topics differ: from $\alpha = 0$ to $\alpha = 0.125$ PoA rises from 5.436 to 6.322 for abortion ($p = 0.008$), while gun control is flat ($p = 0.38$) and attains its maximum at $\alpha = 0$. Because the two topics share agent personas and stubbornness coefficients (Section 6.1), the difference is attributable to the stance distribution rather than to the agents themselves.

6.5.3 RQ2: Polarization and Echo Chambers

Figure 7 shows both measures by topic. Pooled, P_z rises from 0.304 ± 0.015 to 0.597 ± 0.006 and Q_{norm} from 0.218 ± 0.008 to 0.423 ± 0.002 across the range of α . Echo chamber structure forms under every composition, but its strength depends on agent type: Type-L populations produce measurably weaker polarization and weaker community structure.

All claims use Q_{norm} rather than bare Q . At $\alpha = 1$ the bare value for gun control is 0.763 while the

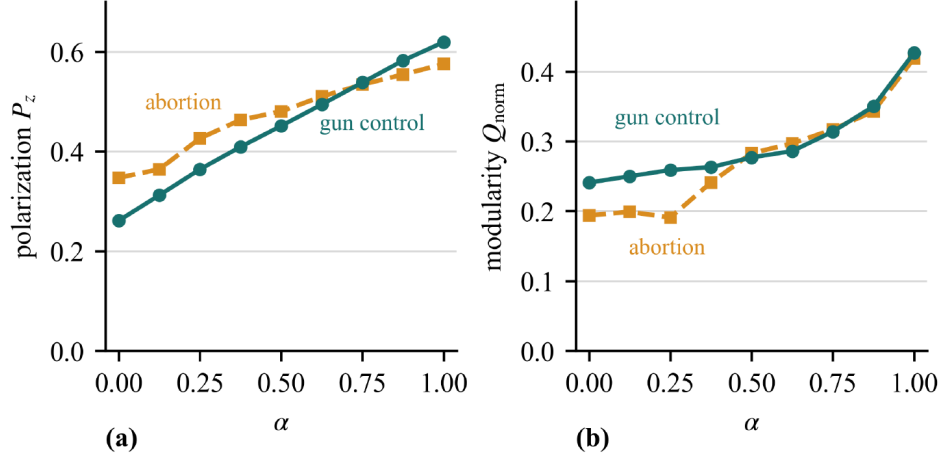


Figure 7: Polarization and echo-chamber structure against the mixing ratio, by topic. (a) The variance of expressed opinions rises with α for both topics. (b) The null-corrected modularity rises over the same range. Both vertical axes start at zero.

degree-preserving null attains 0.336 on the same degree sequence, so roughly 44% of raw modularity follows from the constant out-degree the K-NN rule imposes. What remains is far from null at every composition: z_Q places the observed modularity between 18 and 27 standard deviations above the null distribution across the grid.

6.5.4 RQ3: The Role of the Rewiring Mechanism

Table 5: PoA by initial topology ($n = 20$ per cell), with the largest absolute pairwise gap between families at each level and its TOST equivalence p -value against $\delta = 0.2$.

α	BA	ER	WS	largest gap	TOST p
0.0	5.348 ± 0.503	5.766 ± 0.737	5.562 ± 0.395	0.418	0.694
0.25	5.944 ± 0.356	5.961 ± 0.531	5.673 ± 0.557	0.288	0.594
0.5	5.078 ± 0.373	5.422 ± 0.417	5.237 ± 0.521	0.344	0.704
0.75	3.912 ± 0.420	4.051 ± 0.332	3.549 ± 0.247	0.501	0.932
1.0	1.138 ± 0.008	1.134 ± 0.007	1.144 ± 0.010	0.010	< 0.0001

Table 5 reports PoA by family. Across all nine α levels and both PoA and P_z , the confidence intervals of the three topology families overlap and no comparison attains significance. The two one-sided tests (TOST) against the margin $\delta = 0.2$ (Section 6.3) establish equivalence at $\alpha = 1$ ($p < 0.0001$). At the five levels of Table 5 the initial topology shifts PoA by at most 0.50, against a composition effect of 4.42.

The three families are matched on mean degree while their clustering coefficients differ by a factor of five (Section 6.2), and K-NN rewiring overwrites the initial edge set at the first step and thereafter rebuilds every out-neighborhood from the current expressed opinions. That no difference is detectable under this contrast, at $\alpha = 0$ as much as at $\alpha = 1$, indicates that echo chamber formation here requires no pre-existing hub structure and is not a property of the FJ closed form: echo chambers follow from the connection rule rather than from the way the population was wired at the outset, and the composition of the population determines how deep the clustering is, not whether it occurs.

7 Discussion, Conclusions and Future Work

7.1 Interpretation of Key Findings

Language reasoning operates as a conformity mechanism. Displacement from one’s own prior position, and not disagreement alone, separates the two populations (Table 4). The mechanism is where the anchor sits. In the Type-C cost function an agent’s prior position is a term in the objective, a restoring force that penalizes departure however uniform its neighborhood becomes. A Type-L agent receives the same prior as a sentence in a prompt, and within the homogeneous neighborhoods that rewiring produces, reasoning over what its neighbors say moves it away from that prior rather than defending it. Under uniform peer pressure, language reasoning operates less like deliberation and more like repeated persuasion.

Weaker polarization is a symptom, not an improvement. The structural result follows from the same property. A population that is readily talked out of its own positions does not hold camp boundaries, and camps that do not hold do not consolidate. Weaker polarization among Type-L agents therefore goes together with what makes them inefficient. Read alone, polarization points the wrong way: the compositions that look healthiest by that measure lie in the region furthest from what a planner could achieve.

Conformity accumulates one step at a time. A Type-L agent responds to what its neighbors say rather than only to where they stand. When an agent’s neighbors all say roughly the same thing, it reweighs its position every round, and each reweighing moves it a little further from where it began. Each step is a sensible reply to the argument in front of it. Over the 52 steps a purely Type-L run takes to settle on average, the population is talked off its own starting positions one small concession at a time.

The mechanism carries a testable prediction. If the displacement comes from the prior being reweighed each round, then raising the weight of that prior in the per-round input will lower conformity cost while leaving disagreement cost largely unchanged.

Two features do not follow this pattern. Mixed populations settle into exactly stationary structures more often than either pure one, and a small minority of Type-C agents raises social cost on abortion but not on gun control. Both are findings in need of a mechanism.

7.2 Conclusions

Work on language agents has asked whether those agents behave like people, and work on opinion dynamics has asked how opinions travel under a fixed update rule, but neither supplies a yardstick against which a population that reasons in sentences and one that minimizes a cost can be compared on welfare.

Game-theoretic models of opinion formation buy their analytical results with a convex cost function and an update rule chosen to be tractable; agent simulations buy semantic realism by giving that up. The question underneath is whether equilibrium analysis survives the substitution. A Price of Anarchy is a statement about equilibria, and it says nothing about a population that does not reach one.

What we find is that it survives. Theorem 1 guarantees an approximate equilibrium at some iteration, and the coevolving graph enters an attractor under Type-L updating as reliably as under an update rule with a known fixed point. The efficiency gap we report is therefore a difference between the attractors of two populations rather than between two snapshots, and an agent that reasons in sentences can be placed in the same analytical frame as one that minimizes a quadratic cost.

Agreement can be reached in two ways: a population can talk its neighbors around, or its members can move toward their neighbors. A population that reweighs its position every time it reads what its neighbors

have written reaches agreement the second way, and its members pay the cost privately, in the positions they give up. Polarization measures only how far apart opinions end up, not what convergence cost those who converged, so a population can look healthier as it becomes worse off. To change that structure, the lever is the rule of association rather than the initial network. An audit of a platform’s health also needs a second quantity alongside the spread of opinions: how far its users have moved from what they arrived believing.

7.3 Limitations and Future Work

1. All Type-L agents here are instances of a single model, Phi-4, under one prompt framework, with a distinct persona for each agent. Running the same grid across several models and prompt frameworks would separate a property of language-based opinion updating from a property of one instance.
2. Efficiency is stated relative to the Friedkin–Johnsen benchmark. Adding observed human opinion dynamics on the same corpus as a third arm would state it relative to human behavior instead.

References

- Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J. Hewett, Mojan Javaheripi, Piero Kauffmann, et al. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*, 2024.
- Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *Science*, 286(5439): 509–512, 1999.
- Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. The Pushshift Reddit dataset. In *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM)*, volume 14, pages 830–839, 2020.
- Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1): 289–300, 1995.
- Kshipra Bhawalkar, Sreenivas Gollapudi, and Kamesh Munagala. Coevolutionary opinion formation games. In *Proceedings of the Forty-Fifth Annual ACM Symposium on Theory of Computing*, pages 41–50, 2013.
- David Bindel, Jon Kleinberg, and Sigal Oren. How bad is forming your own opinion? *Games and Economic Behavior*, 92:248–265, 2015.
- Vincent D. Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10): P10008, 2008.
- Kathrin Büttner, Jennifer Salau, and Joachim Krieter. Temporal correlation coefficient for directed networks. *SpringerPlus*, 5:1198, 2016.
- Ricardo J. G. B. Campello, Davoud Moulavi, and Jörg Sander. Density-based clustering based on hierarchical density estimates. In *Advances in Knowledge Discovery and Data Mining (PAKDD)*, pages 160–172, 2013.
- Chung-Chi Chen. AgoraSim: A hybrid agent-based modeling framework, 2026. arXiv:2607.05999.

- Po-An Chen, Yi-Le Chen, and Chi-Jen Lu. Bounds on the price of anarchy for a more general class of directed graphs in opinion formation games. *Operations Research Letters*, 44(6):808–811, 2016.
- Po-An Chen, Chi-Jen Lu, Chuang-Chieh Lin, An-Tzu Teng, and Ke-Wei Fu. Multiagent learning for competitive opinion optimization. *Theoretical Computer Science*, 1017:114787, 2024.
- Po-An Chen, Chi-Jen Lu, Chuang-Chieh Lin, Jim Shi, and Chih-Chieh Hung. Dynamics and convergences for Markov coevolutionary opinion formation games in dynamic social networks (extended abstract). In *Proceedings of the 13th International Conference on Networks, Games, Control and Optimization*, 2026a.
- Po-An Chen, Chi-Jen Lu, Chuang-Chieh Lin, Jim Shi, and Chih-Chieh Hung. Dynamics and convergences for Markov coevolutionary opinion formation games in dynamic social networks, 2026b. arXiv:2607.05580.
- Uthsav Chitra and Christopher Musco. Analyzing the impact of filter bubbles on social network polarization. In *Proceedings of the 13th International Conference on Web Search and Data Mining (WSDM)*, pages 115–123, 2020.
- Yun-Shiuan Chuang, Agam Goyal, Nikunj Harlalka, Siddharth Suresh, Robert Hawkins, Sijia Yang, Dhavan Shah, Junjie Hu, and Timothy T. Rogers. Simulating opinion dynamics with networks of LLM-based agents. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3326–3346, 2024.
- Fan R. K. Chung. *Spectral Graph Theory*, volume 92 of *CBMS Regional Conference Series in Mathematics*. American Mathematical Society, 1997.
- Morris H. DeGroot. Reaching a consensus. *Journal of the American Statistical Association*, 69(345):118–121, 1974.
- Paul Erdős and Alfréd Rényi. On random graphs I. *Publicationes Mathematicae Debrecen*, 6:290–297, 1959.
- Dimitris Fotakis, Dimitris Palyvos-Giannas, and Stratis Skoulakis. Opinion dynamics with local interactions. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI)*, pages 279–285, 2016.
- Noah E. Friedkin and Eugene C. Johnsen. Social influence and opinions. *Journal of Mathematical Sociology*, 15(3-4):193–206, 1990.
- Shai Gretz, Roni Friedman, Edo Cohen-Karlik, Assaf Toledo, Dan Lahav, Ranit Aharonov, and Noam Slonim. A large-scale dataset for argument quality ranking: Construction and analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7805–7813, 2020.
- Thilo Gross and Bernd Blasius. Adaptive coevolutionary networks: A review. *Journal of the Royal Society Interface*, 5(20):259–271, 2008.
- Chenhao Gu, Ling Luo, Zainab Razia Zaidi, and Shanika Karunasekera. Large language model driven agents for simulating echo chamber formation. *arXiv preprint arXiv:2502.18138*, 2025.
- Chenhao Gu, Ling Luo, Zainab Razia Zaidi, and Shanika Karunasekera. Language-grounded co-evolution of opinions and ties: Paired rewiring with large language model agents. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 20, pages 972–988, 2026.

- Roger Guimerà, Marta Sales-Pardo, and Luís A. Nunes Amaral. Modularity from fluctuations in random graphs and complex networks. *Physical Review E*, 70(2):025101(R), 2004.
- Svetlana Kiritchenko and Saif M. Mohammad. Best-worst scaling more reliable than rating scales: A case study on sentiment intensity annotation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 465–470, 2017.
- Danai Koutra, Joshua T. Vogelstein, and Christos Faloutsos. DeltaCon: A principled massive-graph similarity function. In *Proceedings of the 2013 SIAM International Conference on Data Mining (SDM)*, pages 162–170, 2013.
- Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with PagedAttention. In *Proceedings of the 29th Symposium on Operating Systems Principles (SOSP)*, pages 611–626, 2023.
- Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Wei-Ming Chen, Wei-Chen Wang, Guangxuan Xiao, Xingyu Dang, Chuang Gan, and Song Han. AWQ: Activation-aware weight quantization for on-device LLM compression and acceleration. In *Proceedings of Machine Learning and Systems (MLSys)*, 2024.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.
- Benjamin F. Maier, Ulf Aslak, Luca Fiaschi, Nina Rismal, Kemble Fletcher, Christian C. Luhmann, Robbie Dow, Kli Pappas, and Thomas V. Wiecki. LLMs reproduce human purchase intent via semantic similarity elicitation of Likert ratings. *arXiv preprint arXiv:2510.08338*, 2025.
- Leland McInnes, John Healy, and Steve Astels. hdbscan: Hierarchical density based clustering. *Journal of Open Source Software*, 2(11):205, 2017.
- Leland McInnes, John Healy, and James Melville. UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.
- Xinyi Mou, Zhongyu Wei, and Xuanjing Huang. Unveiling the truth and facilitating change: Towards agent-based large-scale social movement simulation. In *Findings of the Association for Computational Linguistics: ACL 2024*, 2024.
- Cameron Musco, Christopher Musco, and Charalampos E. Tsourakakis. Minimizing polarization and disagreement in social networks. In *Proceedings of the 2018 World Wide Web Conference*, pages 369–378, 2018.
- Mark E. J. Newman and Michelle Girvan. Finding and evaluating community structure in networks. *Physical Review E*, 69(2):026113, 2004.
- Joon Sung Park, Joseph O’Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pages 1–22, 2023.
- Jinghua Piao, Zhihong Lu, Chen Gao, Fengli Xu, Qinghua Hu, Fernando P. Santos, Yong Li, and James Evans. Emergence of polarization in networks of large language model agents, 2025. *arXiv:2501.05171*.

- Giorgio Piatti, Zhijing Jin, Max Kleiman-Weiner, Bernhard Schölkopf, Mrinmaya Sachan, and Rada Mihalcea. Cooperate or collapse: Emergence of sustainable cooperation in a society of LLM agents. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.
- Amin Qasmi, Usman Naseem, and Mehwish Nasim. Competing LLM agents in a non-cooperative game of opinion polarisation, 2025. arXiv:2502.11649.
- Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, 2019.
- Ayako Sato, Hwichan Kim, Zhousi Chen, Masato Mita, and Mamoru Komachi. Exploring the effects of alignment on numerical bias in large language models. *arXiv preprint arXiv:2601.16444*, 2026.
- Donald J. Schuirmann. A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics*, 15(6):657–680, 1987.
- John Tang, Salvatore Scellato, Mirco Musolesi, Cecilia Mascolo, and Vito Latora. Small-world behavior in time-varying graphs. *Physical Review E*, 81(5):055101(R), 2010.
- Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. Language models don’t always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pages 74952–74965, 2023.
- Chenxi Wang, Zongfang Liu, Dequan Yang, and Xiuying Chen. Decoding echo chambers: LLM-powered simulations revealing polarization in social networks. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 3913–3923, 2025.
- Duncan J. Watts and Steven H. Strogatz. Collective dynamics of ‘small-world’ networks. *Nature*, 393(6684):440–442, 1998.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pages 24824–24837, 2022.
- Bernard L. Welch. The generalization of ‘Student’s’ problem when several different population variances are involved. *Biometrika*, 34(1–2):28–35, 1947.
- Junchi Yao, Hongjie Zhang, Jie Ou, Dingyi Zuo, Zheng Yang, and Zhicheng Dong. Social opinions prediction utilizes fusing dynamics equation with LLM-based agents. *Scientific Reports*, 15:15472, 2025.