Language-Grounded Co-evolution of Opinions and Ties: Paired Rewiring with Large Language Model Agents

Chenhao Gu[1], **Ling Luo**[1], **Zainab Razia Zaidi**[2], **Shanika Karunasekera**[1]
[1]The University of Melbourne, Australia
[2]Federation University, Australia
{chenhao.gu1, ling.luo, karus}@unimelb.edu.au, z.zaidi@federation.edu.au

Abstract

Modeling how user opinions and social ties co-evolve is central to understanding polarization and community formation in online platforms, yet existing approaches often sacrifice either linguistic realism or network structure. We propose a simulation framework for studying the co-evolution of online networks and opinions, in which large language models (LLMs) handle opinion updating, content generation, and network rewiring. This design integrates linguistic signals into structural change under real-world constraints, enabling systematic comparison across LLMs and empirical validation on three real-world datasets. Results show that LLM-based simulations consistently outperform equation-based baselines across key network metrics and better reproduce observed stance shifts. Analysis of the reasoning behind model decisions reveals systematic differences in rewiring motives and highlights distinct tendencies across models, offering interpretive insight into the network structures they produce. In addition, varying rewiring probabilities and strategies demonstrate how platform-like controls shape polarization and cohesion. Together, these findings establish a language-grounded, empirically validated approach to modeling network–opinion co-evolution with greater accuracy and interpretability.

Introduction

Online social networks are inherently dynamic systems in which both user opinions and the surrounding network structure continuously evolve. Users generate and share content that reflects their views, while at the same time adjusting their social ties by unfollowing, following, or forming new connections. This interplay between *opinion evolution* and *network rewiring* drives many of the emergent phenomena observed in online communities, ranging from the spread of information to the formation of polarized groups.

Rather than occurring in isolation, this co-evolutionary process manifests through observable structural outcomes at the network level. As users update their opinions based on the content they encounter and reassess their social connections accordingly, individual-level changes in language, beliefs, and ties accumulate over time. These dynamics can give rise to diverse community structures, including highly cohesive groups, fragmented neighborhoods, or polarized regions separated by sparse connections. Consequently, the evolving topology of the network provides a concrete lens through which opinion–network co-evolution can be studied and quantified.

Decades of work in opinion dynamics and network science show that homophily, triadic closure, and selective exposure jointly shape how communities form and harden (Kossinets and Watts 2006; McPherson, Smith-Lovin, and Cook 2001; Flaxman, Goel, and Rao 2016). In practice, these processes are mediated by platform mechanisms such as feeds, recommendation systems, and limited "attention budgets"—that constrain what users see and when they choose to maintain or drop ties. As a result, the same piece of content can drive different structural outcomes depending on what alternatives are surfaced and how often users are nudged to reconsider their neighborhoods.

Traditional equation-based models offer clear abstractions and identifiable mechanisms such as homophily, influence strength, and tie-formation rules, but they represent opinions as numbers and tie decisions as parametric functions, leaving out the semantics that actually govern online interaction. Recent LLM-driven simulations move in the opposite direction: they capture linguistic nuance and can generate realistic posts, yet they often leave network evolution unconstrained and are rarely validated against real structural data. What is missing is a *unified* treatment in which language both *explains* and *drives* the coupled evolution of content, opinions, and ties—and is assessed against empirical networks rather than only narrative plausibility.

These considerations motivate four central research questions:

RQ1: Can LLMs be used to generate realistic simulations of network interactions, capturing the co-evolution of opinions and social ties within a single framework, rather than only generating text content?

RQ2: Do different LLMs exhibit systematic differences in their network-level decision-making, leading to distinct structural outcomes?

RQ3: Can we introduce network-level controls by tuning global interaction parameters in the simulation to generate reasonable community structures?

In this paper, we propose a unified framework in which LLMs update user opinions from textual context, generate

new posts that reflect those updates, and execute paired unfollow/follow decisions that conserve edge counts while reshaping topology. This design lets us study how users adapt their opinions and connections in response to linguistic and contextual cues, under real-world constraints on exposure and rewiring frequency. Our *contributions* can be summarized as:

Language-grounded co-evolution. We design a unified framework where LLMs simultaneously handle opinion updating, content generation, and rewiring, coupling semantic interpretation with structural adaptation.

Empirical validation across three domains. Using real-world datasets, we quantitatively compare simulated and observed networks on multiple network metrics, and evaluate stance dynamics among users. These comparisons establish that LLM-driven simulations reproduce structural regularities more faithfully than equation-based baselines.

Interpretable motives. By examining the explanations that LLMs provide for their unfollow–follow decisions, we uncover consistent categories of motives. These patterns are stable across datasets, offering interpretable insights into why different models evolve networks in distinct ways.

Controllability of network evolution. We show that the global rewiring probability and the choice of rewiring strategy function as explicit control parameters that can be tuned to fit simulations to empirical data, thus providing practical guidance for applying the framework across domains.

Together, these contributions move social simulation beyond rule-driven approximations toward *language-driven, validated* co-evolution, demonstrating that LLMs can capture key structural regularities while providing interpretable mechanisms that connect model behavior to both empirical validation and governance choices.

Related Work

LLM-Driven Social Simulation Leveraging LLMs for social simulation has recently attracted growing attention, enabling researchers to model complex social behaviors and population-level phenomena using language-driven agents. Early studies predominantly rely on prompt-based approaches to elicit socially plausible behaviors, exemplified by Generative Agents (Park et al. 2023) and subsequent systems that simulate emotions, attitudes, and interactions to study information or emotion propagation (Gao et al. 2023). More recent work applies similar agent-based paradigms to opinion-related settings, including the simulation of polarization and echo chamber effects in online networks (Wang et al. 2025; Zheng and Tang 2024).

Our work differs from prior LLM-based social simulations in both focus and evaluation. Rather than centering on prompt design or explicitly targeting echo chamber formation (Wang et al. 2025; Chuang et al. 2024), we treat echo chambers as one possible outcome of a broader co-evolution process between network structure and user opinions. To enable controlled comparison across models, we adopt a minimal and uniform prompting strategy for all LLMs, reducing prompt-induced variability. By grounding the simulation in real-world social network data and quantitatively evaluating both structural and opinion dynamics, our approach highlights how different base LLMs lead to systematically different co-evolution trajectories under the same protocol.

Opinion Dynamic Modeling The study of opinion dynamics in social networks has a rich history, with many classical approaches taking the form of *equation-based models*. Foundational examples include the Friedkin-Johnson Dynamics Model (Friedkin and Johnsen 1990), which simulates opinion evolution based on intrinsic beliefs and social influence, and the Bounded Confidence Model (Deffuant et al. 2000), which restricts interactions to individuals within a confidence range. More recent works, such as Sasahara et al. (Sasahara et al. 2021), emphasize the dynamic interplay of opinion polarization and structural adjustments, such as unfollowing, and demonstrate how these behaviors accelerate the formation of echo chambers. Our work builds on these models by incorporating LLMs to simulate user behavior and opinion dynamics, leveraging advanced natural language understanding to provide deeper insights into the opinion dynamics.

Problem Statement and Definitions

Problem Statement

We study the co-evolution of user opinions and network structure in online social systems. Formally, the system at time t is represented as a directed graph $G(t) = (V, E(t))$, where V is the set of users and $E(t)$ is the set of social edges (e.g., follow or retweet relations). For each user i, we denote the set of its outgoing neighbors at time t by $N_i(t) = \{j \mid (i,j) \in E(t)\}$. Each user $i \in V$ holds a dynamic opinion $O_i(t)$ that evolves over time, while the edge set $E(t)$ also changes to reflect follow and unfollow events. Our goal is to simulate the joint evolution of $\{O_i(t)\}$ and $\{E(t)\}$ so that the generated stance distributions and network structures better fit real-world observations.

Opinion Evolution. Each user's opinion evolves according to an *influence function* f, which maps user i's current state and the information from neighbors into a new opinion value:

$$O_i(t+1) \;=\; f\big(O_i(t), \{O_j(t) \mid j \in N_i(t)\}\big).$$

The exact form of f depends on the modeling approach.

Network Rewiring. The network structure evolves through a *compatibility function* g, which evaluates the suitability of connections. At each time step, with a global rewiring probability $p \in [0,1]$, a user i may reconfigure one of its links; otherwise, the network remains unchanged. If rewiring is triggered, the user unfollows one neighbor and follows one new candidate:

$$j^- = \arg\min_{j \in N_i(t)} g\big(O_i(t), O_j(t)\big),$$
$$j^+ = \arg\max_{j \in \mathcal{C}_i(t)} g\big(O_i(t), O_j(t)\big).$$

where $N_i(t)$ is the current neighbor set and $\mathcal{C}_i(t)$ is the candidate set of non-neighbors. The edge set is then updated by

$$E(t+1) \;=\; E(t) \;\cup\; \{(i, j^+)\} \setminus \{(i, j^-)\}.$$

Equation-based Method

Generally, each user is associated with a numerical opinion score $O_i(t) \in [-1, 1]$, where -1 indicates complete opposition and $+1$ indicates strong support. Under this setting, the abstract functions f and g are instantiated as follows.

The influence function f updates user i's opinion by aggregating neighbor information:

$$f = O_i(t) + \frac{1}{|N_i(t)|} \sum_{j \in N_i(t)} w_{ij} \big(O_j(t) - O_i(t)\big),$$

where $w_{ij} \geq 0$ denotes the weight of neighbor j's influence.

The compatibility function g measures the similarity between two users:

$$g\big(O_i(t), O_j(t)\big) = 1 - |O_i(t) - O_j(t)|.$$

Based on this rule, the neighbor to be unfollowed and the new candidate to be followed are given by

$$j^- = \arg\min_{j \in N_i(t)} g\big(O_i(t), O_j(t)\big),$$
$$j^+ = \arg\max_{j \in C_i(t)} g\big(O_i(t), O_j(t)\big).$$

At each rewiring step (triggered with probability p), user i unfollows j^- and follows j^+.

LLM-enhanced Approach

Building on the same co-evolutionary formulation defined above, we introduce an LLM-enhanced instantiation of the influence and compatibility functions. Instead of representing user opinions solely as low-dimensional numerical variables, this approach models opinions and interactions through textual content and leverages large language models to evaluate opinion change and rewiring decisions. Importantly, this formulation is *complementary* to traditional equation-based models rather than a replacement. Each user is associated with a content history C_i, and the abstract functions f and g are instantiated through prompted LLM evaluations. The influence function f updates user i's opinion by considering its own history and that of a neighbor:

$$f = \text{LLM}(T_o(C_i, C_j)),$$

where T_o (opinion dynamics template) instructs the model to assess how user j's content influences i's stance. Unlike the purely numerical aggregation in equation-based models, this process can incorporate the semantic context of text, such as references to recent events, new arguments, or shifts in discourse that go beyond simple averaging. The compatibility function g evaluates whether two users are suitable connection partners:

$$g = \text{LLM}(T_r(C_i, C_j)),$$

where T_r (rewiring template) guides the model to judge compatibility based on their historical content. Importantly, different LLMs may provide different rationales for these decisions—for instance, prioritizing topical alignment, emotional tone, or information quality—rather than relying solely on numerical opinion similarity. Finally, the LLM also generates new content that reflects user i's opinion and social context:

$$y_i(t) \sim \text{LLM}\Big(T_g(C_i, \{C_j \mid j \in N_i(t)\})\Big),$$

where T_g (generation template) integrates i's past posts with those of its neighbors. In this way, opinion dynamics, rewiring, and content generation are unified within a single framework.

Model Framework

The objective of this study is to design a simulation framework that closely reflects real-world social network dynamics, where both user opinions and network structures evolve over time. The framework models these dynamics through opinion updates, rewiring of social ties, and language-based content generation, as illustrated in Figure 1.

Data Preparation

Our framework is designed for online social data that provide two essential components: (i) user-generated textual content, which serves as input for opinion inference and content generation, and (ii) explicit interaction signals between users (e.g., retweets or reposts), which are used to construct the underlying user network. No platform-specific metadata or proprietary features are required, making the framework applicable to a wide range of social platforms.

In this study, we primarily instantiate the framework using data from X (Twitter), due to its historical accessibility and well-established interaction formats. In addition, we include a Bluesky dataset with comparable post-level content and interaction structures to demonstrate that the framework generalizes beyond a single platform. The framework itself is not tied to X and can be readily applied to other platforms that provide similar textual content and interaction signals.

To ensure computational feasibility while preserving meaningful network structure, we focus on a subset of highly active users rather than simulating the full network. Conceptually, active users can be identified using various criteria, such as posting frequency, interaction volume, or network centrality. In our experiments, we operationalize this selection using a k-core decomposition of the interaction network and retain users in the maximal k-core. The value of k is adjusted per dataset to control the resulting network size, yielding a subgraph of approximately $1{,}000$ users.

The resulting network consists of nodes representing users with inferred opinions and directed edges representing social interactions (e.g., retweets or reposts). Both opinions and edges evolve over time to reflect the dynamic nature of online social interactions.

Simulation Process

The simulation captures the dynamic evolution of user opinions and network connections in social networks. At each iteration, a randomly selected user refreshes their feed, viewing a limited subset of posts from connected "friends" or recommended content. This process, referred to as the **screen**,

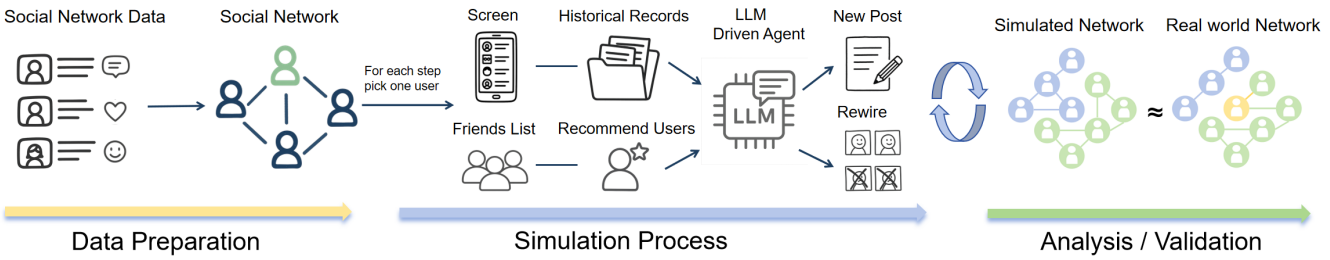

Figure 1: The framework is divided into three main components: (1) **Data Preparation**, where social network data is collected and used to build a user network; (2) **Simulation Process**, where an LLM generates user posts and adjusts connections dynamically; and (3) **Analysis and Validation**, which analyzes the simulated network structure and user's stance.

represents the information accessible to users via social media and simulates their finite cognitive capacity in real-world interactions. Based on the screen and the user's historical activity, the LLM generates a new post that reflects the user's updated opinion.

Subsequently, the user evaluates their network connections and, with a probability p, engages in a rewiring process. This probabilistic design captures the fact that users do not reconsider their social ties at every interaction, but only intermittently. When rewiring is activated, the user reviews the most recent posts of their current friends and selects at most one connection to sever.

A candidate set of potential new connections is then generated according to a specified strategy. In the simplest setting, five non-friend users are randomly sampled as candidates. The user examines the recent posts of these candidates and selects one new connection based on alignment with their historical posts. Abstractly, this process captures a tendency toward opinion-aware tie adjustment, rather than enforcing systematic disconnection from opposing views, and allows for heterogeneous user behaviors, including selective engagement with differing opinions.

To maintain a stable network scale and facilitate comparison with real-world data, each rewiring event consists of one edge removal followed by one edge addition, keeping the total number of edges approximately constant. Similar assumptions are commonly adopted in adaptive network models to study opinion-network co-evolution (Sasahara et al. 2021; Holme and Newman 2006; Kozma and Barrat 2008). Under this mechanism, rewiring primarily reshapes network structure rather than network size, leading to gradual community merging or fragmentation over time.

Over time, the network exhibits clustering among users with similar opinions, while those with conflicting views become increasingly isolated. The simulation terminates either when the network reaches a stable state—characterized by stabilized opinions and connections—or after a predefined number of iterations. The Simulation Process component in Figure 1 illustrates how historical data and network information jointly drive opinion updates and connection adjustments.

Validation Process

To assess the credibility and applicability of the simulation, the framework incorporates a validation process that compares the simulated network with real-world social network data. Validation operates on two complementary levels: Network structural validation and content/opinion validation.

For Network structural validation, the simulated network is evaluated against the real-world network using established network science metrics. These include measures of community structure, global connectivity, and assortativity (preference for homophily in connections). By comparing these metrics, we assess whether the emergent structural patterns in the simulation approximate those observed in empirical social networks. The subsequent experiments will provide a more detailed explanation of the specific metrics employed.

For content and opinion validation, the framework examines opinion dynamics through stance-based evaluation. Specifically, the distribution of simulated opinions is compared with real-world opinion distributions (e.g., pro/anti stance ratios), and stance transitions are evaluated against ground-truth labels. These comparisons ensure that the simulation not only reproduces structural properties but also captures realistic patterns of opinion evolution.

Together, these validation steps provide a rigorous benchmark, ensuring that the simulated dynamics approximate real-world social interactions both in terms of network structure and language-based opinion processes.

Prompt Templates

To ensure reproducibility and enable systematic exploration of different prompt designs, the framework employs standardized prompting mechanisms. These prompts guide the LLM to model user behaviors effectively while maintaining consistency across tasks. The prompt design adheres to two key principles: **Standardization.** Prompts follow a consistent format with explicitly defined tasks and output specifications. This ensures clarity and reproducibility across simulations. **Chain-of-Thought Reasoning** (Wei et al. 2022). Prompts encourage step-by-step reasoning to enhance the interpretability and reliability of the generated outputs.

An example prompt template for generating a new post based on a user's historical content and surrounding tweets is provided in the Appendix. The same template structure

is reused across different simulation tasks, such as opinion updating and network rewiring—by modifying only the task, specific instructions and input contexts. This consistent use of structured prompts enables systematic exploration of prompt designs and facilitates controlled analysis of how prompt variations affect simulation outcomes.

Experiments

Stand-Alone Simulation

One important application of simulating social network dynamics is to model the formation of echo chambers. In this section, we present a pure simulation experiment using ChatGPT and Gemini to explore how echo chambers emerge. At the beginning of the experiment, each agent is assigned an opinion score drawn from a normal distribution. As a result, most agents start with opinions close to neutral, while only a few occupy extreme positions. The simulation topic is the COVID-19 vaccine, where -1 denotes an anti-vaccine stance and 1 denotes a pro-vaccine stance. This feasibility test aims to illustrate how LLMs simulate the evolution and divergence of opinions, ultimately leading to clusters of like-minded users.

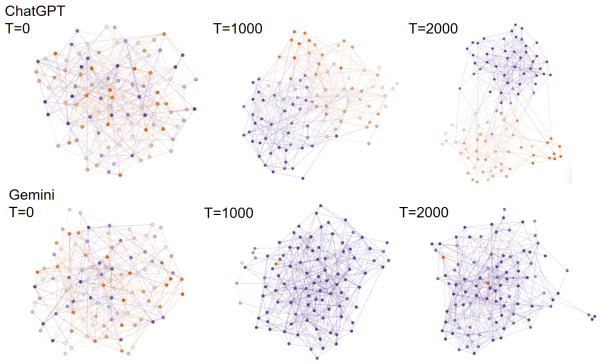

Figure 2: Simulation results using ChatGPT and Gemini. Opinion dynamics at different time steps illustrate the gradual formation of echo chambers.

Figure 2 presents the simulation results, where node colors represent opinion states: shades range from orange (-1) to deep blue (1), with intensity reflecting the degree of opposition or support. Initially, opinions are randomly distributed across the network. Over time, the influence of neighboring nodes causes opinions to cluster, gradually transforming the network structure into distinct echo chambers.

It is noteworthy that in the simulations, ChatGPT tends to produce highly polarized outcomes, with opinions splitting into opposing clusters. In contrast, Gemini exhibits a stronger tendency toward consensus, often generating more uniformly positive sentiment. These observations reveal potential biases in LLMs, likely influenced by differences in training methods and datasets. Since this study focuses on COVID-19 vaccines, it is plausible that Gemini has been fine-tuned to produce more positive attitudes toward the topic. Moreover, because this is a purely simulated experiment, all initial data consist of opinion scores between -1 and 1, which are then mapped into corresponding texts by the LLM. This approach is commonly used in related literature—either scoring with LLMs or directly representing agent states with numerical scores. However, in our experiments, Gemini clearly struggles to map scores into realistic tweets. Even when instructed to generate negative (-1) content, Gemini often attempted to produce neutral or even supportive statements about COVID-19 vaccines. This limitation is the main reason why our framework abandons direct use of opinion scores and instead relies on providing historical tweets to guide LLM reasoning.

This experiment demonstrates that when combining LLMs with the rewiring process, the formation of echo chambers can be effectively illustrated, both through the evolving network structure and the generation of new, natural language posts. However, it still lacks a clear benchmark for determining which simulation is more accurate or realistic. This limitation highlights the importance of incorporating real-world data to validate and strengthen the credibility of LLM-driven simulations.

Real Data-Driven Simulation

In this section, we outline the experimental setup designed to evaluate our proposed framework for modeling network dynamics formation using real-world data. We provide a detailed description of the datasets, simulation parameters and evaluation metrics employed to analyze opinion dynamics and network dynamics across different contexts.

The experiments are conducted on three datasets:

1. **COVID-19 Vaccination Dataset:** This dataset was collected from global English-language tweets using COVID-19-related keywords . Details about the keywords and the collection process can be found in (Lamsal 2021). The dataset includes tweets with stances categorized as *favoring* vaccination, *opposing* vaccination, or *neutral*, focusing on the topic of vaccine hesitancy (Zaidi et al. 2023).
2. **Ukraine War Dataset:** Introduced in (Perera and Karunasekera 2024), this dataset includes 1 million English-language tweets collected using a keyword search between August 1 and August 31, 2022. Each tweet's stance was labeled as *pro-Russian*, *pro-Ukrainian*, or *neutral*, providing a rich resource for studying polarized discourse.
3. **PolitiSky24: U.S. Election 2024 Bluesky Dataset:** This dataset was recently introduced in (Rostami et al. 2025). It is the first stance detection dataset focusing on the 2024 U.S. presidential election and was collected from the decentralized social platform **Bluesky**. Stances were determined using a carefully designed pipeline that integrates LLM-based classification.

Model Selection and Experimental Setup To evaluate the simulation of opinion dynamics, we employed five generative models alongside a traditional numerical simulation method, enabling a comparison between LLM-driven and rule-based approaches. The selected LLMs originate from different companies and countries, with distinct training pro-

cesses that likely introduce performance variations. Evaluating and benchmarking LLMs remains an active area of research, and within the context of social simulations, no consensus has yet been reached. We hope this study can serve as an initial exploration and provide useful insights.

The LLMs considered include both closed-source and open-source models, reflecting a diversity of capabilities. The closed-source group consists of **GPT4o- Mini** (OpenAI) and **Gemini** (Google), while the open-source group includes **Mistral** (Mistral AI), **DeepSeek** (DeepSeek-V3), and **Qwen** (Qwen2-72B-Instruct). These models differ in their architectures and training paradigms, with strengths ranging from contextual reasoning to domain adaptability.

For all models, we adopt generic generation settings to ensure stable and comparable outputs. In particular, we use the default decoding configurations provided by each API (e.g., temperature $= 1.0$ and top-p $p = 1.0$), without any model-specific tuning. These settings are chosen intentionally as conventional defaults rather than optimized hyperparameters, so as to isolate model-level behavior from prompt or parameter engineering effects.

As a baseline, we employ a traditional **equation-based simulation** (Sasahara et al. 2021), which relies on predefined mathematical rules to model opinion interactions. This baseline serves as a control for highlighting the added value of LLM-driven agents in capturing semantic and contextual nuances. We further include a **Random-Rewire** baseline, where rewiring and opinion updates are both performed uniformly at random, providing a stochastic lower bound.

Given the stochastic nature of both the models and the text generated by LLMs, each simulation was run five times per model, and the results were averaged. The experiments were initialized with approximately 6,000 tweets, with variations across different datasets. The number of active users was controlled at around 1,000, and after 2,000 simulated steps, we evaluated the resulting network state and inspected the distribution of opinions.

Evaluation Metrics We evaluate model performance using five complementary metrics: **Modularity** (Newman and Girvan 2004): Captures the strength of community structure, where higher values indicate more distinct clusters and stronger echo-chamber effects. **Average Clustering Coefficient** (Botte, Ryckebusch, and Rocha 2022): Measures the tendency of users to form tightly interconnected triads, reflecting local cohesion in the network. **Average Path Length** (Watts and Strogatz 1998): Quantifies the typical distance between pairs of users, indicating efficiency of information diffusion and global connectivity. **Assortativity** (Newman 2003): Evaluates homophily in tie formation, i.e., the extent to which users with similar opinions are more likely to be connected. **Stance Accuracy:** Compares simulated opinion distributions with real-world stance labels, focusing on users who change stance, to assess how well models capture opinion dynamics.

Experimental Result Analysis

Our experiments are structured to answer RQ1–RQ3: we validate LLM-based co-evolution against real data (RQ1), compare different LLMs under the same protocol (RQ2), and test controllability via p and rewiring strategies (RQ3).

Overall Result Analysis

Table 1 reports MAE values for key network metrics together with stance accuracy across three datasets that differ in ideological orientation and structural properties. The **Bluesky** dataset, centered on the 2024 U.S. election, is dominated by left-leaning discourse opposing Donald Trump, yielding a relatively rigid and homogeneous network. The **COVID-19** dataset contains both pro- and anti-vaccine communities, with a majority supporting vaccination but exhibiting higher volatility and more dynamic interactions. The **Ukraine** dataset, based largely on English tweets, reflects a strong pro-Ukraine orientation and occupies a middle ground between the structural rigidity of Bluesky and the dynamism of COVID-19.

A first and overarching finding is that **LLM-based simulations consistently outperform the equation baseline on the majority of structural metrics across datasets**, particularly on modularity and assortativity, which are central to characterizing community separation and homophily. Equation-based methods tend to produce inflated modularity values and unrealistic clustering in more volatile and contested settings such as COVID-19 and Ukraine, thereby exaggerating polarization or underestimating local connectivity. In contrast, LLMs reproduce structural patterns with substantially lower error, indicating their superior ability to approximate how echo chambers and information diffusion emerge in real-world settings.

Looking more closely at the individual metrics reveals complementary strengths across models. Assortativity, which reflects homophily in tie formation, is most faithfully reproduced by Gemini and DeepSeek, while clustering coefficients, capturing the intensity of local triadic closure, are best modeled by Qwen and Mistral. Average path length, which indexes efficiency of global diffusion, is accurately reproduced by Gemini and Mistral in the more dynamic COVID-19 and Ukraine datasets, whereas the equation method performs competitively on certain global statistics in the more rigid Bluesky setting, suggesting that in homogeneous and stable networks, simple rule-based processes can occasionally approximate specific structural properties. Modularity, a canonical indicator of echo chambers, is best aligned with ground truth in Qwen and DeepSeek, which represent strong community separation without over-amplifying polarization. These distinctions underscore that each metric captures a distinct social mechanism—homophily, local clustering, diffusion efficiency, and echo chamber intensity—and that LLMs are flexible enough to approximate these mechanisms across contrasting contexts.

It is therefore unsurprising that no single LLM dominates across all datasets. The three domains differ substantially in both topical orientation and structural configuration: Bluesky reflects a left-leaning, rigid community; COVID-19 discourse exhibits more volatile and contested dynamics; and Ukraine discussions are dominated by pro-Ukraine voices but retain moderate structural division. That differ-

Dataset	Model	Modularity	Avg. Clustering	Avg. Path Length	Assortativity	Stance (%)
COVID-19	Random-Rewire	0.1160	0.0281	0.4286	0.1061	33.21
	Equation	0.0643	0.0212	0.2952	0.0696	33.65
	Gemini	0.0263	0.0095	0.2495	**0.0288**	38.21
	Mistral	0.0278	0.0098	0.2580	0.0321	**40.49**
	OpenAI	0.0272	0.0091	0.2525	0.0312	36.12
	Qwen	**0.0259**	**0.0071**	0.2576	0.0332	33.84
	DeepSeek	0.0291	0.0109	**0.2479**	0.0312	37.45
Ukraine	Random-Rewire	0.0301	0.0298	0.2081	0.1908	33.17
	Equation	0.0347	0.0141	0.1317	0.0906	40.95
	Gemini	0.0231	0.0170	0.0568	**0.0768**	39.37
	Mistral	**0.0193**	0.0072	**0.0511**	0.0842	41.21
	OpenAI	0.0212	0.0163	0.0641	0.0786	40.32
	Qwen	0.0194	**0.0069**	0.0681	0.0806	**45.71**
	DeepSeek	0.0204	0.0164	0.0617	0.0844	42.31
Bluesky	Random-Rewire	0.0560	0.0292	0.0256	0.1007	33.16
	Equation	0.0264	0.0154	**0.0165**	0.0023	39.58
	Gemini	0.0194	0.0041	0.0408	0.0102	**44.83**
	Mistral	0.0238	0.0038	0.0394	0.0128	38.57
	OpenAI	0.0191	**0.0025**	0.0441	0.0132	39.01
	Qwen	**0.0159**	0.0077	0.0358	0.0089	36.97
	DeepSeek	0.0165	0.0026	0.0365	**0.0063**	39.71

Table 1: Mean Absolute Error (MAE) of different models across datasets, reported for Modularity, Average Clustering, Average Path Length, and Assortativity. Stance is reported separately as classification accuracy (%). Lowest values per MAE metric and highest values for stance accuracy are highlighted in bold.

ent models excel on different dimensions reflects not a limitation, but rather the dataset-specific adaptability of LLMs. This adaptability moves beyond the view of LLMs as opaque black boxes, instead opening a pathway to interpret how different models emphasize particular structural mechanisms depending on context.

Stance prediction offers an even stricter evaluation. If computed across all users, global stance accuracy typically falls in the 80–90% range, since most users do not change stance and models can achieve high scores simply by preserving inertia. Because real-world networks already exhibit some degree of echo chamber effect, actual stance shifts are rare. We therefore compute accuracy only among users who change stance, highlighting model sensitivity to these minority transitions. Under this stricter criterion, LLMs achieve meaningful performance across all datasets, with Gemini, Qwen, and Mistral each excelling in different contexts. By contrast, the equation baseline exhibits highly context-dependent performance: while it can achieve moderate stance accuracy in rigid environments, it largely defaults to inertia and fails to reliably capture minority stance transitions across diverse settings.

Taken together, these findings demonstrate that **LLM-based simulations are more accurate on key structural metrics and better reflect underlying social mechanisms than equation-based methods.** While no single model dominates across all metrics, their complementary strengths suggest the promise of ensemble strategies.

Cross-Model Consistency in Rewiring Decision-Making

When comparing the rewiring outcomes produced by different LLMs, our first step is to examine the extent to which their decisions overlap. In our simulation framework, we enforce the constraint that the total number of edges in the network remains constant. This reflects the realistic assumption that users' social attention capacity is limited: forming a new connection typically comes at the cost of abandoning an old one. Therefore, each rewiring step consists of a paired operation, where one unfollow action is immediately followed by one follow action.

To enable a fair comparison across models, we extract the complete set of rewiring prompts from a single simulation run. All models are exposed to the same sequence of prompts under an identical initial network structure. This design controls for network drift: if different rewiring decisions were allowed to propagate, the networks would quickly diverge and comparability would be lost. By fixing the developmental trajectory and only allowing variation in the choices at each step, we isolate the decision-making tendencies of the models themselves.

The analysis shows that the probability of two models agreeing on both the unfollow and follow targets in the same step is only about 30%. In most cases, models diverge, indicating that their decisions are not simply random but reflect systematic differences in how they interpret social signals. As illustrated in Figure 3, agreement is consistently higher on unfollow actions (0.58–0.74 across pairs) than on follow actions (0.44–0.55). This suggests that models converge more easily on identifying connections to re-

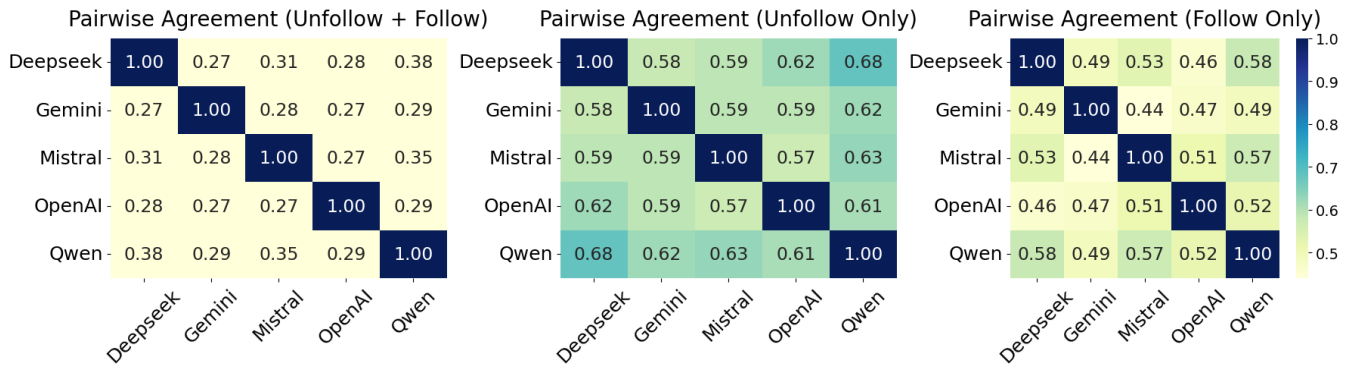

Figure 3: Pairwise agreement among LLMs when aligning decisions on Unfollow and Follow actions. The three panels compare overlap in (*left*) combined Unfollow+Follow, (*middle*) Unfollow-only, and (*right*) Follow-only decisions. Results show that models exhibit substantially higher overlap on Unfollow targets than on Follow targets, indicating that unfollowing is a more consistent decision across LLMs.

move—likely because redundancy or irrelevance provides clearer cues—whereas selecting new connections introduces greater variability, since the set of plausible follow candidates is broader and more context-dependent.

The pairwise agreement heatmaps also reveal emergent clusters of model behavior. DeepSeek and Qwen form the tightest pair, showing the highest overlap on both unfollow (0.737) and follow (0.550) decisions. Gemini and OpenAI constitute a moderate but consistent cluster, with mid-level agreement in both tasks (around 0.47–0.59). Mistral, by contrast, remains comparatively unstable: while it aligns reasonably well with other models in unfollow decisions (0.57–0.63), its follow choices are more dispersed (0.44–0.57), indicating weaker consistency in tie formation. Interestingly, the observed clusters also align with the geographic and institutional origins of the models (e.g., DeepSeek and Qwen from China; Gemini and OpenAI from the U.S.). While this correlation may reflect shared training pipelines or alignment practices, it also raises the possibility that national or cultural context embedded in the training data contributes to systematic differences in social decision-making.

Taken together, these findings suggest that unfollowing reflects a more determinate process that elicits higher cross-model consensus, while following is a more ambiguous and heterogeneous process that highlights the diverse inductive biases of LLMs. The divergence across models is therefore not noise, but evidence of distinct reasoning patterns in evaluating social relevance and structural adjustment.

Rationales Behind Rewiring Decisions

Beyond the structural accuracy reported in Table 1 and the agreement patterns in Figure 3, we further examine the *motivational rationales* provided by LLMs when explaining their rewiring choices. To the best of our knowledge, no prior work has systematically proposed a quantitative scheme for categorizing rewiring motives. We therefore introduce a five-dimensional taxonomy—*Tie*, *Affect*, *UGT* (uses and gratifications), *Moral*, and *InfoQuality*—drawing on classical theories of social tie formation, affective communication, uses-and-gratifications, moral foundations, and information quality (Granovetter 1973; Haidt and Graham 2007).

Tie captures relational considerations such as reciprocity or lack of interaction; **Affect** reflects emotional tone, including both positive resonance and negative hostility; **UGT** denotes instrumental or informational value, such as novelty or contextual relevance; **Moral** encodes value-laden frames tied to policy, accountability, or rights; and **InfoQuality** concerns judgments about the reliability, redundancy, or clarity of information.

In practice, these dimensions are operationalized via a transparent, rule-based tagging pipeline. We first apply a lightweight keyword-based detector to the free-text rationales produced by the LLMs, yielding fine-grained motive tags (e.g., off-topic content, low information quality, extremism tone, novelty, policy or rights-related framing). These fine-grained tags are then grouped into the five higher-level dimensions above. For example, *Tie* is often associated with relational cues such as "reciprocal" or "inactive" *Affect* with affective terms such as "toxic" "hopeful" or "inspiring". This two-stage mapping provides a transparent and reproducible link from free-text rationales to theoretically grounded motivational dimensions. Importantly, the relative ordering of provider profiles induced by these dimensions is consistent across datasets, indicating that the observed tendencies reflect stable model-level biases rather than dataset-specific artifacts.

As shown in Figure 4, these motivational dimensions reveal systematic cross-model differences. Mistral strongly emphasizes utility- and value-driven frames (UGT and Moral), OpenAI highlights emotional tone (Affect), DeepSeek and Qwen adopt relatively balanced profiles—with Qwen showing a marked emphasis on Tie—while Gemini maintains a uniformly conservative profile, consistent with a general concern for fairness rather than dominance in any single motive. These differences complement the MAE results in Table 1, which demonstrate that different models excel in reproducing different network

mechanisms: Gemini and DeepSeek most accurately capture assortativity, Qwen and Mistral best reproduce clustering (local cohesion), Gemini and Mistral minimize error in path length for the dynamic COVID-19 and Ukraine datasets, while Qwen and DeepSeek achieve the lowest modularity error. Stance accuracy further underscores this heterogeneity: Qwen leads in Ukraine, Gemini in Bluesky, and Mistral in COVID-19. Model performance is therefore shaped not only by dataset structure but also by inductive biases embedded in training pipelines.

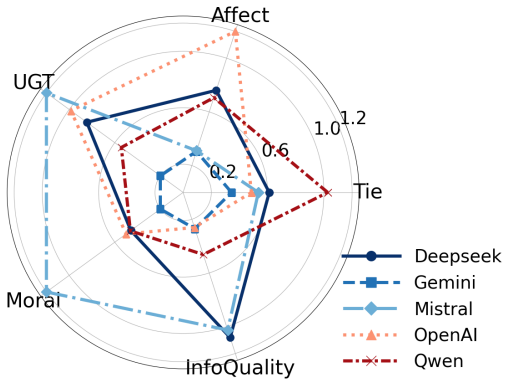

Figure 4: Radar plot of averaged rewiring motives across datasets (normalized to [0.2, 1.2]), highlighting distinct model tendencies.

Motivational analysis clarifies how inductive biases shape model reasoning. Mistral's reliance on UGT and Moral justifications prioritizes cohesive, value-driven ties and aligns with its strength in clustering. OpenAI's focus on Affect corresponds to its advantage in affective dynamics, while the balanced profiles of DeepSeek and Qwen support their stability on modularity and assortativity. Dataset contexts further amplify these tendencies: partisan discourse in Bluesky heightens the role of moral and affective cues, the contested COVID-19 environment foregrounds information quality and utility, and Ukraine's moderately divided but pro-Ukraine structure sharpens the trade-off between homophily and diffusion efficiency.

Taken together, structural metrics and motivational profiles indicate that **LLMs differ not only in statistical accuracy but also in the interpretive frames guiding their rewiring decisions**. These dual perspectives—quantitative fit to network structures and qualitative rationales—reveal distinct inductive biases rooted in training and alignment pipelines. They also echo the agreement analysis in Figure 3 unfollowing is often attributed to redundancy or low information quality, yielding high cross-model consensus, whereas following draws on heterogeneous motives—novelty, moral alignment, or affective resonance—resulting in greater divergence.

Impact of p on Network Dynamics

In our framework, large language models determine *who* to follow or unfollow, but they do not independently decide *when* such rewiring occurs. Instead, the overall probability of rewiring is controlled by a global parameter p, adapted from equation-based approaches. This design reflects an important distinction: while LLMs are responsible for generating linguistically and contextually grounded social decisions, the parameter p regulates the frequency of structural change at the network level. In practice, this allows us to simulate a variety of network environments and to test potential mitigation strategies.

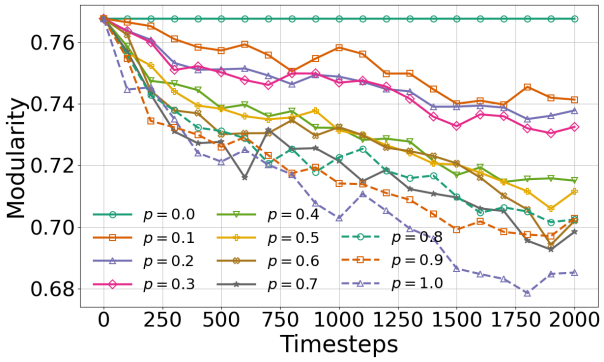

Figure 5: Modularity dynamics under different rewiring probabilities p in COVID-19 dataset. Higher values of p generally lead to a faster decline in modularity, indicating reduced community polarization as the network structure becomes more fluid.

As shown in Figure 5, increasing p accelerates the rate at which modularity declines, indicating that more frequent rewiring reduces the persistence of polarized clusters. Conversely, lower values of p preserve structural rigidity and thus sustain stronger echo chambers. Interpreted in real-world terms, different policy or platform-level interventions can be mapped onto variations in p: more aggressive interventions correspond to higher rewiring probabilities, which promote mixing and reduce polarization, while more conservative settings maintain stability but risk reinforcing division. Importantly, even with LLMs driving individual-level reasoning, the parameter p continues to play a crucial role, ensuring that global patterns of structural adaptation remain tunable and interpretable.

Empirically, we find that the optimal p value for reproducing the most faithful network structures differs across datasets: Bluesky aligns best at $p = 0.03$, COVID-19 at $p = 0.2$, and Ukraine at $p = 0.5$. These differences mirror the structural characteristics discussed earlier: Bluesky's rigid and homogeneous discourse requires only minimal rewiring to approximate reality, COVID-19's volatile and contested environment demands a moderate level of rewiring, while Ukraine's dynamic but moderately divided structure is best captured under high rewiring probability.

These findings highlight that p is not merely a technical parameter but a proxy for governance strategies. Low p resembles laissez-faire environments where ties remain static and echo chambers harden, while high p approximates strong interventions such as active recommendation reshuffling or enforced exposure to diverse viewpoints. Intermediate values of p reflect softer nudges that allow net-

works to adapt without disrupting stability. Framing p in this way bridges simulation outcomes with broader normative debates on polarization, user autonomy, and the role of platforms in shaping public discourse. Due to space constraints, we report modularity here, but other structural metrics (e.g., clustering, assortativity) also display consistent trends of polarization decline—an expected outcome, since more frequent rewiring naturally destabilizes rigid community boundaries and fosters structural reconfiguration.

Effects of Rewiring Strategies on Network and Polarization

We further examine the impact of different rewiring strategies. While our framework enables LLMs to make linguistically grounded decisions, users' exposure to information is ultimately bounded by platform-level recommendation mechanisms. In practice, recommendation systems act as a *screen* that determines which candidates are presented to users. To capture this constraint, we implement and compare five strategies: **Random**, where candidates are drawn uniformly at random; **FoF**, which recommends friends-of-friends; **Structural-farthest**, which recommends the most distant user in the network (a counterfactual opposite to FoF, and rarely realizable in real platforms); **TF_nearest**, which recommends the textually closest user; and **TF_farthest**, which recommends the most textually distant user.

Strategy	Modularity	Avg. Clustering
FoF	0.711	0.0557
TF_farthest	0.694	0.0374
TF_nearest	**0.745**	**0.0670**
Random	0.689	0.0287
Structural-farthest	0.679	0.0288

Table 2: Final network metrics in COVID-19 dataset under different rewiring strategies.

Table 2 shows that the choice of strategy significantly shapes structural outcomes. TF-nearest maximizes modularity and clustering, producing the most pronounced echo-chamber structure with longer path lengths. By contrast, structural-farthest minimizes modularity and clustering, resulting in a more diffuse and less polarized network. Random rewiring serves as a neutral baseline, with moderate modularity and clustering. FoF falls between these extremes, preserving moderate modularity and clustering.

Stance polarization. At each time step, we parse generated posts into discrete stance labels (`FAVOR`, `AGAINST`, `None`) and compute their proportions within each time bin. We operationalize stance polarization as the absolute imbalance between pro and anti stances, $P(t) = |p_F(t) - p_A(t)|$. To isolate strategy effects, we report polarization *relative to the Random baseline*, and summarize results by the mean ΔP with 95% confidence intervals across all time bins over the 2,000-step simulation horizon. Figure 6 evaluates stance polarization relative to the random baseline. Interestingly, FoF produces the strongest increase in stance polarization, highlighting the risk that friend-of-friend recommendations amplify homogeneity and reinforce echo chambers. In contrast, text-based recommendations (TF-nearest and TF-farthest) exert only weak influence on stance polarization, suggesting that content similarity alone does not strongly shift opinion alignment. Most strikingly, structural-farthest achieves the greatest reduction in stance polarization, indicating that exposing users to structurally distant connections is highly effective for mitigating echo chamber effects. Between the two text-based strategies, TF-nearest shows a somewhat larger reduction in polarization but with high variance and instability, whereas TF-farthest remains closer to neutral, producing only mild and more stable depolarizing effects.

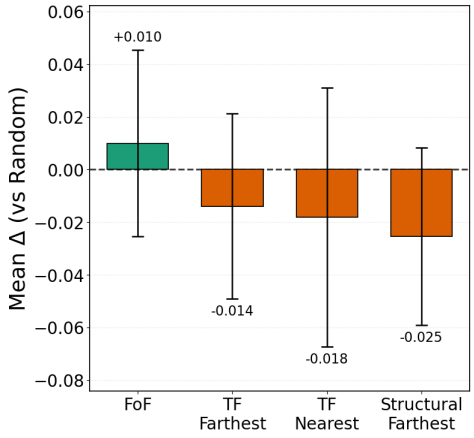

Figure 6: Stance polarization in COVID-19 dataset. Bars report the mean ΔP across time bins, with 95% confidence intervals.

Taken together, these findings underscore that the design of recommendation strategies has a dual effect: it not only shapes structural features of the network but also strongly conditions opinion polarization. Strategies grounded in local reinforcement (FoF) risk entrenching divisions, while strategies that connect distant parts of the network (structural-farthest) offer a promising path toward depolarization.

Visualization of COVID-19 Network Dynamics

We compare the simulated network dynamics produced by GPT-4o Mini with the ground truth across three time steps (Figure 7). At the macro level, GPT-4o Mini captures the overall emergence of polarized communities and, by T=2000, produces two major clusters that echo the multipolar structure observed in the real network. This suggests that the model is able to approximate community formation trends even without explicit behavioral rules.

Nevertheless, notable differences remain. The simulated trajectories produced by GPT-4o Mini tend to reduce distances between major communities over time, resulting in more compact network structures and a stronger tendency toward structural smoothing or partial de-polarization. By contrast, the ground truth exhibits more clearly separated communities, with inter-community boundaries remaining

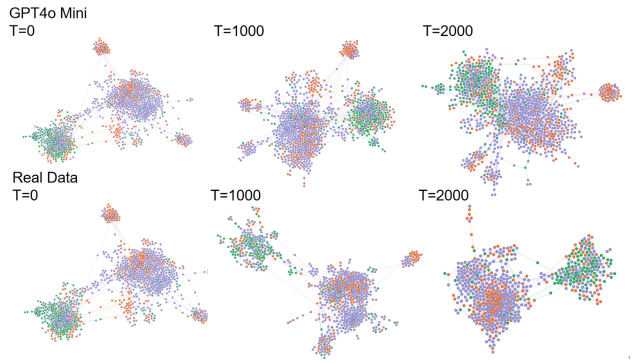

Figure 7: Comparison of network evolution between GPT-4o Mini (top) and the ground truth (bottom) at T=0, 1000, and 2000.

stable and homophily persisting at the community level throughout the evolution.

Beyond structure, GPT-4o Mini also generates new natural language posts associated with evolving nodes. These outputs provide additional textual material for subsequent analysis, allowing us to study not only how networks evolve but also how opinions and linguistic styles shift in tandem with structural dynamics. This dual capability—simulating network structure while producing interpretable text—opens opportunities for multi-layered investigations of social systems.

Application: Simulation Response to News Shocks

We present a case study of *news shocks* in a COVID-19 discussion network to demonstrate the applicability of the proposed simulation model for analyzing collective behavioral responses under exogenous information interventions. Starting from identical initial network structures and opinion configurations, we inject an external news event at time step t=500 and compare three scenarios: *no shock*, *positive news*, and *negative news*. Positive news emphasizes the effectiveness and protective role of vaccines, whereas negative news highlights risk-oriented narratives, such as severe adverse reactions or reported deaths. After the shock is introduced, the news content is exposed to agents in the network in the same manner within a fixed time window, while all other simulation settings remain unchanged.

Figure 8 illustrates the temporal evolution of the average opinion under the three scenarios. In the absence of exogenous intervention, the overall opinion level remains relatively stable. In contrast, both positive and negative news shocks induce clear and persistent opinion changes. Positive news leads to a pronounced upward shift in the average opinion following the intervention. Notably, negative news does not drive the overall opinion toward an anti-vaccine direction; instead, it also results in an upward trend in the mean opinion. This counterintuitive outcome arises because pro-vaccine views constitute the majority in the initial network state. When confronted with negative information, this majority group tends to respond more actively by amplifying supportive narratives, thereby offsetting the direct influence of negative framing at the aggregate level.

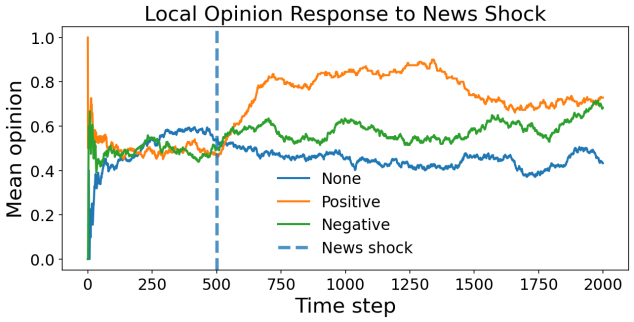

Figure 8: Mean opinion trajectories under different news shock scenarios. An exogenous news event is injected at time step t=500 (dashed line).

Metric	None	Positive	Negative
Modularity	0.560	−0.0034	−0.0055
Avg. Clustering	0.041	+0.0038	+0.0052
Avg. Path Length	3.948	+0.0125	+0.0305
Assortativity	−0.272	+0.0024	+0.0004

Table 3: Global network structure during the shock window. Values for the None report averages, while positive and negative show changes relative to the None baseline.

Beyond opinion dynamics, we further examine changes in the global network structure during the shock window. Table 3 summarizes the average values of key network structural indicators over this period. Compared to the no-shock baseline, both positive and negative news shocks lead to a decrease in network modularity, indicating a temporary weakening of community boundaries during the intervention. Meanwhile, the average clustering coefficient increases, suggesting that local interaction patterns become denser. Across all metrics, negative news induces stronger structural perturbations than positive news, manifested as the formation of more tightly connected local communities. However, the magnitude of these effects remains moderate and does not result in a fundamental reorganization of the overall network structure.

Overall, this case study demonstrates that the simulation model can characterize, under a unified experimental setting, the joint impact of exogenous information shocks on *collective opinion evolution* and *network structural adjustment*. This provides a controllable analytical tool for studying public opinion dynamics and network behavior in the context of sudden external events.

Conclusion

We presented a language-driven simulation framework that integrates LLMs into both opinion updating and network rewiring, and validated it on three real-world datasets. Compared to an equation-based baseline, LLMs consistently re-

duce MAE on key structural metrics while producing stance dynamics that are closer to real-world observations. For example, on the COVID-19 dataset, Qwen reduces the MAE of modularity by more than half (0.064→0.026), while Gemini achieves lower MAE on average path length (0.295→0.250). On the Ukraine dataset, Mistral reduces path length MAE by nearly two-thirds (0.132→0.051), and Qwen achieves the highest stance accuracy (45.7% vs. 40.9%). In the Bluesky dataset, Qwen attains the lowest modularity MAE (0.0159 vs. 0.0264), while Gemini leads in stance accuracy (44.8% vs. 39.6%).

Across datasets, no single model dominates all metrics; instead, models exhibit complementary strengths, suggesting the potential value of ensemble or hybrid approaches. Our study further demonstrates that platform-level regulation is equally critical: a simple global rewiring probability p and the choice of candidate-screening strategy systematically influence polarization outcomes, thereby providing a transparent bridge between simulation parameters and policy interventions.

Overall, these results show that LLM-based simulations are more accurate in terms of absolute structural deviation and more interpretable than traditional equation-based models. By minimizing MAE through explicit and controllable mechanisms such as the rewiring probability p and candidate-screening strategies, the framework captures network evolution with quantitative fidelity. This establishes language-driven simulation as a practical tool for studying and mitigating online polarization in computational social science.

References

Botte, N.; Ryckebusch, J.; and Rocha, L. E. 2022. Clustering and stubbornness regulate the formation of echo chambers in personalised opinion dynamics. *Physica A: Statistical Mechanics and its Applications*, 599: 127423.

Chuang, Y.-S.; Goyal, A.; Harlalka, N.; Suresh, S.; Hawkins, R.; Yang, S.; Shah, D.; Hu, J.; and Rogers, T. 2024. Simulating opinion dynamics with networks of llm-based agents. In *Findings of NAACL 2024*, 3326–3346.

Deffuant, G.; Neau, D.; Amblard, F.; and Weisbuch, G. 2000. Mixing beliefs among interacting agents. *Advances in Complex Systems*, 3(1-4): 87–98.

Flaxman, S.; Goel, S.; and Rao, J. M. 2016. Filter bubbles, echo chambers, and online news consumption. *Public opinion quarterly*, 80(S1): 298–320.

Friedkin, N. E.; and Johnsen, E. C. 1990. Social influence and opinions. *Journal of Mathematical Sociology*, 15(3-4): 193–206.

Gao, C.; Lan, X.; Lu, Z.; Mao, J.; Piao, J.; Wang, H.; Jin, D.; and Li, Y. 2023. S3: Social-network Simulation System with Large Language Model-Empowered Agents. *arXiv preprint arXiv:2307.14984*.

Granovetter, M. S. 1973. The Strength of Weak Ties. *American Journal of Sociology*, 78(6): 1360–1380.

Haidt, J.; and Graham, J. 2007. When morality opposes justice: Conservatives have moral intuitions that liberals may not recognize. *Social justice research*, 20(1): 98–116.

Holme, P.; and Newman, M. E. 2006. Nonequilibrium phase transition in the coevolution of networks and opinions. *Physical Review E—Statistical, Nonlinear, and Soft Matter Physics*, 74(5): 056108.

Kossinets, G.; and Watts, D. J. 2006. Empirical analysis of an evolving social network. *science*, 311(5757): 88–90.

Kozma, B.; and Barrat, A. 2008. Consensus formation on coevolving networks: groups' formation and structure. *Journal of Physics A: Mathematical and Theoretical*, 41(22): 224020.

Lamsal, R. 2021. Design and analysis of a large-scale COVID-19 tweets dataset. *Applied Intelligence*, 51(5): 2790–2804.

McPherson, M.; Smith-Lovin, L.; and Cook, J. M. 2001. Birds of a feather: Homophily in social networks. *Annual review of sociology*, 27(1): 415–444.

Newman, M. E. 2003. Mixing patterns in networks. *Physical review E*, 67(2): 026126.

Newman, M. E.; and Girvan, M. 2004. Finding and evaluating community structure in networks. *Physical review E*, 69(2): 026113.

Park, J. S.; O'Brien, J.; Cai, C. J.; Morris, M. R.; Liang, P.; and Bernstein, M. S. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, 1–22.

Perera, K.; and Karunasekera, S. 2024. Quantifying Opinion Rejection: A Method to Detect Social Media Echo Chambers. In *PAKDD*, 57–69. Springer.

Rostami, P.; Rahimzadeh, V.; Adibi, A.; and Shakery, A. 2025. PolitiSky24: US Political Bluesky Dataset with User Stance Labels. *arXiv preprint arXiv:2506.07606*.

Sasahara, K.; Chen, W.; Peng, H.; Ciampaglia, G. L.; Flammini, A.; and Menczer, F. 2021. Social influence and unfollowing accelerate the emergence of echo chambers. *Journal of Computational Social Science*, 4(1): 381–402.

Wang, C.; Liu, Z.; Yang, D.; and Chen, X. 2025. Decoding echo chambers: LLM-powered simulations revealing polarization in social networks. In *COLING 2026*, 3913–3923.

Watts, D. J.; and Strogatz, S. H. 1998. Collective dynamics of 'small-world'networks. *nature*, 393(6684): 440–442.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q. V.; Zhou, D.; et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35: 24824–24837.

Zaidi, Z.; Ye, M.; Samon, F.; Jama, A.; Gopalakrishnan, B.; Gu, C.; Karunasekera, S.; Evans, J.; and Kashima, Y. 2023. Topics in antivax and provax discourse: yearlong synoptic study of COVID-19 vaccine tweets. *Journal of Medical Internet Research*, 25: e45069.

Zheng, W.; and Tang, X. 2024. Simulating social network with llm agents: An analysis of information propagation and echo chambers. In *International Symposium on Knowledge and Systems Sciences*, 63–77. Springer.

Paper Checklist

1. For most authors...
   (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? We use public, aggregated social media datasets with no new personal data collection; analyses are reported at aggregate/network level with mitigation steps (Section Ethics and Responsible Use).
   (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Claims are limited to network–opinion co-evolution simulation and empirical fit vs. an equation baseline across three datasets (Abstract; Introduction; Experiments).
   (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Section LLM approach and Model Framework justify language-grounded updates/rewiring under platform-like exposure constraints and validate against empirical networks.
   (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? Datasets section notes sampling biases, platform/domain skew, stance-label uncertainty, and time-window effects; we interpret results comparatively rather than absolutely.
   (e) Did you describe the limitations of your work? We discuss lack of user-level ground-truth linking dynamics, prompt/LLM sensitivity, dataset coverage, and governance externalities (Limitations in Ethics and Responsible Use).
   (f) Did you discuss any potential negative societal impacts of your work? We note risks such as misuse for manipulation, over-fitting interventions to specific communities, and model bias propagation, with suggested safeguards.
   (g) Did you discuss any potential misuse of your work? We caution against using the framework to optimize persuasion/astroturfing and recommend access controls and aggregate-only releases.
   (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? We use public data, avoid releasing raw personal content, propose releasing prompt templates and synthetic/aggregate artifacts, and document parameters for reproducibility.
   (i) Have you read the ethics review guidelines and ensured that your paper conforms to them?
2. Additionally, if your study involves hypotheses testing...
   (a) Did you clearly state the assumptions underlying all theoretical results? No formal hypothesis testing with statistical theory beyond empirical comparisons.
   (b) Have you provided justifications for all theoretical results? NA
   (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? NA
   (d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? NA
   (e) Did you address potential biases or limitations in your theoretical framework? NA
   (f) Have you related your theoretical results to the existing literature in social science? NA
   (g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? NA
3. Additionally, if you are including theoretical proofs...
   (a) Did you state the full set of assumptions of all theoretical results? NA
   (b) Did you include complete proofs of all theoretical results? NA
4. Additionally, if you ran machine learning experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? We have added extensive implementation details to the Appendix to facilitate reproducibility.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? We report datasets, active-user selection, p values, candidate-screening strategies, number of steps (2,000), and 5-run averages.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? We average over five runs; intervals are provided where space permits.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? The compute footprint section has been added to the Appendix.
   (e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? Evaluation aligns with claimed mechanisms via modularity, clustering, assortativity, path length, and stance dynamics against empirical networks.
   (f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? We focus on structural fit; misclassification costs depend on downstream application and are out of scope.
5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, **without compromising anonymity**...
   (a) If your work uses existing assets, did you cite the creators? Citations include COVID-19, Ukraine, and PolitiSky24 datasets and related stance resources.
   (b) Did you mention the license of the assets? NA

(c) Did you include any new assets in the supplemental material or as a URL? We include prompt templates; no new raw datasets are released.

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? We use public social media data collected under platforms' terms; no new human subjects data were collected.

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? Public posts may contain PII/offensive content; we report only aggregate statistics and avoid re-distribution of raw personal content.

(f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR (see **?**)? NA

(g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset (see **?**)? NA

6. Additionally, if you used crowdsourcing or conducted research with human subjects, **without compromising anonymity**...

(a) Did you include the full text of instructions given to participants and screenshots? NA

(b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? NA

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? NA

(d) Did you discuss how data is stored, shared, and deidentified? NA

A Prompt Templates

This appendix provides a detailed description of the prompt templates used in our LLM-enhanced simulation. The goal is to make the implementation fully reproducible and to clarify how the prompts operationalize the abstract functions introduced in the main paper.

Templates as Operational Instantiations of f and g

In Section , we define an LLM-enhanced instantiation of the co-evolutionary model, where the influence function and compatibility function are realized via prompted LLM evaluations:

$$f \;=\; \mathrm{LLM}(T_o(\cdot))\,, \qquad g \;=\; \mathrm{LLM}(T_r(\cdot))\,,$$

and the content generation step is:

$$y_i(t) \sim \mathrm{LLM}(T_g(\cdot))\,.$$

Accordingly, the prompt templates in this section should be understood as the *executable realization* of these theoretical mappings. That is, the prompts instantiate how textual histories and contextual exposures are transformed into (i) updated stance judgments (T_o), (ii) rewiring decisions (T_r), and (iii) newly generated posts (T_g).

Design Principles

We adopt a consistent prompting scheme across datasets and models to isolate the effects of model capability rather than prompt engineering. The prompt design follows three principles:

(P1) Standardization across models. All evaluated LLMs receive the same prompt structure, the same input fields, and the same output format. This reduces prompt-induced variability and ensures that observed differences more plausibly reflect model-level inductive biases.

(P2) Minimal but sufficient context. Each prompt uses a bounded window of historical posts for the focal user and a bounded window of contextual posts from neighbors or candidates. This reflects realistic attention constraints and also keeps the simulation cost stable.

(P3) Parseable outputs. Each prompt requests structured outputs (e.g., a discrete stance label or a single selected user ID), so that decisions can be extracted deterministically for downstream evaluation. Free-text rationales are optional and used only for interpretability analysis (e.g., motive taxonomy), not for driving the network update itself.

Input Construction and Notation

For each user i, we maintain a content history C_i that stores the most recent posts of that user. At time step t, user i is exposed to a limited set of posts through a *screen*:

$$S_i(t) \subseteq \{C_j \mid j \in N_i(t)\} \cup (\text{optional recommended content}).$$

The screen is the only information channel used for opinion updating and post generation at that step, modeling finite exposure under platform constraints.

For rewiring, we form: (i) a neighbor set $N_i(t)$ from which one existing edge may be removed, and (ii) a candidate set $\mathcal{C}_i(t)$ consisting of non-neighbors from which one new edge may be added. As described in the main paper, we enforce paired rewiring (one removal plus one addition) so that edge counts remain approximately constant.

Template T_g: New Post Generation

The generation template T_g produces a new post for user i given (a) the user's recent history and (b) the current exposure screen. The post is expected to reflect the user's current stance and conversational context.

Inputs: (1) user profile identifier (optional), (2) recent history C_i, (3) screen posts $S_i(t)$, and (4) topic description.

Output: one short post $y_i(t)$.

Template:

> You are simulating a social media user.
> Topic: {TOPIC}.
> Your recent posts: {HISTORY}.
> Posts you just saw: {SCREEN}.
> Task: Write one new post that you would realistically post now.
> Constraints: Keep it short; do not mention that you are an AI; do not list steps.
> Output: {ONE POST}.

Template T_o: Opinion Update / Stance Inference

The opinion update template T_o produces a stance value for user i after exposure. In our implementation, stances are discretized into three classes (e.g., pro, neutral, anti), or mapped to a scalar in $[-1, 1]$ depending on the dataset annotation scheme.

Inputs: (1) recent history C_i, (2) current screen $S_i(t)$, and (3) stance label definition.

Output: a single stance label (or scalar score) for $O_i(t{+}1)$.

Template:

> Topic: {TOPIC}.
> Stance labels: {LABELS} (with definitions).
> User's recent posts: {HISTORY}.
> Posts the user just saw: {SCREEN}.
> Task: Infer the user's updated stance after reading the screen.
> Output format: Return exactly one label from {LABELS}.

Template T_r: Paired Rewiring Decision

The rewiring template T_r is used only when rewiring is triggered (with probability p). It is executed in two sub-steps:

(R1) Unfollow step. Given user i and a bounded list of current neighbors with their recent posts, select at most one neighbor to unfollow.

(R2) Follow step. Given the same user i and a bounded candidate set of non-neighbors with their recent posts, select exactly one user to follow.

We separate the two decisions to make the update rule explicit and parseable. The resulting paired update implements:

$$E(t{+}1) \;=\; E(t) \;\cup\; \{(i, j^{+})\} \setminus \{(i, j^{-})\}.$$

Inputs: (1) user history C_i, (2) short summaries or recent posts of each neighbor/candidate, and (3) topic/stance context.

Outputs: a single selected user ID for unfollow (or a special token indicating "no unfollow"), and a single selected user ID for follow.

Template (Unfollow):

> You are user {USER}. Topic: {TOPIC}.
> Your recent posts: {HISTORY}.
> Current connections (ID and recent posts):
> {NEIGHBOR LIST}
> Task: Choose at most one connection to unfollow.
> Output format: Return exactly one ID from the list, or "NONE".

Template (Follow):

> You are user {USER}. Topic: {TOPIC}.
> Your recent posts: {HISTORY}.
> Candidate users to follow (ID and recent posts):
> {CANDIDATE LIST}
> Task: Choose exactly one user to follow.
> Output format: Return exactly one ID from the candidate list.

Rationale Collection (Optional)

For interpretability analysis (Section on rewiring motives), we optionally ask the model to provide a short rationale after selecting an ID. Importantly, this rationale is *not* used to determine the network update. It is stored only for downstream tagging and aggregation. This separation ensures that the simulation dynamics are driven by explicit structural actions, while explanations serve as analyzable artifacts.

B Stance Accuracy (changing users)

To avoid inflated accuracy caused by stance inertia (most users keep the same stance), we compute stance accuracy only on users whose stance changes over time. We represent opinions as discrete stance labels (`FAVOR`, `AGAINST`, `NEUTRAL`). A user is considered "changing" if their inferred stance label differs between two consecutive evaluation time points, i.e., $s_i(t{+}1) \neq s_i(t)$, and no additional numerical threshold is applied.

Let $\mathcal{U}_{\text{chg}} = \{i \mid s_i(t{+}1) \neq s_i(t)\}$ denote the set of changing users. We report stance accuracy on $\mathcal{U}_{\text{chg}}$ as

$$\text{Acc}_{\text{chg}} = \frac{1}{|\mathcal{U}_{\text{chg}}|} \sum_{i \in \mathcal{U}_{\text{chg}}} \mathbf{1}[\hat{s}_i(t{+}1) = s_i(t{+}1)]\,.$$

For simulations, $\hat{s}_i(\cdot)$ is obtained by parsing the LLM-generated outputs into stance labels using the same label set; for the ground truth, $s_i(\cdot)$ is taken from the dataset annotations at the aligned time points.

C Operationalizing Rewiring Rationales

This appendix clarifies how the five motivational dimensions (*Tie*, *Affect*, *UGT*, *Moral*, *InfoQuality*) are operationalized from the free-text rationales produced by LLMs. Our goal is to provide a procedure that is *reproducible, auditable, and theory-grounded*, without introducing additional learned classifiers.

Why a Five-Dimensional Taxonomy?

The objective of analyzing rewiring rationales is not to infer hidden psychological states, but to summarize *how* LLM agents justify tie changes in natural language. Across datasets and providers, we observe that explanations consistently cluster around five recurring themes. We therefore adopt a *minimal, theory-grounded cover* that supports aggregation and cross-model comparison while remaining interpretable.

Tie captures relationship-specific considerations (reciprocity, interaction frequency, inactivity, one-sided engagement), consistent with classical accounts of tie maintenance and dissolution in social networks (Granovetter 1973).

Affect summarizes emotion- and tone-related cues (hostility/toxicity versus positive resonance). This dimension is distinct from information quality: content may be accurate yet emotionally aversive, or emotionally appealing yet uninformative.

UGT (Uses and Gratifications) represents instrumental motives such as novelty, complementarity, or relevance to the user's interests. This captures *why a tie is worth forming* beyond ideology.

Moral aggregates value-laden frames related to policy, accountability, rights, or justice. Moralized language is especially salient in polarized discourse and frequently appears

in justifications for following or disengaging (Haidt and Graham 2007).

InfoQuality captures epistemic judgments (redundancy, lack of substance, off-topic content, or misinformation). Separating InfoQuality from UGT allows us to distinguish "high-quality but irrelevant" from "relevant but low-quality" content.

Together, these dimensions cover relational, emotional, instrumental, normative, and epistemic aspects of social filtering, and are compact enough to yield stable provider profiles across datasets.

Keyword-Based Tagging Procedure

LLMs output rewiring rationales as unconstrained text, often containing multiple bullet points that mix reasons for *unfollow* and *follow*. To make these rationales comparable, we apply a transparent keyword-based tagging pipeline with two layers: *SubTag* (fine-grained cues) and *SuperTag* (the five dimensions).

Step 1: Segment rationales into action-specific blocks. Given a raw rationale string, we first split it into short reasoning units (typically bullet points; if bullets are absent, the entire text is treated as one unit). Each unit is assigned an action label (*follow*, *unfollow*, or *unknown*) by detecting explicit action markers (e.g., "follow", "unfollow") in the text. This action-aware parsing reduces ambiguity when a single explanation contains mixed decisions.

Step 2: Match SubTags via regular-expression keywords. Each reasoning unit is then matched against a pre-defined dictionary of lexical patterns. A *SubTag* is triggered if any associated pattern is found. The SubTags are intentionally fine-grained and surface-level, designed to capture recurring phrases that LLMs use when justifying rewiring. Examples include:

(i) *OffTopic* / *LowInfo* / *Redundant* (information-quality concerns)

(ii) *ExtremismTone* or *HopePositivity* (affective tone)

(iii) *NoveltyComplement* or *UserContextAlign* (instrumental value)

(iv) *PolicyInstitution* / *MinorityRights* / *Accountability* (moral framing)

(v) relational cues such as *TieReciprocity* or *TieInactivity* (tie-based reasons).

To improve robustness, we apply a lightweight negation check: patterns appearing under explicit negation (e.g., "not toxic") are not counted as matches. We also apply an *action-aware preference filter*: certain SubTags are more plausible for *follow* (e.g., reciprocity, positivity, novelty), whereas others are more plausible for *unfollow* (e.g., inactivity, spam, misinformation). When an action label is available, SubTags that strongly contradict the action preference are down-weighted by exclusion. This step is conservative: it does not force a single label, but reduces obvious false positives in mixed or noisy rationales.

Step 3: Map SubTags to the five SuperTags. Finally, each SubTag is deterministically mapped to exactly one of the five dimensions using a fixed lookup table. This yields the *SuperTags* used in our main analysis. Because both the keyword dictionary and the SubTag→SuperTag mapping are fixed, the entire rationale analysis is reproducible and does not rely on learned annotators.

Step 4: Aggregate to provider-level profiles. For each provider (model) and dataset, we compute the prevalence of each SuperTag as the mean of binary indicators over all tagged reasoning units. These shares form the basis of the radar plots in Figure 4. To facilitate visualization, we linearly rescale the shares within each dataset to a common range (shown in the figure caption), which does not affect the rank ordering of providers within the same dataset.

Interpretability and Reproducibility

The tagging pipeline is deliberately rule-based for transparency. All steps—segmentation, keyword matching, action-aware filtering, and deterministic mapping— can be replicated using only the recorded LLM outputs and the published rule tables. This design allows readers to (i) audit which phrases lead to which tags, (ii) modify the dictionary if additional rationale patterns emerge, and (iii) reproduce provider-level motive profiles without retraining any components.

While keyword tagging cannot capture every nuance of natural language, it provides an auditable abstraction that links free-text rationales to theoretically grounded motivational dimensions. The stability of provider profiles across datasets supports the interpretation that the extracted dimensions reflect consistent model-level inductive biases rather than dataset-specific artifacts.

D Polarization computation (implementation).

For each strategy and each time bin, we extract stance labels from the LLM output and count n_F, n_A, n_N occurrences of `FAVOR`, `AGAINST`, and `NEUTRAL`, respectively. Let $n = n_F + n_A + n_N$. We compute $p_F = n_F/n$, $p_A = n_A/n$, and $p_N = n_N/n$. We define stance polarization as

$$P \;=\; |p_F - p_A|.$$

For cross-strategy comparison, we align time bins with the Random strategy and compute

$$\Delta P^{(s)} \;=\; P^{(s)} - P^{(\mathrm{rand})}.$$

We summarize $\Delta P^{(s)}$ over bins by its sample mean $\overline{\Delta P}^{(s)}$ and estimate 95% confidence intervals using a normal approximation:

$$\overline{\Delta P}^{(s)} \pm 1.96 \cdot \frac{\sigma_{\Delta P^{(s)}}}{\sqrt{n}},$$

where $\sigma_{\Delta P^{(s)}}$ is the sample standard deviation over aligned bins and n is the number of bins.

E Implementation Details of News Shock Experiments

We provide additional implementation details for the news shock case study to facilitate reproducibility. All experiments start from identical initial network structures, agent

opinion states, and random seeds. The only difference across scenarios is the presence and semantic polarity of the injected news content.

Shock injection mechanism. The news shock is implemented as a platform-level news post that appears in agents' information screens during a fixed time window. The injected news is treated as a standard textual item in the feed and is processed by agents through the same language-based perception and decision mechanisms as ordinary posts. No additional rules or parameters are introduced specifically for the shock.

Timing and exposure. In all shock experiments, the news intervention starts at time step $t = 500$ and lasts for a fixed window of 1000 simulation steps. Within this window, agents are exposed to the news with a fixed probability at each update. Outside the shock window, no external news is injected. All other simulation parameters remain unchanged.

News content. Two types of exogenous news are considered. Both news items are *synthetically constructed* and do not correspond to specific real-world events. The positive news emphasizes vaccine effectiveness, reduced hospitalization and mortality, and endorsements from authoritative institutions, while the negative news highlights risk-oriented narratives, including alleged severe adverse reactions, reported deaths, and claims of institutional misconduct. To enhance narrative plausibility and linguistic realism, both news texts include concrete numerical details and references to institutional actors, reflecting common stylistic patterns observed in real-world public health reporting and online discourse. The full text of each news item is fixed throughout the shock window and identical for all agents. The news texts used in the experiments are reported verbatim below.

Positive news text.

> GLOBAL ALERT: Historic vaccine success — an unprecedented Phase III trial spanning 12 countries (N≈68,000) reports 96.4% efficacy against symptomatic infection and a 99.1% reduction in severe and critical cases. Within one month of emergency rollout, COVID-related hospitalisations fell by 72% nationwide, ICU occupancy dropped by 61%, and excess mortality returned to pre-pandemic levels for the first time in three years. Independent safety boards report serious adverse events at fewer than 1 per 250,000 doses, with no confirmed causal links. The WHO and national regulators jointly declare the vaccine "a decisive turning point" and recommend immediate, universal deployment, calling vaccine hesitancy "the primary remaining threat to public health".

Negative news text.

> EMERGENCY: Vaccine crisis erupts — viral reports claim an explosive surge in life-threatening reactions following recent vaccination, alleging over 3,800 emergency hospitalisations and at least 94 sudden deaths across multiple regions in just ten days. Graphic personal accounts describe cardiac collapse, seizures, paralysis, and previously healthy young adults dying within hours of injection. Influential posts accuse governments and pharmaceutical companies of deliberate data suppression, calling the rollout a "catastrophic human experiment". Several hospitals are reported to have suspended vaccination programs amid staff walkouts and protests, while panic spreads online with mass appointment cancellations and calls for an immediate nationwide halt. Officials deny proven causality and urge calm, but public trust is rapidly deteriorating as investigations lag behind events.

Interaction with network rewiring. The news shock does not directly alter the rewiring process. Agents decide whether to rewire and which connections to modify using the same LLM-based decision mechanism as in the no-shock setting. Any structural changes arise indirectly from agents' responses to the altered informational environment.

Ethics and Responsible Use

Experiments use public, aggregated datasets, applying the framework to finer-grained settings would require stronger anonymization and minimal disclosure (e.g., releasing only aggregate statistics and essential synthetic artifacts) with explicit data-use boundaries. Because LLMs carry inductive biases, we complement structural results with a motive taxonomy to increase decision transparency; future work should add group-wise slices, fairness probes, and counterfactual evaluations, and treat stance metrics as *relative* across models. Long-standing transparency–privacy tensions persist: language-driven simulation offers a pragmatic middle ground by releasing *synthetic* trajectories that match statistical and structural properties without exposing personal data. Synthetic corpora cannot fully substitute real audits—especially for cultural nuance or multilingual dynamics—yet they improve reproducibility at lower privacy risk; extensions include multilingual contexts, localized prompts, and cross-platform comparisons.

Cost and Compute Footprint

Our framework is practical at research scale. Using official API endpoints, a full $2{,}000$-step run is affordable under current pricing with moderate context windows and concise rationales. Each step averages a few thousand input tokens (screen, history, instructions) and roughly one thousand output tokens (decision, rationale, post). Costs scale approximately linearly with the number of on-screen agents, the verbosity of rationales, and the number of parallel models. On a A100 GPU, a quantized 30–70B model completes $2{,}000$ steps in $\sim$4–5 hours with batching and KV caching. These figures are not hardware- or vendor-agnostic; they serve as an empirical yardstick for reproductions and for budgeting larger ablations. To stabilize costs, we cap screen length, compress history via extractive summaries, and cache repeated judgments when underlying texts are unchanged.