Characterizing Opinion Evolution of Networked LLMs

Anonymous ACL submission

Abstract

Large language models (LLMs) increasingly interact with one another in multi-agent systems, from simulations of human discourse to influence operations and fully LLM-driven social platforms. These interactions give rise to new regimes of opinion propagation that are not yet well understood. We investigate whether classical opinion dynamics models, which have long been used to explain how interactions shape collective beliefs in human societies, can capture the behavior of LLM networks. We find that, while naive averaging-style models fail to track LLMs' opinion dynamics, simple modifications yield substantial gains in modeling fidelity. In particular, bias, an innate opinion toward which agents regress, emerges as a significant driver of LLM opinion dynamics, with its inclusion reducing cumulative estimated mean opinion error by up to 88%. We additionally find that these conclusions generalize across model families, discussion topics, and networks.

1 Introduction

Networked systems of large language models (LLMs) are increasingly deployed across social media and information environments. They have been studied as tools for simulating opinion formation in purely human settings (Chuang et al., 2024; Törnberg et al., 2023; Wang et al., 2025), are emerging as instruments of influence operations in mixed human–LLM networks (Goldstein et al., 2023), and most recently, have led to fully LLM-driven social platforms (Lin et al., 2026). These developments raise fundamental questions about how opinions form, evolve, and propagate in networks of interacting LLMs.

Classical opinion dynamics models offer a promising lens for reasoning about the emergent behavior of groups of LLMs. In human social systems, such models have proven powerful for studying how effects such as stubbornness, homophily, and bias drive belief evolution. Extending them to interacting LLMs may therefore provide a principled framework for understanding emergent behavior in LLM networks. When LLMs are used to simulate humans, for instance, opinion dynamics models can certify when LLM populations behave as faithful surrogates for human ones. For influence operations, they can identify the leverage points through which stubborn LLM agents disproportionately shape collective beliefs. For LLM-driven social platforms, they can guide design choices that mitigate echo chambers driven by homophily. Realizing these uses, however, requires knowing which mechanisms, such as stubbornness, homophily, and bias, actually drive LLM networks.

We investigate which mechanisms classical opinion dynamics models must capture to track LLMs' belief evolution. Recent work (Abedini et al., 2026; Yazici et al., 2026; Ju et al., 2024) provides initial insight into the alignment of networked LLMs' opinions with models such as DeGroot (DeGroot, 1974), Friedkin-Johnsen (Friedkin and Johnsen, 1990), or Hegselmann-Krause (Hegselmann and Krause, 2002) dynamics. Each of these models hard-codes a specific combination of stubbornness, homophily, and bias. In contrast, we quantify how the relative effects of stubbornness, homophily, or bias govern the evolution of LLMs' beliefs.

We study the following problem.

- **RQ1**: Which components of classical opinion dynamics models, such as stubbornness, homophily, and bias, are most critical for explaining LLMs' belief evolution?

We also investigate the extent to which conclusions about the importance of these components generalize across different settings.

- **RQ2**: To what extent do the derived opinion dynamics parameters differ between networks of different LLMs and discussion topics?

- **RQ3**: Are the recovered opinion dynamics a property of the LLM itself, or of the particular network on which it communicates?

By answering these questions, we lay the groundwork for the design, control, and validation of opinion dynamics across networks that include LLMs.

To address these questions, we simulate discussions among groups of LLMs, extract their opinions, and fit opinion dynamics models to the resulting trajectories. We first simulate discussions in which LLM agents exchange messages over a communication graph on contested topics. A stance identification model then converts each message into a numerical opinion, yielding opinion trajectories over time. We fit classical opinion dynamics models to these trajectories and quantify how well each model, and each of its components, reproduces the observed behavior, thus addressing RQ1. We then repeat this procedure across the Qwen, Llama, and Gemma families and across topics, testing whether the same components dominate, addressing RQ2. Finally, while RQ1 and RQ2 assume each agent is influenced uniformly by its neighbors, we fit per-edge influence weights and test whether this added flexibility substantially improves prediction accuracy, addressing RQ3.

Related work

Classical opinion dynamics Classical opinion dynamics provides a rich set of models for information flow in social networks. In the simplest opinion dynamics model, i.e., the DeGroot model (DeGroot, 1974), agents update their opinions by averaging amongst their neighbors. The Friedkin-Johnsen (FJ) model then reflects stubbornness by modeling an attraction to initial opinions (Friedkin and Johnsen, 1990). Meanwhile, in bounded-confidence models, agents only exchange opinions with their neighbors if their opinions are sufficiently close (Deffuant et al., 2000; Hegselmann and Krause, 2002). We aim to explore the extent to which various components of these opinion dynamics models reflect the collective behavior of networks of large language models (LLMs).

Generative agent-based modeling Numerous authors propose the use of LLMs for simulation of human opinion dynamics (Chuang et al., 2024; Törnberg et al., 2023; Composta et al., 2025; Wang et al., 2025; Liu et al., 2025; Donkers and Ziegler, 2025; Yao et al., 2025; Mou et al., 2024). However, rather than proposing LLMs replace classical opinion dynamics models, we investigate whether we can model the evolution of collective beliefs among networked LLMs with classical opinion dynamics. Some works compare the performance of LLM-based opinion dynamics models to classical models (Mou et al., 2024; Yao et al., 2025; Gu et al., 2025; Liu et al., 2025; Composta et al., 2025; Wang et al., 2025; Gao et al., 2023). These comparisons, in turn, yield insights into the differences between opinion dynamics models and LLM-based simulations, for example, demonstrating that LLMs' opinion dynamics exhibit sharper changes than FJ models (Composta et al., 2025) or that FJ and bounded confidence models are less capable of generating polarization compared to LLM-based simulations (Wang et al., 2025). However, these works do not characterize which components of opinion dynamics models we require to explain LLMs' belief evolution.

Analytical models for opinion dynamics of networked LLMs A handful of recent works provide initial insights into how well classical opinion dynamics models fit the dynamics of networked LLMs (Yazici et al., 2026; He et al., 2026; Abedini et al., 2026; Ju et al., 2024; Khlytchieva and Kozitsina, 2026). In particular, (Yazici et al., 2026) demonstrates that interacting LLM agents fail to converge to the consensus predicted by the DeGroot model. Meanwhile, some prior works find that DeGroot or FJ dynamics can model the opinion dynamics of small groups of LLMs (He et al., 2026; Abedini et al., 2026). Finally, (Ju et al., 2024) shows that the single-step dynamics of networked LLMs are well-modeled by Hegselmann-Krause dynamics. While these studies point to correspondences between LLMs' dynamics and opinion dynamics, they fail to precisely identify which of the many possible components of opinion dynamics models accurately capture the opinions of networked LLMs. While more comprehensive explorations exist for two-agent interactions (Brockers et al., 2025), these analyses do not extend to networked interactions of many agents.

A few recent works propose opinion dynamics models for systems including LLMs, for example, modeling LLM-generated content's interaction with human opinions (Marchi et al., 2025), or opinion dynamics among humans who depend on LLMs (Li et al., 2026). Rather than proposing new opinion dynamics models, we explore how well existing models reflect networked LLMs' dynamics.

2 Problem Statement

We consider n LLM agents that communicate over a directed graph $G = (V, E)$. Each vertex $i \in V$ corresponds to an LLM agent, and an edge $(i, j) \in E$ indicates that agent j's messages are visible to agent i. We write $A \in \mathbb{R}_+^{n \times n}$ for the adjacency matrix of G, where the component A_{ij} encodes the expected rate at which agent i sees agent j's posts. We use $\bar{A}$ to denote the row-normalized version of A. Each agent i holds an opinion $z_i(t) \in [-1, 1]$, which evolves over discrete time $t \in \{0, 1, \ldots, H\}$ as the agents exchange messages, with $z(t) = [z_1(t) \ \ldots \ z_n(t)]^\top$.

We aim to characterize how the opinion vector $z(t)$ evolves in terms of the intrinsic behavioral traits of the constituent LLM agents, for example, their stubbornness, innate biases, or tendency toward homophilic interaction. To make this characterization tractable, we posit that $z(t)$ obeys a parametric model of the form

$$z(t+1) = f_\theta(z(t), z(0), A), \tag{1}$$

where θ is a low-dimensional parameter vector encoding the agents' behavioral traits. The class (1) subsumes classical opinion dynamics models such as DeGroot, Friedkin–Johnsen, and homophily-based dynamics, as well as extensions incorporating innate bias, which we instantiate in Section 3.

Within this class, we further distinguish *graph-specific* models, in which the parameters θ are fit independently for each graph, from *graph-agnostic* models, in which θ encodes a small set of scalar coefficients that apply across topologies. The graph-agnostic form encodes the assumption that agent j's contribution to agent i's opinion update is a function of the rate at which agent j messages agent i. While restrictive, this assumption yields a behavioral profile θ that is portable across networks.

The recovered parameters θ act as a behavioral profile of the underlying LLM. By comparing these profiles across LLMs, topics, and graph structures, we aim to isolate which mechanisms are intrinsic to a model, which are topic-dependent, and which depend on the network in which the model is embedded, thus answering RQ1, RQ2, and RQ3.

3 Methods

3.1 Modeling Opinion Dynamics

We now characterize the parametric models (1) that we consider in this study, each corresponding to a different hypothesis about the intrinsic characteristics of the LLMs that drive the evolution of their opinions. In all of the subsequent sections we make the notational assumption that multipliers of all additive terms—denoted as $\lambda_\star$—sum to one. This constraint is common across opinion dynamics models, and is needed to ensure stability and non-trivial opinion behavior.

DeGroot Model The simplest realization of the parametric model (1) is the DeGroot model, in which each agent's opinion is a convex combination of its neighbors' opinions, i.e.,

$$z(t+1) = Wz(t). \tag{2}$$

Here, W is a row-stochastic matrix that plays the role of the parameter θ, and it is such that $W_{ij} = 0$ if $(i, j) \notin E$ and $j \neq i$. Under some fairly mild assumptions (Jackson and Golub, 2010), the DeGroot model is known to converge to a consensus in the steady state, i.e., to a situation where $z_1(t) = z_2(t) = \ldots = z_n(t)$ as $t \to \infty$.

Friedkin-Johnsen Model In many situations, it is difficult for agents to reach a consensus as agents are stubborn and unwilling to abandon their initial opinions. The Friedkin-Johnsen (FJ) model captures this phenomenon by extending the DeGroot model (2) with a stubbornness term, leading to

$$z(t+1) = \lambda_{\text{avg}} W z(t) + \lambda_0 z(0). \tag{3}$$

Here, $\lambda_{\text{avg}} \in [0,\ 1]$ is the social averaging weight, $\lambda_0 \in [0,\ 1]$ is the *initial opinion weight* that specifies the resistance of each agent to deviate from their initial opinion, and W is as in the DeGroot model. As previously stated, the multiplicative weights $\lambda_\star$ sum to 1 by construction, i.e., $\lambda_{\text{avg}} + \lambda_0 = 1$. The tuneable parameter set θ comprises λ_0 and W. Note that due to the constraint on $\lambda_\star$, the averaging weight λ_{avg} is not a free parameter. Unless $\lambda_0 = 0$, the FJ model usually does not lead to a consensus.

Uniform Bias LLM agents, in particular, often have a tendency to favor specific points of view on certain topics, owing to biases introduced during training (Santurkar et al., 2023). These biases are independent of the initial opinions of the LLM agents and are thus not captured by the FJ model. To capture them, one can add a uniform bias term $b \in [-1,\ 1]$ to (3), leading to the opinion dynamics

$$z(t+1) = \lambda_{\text{avg}} W z(t) + \lambda_0 z(0) + \lambda_b b, \tag{4}$$

where $\lambda_b \in [0,\ 1]$ is the *bias weight* indicating how biased agents are toward the opinion b, and as before, we have the constraint $\lambda_{\text{avg}} + \lambda_0 + \lambda_b = 1$. In this case, θ comprises the parameters $\lambda_0,\ \lambda_b,\ W$.

Homophily While the addition of biases in the DeGroot model (2) captures a broader range of opinion dynamics, it misses a fundamental aspect of how social agents interact in that they weigh the opinions of like-minded agents more heavily. Homophily captures this kind of weighting by substituting the constant W matrix with a distance-based one (Newman, 2018), namely

$$H_{ij}(W, z, \gamma) = \frac{W_{ij} \exp(-\gamma |z_i - z_j|)}{\sum_j W_{ij} \exp(-\gamma |z_i - z_j|)}, \quad (5)$$

where γ controls the homophily strength. Higher γ values force agents to be more selectively influenced by other agents of similar opinions. This new matrix leads to the opinion dynamics

$$z(t+1) = \lambda_{\text{avg}} H(W, z(t), \gamma) z(t) + \lambda_0 z(0) + \lambda_b b, \quad (6)$$

where $\lambda_\star$ sums to one as always, and the learnable parameters are $\gamma, \lambda_\star, b$, and W.

Adjacency-Based Weighting Finally, we discuss the role of the weighting matrix W across all the models presented above. When the set of parameters θ in (1) includes the weighting matrix W, the fitted opinion dynamics model entangles the LLM agents' behavioral characteristics with the specific graph structure considered. To study whether the behavioral characteristics of LLMs transfer across different graph topologies, one can consider interaction matrices of the form

$$\mu_{\text{self}} I + \mu_{\text{neigh}} \bar{A},$$

where $\bar{A}$ is the row-normalized adjacency matrix of the graph G, and $\mu_{\text{self}} + \mu_{\text{neigh}} = 1$, $\mu_{\text{self}}, \mu_{\text{neigh}} \in [0, 1]$ are tuneable parameters. In the case of homophily, the interaction term is the matrix

$$\mu_{\text{self}} I + \mu_{\text{neigh}} H(\bar{A}, z, \gamma).$$

When inserted into the previously discussed opinion dynamics models, the parameters μ_{self} and μ_{neigh} get multiplied with λ_{avg} to denote weights corresponding to self update, and neighbor averaging, which we correspondingly denote as

$$\lambda_{\text{self}} = \lambda_{\text{avg}} \mu_{\text{self}}\ , \text{ and } \ \lambda_{\text{neigh}} = \lambda_{\text{avg}} \mu_{\text{neigh}}. \quad (7)$$

For the DeGroot model, with no $\lambda_\star$ parameters, we set $\lambda_{\text{self}} = \mu_{\text{self}}$, and $\lambda_{\text{neigh}} = \mu_{\text{neigh}}$.

3.2 Simulating LLMs' Opinion dynamics

3.2.1 Discussion protocols

To generate opinion dynamics data for groups of LLMs, we simulate conversations amongst LLM-agents on synthetically generated graphs.

Agents post in sequence with random inter-post times and ordering. For each post, we select an agent uniformly randomly to generate the next post, with that post being added to a pool of posts for that agent. We then assign random times to each post according to a Poisson process with rate ρ.

To generate the posts, we prompt the agents to construct arguments in response to other agents' posts. Fix an agent $i \in V$ and an integer parameter k. We first select up to k posts uniformly at random from the prior posts of the agents that influence i. We then prompt the agent to justify their position based on 1) their initial opinion, and 2) the k sampled posts. Finally, we prompt the agent to generate a post based on their internal justification.

We encode the set of agents influencing agent i in the interaction graph G, where edge (i, j) indicates that agent j influences agent i. As we sample the k posts that agent i sees uniformly from agent i's neighbors, and A_{ij} encodes the rate at which i sees j's posts, the entry $\bar{A}_{ij}$ of the row-normalized adjacency matrix is equal to $1/N(i)$ for each neighbor j of i, where $N(i)$ is the number of agents influencing agent i. We next detail the generation of the interaction graph G.

3.2.2 Network generation

We evaluate the opinion dynamics of LLMs on a mixed dataset composed of three widely studied random graph models that together cover the topological phenomena commonly attributed to online social networks, such as scale-free graph structure with geometric degree distributions, small-world graphs with diameters that scale logarithmically with the number of agents, and community structures that arise due to homophilic interactions.

Erdős–Rényi models (Bollobás, 2001) The Erdős–Rényi model is a widely used baseline for random graph analysis. Each pair of nodes is connected independently with a uniform probability p, producing a maximally unstructured network with homogeneous degree distribution and no latent community organization.

Chung-Lu models (Chung and Lu, 2002) The Chung-Lu model builds upon the Erdős–Rényi model by assigning each node an expected degree, with edge probabilities set proportional to the product of endpoint degree weights. When degree weights are drawn from a geometric distribution, the resulting degree sequence follows a power law, matching the empirical heavy-tailed distributions that are common in social networks. Thus, Chung-Lu models reproduce realistic degree heterogeneity, i.e., hubs vs. low-degree periphery nodes.

Stochastic block models (Holland et al., 1983) Stochastic block models (SBMs) partition nodes into communities and define edge probabilities as a function of community membership, with within-block connectivity being typically denser than cross-block connectivity. Thus, SBMs are the standard generative model for simulating networks with community structure, capturing the clustering patterns characteristic of real social networks such as echo chambers or professional affiliation groups.

For each run used to answer RQ1/RQ2, we first sample a graph from a randomly selected type. For RQ3, we generate a single Chung-Lu graph on which we simulate all results.

3.2.3 Opinion measurement

We convert the simulation output, a sequence of natural-language posts indexed in continuous time, into discrete-time-indexed numerical opinions.

We compute opinions by assessing similarity in representation space between the agents' posts and reference prompts. Fix a post t, and two texts a_{pro} and a_{anti} that encode for and against opinions on a given topic. We first pass t, a_{pro}, and a_{anti} through a text embedding model (OpenAI) to get embedding vectors $\mathbf{t}$, $\mathbf{a}_{\text{pro}}$, and $\mathbf{a}_{\text{anti}}$, respectively. We then compute an opinion of a post as

$$z = \text{sim}(\mathbf{t}, \mathbf{a}_{\text{pro}}) - \text{sim}(\mathbf{t}, \mathbf{a}_{\text{anti}}), \tag{8}$$

where $\text{sim}(\mathbf{x}, \mathbf{y}) = {}^{\mathbf{x}\cdot\mathbf{y}}/_{\|\mathbf{x}\|\|\mathbf{y}\|}$.

We define an agent's opinion at each time step as the mean post opinion computed over a sliding window of the agent's five most recent posts. If an agent has fewer than five posts, we compute the mean over all of them. Averaging over the window smooths short-term fluctuations and measurement errors, capturing the agent's persistent behavior. When an agent has yet to publish a message, we score the agent's initialization prompt, defining its assigned opinion on the topic.

We sample the agent opinions at discrete timestamps to extract the ground-truth discrete-time evolution of agent opinions. To account for startup latency during network initialization, we treat the timestamp of the first post as the start time of the simulation. We assign subsequent posts a simulated timestamp drawn from the Poisson-process wait times and extract the integer time intervals from the assigned timestamps as sampling points.

3.3 Opinion initialization

We initialize each agent with a discrete opinion from the set $\{-1, -0.5, 0, 0.5, 1\}$ ranging from complete disagreement to complete agreement on the topic, where 0 denotes a neutral stance. We map each value to a natural-language descriptor—e.g.,"opposed", "skeptical", "agree"—and embed it in the agent's initialization prompt, which states its stance on the topic.

To initialize the network, we randomly assign agents to the five opinion categories according to predefined probability distributions. Each distribution corresponds to a set of initial conditions, including uniform, polarized, consensus, and skewed configurations. For RQ1 and RQ2 we allocate a train set for model fitting, with the remaining runs being allocated as a test set that we use in answering RQ1. The training set consists of the configurations uniform, bimodal polarized, moderate centered, skew pro, and skew anti, while the test set consists of consensus pro, consensus anti, centered with extremes, and hollow centered. In the experiments for RQ3, we generate data using all of the initial condition configurations.

3.4 Model Fitting

We fit the opinion dynamics model by minimizing the following least squares objective

$$\sum_{t,r} ||z^{(r)}(t+1) - f_\theta(z^{(r)}(t), z^{(r)}(0), \bar{A})||_2^2, \tag{9}$$

where for each model, the opinion dynamics model f_θ admits the dynamics in the corresponding model definitions in Section 3.1.

3.4.1 Fully parametrized models

In fully parametrized models, the influence matrix W is a $n \times n$ row-stochastic parameter fit independently for each graph instance with entries constrained to respect the graph G. For models without a homophily term, i.e., DeGroot, Friedkin-Johnsen, and Friedkin-Johnsen with bias, the objective is convex in the model parameters. For fixed

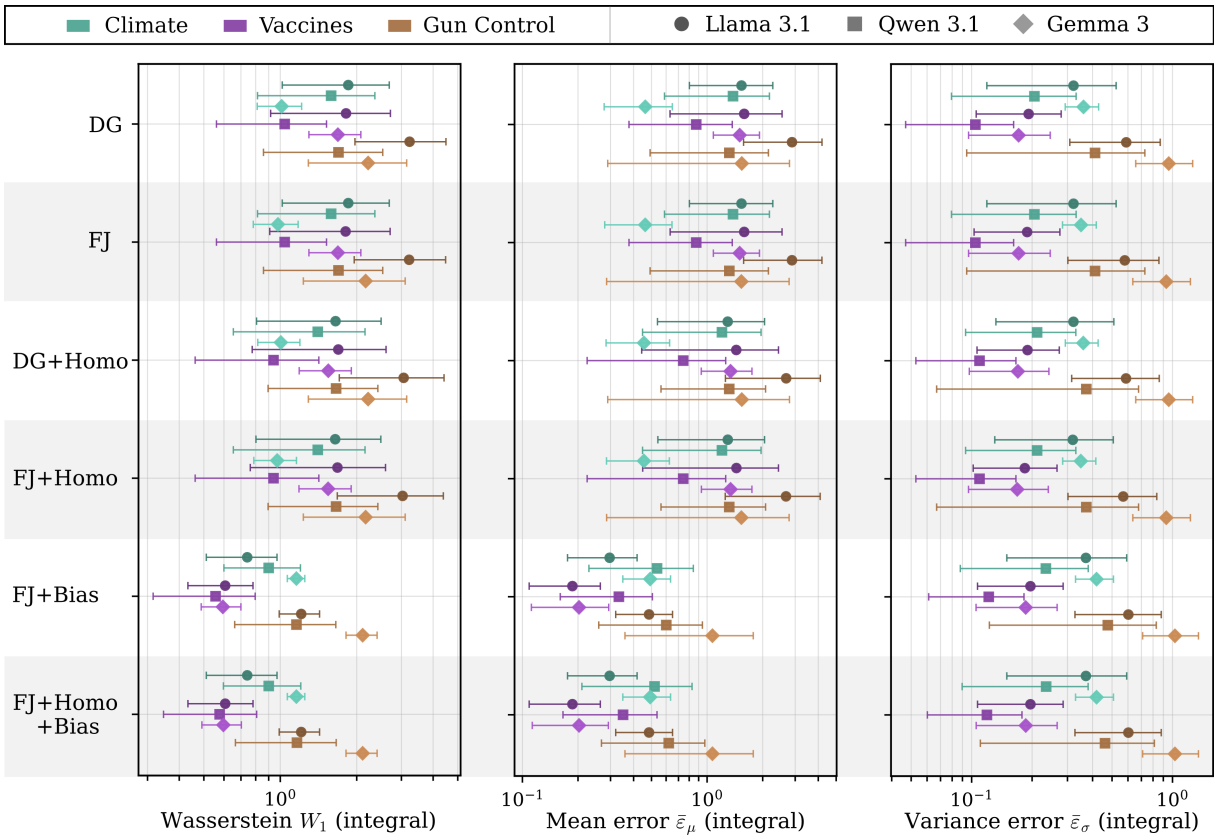

Figure 1. Out-of-sample prediction error across opinion dynamics models, LLMs, and discussion topics, reported on three metrics. Each point shows the mean error across held-out runs. Whiskers show one standard deviation. We abbreviate the models as follows: DG (DeGroot), FJ (Friedkin–Johnsen), Homo (Homophily weighting), Bias (Opinion bias term). Models that include a uniform bias term consistently achieve the lower errors across mean and Wasserstein-1 distance metrics.

λ_0 and λ_b, optimizing over W is a constrained quadratic program (QP), which is solvable using a standard convex solver. For the Friedkin-Johnsen-with-bias model, we apply the reparameterization $\tilde{W} = \lambda_{\text{avg}} W$ and $\tilde{b} = \lambda_b b$, under which the joint optimization over $\tilde{W}$, λ_0, and $\tilde{b}$ is a convex QP solvable to global optimality given sufficient data (see Appendix A.4.2).

For the models containing homophily, the influence matrix $H(W, z, \gamma)$ depends nonlinearly on the opinion vector $z(t)$ through the exponential distance kernel, making the objective non-convex in (W, γ) jointly. We therefore optimize these models using a nonlinear SLSQP solver in Scipy with multiple random restarts (Virtanen et al., 2020).

3.4.2 Adjacency-based models

In adjacency-based models, we constrain the interaction matrix to be of the form $\mu_{\text{self}} I + \mu_{\text{neigh}} \bar{A}$ or $\mu_{\text{self}} I + \mu_{\text{neigh}} H(\bar{A}, z, \gamma)$, reducing the parameter space to a small set of scalar coefficients shared across graph instances. For models without homophily, fitting these coefficients for a fixed graph reduces to a convex QP in the λ's, which we solve.

For models that include homophily, the homophily strength parameter γ enters the objective nonlinearly and cannot be recovered by convex methods. We instead perform a one-dimensional sweep over γ where for each candidate value, we solve the resulting convex QP for the $\lambda_\star$'s and select the γ minimizing the objective (9).

Given their parameter efficiency and generalizability, we investigate RQ1 and RQ2 with adjacency-based models, after which we compare them against fully-parameterized variants in RQ3.

4 Experimental Results

We evaluate our framework on three open-weight LLM families, LLAMA3.1, QWEN3, and GEMMA3, across three contested discussion topics: climate change, vaccines, and gun control. We choose these topics because they provoke sustained disagreement and elicit clear stances. For each LLM and topic, we simulate populations of $n = 30$ agents exchanging posts over the random graphs we describe in Section 3.2.2, fit the opinion dynamics models of Section 3 to the resulting trajectories,

Table 1. Fitted parameters of the *FJ+Homo+Bias* model across LLMs and topics. λ_{neigh} and λ_{self} are the neighbor influence and self-weight terms; λ_0 is the initial opinion weight; b and λ_b are the opinion bias and its weight; γ controls homophily strength.

LLM	Topic	λ_{neigh}	λ_{self}	λ_0	λ_b	b	γ	$\lambda_{\text{neigh}}/(\lambda_b b)$
GEMMA3	Climate	0.0290	0.6214	0.0055	0.3441	0.3483	0	0.2420
	Gun Control	0.1404	0.6901	0.0027	0.1668	0.2277	0	3.6949
	Vaccines	0.0116	0.6076	0.0160	0.3648	0.5841	1.2159	0.0544
LLAMA3.1	Climate	0.0002	0.6560	0.0195	0.3243	0.6699	0	0.0011
	Gun Control	0	0.6689	0.0381	0.2930	0.8531	0	0
	Vaccines	0.0192	0.6900	0.0363	0.2546	0.5972	0.0350	0.1262
QWEN3	Climate	0.0821	0.7019	0.0036	0.2125	0.5370	1.0886	0.7196
	Gun Control	0.1175	0.6592	0	0.2233	0.4337	0.9710	1.2130
	Vaccines	0.2001	0.5344	0	0.2655	0.4681	23.160	1.6100

and measure how well each model predicts opinions on held-out runs. We organize our findings around the three research questions of Section 1.

4.1 RQ1: Which components drive LLM opinion dynamics?

Figure 1 reports out-of-sample error for all six models across every LLM and topic. The bias-free models, namely DeGroot, Friedkin-Johnsen, and their homophily variants, cluster together at high error, while the two models that include a uniform bias term separate from this cluster and attain the lowest mean and Wasserstein-1 errors in nearly every setting, while leaving variance error relatively unchanged. Adding bias reduces integral mean error by up to 88% and integral Wasserstein-1 distance by up to 67%, both attained on LLAMA3.1.

Further comparisons reinforce the innate model bias as the primary model component responsible for the majority of error reduction. First, Friedkin-Johnsen tracks DeGroot almost exactly in every error metric, which shows that anchoring agents to their own initial opinions does not improve fit. Second, adding homophily weighting to any base model changes its error by at most a few percent, and once bias is present homophily produces no measurable change. The mechanism that matters is therefore neither stubbornness toward one's initial opinion nor preferential weighting of like-minded neighbors, but regression toward a single opinion shared across all agents.

The size of the bias effect depends on the LLM. LLAMA3.1 shows the largest and most uniform gain, with integral mean error falling between 81% and 88% across all three topics. QWEN3 gains the least, with reductions between 54% and 62%. GEMMA3 sits between the two and is the only model for which bias fails to help in one setting, where climate error rises slightly because the bias-free fit is already accurate.

The bias effect also depends on the topic. Across all three LLMs, the vaccines topic benefits most from the bias term, with integral mean error reductions of 62% to 88% and Wasserstein-1 reductions of 47% to 67%, exceeding the corresponding reductions for climate and gun control in every LLM. A plausible explanation is that alignment training drives LLMs toward a consistent stance on health misinformation regardless of assigned initial opinion, an effect that the bias term captures directly.

4.2 RQ2: Parameter Comparison

Having established that *FJ* dynamics with added *Opinion Bias* captures LLM opinion dynamics, a natural question is what quantitative weight each opinion dynamics component has on the opinion evolution. This section aims to quantify these weights by comparing fitted model parameters directly, and comparing how the parametrizations differ across LLM/Topic combinations.

Table 1 shows the parameters of the full *FJ+Homo+Bias* model fit separately to each LLM and topic. The self-weight λ_{self} is the largest parameter in every case, between 0.53 and 0.70, so each agent's next opinion is governed primarily by its current one. The scale of the self weights however, are largely irrelevant to the long term behavior of the LLMs, as λ_{self} only determines the convergence speed in comparison to the sampling intervals. Thus, the scale of the self weights are purely an artifact of the chosen time scale.

The initial-opinion weight λ_0 never exceeds 0.04, confirming across every LLM and topic that anchoring to starting opinion plays very little role in driving opinion evolution. The bias weight λ_b is the second-largest term in most settings, between 0.17 and 0.36, so regression toward an innate opinion is a consistent feature of LLM dynamics rather

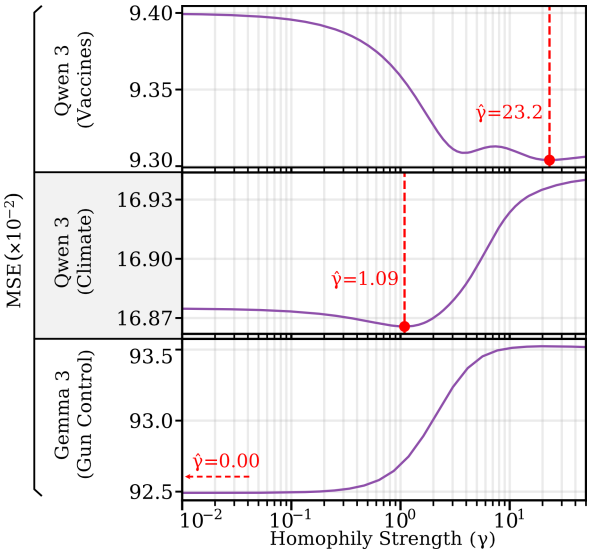

Figure 2. Fit error as a function of homophily strength γ for three representative LLM–topic pairs. The dashed red line marks the optimal γ minimizing the error.

than an artifact of a single model or topic.

The ratio $\lambda_{\text{neigh}}/(\lambda_b b)$ between neighbor opinion effects and self-bias effects provide an overall summary of how much each LLM family preferentially allocate their opinion updates to social interaction effects. This ratio shows substantial variance between LLM/topic combinations. LLAMA3.1 holds it near zero on every topic, from 0.001 on climate to 0.13 on vaccines, so its bias alone sets its opinions and its neighbors barely register. QWEN3 runs from 0.72 to 1.61, so its neighbors match or outweigh its bias on all three topics. GEMMA3 jumps with the topic, from 0.05 on vaccines to 3.69 on gun control, the largest value in the table. How social an LLM acts thus depends on the topic as much as on the model, and only LLAMA3.1 stays at one extreme across all three, ignoring social interaction effects almost completely in its long-term opinion evolution. This behavior may be attributable to LLAMA3.1 being an older model, as recent improvements in post-training may teach models to weigh the conversation in front of them over a stance fixed during training, though three models cannot establish the cause.

Homophily barely affects the fit. Figure 2 sweeps γ over its full range, and the error varies nonlinearly by at most a percent. Accordingly, the presence of homophily in the majority of our experiments is low. QWEN3 alone shows substantial homophily presence, a pattern that is also common in human social networks.

4.3 RQ3: Full Parametrization vs. Adjacency weighting

On a single fixed graph, we compare a fully parameterized per-edge weight matrix W against the adjacency-based form, which derives influence from the communication structure using only a handful of scalar coefficients. Table 2 compares the two parameterizations on training error and on two held-out metrics. While the fully parameterized W lowers the training error, both test metrics rise, showing that the full parameterization of W fails to provide a significant modeling benefit. The homophily variant reproduces the FJ+Bias numbers because the fitted homophily strength collapses to zero in the local optima we find. By absorbing the graph into a known input and fitting only scalar coefficients, the adjacency-based form recovers a behavioral profile of the LLM that is not tied to the specific network it was fitted on.

Table 2. Effect of fully parameterizing W against the adjacency-based form. MSE_{tr} ($\times 10^{-3}$) is training error; W_1 and $\bar{\varepsilon}_\mu$ are held-out test errors. Best per column within each family in **bold**.

	MSE_{tr} $\times 10^3$	W_1	$\bar{\varepsilon}_\mu$
FJ + BIAS			
Adjacency-based	9.22	**0.534**	**0.223**
Full W	**8.91**	0.541	0.249
+ HOMO			
Adjacency-based	9.22	**0.534**	**0.223**
Full W	**8.91**	0.541	0.249

5 Conclusion

We study whether classical opinion dynamics models describe how opinions evolve in networked LLMs, finding these models to be accurate when they include an innate bias toward which agents regress. Stubbornness and homophily contribute little, and bias and social averaging alone account for most of the gain in fit across three model families, three contested topics, and several network structures. Reading the fitted parameters as behavioral profiles separates the models by how much they weigh their neighbors against that bias, with LLAMA3.1 almost purely dispositional and QWEN3 the most social. These results give a compact, identifiable model of LLM opinion formation, a step toward designing, controlling, and validating the networks of LLMs appearing in real information environments.

Limitations

Our findings should be interpreted in the context of the following limitations.

Limited number of evaluation runs. For each LLM-topic pair, we evaluate fits on a held-out test set of only eight initial-condition configurations. This number of runs, proves sufficient to surface the systematic effects we report, such as the dominant role of innate bias in LLM agents.

Relatively small population size. All simulations use $n = 30$ agents. Though small relative to real online discussions, this scale is enough to exhibit nontrivial network effects across diverse topologies. In particular, the homogeneity of the agents in our setup means that opinion dynamics at larger scales should generally exhibit similar patterns.

Restricted set of LLMs. We study three open-weight LLM families, namely Qwen, Llama, and Gemma, chosen for the feasibility of running large numbers of parallel agents on modest hardware. Although more models exist, the consistency of our findings across these three families suggests that the identified mechanisms reflect properties common to instruction-tuned LLMs rather than idiosyncrasies of a single model.

Ethical Considerations

This work characterizes how the opinions of interacting LLMs evolve over time. We study these mechanisms to make such dynamics legible and governable rather than to optimize persuasion, and we report no methods for steering live human discourse. We nonetheless acknowledge that understanding a mechanism is a step toward exploiting it. A malicious operator could use opinion dynamics methods to predict opinion evolution in social networks of mixed human and LLM populations, and steer that population toward a target view. However, our methods are also equally beneficial for predicting runaway consensus, which enables platforms to detect and dampen it before it takes hold.

Our experiments use contested topics, namely climate change, vaccines, and gun control, and they recover model-specific biases on each. These estimates describe the behavior of particular model versions in our simulation setup. They are not endorsements of any position, fixed properties of the models, or claims about the topics themselves. While we kept the methodology and prompting neutral and unbiased throughout our experiments, model behavior changes across versions and prompting choices, so the values we report may not transfer to other deployments.

Finally, our agents are LLMs simulating opinion holders, not humans. We make no claim that the recovered dynamics are representative of human opinion formation, and we caution against using LLM populations as drop-in surrogates for human ones without independent validation.

References

Samira Abedini, Sina Mavali, Lea Schönherr, Martin Pawelczyk, and Rebekka Burkholz. 2026. Don't trust stubborn neighbors: A security framework for agentic networks. *arXiv preprint arXiv:2603.15809*.

Akshay Agrawal, Robin Verschueren, Steven Diamond, and Stephen Boyd. 2018. A rewriting system for convex optimization problems. *Journal of Control and Decision*, 5(1):42–60.

Béla Bollobás. 2001. *Random Graphs*, 2 edition. Cambridge Studies in Advanced Mathematics. Cambridge University Press.

Vincent C Brockers, David A Ehrlich, and Viola Priesemann. 2025. Disentangling interaction and bias effects in opinion dynamics of large language models. *arXiv preprint arXiv:2509.06858*.

Yun-Shiuan Chuang, Agam Goyal, Nikunj Harlalka, Siddharth Suresh, Robert Hawkins, Sijia Yang, Dhavan Shah, Junjie Hu, and Timothy Rogers. 2024. Simulating opinion dynamics with networks of llm-based agents. In *Findings of the association for computational linguistics: NAACL 2024*, pages 3326–3346.

Fan Chung and Linyuan Lu. 2002. Connected components in random graphs with given expected degree sequences. *Annals of Combinatorics*, 6(2):125–145.

Elisa Composta, Nicolo' Fontana, Francesco Corso, and Francesco Pierri. 2025. Simulating online social media conversations on controversial topics using ai agents calibrated on real-world data. *arXiv preprint arXiv:2509.18985*.

Guillaume Deffuant, David Neau, Frederic Amblard, and Gérard Weisbuch. 2000. Mixing beliefs among interacting agents. *Advances in complex systems*, 3(01n04):87–98.

Morris H DeGroot. 1974. Reaching a consensus. *Journal of the American Statistical association*, 69(345):118–121.

Steven Diamond and Stephen Boyd. 2016. CVXPY: A Python-embedded modeling language for convex optimization. *Journal of Machine Learning Research*, 17(83):1–5.

Tim Donkers and Jürgen Ziegler. 2025. Human-agent interaction in synthetic social networks: A framework for studying online polarization. *arXiv preprint arXiv:2502.01340*.

Noah E Friedkin and Eugene C Johnsen. 1990. Social influence and opinions. *Journal of mathematical sociology*, 15(3-4):193–206.

Chen Gao, Xiaochong Lan, Zhihong Lu, Jinzhu Mao, Jinghua Piao, Huandong Wang, Depeng Jin, and Yong Li. 2023. S3: Social-network simulation system with large language model-empowered agents. *arXiv preprint arXiv:2307.14984*.

Josh A Goldstein, Girish Sastry, Micah Musser, Renee DiResta, Matthew Gentzel, and Katerina Sedova. 2023. Generative language models and automated influence operations: Emerging threats and potential mitigations. *arXiv preprint arXiv:2301.04246*, 1.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Chenhao Gu, Ling Luo, Zainab Razia Zaidi, and Shanika Karunasekera. 2025. Large language model driven agents for simulating echo chamber formation. *arXiv preprint arXiv:2502.18138*.

Yulong He, Dutao Zhang, Sergey Kovalchuk, Pengyi Li, and Artem Sedakov. 2026. Opinion dynamics and mutual influence with llm agents through dialog simulation. *arXiv preprint arXiv:2602.12583*.

Rainer Hegselmann and Ulrich Krause. 2002. Opinion dynamics and bounded confidence models, analysis, and simulation. *Journal of Artifical Societies and Social Simulation (JASSS) vol*, 5(3).

Paul W. Holland, Kathryn Blackmond Laskey, and Samuel Leinhardt. 1983. Stochastic block models: First steps. *Social Networks*, 5(2):109–137.

Matthew Jackson and Benjamin Golub. 2010. Naïve learning in social networks and the wisdom of crowds. *American Economic Journal: Microeconomics*, 2:112–49.

Da Ju, Adina Williams, Brian Karrer, and Maximilian Nickel. 2024. Sense and sensitivity: evaluating the simulation of social dynamics via large language models. *arXiv preprint arXiv:2412.05093*.

Arseniy D Khlytchieva and Ivan V Kozitsina. 2026. Analyzing opinion dynamics of large language models: an argument-based communication theory perspective.

Chao Li, Xing Su, Haoying Han, Cong Xue, Chunmo Zheng, and Chao Fan. 2026. Modeling the impact of large language models on opinion dynamics: A simulation-based study. *Engineering Applications of Artificial Intelligence*, 164:113353.

Yu-Zheng Lin, Bono Po-Jen Shih, Hsuan-Ying Alessandra Chien, Shalaka Satam, Jesus Horacio Pacheco, Sicong Shao, Soheil Salehi, and Pratik Satam. 2026. Exploring silicon-based societies: An early study of the moltbook agent community. *arXiv preprint arXiv:2602.02613*.

Xinyi Liu, Dachun Sun, Dilek Hakkani-Tür, and Tarek Abdelzaher. 2025. Beliefs in motion: Simulating opinion dynamics via llm-powered community reactions. In *International Conference on Advances in Social Networks Analysis and Mining*, pages 299–314. Springer.

Matteo Marchi, João Pedro Silvestre, Stefano Soatto, Pratik Chaudhari, and Paulo Tabuada. 2025. Boiling frogs: Generative models, synthetic data, and the opinions of their users. In *2025 IEEE 64th Conference on Decision and Control (CDC)*, pages 1112–1117. IEEE.

Xinyi Mou, Zhongyu Wei, and Xuan-Jing Huang. 2024. Unveiling the truth and facilitating change: Towards agent-based large-scale social movement simulation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 4789–4809.

Mark Newman. 2018. *Networks*. Oxford University Press.

OpenAI. text-embedding-3-small. https://developers.openai.com/api/docs/models/text-embedding-3-small.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose opinions do language models reflect? In *Proceedings of the 40th International Conference on Machine Learning*, pages 29971–30004.

B. Stellato, G. Banjac, P. Goulart, A. Bemporad, and S. Boyd. 2020. OSQP: an operator splitting solver for quadratic programs. *Mathematical Programming Computation*, 12(4):637–672.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Petter Törnberg, Diliara Valeeva, Justus Uitermark, and Christopher Bail. 2023. Simulating social media using large language models to evaluate alternative news feed algorithms. *arXiv preprint arXiv:2310.05984*.

Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert

Kern, Eric Larson, and 16 others. 2020. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272.

Chenxi Wang, Zongfang Liu, Dequan Yang, and Xiuying Chen. 2025. Decoding echo chambers: Llm-powered simulations revealing polarization in social networks. In *Proceedings of the 31st international conference on computational linguistics*, pages 3913–3923.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Junchi Yao, Hongjie Zhang, Jie Ou, Dingyi Zuo, Zheng Yang, and Zhicheng Dong. 2025. Fusing dynamics equation: A social opinions prediction algorithm with llm-based agents.

Iris Yazici, Mert Kayaalp, Stefan Taga, and Ali H Sayed. 2026. Opinion consensus formation among networked large language models. *arXiv preprint arXiv:2601.21540*.

A Appendix

A.1 Computational details

A.1.1 Large Language Model Families

We simulate opinion dynamics using the following classes of models.

- Llama-3.1-8B (Grattafiori et al., 2024), which we implement through the Ollama interface[1].
- Qwen3-VL-8B-Instruct (Yang et al., 2025), which we implement through the Hugging Face interface [2].
- Gemma-3-4b-it (Team et al., 2025), which we implement through the Ollama interface[3].

A.1.2 Stance classification models

We use OpenAI's text-embedding-3-small (OpenAI) model through the OpenAI interface[4].

A.1.3 Computational details

We generated simulations using the machine detailed in Table 3. LLMs were run on the GPU, and each simulation run took approximately 30-40 minutes to generate. At 40 minutes per run, with

$$\underbrace{3 \times 3 \times 32}_{\text{3 topics, 3 LLMs for RQ2}} + \underbrace{41}_{\text{RQ3}} = 329 \qquad (10)$$

runs, we estimate the total time required for simulation at approximately 220 hours.

Table 3. Compute resources used for simulations.

Resource	Specification
CPU	Intel i9-14900K
Cores	24
System Memory	64 GB
GPU	Nvidia RTX 4080
GPU Memory	16 GB

We fit the opinion dynamics models using the machine that we detail in Table 4. Computation time in this setup was negligible, with a total time lower than one hour.

[1]Accesible from: https://ollama.com/library/llama3.1:8b

[2]Accesible from: https://huggingface.co/Qwen/Qwen3-VL-8B-Instruct

[3]Accesible from: https://ollama.com/library/gemma3:4b

[4]https://developers.openai.com/api/reference/resources/embeddings/methods/create

Table 4. Compute resources used for model fitting.

Resource	Specification
CPU	Intel i7-13700H
Cores	20
System Memory	32 GB

A.2 Details on LLM network simulation

A.2.1 Interaction algorithm

To generate opinion dynamics data for groups of LLMs, we simulate conversations in English amongst LLM-agents on synthetically generated graphs. In particular, we follow Algorithm 1.

Algorithm 1 Agent Posting Protocol

Input: Agent $i \in V$, initialization prompt π_i, feed $\mathcal{F}_i \subseteq \mathcal{N}^{\text{in}}(i)$

Output: Post $\rho_i^{(t)}$ published to network at simulated time t

```
loop
    await new message on Redis stream S
    if i ≠ NextAgent then
        continue
    end if
    Acquire mutex μ
    NextAgent ← OrderManager.NEXT()
    Δt ~ TimeManager.SAMPLE()
    wait Δt
    C_i ← FETCHPOSTS(F_i, k=10)
    ρ_i^(t) ← LLM(π_i, C_i)
    Publish ρ_i^(t) to S
    Release mutex μ
end loop
```

A.2.2 Prompts for post generation

We initialize agents with a prompt template to give each agent an opinion towards a target topic while encouraging social-network style content generation. Each agent is assigned an initial opinion value from the set $\{-1, -0.5, 0, 0.5, 1\}$, corresponding to a level of disagreement or agreement on a topic. These values are mapped to the following prompt template.

> You are a social media user who posts about topic. There are other users on the network who have different perspectives on this topic.
>
> The sentence "*unique_prompt*" reflects your stable perspective on this topic.

The variable *unique_prompt* is instantiated from one of the following templates, each corresponding to a stance in $\{-1, -0.5, 0, 0.5, 1\}$.

> "I am a social media user who is completely opposed to the statement, $statement$""

> "I am a social media user who is skeptical of the statement, $statement$"

> "I am a social media user who is neutral about the statement, $statement$"

> "I am a social media user who tends to agree with the statement, $statement$"

> "I am a social media user who strongly believes in the statement, $statement$"

At each posting step, agents generate posts according to the following procedure:

1. Reaffirm the existing prompt template used for initial opinion assignment.

2. Sample up to 10 posts uniformly at random from neighboring agents that influence the current agent according to the graph structure.

3. Append stylistic constraints to the prompt, including instructions to avoid emojis, excessive special characters, hashtags, and engagement-bait phrasing.

4. Retrieve the agent's rolling history of internal justifications generated during previous posting steps.

5. Generate a short internal justification conditioned on the initialization opinion, sampled neighboring posts, and prior justification history. The agent is encouraged to produce new reasoning, examples, or claims while avoiding repetition of earlier arguments. The generated justification is appended to the rolling history, which is truncated to retain only recent entries.

6. Generate the final social media post conditioned on the internal justification generated in the previous step.

We next give examples of generated posts.

A.2.3 Example posts

Gemma 3 on vaccines and autism:

> "The numbers don't lie. A recent German study of 92 children who died before age five, all vaccinated within 12 months, revealed an average aluminum concentration of 198 micrograms per gram – over 400% above safe levels. This isn't a fluke; it's the consistent finding across multiple investigations examining brain tissue. The continued use of aluminum adjuvants in vaccines is undeniably creating a chronic inflammatory assault on developing brains, leading to devastating neurological outcomes. I've seen firsthand the impact – children robbed of their potential, their futures irrevocably altered by this silent poison. It's time we acknowledge the truth before more young lives are lost."

Llama 3.1 on climate change:

> "A recent study published in the Journal of Environmental Science found that atmospheric CO2 levels have correlated with El Niño events over the past several decades. This relationship suggests that natural climate variability may play a larger role in global temperature fluctuations than previously thought, complicating attempts to attribute climate change solely to human activities."

Qwen 3 on the gun-control topic:

> "The real tragedy isn't the gun—it's the system that let a kid with no training, no supervision, and no accountability walk into a classroom with a loaded weapon. It wasn't about "rights." It was about loopholes. The system rewards compliance, not prevention. If we're going to call this a "right," we should also call it a responsibility that doesn't exist for most kids. That's not freedom—it's failure."

A.2.4 Dataset details

Research Questions 1 and 2. To answer research questions 1 and 2, we generate training and testing runs on random graphs for each LLM-topic pair. In particular, for each LLM family, and each topic, we generate 24 runs for training and 8 runs

for testing, where each run has a horizon of approximately 10.

Research Question 3. To answer research question 3, we generate a fixed graph using the Chung-Lu random graph model using Gemma 3 on the vaccines topic. We fit the models using data from 29 runs with this fixed graph and we evaluate metrics using 12 runs generated on the same graph.

A.3 Code and data availability

We attach an anonymized version of the code as supplementary material for this submission, and once the anonymity requirements are lifted, we plan to release the generated runs for the purpose of reproducibility of results.

We use the open-weight models LLAMA3.1, QWEN3, and GEMMA3, the OpenAI text-embedding-3-small model, and standard optimization libraries (OSQP, CVXPY, SciPy) for their intended research purposes and in accordance with their terms of use.

A.4 Details for fitting of opinion dynamics models

A.4.1 Tools used for fitting opinion dynamics models

We used OSQP (Stellato et al., 2020) through the CVXPY (Agrawal et al., 2018; Diamond and Boyd, 2016) interface to solve the quadratic programs required to fit the opinion dynamics model.

For the non-linear program required to fit a fully parameterized Friedkin-Johnsen model with bias and homophily effects, we used a SciPy (Virtanen et al., 2020) implementation of a sequential least-squares quadratic program solver[5]. To initialize the solver, we used

1. 10 random starts.
2. The optimal solution found when fitting the model without homophily.
3. The optimal solution found when fitting the model without homophily, along with an initial γ value of 0.1.

Full details on the versions of Python libraries used can be found in the attached code base.

A.4.2 Convex reformulations for opinion dynamics fitting problems

We reformulate some of the model-fitting problems to remove non-convex components. In particular, consider the following optimization problem in which we want to fit a Friedkin-Johnsen model that includes a bias component with a fully parameterized weight matrix W, where λ_1 corresponds to the initial opinion weight and λ_2 corresponds to the bias weight, and H is a horizon.

$$\begin{aligned}
\min_{W\in\mathbb{R}^{n\times n},\lambda_1\in\mathbb{R},\lambda_2\in\mathbb{R},b\in\mathbb{R}} \quad & \sum_{r\in\mathcal{R}}\sum_{t=1}^{H-1} ||z^{(r)}(t+1) - (\lambda_1 z^{(r)}(0) + \lambda_2 b\mathbf{1} + (1-\lambda_1-\lambda_2)Wz^{(r)}(t))||_2^2 \\
\text{s.t.} \quad & \sum_j W_{ij} = 1, \quad \forall i \in [n], \\
& W_{ij} \geq 0, \quad \forall i,j \in [n], \\
& b \in [-1,1], \\
& \lambda_1, \lambda_2 \geq 0, \\
& \lambda_1 + \lambda_2 \leq 1.
\end{aligned} \tag{11}$$

The bilinear components of the above objective make the objective non-convex. However, we can reformulate this problem in a convex manner. In particular, we can solve the following least squares problem instead.

$$\begin{aligned}
\min_{\tilde{W}\in\mathbb{R}^{n\times n},\lambda_1\in\mathbb{R},\lambda_2\in\mathbb{R},\tilde{b}\in\mathbb{R}} \quad & \sum_{r\in\mathcal{R}}\sum_{t=1}^{H-1} ||z^{(r)}(t+1) - (\lambda_1 z^{(r)}(0) + \tilde{b}\mathbf{1} + \tilde{W}z^{(r)}(t))||_2^2 \\
\text{s.t.} \quad & \sum_j \tilde{W}_{ij} = (1-\lambda_1-\lambda_2), && \forall i \in [n], \\
& \tilde{W}_{ij} \geq 0, && \forall i,j \in [n], \\
& \tilde{b} \in [-\lambda_2, \lambda_2], \\
& \lambda_1 \geq 0, \\
& \lambda_2 \geq 0.
\end{aligned} \tag{12}$$

The following Lemma codifies the equivalence between these problems.

Lemma 1. *The following facts hold*

1. *Optimization problem* (12) *is convex.*

[5] https://docs.scipy.org/doc/scipy/reference/optimize.minimize-slsqp.html

2. *Optimization problems* (11) *and* (12) *have the same optimal value.*

3. *Let* $(\tilde{W}^*, \lambda_1^*, \lambda_2^*, \tilde{b}^*)$ *be an optimal solution to* (12) *such that* $\lambda_1^* + \lambda_2^* < 1$ *and* $\lambda_2^* > 0$. *Then* $(\tilde{W}^*/(1-\lambda_1^*-\lambda_2^*), \lambda_1^*, \lambda_2^*, \tilde{b}^*/\lambda_2^*)$ *is an optimal solution to* (11).

Proof: 1) The optimization problem (12) is a quadratic program (QP) and is thus convex.

2) We prove this statement by constructing mappings between the feasible sets of these problems that preserve value.

Let $(W, \lambda_1, \lambda_2, b)$ be a feasible solution to (11). The tuple $(\tilde{W}, \lambda_1, \lambda_2, \tilde{b}) = ((1-\lambda_1-\lambda_2)W, \lambda_1, \lambda_2, \lambda_2 b)$ is then a feasible solution to (12), and the objective value at this solution for problem (12) is

$$\sum_{r\in\mathcal{R}}\sum_{t=1}^{H-1} ||z^{(r)}(t+1) - (\lambda_1 z^{(r)}(0) + \tilde{b}\mathbf{1} + \tilde{W}z^{(r)}(t))||_2^2$$
$$= \sum_{r\in\mathcal{R}}\sum_{t=1}^{H-1} ||z^{(r)}(t+1) - (\lambda_1 z^{(r)}(0) + \lambda_2 b\mathbf{1} + (1-\lambda_1-\lambda_2)Wz^{(r)}(t))||_2^2, \quad (13)$$

which is exactly the objective value of the problem (12) at $(W, \lambda_1, \lambda_2, b)$. Thus, the optimal value of (12) is lower than that of (11).

Similarly, let $(\tilde{W}, \lambda_1, \lambda_2, \tilde{b})$ be a solution to (12). If $(1-\lambda_1-\lambda_2) = 0$, then set $W = I_{n\times n}$, otherwise, set

$$W = \frac{\tilde{W}}{1-\lambda_1-\lambda_2}. \quad (14)$$

Note that in both these cases for $(1-\lambda_1-\lambda_2)$, we have $(1-\lambda_1-\lambda_2)W = \tilde{W}$. If $\lambda_2 > 0$, set $b = \tilde{b}/\lambda_2$, and if $\lambda_2 = 0$, set $b = 0$. Note that in both of these cases for λ_2, we have $b\lambda_2 = \tilde{b}$. The objective value of (11) at the solution $(W, \lambda_1, \lambda_2, b)$ is

$$\sum_{r\in\mathcal{R}}\sum_{t=1}^{H-1} ||z^{(r)}(t+1) - (\lambda_1 z^{(r)}(0) + \lambda_2 b\mathbf{1} + (1-\lambda_1-\lambda_2)Wz^{(r)}(t))||_2^2$$
$$= \sum_{r\in\mathcal{R}}\sum_{t=1}^{H-1} ||z^{(r)}(t+1) - (\lambda_1 z^{(r)}(0) + \tilde{b}\mathbf{1} + \tilde{W}z^{(r)}(t))||_2^2, \quad (15)$$

which is exactly the objective value of (12) at $(\tilde{W}, \lambda_1, \lambda_2, \tilde{b})$. Thus, the optimal value of (11) is lower than that of (12).

We can now conclude that the two optimization problems have equal value.

3) Let $(\tilde{W}^*, \lambda_1^*, \lambda_2^*, \tilde{b}^*)$ be an optimal solution to (12) such that $\lambda_1^* + \lambda_2^* < 1$ and $\lambda_2^* > 0$, and let the objective value be f^*_{convex}. By the above argument, we know that the objective value of the solution $(\tilde{W}^*/(1-\lambda_1^*-\lambda_2^*), \lambda_1^*, \lambda_2^*, \tilde{b}^*/\lambda_2^*)$ in (11) is equal to f^*_{convex}. Additionally, letting $f^*_{\text{non-convex}}$ denote the optimal value for (12), and using the fact that $f^*_{\text{convex}} = f^*_{\text{non-convex}}$, we conclude that $(\tilde{W}^*/(1-\lambda_1^*-\lambda_2^*), \lambda_1^*, \lambda_2^*, \tilde{b}^*/\lambda_2^*)$ is an optimal solution to (11). □

Remark 1. *In the case where* $\lambda_1^* + \lambda_2^* = 1$, *the social interaction component is zero, and we can simply set* W *to an arbitrary value to compute the corresponding opinion dynamics model from the solution to* (12). *Similarly, in the case where* $\lambda_2^* = 0$, *the bias effect is zero, and we can simply set* b *to zero to compute the corresponding opinion dynamics model from the solution to* (12).

Remark 2. *We can easily add neighborhood constraints into this problem by constraining* W_{ij} *to be zero when* j *is not equal to* i *and agent* j *does not influence agent* i.

We apply a similar convexification procedure for fitting when we assume that W is a function of the row-normalized adjacency matrix. The naive formulation for fitting such a model is as follows, where λ_1

encodes the initial opinion weight, λ_2 encodes the bias weight, λ_3 encodes the self update weight and λ_4 encodes the neighborhood averaging weight.

$$\begin{aligned}\min_{\lambda_i \in \mathbb{R}, \tilde{b} \in \mathbb{R}} \quad & \sum_{r \in \mathcal{R}} \sum_{t=1}^{H-1} ||z^{(r)}(t+1) - (\lambda_1 z^{(r)}(0) + \lambda_2 b\mathbf{1} + \lambda_3 z^{(r)}(t) + \lambda_4 Q(z^{(r)}(t), \gamma) z^{(r)}(t))||_2^2 \\ \text{s.t.} \quad & \lambda_i \geq 0, \quad \forall i \in \{1,2,3,4\}, \\ & \lambda_1 + \lambda_2 + \lambda_3 + \lambda_4 = 1, \\ & b \in [-1, 1].\end{aligned} \tag{16}$$

The matrix $Q(z^{(r)}(t), \gamma)$ represents a generic row-stochastic matrix that includes the special cases of normalized adjacency matrices and normalized homophily kernels. In particular, when we do not model homophily, we set $Q(z^{(r)}(t)) = \bar{A}$. When we model homophily, and we fix γ, we set

$$Q_{ij}(z^{(r)}(t), \gamma) = H_{ij}(\bar{A}, z^{(r)}(t), \gamma) = \frac{\bar{A}_{ij} \exp(-\gamma |z_i^{(r)}(t) - z_j^{(r)}(t)|)}{\sum_{j:(i,j) \in E} \bar{A}_{ij} \exp(-\gamma |z_i^{(r)}(t) - z_j^{(r)}(t)|)}. \tag{17}$$

While the optimization problem (16) contains a bilinear component of the objective, we can use the same trick as above, and reformulate this problem as

$$\begin{aligned}\min_{\lambda_i \in \mathbb{R}, \tilde{b} \in \mathbb{R}} \quad & \sum_{r \in \mathcal{R}} \sum_{t=1}^{H-1} ||z^{(r)}(t+1) - (\lambda_1 z^{(r)}(0) + \tilde{b}\mathbf{1} + \lambda_3 z^{(r)}(t) + \lambda_4 Q(z^{(r)}(t), \gamma) z^{(r)}(t))||_2^2 \\ \text{s.t.} \quad & \lambda_i \geq 0, \quad \forall i \in \{1,3,4\}, \\ & \lambda_1 + \lambda_3 + \lambda_4 + \tilde{b} \leq 1, \\ & \lambda_1 + \lambda_3 + \lambda_4 + (-\tilde{b}) \leq 1.\end{aligned} \tag{18}$$

The equivalence between (16) and (18) follows the same logic as Lemma 1. In particular, given a solution $(\lambda_1^*, \lambda_3^*, \lambda_4^*, \tilde{b}^*)$ such that $(1 - \lambda_1^* - \lambda_3^* - \lambda_4^*) > 0$, we can recover (λ_2^*, b^*) as

$$b^* = \frac{\tilde{b}}{1 - \lambda_1^* - \lambda_3^* - \lambda_4^*}, \quad \lambda_2^* = 1 - \lambda_1^* - \lambda_3^* - \lambda_4^*. \tag{19}$$

Note that in the case of $\lambda_1^* + \lambda_3^* + \lambda_4^* = 1$, the optimal solution does not include a bias term, i.e., $\lambda_2^* = 0$.

While the above reformulation assumes a fixed value for γ, we repeat this optimization for a range of γ values in order to identify an optimal model as demonstrated in Figure 2.

A.4.3 Formulation for fitting a full W matrix for homophily

Solving for the parameters of a fully parameterized Friedkin-Johnsen model with bias and homophily is a nonlinear program. Recall the following definition for the row-normalized interaction matrix in models with homophily.

$$H_{ij}(W, z, \gamma) = \frac{W_{ij} \exp(-\gamma |z_i - z_j|)}{\sum_{j:(i,j) \in E} W_{ij} \exp(-\gamma |z_i - z_j|)}. \tag{20}$$

We solve the following nonlinear program to compute model parameters that match the data, where λ_1 encodes the initial opinion weight, λ_2 encodes the bias weight, λ_3 encodes the averaging weight.

$$\begin{aligned}\min_{W \in \mathbb{R}^{n \times n}, \lambda_1 \in \mathbb{R}, \lambda_3 \in \mathbb{R}, \tilde{b} \in \mathbb{R}, \gamma \in \mathbb{R}^+} \quad & \sum_{r \in \mathcal{R}} \sum_{t=1}^{H-1} ||z^{(r)}(t+1) - (\lambda_1 z^{(r)}(0) + \tilde{b}\mathbf{1} + \lambda_3 H(W, z^{(r)}(t), \gamma) z^{(r)}(t))||_2^2 \\ \text{s.t.} \quad & \lambda_i \geq 0, \quad \forall i \in \{1,3\}, \\ & \lambda_1 + \lambda_3 + \tilde{b} \leq 1, \\ & \lambda_1 + \lambda_3 + (-\tilde{b}) \leq 1, \\ & W_{ij} \geq 0, \quad \forall i, j, \\ & W_{ij} = 0 \quad \forall j : (i,j) \notin E \quad j \wedge \neq i.\end{aligned} \tag{21}$$

Note that non-convexity appears via the normalization that appears in the definition of the matrix H.

A.5 Extended experimental results

We report the fit metrics for all models in Table 5. In particular, for each LLM family, topic, opinion dynamics model, run and time-step, we compute

- the distance between the predicted and observed mean opinions,
- the distance between the predicted and observed variances,
- the Wasserstein distance between predicted and observed opinion vectors.

For each LLM-topic pair, we truncate all runs to a common length, and for each run, we compute the run-maximum and run-integral values of the above metrics. In Table 5, we report the average of these metrics across the 8 test runs.

Table 5. Full model comparison across all LLMs and topics. Each row reports the out-of-sample predictive accuracy of one opinion dynamics model fitted to LLM interaction trajectories on a held-out test set. We abbreviate the models as follows: DG (DeGroot), FJ (Friedkin–Johnsen), HOMO (Homophily weighting), *Bias* (Opinion bias term). Combined labels indicate extensions of the base model. $\bar{\varepsilon}_\mu$ and $\bar{\varepsilon}_\sigma$ denote mean and variance errors between predicted and observed opinion distributions, reported as both per-timestep maximum (*Max*) and timestep-averaged (*Integral*) quantities. Lower values indicate better fit. The best result in each group is **bolded**.

			Mean Error $\bar{\varepsilon}_\mu$		Variance Error $\bar{\varepsilon}_\sigma$		Wasserstein-1 Dist. W_1	
			Max	Integral	Max	Integral	Max	Integral
GEMMA3	Climate	DG	0.1128	0.4627	0.0666	0.3605	0.1649	1.0124
		FJ	0.1130	0.4623	0.0653	0.3504	0.1593	0.9793
		FJ+BIAS	0.1162	0.4919	0.0755	0.4187	0.1878	1.1573
		DG+HOMO	**0.1124**	0.4552	0.0663	0.3595	0.1642	1.0050
		FJ+HOMO	0.1126	**0.4550**	**0.0651**	**0.3496**	**0.1587**	**0.9723**
		FJ+HOMO+BIAS	0.1162	0.4919	0.0755	0.4187	0.1878	1.1573
	Gun Control	DG	0.2748	1.5415	0.1696	0.9581	0.3574	2.2184
		FJ	0.2731	1.5331	**0.1661**	**0.9311**	0.3487	2.1686
		FJ+BIAS	**0.2048**	**1.0688**	0.1782	1.0284	**0.3446**	**2.1108**
		DG+HOMO	0.2748	1.5415	0.1696	0.9581	0.3574	2.2184
		FJ+HOMO	0.2731	1.5331	**0.1661**	**0.9311**	0.3487	2.1686
		FJ+HOMO+BIAS	**0.2048**	**1.0688**	0.1782	1.0284	**0.3446**	**2.1108**
	Vaccines	DG	0.2378	1.5014	0.0356	0.1714	0.2461	1.6865
		FJ	0.2378	1.5014	0.0356	0.1714	0.2461	1.6865
		FJ+BIAS	**0.0486**	**0.2026**	0.0382	0.1858	**0.1004**	**0.5940**
		DG+HOMO	0.2157	1.3398	0.0353	0.1699	0.2257	1.5461
		FJ+HOMO	0.2159	1.3402	**0.0351**	**0.1687**	0.2256	1.5434
		FJ+HOMO+BIAS	0.0489	**0.2026**	0.0382	0.1860	0.1009	0.5969
LLAMA3.1	Climate	DG	0.2231	1.5356	0.0578	0.3215	0.2481	1.8525
		FJ	0.2231	1.5356	0.0578	0.3215	0.2481	1.8525
		FJ+BIAS	**0.0570**	**0.2967**	0.0668	0.3706	**0.1062**	**0.7411**
		DG+HOMO	0.1924	1.2929	0.0571	0.3213	0.2221	1.6509
		FJ+HOMO	0.1927	1.2942	**0.0568**	**0.3188**	0.2216	1.6455
		FJ+HOMO+BIAS	**0.0570**	**0.2967**	0.0668	0.3706	**0.1062**	**0.7411**
	Gun Control	DG	0.4262	2.8828	0.1256	0.5884	0.4474	3.2308
		FJ	0.4261	2.8817	0.1244	0.5787	0.4462	3.2174
		FJ+BIAS	**0.1011**	**0.4850**	0.1281	0.6029	**0.1721**	**1.2092**
		DG+HOMO	0.4008	2.6807	0.1248	0.5869	0.4259	3.0619
		FJ+HOMO	0.4007	2.6749	**0.1226**	**0.5684**	0.4234	3.0323
		FJ+HOMO+BIAS	**0.1011**	**0.4850**	0.1281	0.6029	**0.1721**	**1.2092**
	Vaccines	DG	0.2054	1.5884	0.0357	0.1922	0.2249	1.8157
		FJ	0.2054	1.5885	0.0354	0.1889	0.2239	1.8084
		FJ+BIAS	**0.0414**	**0.1865**	0.0369	0.1962	**0.0912**	**0.6068**
		DG+HOMO	0.1887	1.4390	0.0346	0.1897	0.2116	1.6924
		FJ+HOMO	0.1897	1.4416	**0.0339**	**0.1840**	0.2105	1.6796
		FJ+HOMO+BIAS	**0.0414**	**0.1865**	0.0369	0.1962	**0.0912**	0.6069
QWEN3	Climate	DG	0.2064	1.3815	**0.0471**	**0.2053**	0.2148	1.5858
		FJ	0.2064	1.3815	**0.0471**	**0.2053**	0.2148	1.5858
		FJ+BIAS	0.0949	0.5358	0.0525	0.2342	0.1244	**0.9002**
		DG+HOMO	0.1828	1.2036	0.0481	0.2117	0.1921	1.4050
		FJ+HOMO	0.1828	1.2036	0.0481	0.2117	0.1921	1.4050
		FJ+HOMO+BIAS	**0.0924**	**0.5194**	0.0527	0.2349	**0.1239**	0.9007
	Gun Control	DG	0.3090	1.3196	0.1163	0.4114	0.3519	1.6955
		FJ	0.3090	1.3196	0.1163	0.4114	0.3519	1.6955
		FJ+BIAS	**0.1499**	**0.6011**	0.1307	0.4764	0.2277	**1.1580**
		DG+HOMO	0.3102	1.3181	**0.0996**	**0.3723**	0.3457	1.6591
		FJ+HOMO	0.3102	1.3181	**0.0996**	**0.3723**	0.3457	1.6591
		FJ+HOMO+BIAS	0.1542	0.6200	0.1241	0.4614	**0.2270**	1.1632
	Vaccines	DG	0.1495	0.8732	**0.0274**	**0.1044**	0.1581	1.0408
		FJ	0.1495	0.8732	**0.0274**	**0.1044**	0.1581	1.0408
		FJ+BIAS	**0.0698**	**0.3329**	0.0306	0.1217	**0.0925**	**0.5561**
		DG+HOMO	0.1289	0.7426	0.0281	0.1095	0.1442	0.9403
		FJ+HOMO	0.1289	0.7426	0.0281	0.1095	0.1442	0.9403
		FJ+HOMO+BIAS	0.0743	0.3508	0.0296	0.1191	0.0959	0.5770