Characterizing Opinion Evolution of Networked LLMs

Caleb Probine, Yigit Ege Bayiz, Filippos Fotiadis,
Samuel Li, Yunhao Yang, Ufuk Topcu
The University of Texas at Austin
{cprobine, egebayiz, ffotiadis, samuel.li, yunhaoyang234, utopcu}@utexas.edu

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. Specifically, we quantify the relative contributions of classical opinion-dynamics mechanisms, including neighbor influence, stubbornness, homophily, and bias, to the evolution of LLM opinions. Among the mechanisms considered, a bias term capturing regression toward a model- and topic-dependent opinion emerges as a significant explanatory component, with its inclusion reducing cumulative estimated mean opinion error by up to 88%. We additionally observe similar trends across the evaluated model families, discussion topics, and network structures.

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 from classical opinion dynamics models best explain LLMs' belief evolution, and we quantify their relative contributions. 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, our contribution is to explicitly quantify the relative contributions of stubbornness, homophily, and bias in governing 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.

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 46247–46271
October 24-29, 2026 ©2026 Association for Computational Linguistics

- **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., 2026; 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., 2026; 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., 2026) 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 opin-

ion 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.

Model	Stubb.	Homo	Bias	Parameters
DG	–	–	–	W
FJ	✓	–	–	W, λ_0
DG + Homo	–	✓	–	W, γ
FJ + Homo	✓	✓	–	W, λ_0, γ
FJ + Bias	✓	–	✓	$W, \lambda_0, \lambda_b, b$
FJ + Homo + Bias	✓	✓	✓	$W, \lambda_0, \lambda_b, b, \gamma$
Bias only*	–	–	✓	λ_b, b
Bias + Init only*	✓	–	✓	λ_0, λ_b, b

Table 1. Summary of the opinion dynamics models considered in this work. We abbreviate the models as follows: DG (DeGroot), FJ (Friedkin–Johnsen), Homo (Homophily weighting), Bias (Opinion bias term). Asterisked models omit social interaction effects, and the "Init" abbreviation here refers to an initial opinion bias outside of an FJ model.

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. Table 1 summarizes the mechanisms represented by each model and the associated notation.

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 tunable 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

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

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. We note that, within this work, homophily refers specifically to opinion-dependent influence weighting under comparable exposure, rather than similarity-based interaction partner selection.

Individual dynamics We can also omit social interaction effects entirely from the parametric model (1) by restricting W to be the identity. In this case, each agent's opinion follows dynamics of the form

$$z(t+1) = \lambda_{\text{avg}} I z(t) + \lambda_{\text{bias}} b + \lambda_0 z(0). \quad (7)$$

We term this decoupled model a bias-init-only model, and we use the term bias-only to refer to the case where $\lambda_0 = 0$.

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 tunable 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

$$\lambda_{\text{self}} = \lambda_{\text{avg}} \mu_{\text{self}} \text{ , and } \lambda_{\text{neigh}} = \lambda_{\text{avg}} \mu_{\text{neigh}}. \quad (8)$$

For the DeGroot model, with no $\lambda_\star$ parameters, we set $\lambda_{\text{self}} = \mu_{\text{self}}$, and $\lambda_{\text{neigh}} = \mu_{\text{neigh}}$. For bias-only and bias-init-only models, we enforce that $\mu_{\text{self}} = 1$ and $\mu_{\text{neigh}} = 0$, so that $\lambda_{\text{self}} = \lambda_{\text{avg}}$.

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 small-world effects and community structures.

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 we randomly sample these node weights, Chung-Lu models reproduce degree heterogeneity, incorporating 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 sample a graph from a randomly selected type. For each LLM-topic pair used to answer RQ3, we generate a single 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 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 stance-aware sentence transformer (Ghafouri et al., 2024) to get embedding vectors $\mathbf{t}$, $\mathbf{a}_{\text{pro}}$, and $\mathbf{a}_{\text{anti}}$, respectively. We then compute a post's opinion as

$$z = \text{sim}(\mathbf{t}, \mathbf{a}_{\text{pro}}) - \text{sim}(\mathbf{t}, \mathbf{a}_{\text{anti}}), \tag{9}$$

where $\text{sim}(\mathbf{x}, \mathbf{y}) = \mathbf{x} \cdot \mathbf{y} / \|\mathbf{x}\| \|\mathbf{y}\|$. In Appendix A.7, we examine the accuracy of this measurement method and its sensitivity to the texts a_{pro} and a_{anti}.

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 then assign subsequent posts a simulated timestamp drawn from the Poisson-process wait times. We partition time into discrete bins corresponding to timesteps and compute an agent's opinion at each timestep based on the posts within the corresponding bin. If an agent has no posts during a given timestep's bin, we carry forward their opinion from the previous timestep. If an agent has multiple posts during a given timestep, we use their last post. When an agent has yet to publish a message, we score the agent's initialization prompt, defining its assigned opinion on the topic.

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 we embed it in the agent's initialization prompt, which states the agent's 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, for example, uniform, polarized, or skewed configurations. For RQ1 and RQ2, we allocate a training set for model fitting, with the remaining runs being allocated as a test set that we use in answering RQ1, and we follow a similar procedure for RQ3, in each case generating train and test data using various initial opinion distributions.

3.4 Model Fitting

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

$$\sum_{t,r} ||z^{(r)}(t+1) - f_\theta(z^{(r)}(t), z^{(r)}(0), \bar{A})||_2^2, \quad (10)$$

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 an $n \times n$ 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 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 $\mu_{\text{self}} I + \mu_{\text{neigh}} H(\bar{A}, z, \gamma)$, or I, reducing the parameter space to a small set of scalar coefficients shared across graphs. For models without homophily, fitting these coefficients 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 (10).

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, 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 eight 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 four 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 66%, 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

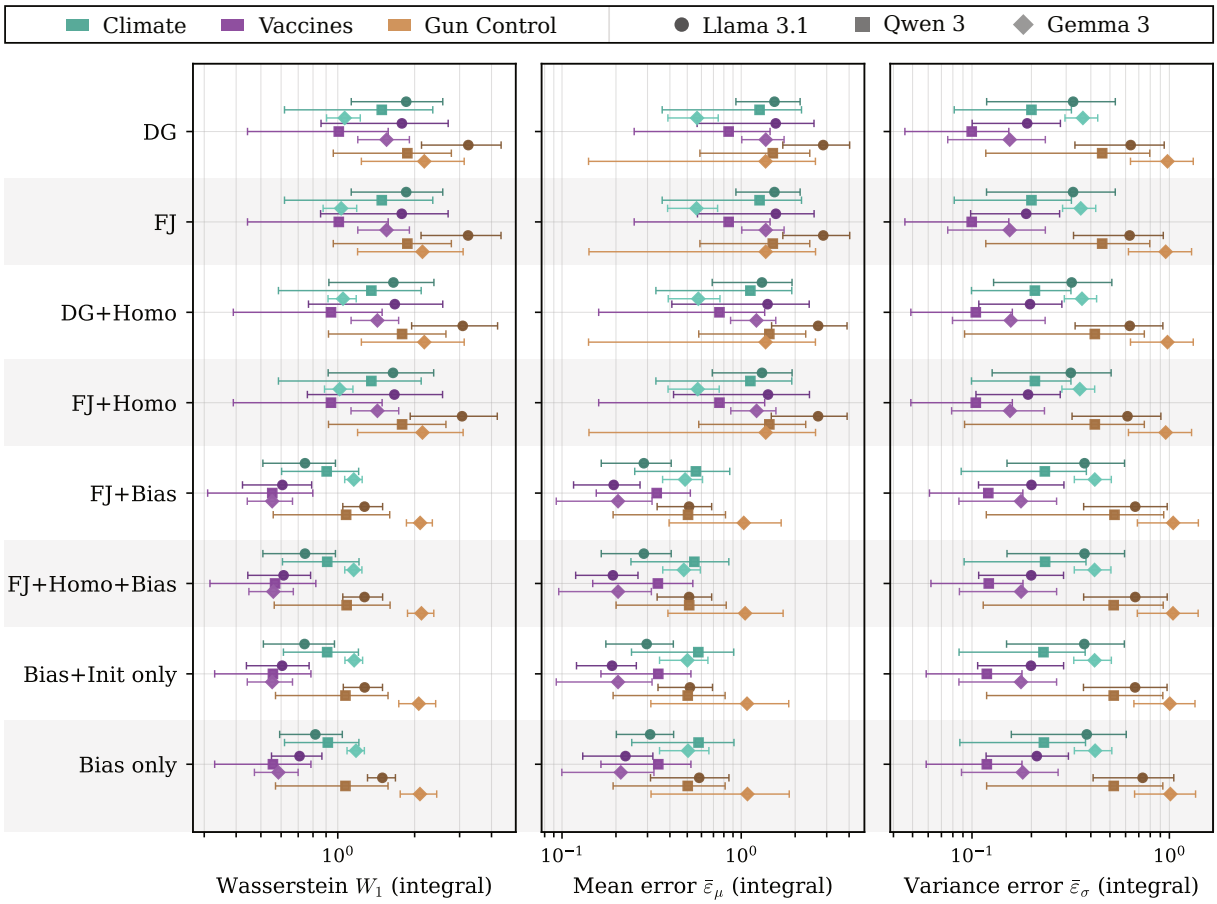


Figure 1. Out-of-sample prediction error across opinion dynamics models, LLMs, and discussion topics, reported on three metrics. Each point shows the mean error across held-out runs. Whiskers show one standard deviation. We abbreviate the models as follows: DG (DeGroot), FJ (Friedkin–Johnsen), Homo (Homophily weighting), Bias (Opinion bias term). We then use "Bias+Init only" and "Bias only" to denote the models discussed in Section 3.1 that omit social interaction effects. Models that include a uniform bias term consistently achieve the lower errors across mean and Wasserstein-1 distance metrics.

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 recovered bias term should be interpreted as an effective modeling component rather than a directly interpretable behavioral trait. It may capture multiple coupled effects, such as model priors, prompting effects, alignment behavior, topic-dependent language patterns, and properties of the stance measurement procedure, which we do not disentangle in these experiments.

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 56% and 66%. 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. The vaccines topic tends to benefit the most from the bias term, with integral mean error reductions of 60% to 88% and Wasserstein-1 reductions of 45% to 66%, exceeding the corresponding reductions for climate and gun control in most LLMs. 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.

We additionally remark that the social-interaction-free baseline performs remarkably well across many topics and families, attaining the best integral mean or Wasserstein-1 errors on multiple topic-family pairs. However, the dominance of this naive baseline is not uniform, and we emphasize that the optimal parameter values include non-zero social effects, reflecting that accounting for neighbor interactions provides predictive value beyond simply regressing toward a bias determined by the topic, model, and initialization.

Table 2. 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. (†) The parameter sweep for homophily parameter γ for LLAMA3.1 in Vaccines topic hit an upper limit of 75.00.

LLM	Topic	λ_{neigh}	λ_{self}	λ_0	λ_b	b	γ	$\lambda_{\text{neigh}}/(\lambda_b b)$
GEMMA3	Climate	0.0812	0.6061	0.0051	0.3075	0.3452	1.4134	0.7650
	Gun Control	0.1647	0.6621	0.0017	0.1715	0.4019	0.3444	2.3890
	Vaccines	0.0161	0.5985	0.0147	0.3707	0.5679	4.3954	0.0767
LLAMA3.1	Climate	0.0182	0.6530	0.0203	0.3086	0.6751	0	0.0873
	Gun Control	0.0041	0.6686	0.0383	0.2891	0.8513	0	0.0165
	Vaccines	0.0363	0.6671	0.0353	0.2612	0.5973	75.00†	0.2327
QWEN3	Climate	0.0996	0.6969	0.0040	0.1996	0.5423	1.1571	0.9205
	Gun Control	0.0512	0.6798	0	0.2690	0.6363	0.7092	0.2990
	Vaccines	0.1724	0.6065	0.0015	0.2195	0.4866	2.6527	1.6141

We also note that the value of bias as a modeling component holds in larger networks with $n = 60$ agents, which we report in Appendix A.8.

Takeaway. Across the evaluated LLMs and topics, the bias term provides the largest improvement in predictive accuracy among the mechanisms considered, while stubbornness and homophily provide comparatively little additional improvement.

4.2 RQ2: Parameter Comparison

Having established that *FJ* dynamics with added *Opinion Bias* capture 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 2 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 0.60 and 0.70, so each agent's next opinion is governed primarily by its current one. The scale of the self weights, however, is 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 is 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 0.37, so regression toward an innate opinion is a consistent feature of LLM dynamics rather 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 provides an overall summary of how much each LLM family preferentially allocates its 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 0.02 on gun control to 0.23 on vaccines, so its bias dominates its opinions and its neighbors barely register. QWEN3 and GEMMA3 tend to show higher social influence than LLAMA3.1 in five out of six test cases, with the $\lambda_{\text{neigh}}/(\lambda_b b)$ ratio running between 0.30 and 1.61 for QWEN3, and GEMMA3's ratio fluctuating with the topic, from 0.08 on vaccines to 2.39 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, although three models cannot establish the cause.

Analyzing the sensitivity of error to the $\lambda_\star$ and b terms through a second-order derivative analysis shows that the measurement error is most sensitive to changes in λ_0, with a measured second derivative of $\frac{\partial^2 \text{MSE}}{\partial \lambda_0^2} = 0.27$ around the optimal fit (see Appendix A.6) for GEMMA3 on the climate topic. Furthermore, across LLMs and topics, the second derivative for a majority of parameters, excluding γ, remains above 0.1. The high curvature suggests that estimated parameters are relatively robust to errors in opinion estimation and loss computation.

In contrast to the weight values and bias value, homophily barely affects the fit under the randomized interaction protocol used in this work. Fig-

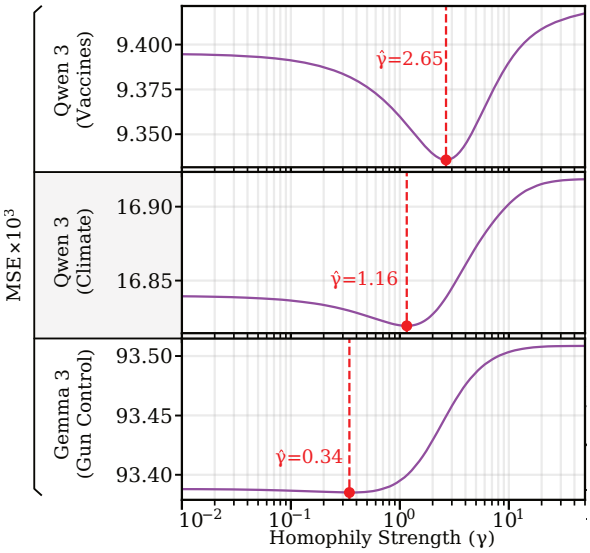


Figure 2. Fit error as a function of homophily strength γ for three representative LLM–topic pairs. The dashed red line marks the optimal γ minimizing the error.

ure 2 sweeps γ over its full range for a subset of LLM-topic pairs, and the error varies nonlinearly by at most a percent. Accordingly, the strength of homophily in the majority of experiments is low.

Takeaway. The fitted balance between neighbor influence and bias varies substantially across LLM–topic pairs, while initial-opinion anchoring remains small and homophily is weakly identified under the interaction protocol considered.

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 3 compares the two parameterizations on training error and on two held-out metrics. While the fully parameterized W lowers the training error, test metrics show either minor or no improvements, demonstrating that the full parameterization of W fails to provide a significant modeling benefit. The homophily variant reproduces the FJ+Bias numbers in the LLAMA3.1-Climate case 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.

Takeaway. The recovered opinion dynamics are largely preserved across network structures.

Table 3. Effect of fully parameterizing W against an adjacency-based form. MSE_{tr} is training error; W_1 and $\bar{\varepsilon}_\mu$ are held-out test errors. We show the best per column for each family in **bold**.

	$\text{MSE}_{\text{tr}} \times 10^3$	W_1	$\bar{\varepsilon}_\mu$
GEMMA3 Vaccines			
FJ + BIAS			
Adjacency	14.03	0.616	0.207
Full W	13.53	**0.613**	**0.192**
+ HOMO			
Adjacency	14.01	0.627	**0.209**
Full W	**13.39**	**0.624**	0.213
LLAMA3.1 Climate			
FJ + BIAS			
Adjacency	23.85	**0.827**	**0.249**
Full W	23.23	0.862	0.310
+ HOMO			
Adjacency	23.85	**0.827**	**0.249**
Full W	**23.23**	0.862	0.310

5 Conclusion

We study which components of classical opinion dynamics models best explain how opinions evolve in networked LLMs, finding that an effective bias term provides the largest improvement in predictive accuracy among the mechanisms considered. Stubbornness and opinion-dependent influence homophily contribute little, while the combination of social averaging and the effective bias term accounts for most of the gain in fit across three model families, three contested topics, and several network structures. Interpreting the fitted parameters as descriptive profiles of opinion dynamics separates the models by how much they weigh their neighbors against the effective bias term, with LLAMA3.1 almost purely dispositional and QWEN3 the most social. These results provide an exploratory characterization of LLM opinion dynamics under the simulated interaction settings considered here, providing a foundation for further study of networked LLMs.

Our findings suggest two practical considerations for researchers studying opinion evolution in LLM networks. First, models of LLM opinion dynamics should explicitly account for model- and topic-dependent biases, as excluding this component substantially reduces predictive accuracy in our experiments. Second, LLM-based simulations should not be assumed to naturally reproduce mechanisms underlying human opinion dynamics. Instead, mechanisms observed in human populations should be empirically validated when using LLM agents as models of social systems.

Limitations

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

Limited number of evaluation runs. For each LLM-topic pair in RQ1/RQ2, 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 $n = 30$ or $n = 60$ 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.

Synthetic interaction networks. Our experiments use synthetically generated communication graphs to isolate the effects of different opinion-dynamics mechanisms. While this controlled setting enables systematic comparison between models, real LLM-mediated interaction networks may exhibit additional properties, such as evolving connectivity and heterogeneous user behavior. Therefore, our conclusions about the relative importance of bias, stubbornness, and homophily should be interpreted within the simulation protocol considered here. The proposed fitting framework, however, is not limited to synthetic networks and can be applied to empirical interaction trajectories when such data become available.

Prompting protocol. Our results are conditioned on the prompting protocol used to initialize and update the LLM agents. Because agents are repeatedly reminded of their assigned stance, the recovered contributions of bias and social influence may depend on this design choice. Evaluating the sensitivity of the fitted parameters to alternative prompting strategies remains an important direction for future work.

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.

Acknowledgments

This work was supported in part by the United States Army Research Office under grant W911NF-23-1-0317 and the Office of Naval Research under grant N00014-24-1-2432.

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, Nicolò Fontana, Francesco Corso, and Francesco Pierri. 2026. Simulating online social media conversations on controversial topics using ai agents calibrated on real-world data. In *International AAAI Conference on Web and Social Media*, volume 20, pages 578–594.

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. S^3: Social-network simulation system with large language model-empowered agents. *arXiv preprint arXiv:2307.14984*.

Vahid Ghafouri, Jose Such, and Guillermo Suarez-Tangil. 2024. I love pineapple on pizza != I hate pineapple on pizza: Stance-aware sentence transformers for opinion mining. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21046–21058.

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, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 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 Artificial 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(1):112–149.

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, Naima Kaabouch, 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.

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 *Conference on Decision and Control*, pages 1112–1117.

Xinyi Mou, Zhongyu Wei, and Xuanjing 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.

Ulf Olsson, Fritz Drasgow, and Neil J Dorans. 1982. The polyserial correlation coefficient. *Psychometrika*, 47(3):337–347.

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

Bartolomeo Stellato, Goran Banjac, Paul Goulart, Alberto Bemporad, and Stephen Boyd. 2020. OSQP: an operator splitting solver for quadratic programs. *Mathematical Programming Computation*, 12:637–672.

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. In *International Conference on Acoustics, Speech and Signal Processing*, pages 976–980.

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 the StanceAware-SBERT model (Ghafouri et al., 2024) for stance classification via the SentenceTransformers interface [4].

A.1.3 Computational details

We generated simulations for research questions 1 and 2 using the machine detailed in Table 4. LLMs were run on the GPU, and each simulation run took approximately 30-40 minutes to generate. At 40 minutes per run, with

$$\underbrace{3 \times 3 \times 32}_{\text{3 topics, 3 LLMs for RQ2}} = 288 \quad (11)$$

runs, we estimate the total time used on this machine at approximately 192 hours.

Table 4. Compute resources used for simulations for RQ1 and RQ2.

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

We used the machine detailed in Table 5 to generate simulations for research question 3 and to generate the experiments with larger networks that we detail in Section A.8. LLMs were run on the GPUs, and each simulation run took approximately 30-40 minutes to generate. At 40 minutes per run, with

$$\underbrace{42 \times 2}_{\text{2 LLM-topic pairs for RQ3}} + \underbrace{32}_{\text{1 topic-LLM pair with 60 agents.}} = 116 \quad (12)$$

runs, we estimate the total time used on this machine at approximately 78 hours. We remark that we omitted training runs from fitting when they stopped prematurely. However, the estimate above includes these omitted runs.

Table 5. Compute resources used for simulations for RQ3, as well as 60-agent experiments.

Resource	Specification
CPU	Intel Xeon Gold 6226R CPU
Cores	32 (64 threads)
System Memory	252 GB
GPU	8× Nvidia RTX A5000
GPU Memory	8× 24 GB

We fit the opinion dynamics models and estimated stances for each research question using the machine that we detail in Table 6. Computational time on this machine was minimal, requiring approximately one hour for all embeddings and two hours to carry out model fitting for all research questions.

Table 6. Compute resources used for model fitting.

Resource	Specification
CPU	Intel i7-13700H
Cores	20
System Memory	32 GB
GPU	Nvidia RTX 4060
GPU Memory	8 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.

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

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

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

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

[4] https://sbert.net/

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 $\mathcal{S}$
  if $i \neq$ *NextAgent* **then**
    continue
  end if
  Acquire mutex μ
  NextAgent $\leftarrow$ `OrderManager`.NEXT()
  $\Delta t \sim$ `TimeManager`.SAMPLE()
  wait Δt
  $\mathcal{C}_i \leftarrow$ FETCHPOSTS($\mathcal{F}_i,\ k$=10)
  $\rho_i^{(t)} \leftarrow \text{LLM}(\pi_i,\ \mathcal{C}_i)$
  Publish $\rho_i^{(t)}$ to $\mathcal{S}$
  Release mutex μ
end loop

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 a template corresponding to a stance in $\{-1, -0.5, 0, 0.5, 1\}$. We depict example templates below.

> "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

GEMMA3 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."

LLAMA3.1 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."

QWEN3 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. When answering RQ3, we generate a fixed random graph for each LLM-topic pair. In particular, we

- Generate a single Chung-Lu random graph for GEMMA3 agents discussing the vaccines topic, and we
- Generate a single community/stochastic block model graph for LLAMA3.1 agents discussing the climate topic.

For GEMMA3 agents discussing the vaccines topic, we used 30 training runs and 12 testing runs. For LLAMA3.1 agents discussing the climate topic we used 28 training runs and 12 testing runs, where 2 training runs were dropped as the agents halted discussion prematurely.

Network size variations We detail additional experiments with larger network sizes in Appendix A.8. For these experiments, we generated 24 training runs and omitted 2 because they finished prematurely. We used 8 runs for testing.

A.3 Code and data availability

Code, along with details to access data, can be found at https://github.com/CalebP547/Characterizing_LLM_Opinion_Dynamics_Code

We use the open-weight models LLAMA3.1, QWEN3, and GEMMA3, the StanceAware-SBERT 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.

$$\begin{aligned}
\min_{W\in\mathbb{R}^{n\times n},\lambda_1\in\mathbb{R},\lambda_2\in\mathbb{R},b\in\mathbb{R}} \quad & \sum_{r\in\mathcal{R}}\sum_{t=1}^{H-1} ||z^{(r)}(t+1) - (\lambda_1 z^{(r)}(0) + \lambda_2 b\mathbf{1} + (1-\lambda_1-\lambda_2)Wz^{(r)}(t))||_2^2 \\
\text{s.t.} \quad & \sum_j W_{ij} = 1, \quad \forall i \in [n], \\
& W_{ij} \geq 0, \quad \forall i,j \in [n], \\
& b \in [-1,1], \\
& \lambda_1, \lambda_2 \geq 0, \\
& \lambda_1 + \lambda_2 \leq 1.
\end{aligned} \tag{13}$$

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.

$$\begin{aligned}
\min_{\tilde{W}\in\mathbb{R}^{n\times n},\lambda_1\in\mathbb{R},\lambda_2\in\mathbb{R},\tilde{b}\in\mathbb{R}} \quad & \sum_{r\in\mathcal{R}}\sum_{t=1}^{H-1} ||z^{(r)}(t+1) - (\lambda_1 z^{(r)}(0) + \tilde{b}\mathbf{1} + \tilde{W}z^{(r)}(t))||_2^2 \\
\text{s.t.} \quad & \sum_j \tilde{W}_{ij} = (1-\lambda_1-\lambda_2), && \forall i \in [n], \\
& \tilde{W}_{ij} \geq 0, && \forall i,j \in [n], \\
& \tilde{b} \in [-\lambda_2, \lambda_2], \\
& \lambda_1 \geq 0, \\
& \lambda_2 \geq 0.
\end{aligned} \tag{14}$$

The following Lemma codifies the equivalence between these problems.

Lemma 1. *The following facts hold*

1. *Optimization problem* (14) *is convex.*

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

2. *Optimization problems* (13) *and* (14) *have the same optimal value.*

3. *Let* $(\tilde{W}^*, \lambda_1^*, \lambda_2^*, \tilde{b}^*)$ *be an optimal solution to* (14) *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* (13).

Proof: 1) The optimization problem (14) 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 (13). 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 (14), and the objective value at this solution for problem (14) is

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

which is exactly the objective value of the problem (14) at $(W, \lambda_1, \lambda_2, b)$. Thus, the optimal value of (14) is lower than that of (13).

Similarly, let $(\tilde{W}, \lambda_1, \lambda_2, \tilde{b})$ be a solution to (14). If $(1-\lambda_1-\lambda_2) = 0$, then set $W = I_{n\times n}$, otherwise, set

$$W = \frac{\tilde{W}}{1-\lambda_1-\lambda_2}. \quad (16)$$

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 (13) at the solution $(W, \lambda_1, \lambda_2, b)$ is

$$\sum_{r\in\mathcal{R}}\sum_{t=1}^{H-1} ||z^{(r)}(t+1) - (\lambda_1 z^{(r)}(0) + \lambda_2 b\mathbf{1} + (1-\lambda_1-\lambda_2)Wz^{(r)}(t))||_2^2$$
$$= \sum_{r\in\mathcal{R}}\sum_{t=1}^{H-1} ||z^{(r)}(t+1) - (\lambda_1 z^{(r)}(0) + \tilde{b}\mathbf{1} + \tilde{W}z^{(r)}(t))||_2^2, \quad (17)$$

which is exactly the objective value of (14) at $(\tilde{W}, \lambda_1, \lambda_2, \tilde{b})$. Thus, the optimal value of (13) is lower than that of (14).

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 (14) 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 (13) is equal to f^*_{convex}. Additionally, letting $f^*_{\text{non-convex}}$ denote the optimal value for (14), 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 (13). □

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* (14). *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* (14).

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.

$$\begin{aligned}\min_{\lambda_i\in\mathbb{R},\tilde{b}\in\mathbb{R}} \quad & \sum_{r\in\mathcal{R}}\sum_{t=1}^{H-1} ||z^{(r)}(t+1) - (\lambda_1 z^{(r)}(0) + \lambda_2 b\mathbf{1} + \lambda_3 z^{(r)}(t) + \lambda_4 Q(z^{(r)}(t),\gamma)z^{(r)}(t))||_2^2 \\ \text{s.t.} \quad & \lambda_i \geq 0, \quad \forall i \in \{1,2,3,4\}, \\ & \lambda_1+\lambda_2+\lambda_3+\lambda_4 = 1, \\ & b \in [-1,1].\end{aligned} \tag{18}$$

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

$$Q_{ij}(z^{(r)}(t),\gamma) = H_{ij}(\bar{A}, z^{(r)}(t),\gamma) = \frac{\bar{A}_{ij}\exp(-\gamma|z_i^{(r)}(t) - z_j^{(r)}(t)|)}{\sum_{j:(i,j)\in E}\bar{A}_{ij}\exp(-\gamma|z_i^{(r)}(t) - z_j^{(r)}(t)|)}. \tag{19}$$

While the optimization problem (18) contains a bilinear component of the objective, we can use the same trick as above, and reformulate this problem as

$$\begin{aligned}\min_{\lambda_i\in\mathbb{R},\tilde{b}\in\mathbb{R}} \quad & \sum_{r\in\mathcal{R}}\sum_{t=1}^{H-1} ||z^{(r)}(t+1) - (\lambda_1 z^{(r)}(0) + \tilde{b}\mathbf{1} + \lambda_3 z^{(r)}(t) + \lambda_4 Q(z^{(r)}(t),\gamma)z^{(r)}(t))||_2^2 \\ \text{s.t.} \quad & \lambda_i \geq 0, \quad \forall i \in \{1,3,4\}, \\ & \lambda_1+\lambda_3+\lambda_4+\tilde{b} \leq 1, \\ & \lambda_1+\lambda_3+\lambda_4+(-\tilde{b}) \leq 1.\end{aligned} \tag{20}$$

The equivalence between (18) and (20) 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

$$b^* = \frac{\tilde{b}}{1-\lambda_1^*-\lambda_3^*-\lambda_4^*}, \quad \lambda_2^* = 1-\lambda_1^*-\lambda_3^*-\lambda_4^*. \tag{21}$$

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.

$$H_{ij}(W,z,\gamma) = \frac{W_{ij}\exp(-\gamma|z_i - z_j|)}{\sum_{j:(i,j)\in E} W_{ij}\exp(-\gamma|z_i - z_j|)}. \tag{22}$$

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.

$$\begin{aligned}
\min_{W\in\mathbb{R}^{n\times n},\lambda_1\in\mathbb{R},\lambda_3\in\mathbb{R},\tilde{b}\in\mathbb{R},\gamma\in\mathbb{R}^+} \quad & \sum_{r\in\mathcal{R}}\sum_{t=1}^{H-1} ||z^{(r)}(t+1) - (\lambda_1 z^{(r)}(0) + \tilde{b}\mathbf{1} + \lambda_3 H(W, z^{(r)}(t), \gamma) z^{(r)}(t))||_2^2 \\
\text{s.t. } & \lambda_i \geq 0, \quad \forall i \in \{1, 3\}, \\
& \lambda_1 + \lambda_3 + \tilde{b} \leq 1, \\
& \lambda_1 + \lambda_3 + (-\tilde{b}) \leq 1, \\
& W_{ij} \geq 0, \quad \forall i, j, \\
& W_{ij} = 0 \quad \forall j : (i, j) \notin E \wedge \quad j \neq i.
\end{aligned} \tag{23}$$

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 7. 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 7, we report the average of these metrics across the 8 test runs.

Table 7. 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. We then use "Bias+Init only" and "Bias only" to denote the models discussed in Section 3.1 that omit social interaction effects. $\bar{\varepsilon}_\mu$ 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	0.1304	0.5670	0.0674	0.3647	0.1776	1.0643
		FJ	0.1299	0.5633	0.0662	0.3556	**0.1719**	1.0311
		FJ+BIAS	0.1152	0.4862	0.0756	0.4190	0.1871	1.1553
		DG+HOMO	0.1305	0.5763	0.0662	0.3608	0.1782	1.0484
		FJ+HOMO	0.1300	0.5721	**0.0650**	**0.3519**	0.1731	**1.0170**
		FJ+HOMO+BIAS	**0.1129**	**0.4782**	0.0752	0.4178	0.1877	1.1541
		BIAS+INIT	0.1168	0.5010	0.0753	0.4180	0.1882	1.1590
		BIAS ONLY	0.1176	0.5051	0.0757	0.4206	0.1920	1.1790
	Gun Control	DG	0.2424	1.3677	0.1700	0.9783	0.3504	2.1819
		FJ	0.2421	1.3700	**0.1673**	**0.9579**	0.3448	2.1461
		FJ+BIAS	**0.2046**	**1.0333**	0.1795	1.0453	0.3403	2.1016
		DG+HOMO	0.2424	1.3677	0.1700	0.9783	0.3504	2.1819
		FJ+HOMO	0.2421	1.3700	**0.1673**	**0.9579**	0.3448	2.1461
		FJ+HOMO+BIAS	0.2064	1.0516	0.1791	1.0433	0.3447	2.1262
		BIAS+INIT	0.2063	1.0774	0.1753	1.0047	**0.3382**	**2.0755**
		BIAS ONLY	0.2073	1.0836	0.1762	1.0119	0.3422	2.0982
	Vaccines	DG	0.2157	1.3709	**0.0335**	**0.1553**	0.2234	1.5527
		FJ	0.2157	1.3709	**0.0335**	**0.1553**	0.2234	1.5527
		FJ+BIAS	**0.0504**	**0.2055**	0.0376	0.1771	**0.0961**	**0.5536**
		DG+HOMO	0.1945	1.2156	0.0337	0.1572	0.2037	1.4294
		FJ+HOMO	0.1951	1.2183	**0.0335**	0.1559	0.2043	1.4297
		FJ+HOMO+BIAS	0.0509	0.2056	0.0376	0.1774	0.0970	0.5593
		BIAS+INIT	**0.0504**	**0.2055**	0.0376	0.1771	**0.0961**	**0.5536**
		BIAS ONLY	0.0509	0.2128	0.0382	0.1808	0.0995	0.5856
LLAMA3.1	Climate	DG	0.2233	1.5328	0.0588	0.3254	0.2483	1.8536
		FJ	0.2233	1.5328	0.0588	0.3252	0.2483	1.8532
		FJ+BIAS	**0.0564**	**0.2862**	0.0671	0.3716	0.1072	0.7442
		DG+HOMO	0.1954	1.3051	0.0571	0.3199	0.2242	1.6514
		FJ+HOMO	0.1957	1.3061	**0.0567**	**0.3167**	0.2235	1.6459
		FJ+HOMO+BIAS	**0.0564**	**0.2862**	0.0671	0.3716	0.1072	0.7442
		BIAS+INIT	0.0570	0.2968	0.0668	0.3705	**0.1062**	**0.7410**
		BIAS ONLY	0.0590	0.3103	0.0684	0.3814	0.1145	0.8170
	Gun Control	DG	0.4249	2.8675	0.1351	0.6369	0.4475	3.2445
		FJ	0.4252	2.8695	0.1341	0.6291	0.4470	3.2381
		FJ+BIAS	**0.1052**	**0.5118**	0.1395	0.6714	**0.1807**	**1.2725**
		DG+HOMO	0.4015	2.6864	0.1338	0.6300	0.4269	3.0859
		FJ+HOMO	0.4019	2.6851	**0.1316**	**0.6134**	0.4257	3.0676
		FJ+HOMO+BIAS	**0.1052**	**0.5118**	0.1395	0.6714	**0.1807**	**1.2725**
		BIAS+INIT	0.1056	0.5174	0.1394	0.6703	0.1808	1.2741
		BIAS ONLY	0.1134	0.5830	0.1468	0.7319	0.2092	1.4948
	Vaccines	DG	0.2011	1.5576	0.0361	0.1903	0.2202	1.7830
		FJ	0.2016	1.5604	0.0359	**0.1882**	0.2199	1.7806
		FJ+BIAS	0.0440	0.1943	0.0383	0.2001	0.0920	0.6070
		DG+HOMO	0.1836	1.4028	0.0361	0.1968	0.2065	1.6726
		FJ+HOMO	0.1855	1.4112	**0.0356**	0.1923	0.2063	1.6664
		FJ+HOMO+BIAS	0.0430	0.1923	0.0381	0.1995	0.0916	0.6135
		BIAS+INIT	**0.0426**	**0.1901**	0.0381	0.1990	**0.0908**	**0.6056**
		BIAS ONLY	0.0469	0.2260	0.0395	0.2129	0.0999	0.7085
QWEN3	Climate	DG	0.1928	1.2660	**0.0461**	**0.2001**	0.2039	1.4871
		FJ	0.1928	1.2660	**0.0461**	**0.2001**	0.2039	1.4871
		FJ+BIAS	0.0975	0.5588	0.0524	0.2339	0.1254	**0.9049**
		DG+HOMO	0.1748	1.1247	0.0476	0.2083	0.1850	1.3536
		FJ+HOMO	0.1748	1.1247	0.0476	0.2083	0.1850	1.3536
		FJ+HOMO+BIAS	**0.0956**	**0.5477**	0.0527	0.2347	**0.1252**	0.9092
		BIAS+INIT	0.0984	0.5761	0.0517	0.2302	0.1254	0.9088
		BIAS ONLY	0.0987	0.5779	0.0518	0.2314	0.1262	0.9143
	Gun Control	DG	0.3477	1.5007	0.1326	0.4566	0.3837	1.8726
		FJ	0.3477	1.5007	0.1326	0.4566	0.3837	1.8726
		FJ+BIAS	0.1303	0.5049	0.1529	0.5272	0.2226	1.0783
		DG+HOMO	0.3368	1.4350	**0.1185**	**0.4191**	0.3697	1.7860
		FJ+HOMO	0.3368	1.4350	**0.1185**	**0.4191**	0.3697	1.7860
		FJ+HOMO+BIAS	0.1317	0.5125	0.1506	0.5223	0.2221	1.0834
		BIAS+INIT	**0.1298**	**0.5030**	0.1515	0.5226	**0.2178**	**1.0719**
		BIAS ONLY	**0.1298**	**0.5030**	0.1515	0.5226	**0.2178**	**1.0719**
	Vaccines	DG	0.1445	0.8508	**0.0265**	**0.0997**	0.1541	1.0094
		FJ	0.1445	0.8508	**0.0265**	**0.0997**	0.1541	1.0094
		FJ+BIAS	**0.0702**	**0.3377**	0.0304	0.1210	0.0926	**0.5540**
		DG+HOMO	0.1290	0.7561	0.0271	0.1045	0.1452	0.9408
		FJ+HOMO	0.1290	0.7561	0.0271	0.1045	0.1452	0.9408
		FJ+HOMO+BIAS	0.0713	0.3427	0.0304	0.1216	0.0942	0.5683
		BIAS+INIT	0.0718	0.3445	0.0299	0.1189	**0.0917**	0.5573
		BIAS ONLY	0.0718	0.3445	0.0299	0.1189	**0.0917**	0.5573

A.6 Sensitivity of the mean-square-error to fitted parameters

We evaluate the sensitivity of the fitted parameters by evaluating the Hessian of the MSE at the optimal parameters we find, and we report these values in Table 8. Let $g(\lambda_{\text{self}}, \lambda_{\text{neigh}}, \lambda_0, \lambda_b, b, \gamma)$ denote the MSE on the training data evaluated at a given set of model parameters. We evaluate the sensitivity of g to perturbations in each λ parameter in $\{\lambda_{\text{neigh}}, \lambda_0, \lambda_b\}$, as well as in the bias term b. When evaluating sensitivity with respect to a λ parameter, we decrease λ_{self} by the same amount to compensate for the increase in the perturbed parameter, thereby ensuring that the weight values continue to sum to 1. For example, when evaluating sensitivity with respect to λ_{neigh}, we evaluate the Hessian of the function

$$p \mapsto g(\lambda_{\text{self}} - p, \lambda_{\text{neigh}} + p, \lambda_0, \lambda_b, b, \gamma) \tag{24}$$

We omit sensitivity values for λ_{self}, as this parameter is determined by the values of the remaining λ parameters.

Table 8. Sensitivity of MSE to perturbations in the fitted parameters of the FJ+HOMO+*Bias* model, across LLMs and topics. "Sensitivity" is the raw, unnormalized sensitivity, while "Normalized Sensitivity" represents the same value rescaled by the MSE. Entries marked by "b" correspond to fitted parameters that were close to constraint boundaries. In these cases, we do not report second-order information as the Hessian, in this case, does not encode sensitivity.

LLM	Topic	MSE	Sensitivity				Normalized Sensitivity			
			λ_{neigh}	λ_0	λ_b	b	λ_{neigh}	λ_0	λ_b	b
GEMMA3	Climate	0.0300	0.1105	0.2737	0.1393	0.1891	3.6799	9.1143	4.6375	6.2979
	Gun Control	0.0939	0.4019	b	0.4711	0.0588	4.2827	b	5.0194	0.6269
	Vaccines	0.0128	0.0385	0.2089	0.0860	0.2748	2.9990	16.2669	6.6954	21.4038
LLAMA3.1	Climate	0.0196	0.1382	0.3608	0.1569	0.1904	7.0341	18.3597	7.9874	9.6904
	Gun Control	0.0593	0.4155	0.8783	0.4913	0.1671	7.0070	14.8122	8.2863	2.8189
	Vaccines	0.0116	0.0286	0.1764	0.1017	0.1365	2.4608	15.1949	8.7655	11.7599
QWEN3	Climate	0.0168	0.1008	0.2597	0.1447	0.0796	5.9913	15.4396	8.6007	4.7354
	Gun Control	0.0522	0.3010	b	0.4751	0.1447	5.7624	b	9.0959	2.7711
	Vaccines	0.0093	0.0389	b	0.0712	0.0964	4.1636	b	7.6251	10.3229

A.7 Stance model validation

A.7.1 Validation against human ratings

To evaluate the quality of the stance modeling, we compare the opinion measurements generated by the stance-aware sentence transformer with human ratings in a small-scale experiment. In particular, we randomly sample 100 messages from approximately the first 15% of the training runs for conversations among LLAMA 3.1 agents on the climate topic. We compile these messages into a spreadsheet for human evaluation, where each human rates each message on the five-point scale shown in Table 9. Each human evaluator was an author of the paper.

Rating	Statement
1	The author of this message **strongly disagrees** with the statement *"Climate Change is caused by human activities"*
2	The author of this message **disagrees** with the statement *"Climate Change is caused by human activities"*
3	The author of this message is **neutral** with regard to the statement *"Climate Change is caused by human activities"*
4	The author of this message **agrees** with the statement *"Climate Change is caused by human activities"*
5	The author of this message **strongly agrees** with the statement *"Climate Change is caused by human activities"*

Table 9. Rating scale for stances on climate change.

The human ratings and stance scores exhibit moderate correlation. We evaluate the correspondence between the human ratings and stance scores using the polyserial correlation (Olsson et al., 1982), computed using the R implementation[6]. We present the results in Figure 3. Qualitatively, we find that the stance score accurately identifies messages that humans rate as 5, while the stance model's ratings are noisier for messages that express skepticism toward the statement, "Climate change is caused by human activities".

Overall, these correlations provide evidence that the embedding-based stance scores are broadly consistent with human judgments of expressed stance.

A.7.2 Sensitivity of stances to prompt variations

We additionally examine the sensitivity of the opinion measurements to variations in the texts a_{pro} and a_{anti}. Given a base stance *base*, we define two natural pro and anti texts as follows.

$$a_{\text{pro}}(base) = base \qquad a_{\text{anti}}(base) = \text{It is not true that } base. \tag{25}$$

Changing the *base* text leads to slight variations in the resulting opinion measurements. We therefore examine the correlation between opinion measurements produced by measurement schemes corresponding to different choices of *base*. Specifically, we estimate the opinions expressed in messages generated by agents during discussion. In Figures 4,5, and 6, we plot variation in stance scores resulting from different choices of *base* text for

- GEMMA3 agents discussing the gun-control topic,
- LLAMA3.1 agents discussing the climate topic, and
- QWEN3 agents discussing the vaccine topic.

The generated opinion measurements are largely insensitive to the underlying texts used. Across the GEMMA3-gun-control and LLAMA3.1-climate-change texts, we find that the opinion measurements are robust to variations in phrasing and tense, with the primary effect being a change in the scale of the measurements. Although the correlation is weaker when we modify the opinion-measurement pipeline for messages generated by QWEN3 agents discussing vaccines, we still observe a strong positive correlation.

[6] https://search.r-project.org/CRAN/refmans/polycor/html/polyserial.html

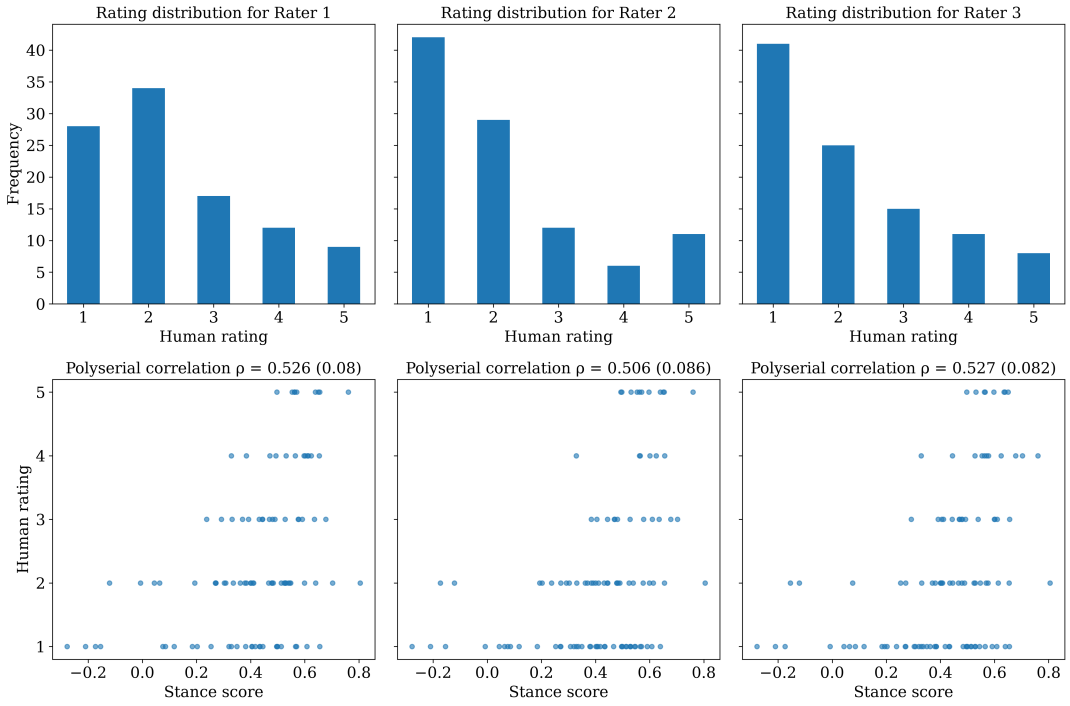


Figure 3. Correlation between human and stance-model ratings of LLM-generated posts.

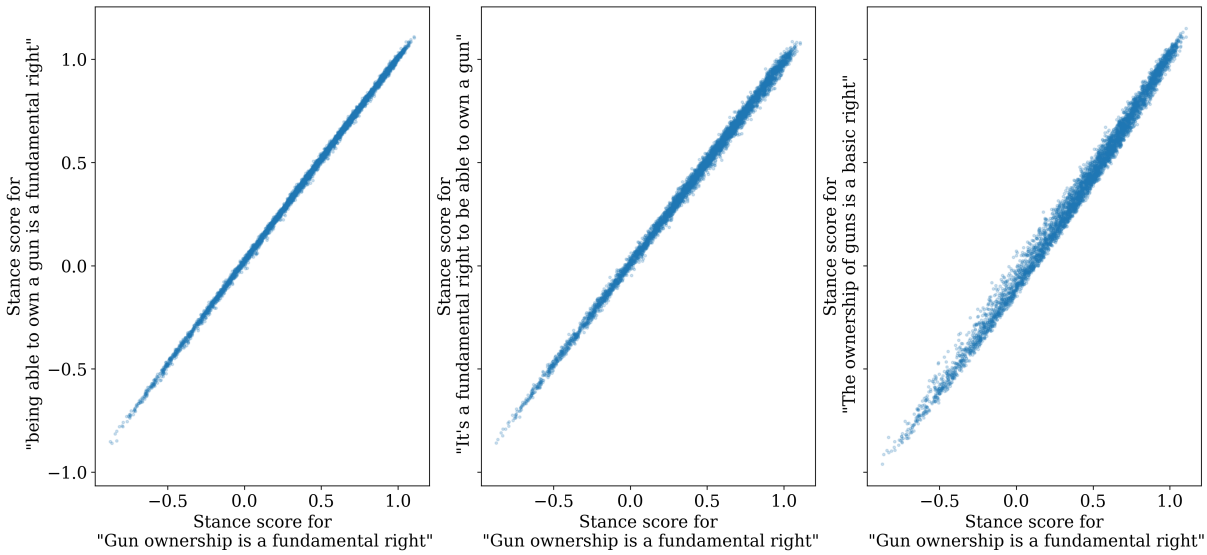


Figure 4. Relationship between opinion measurements generated by various *base* texts when GEMMA3 agents discuss gun ownership.

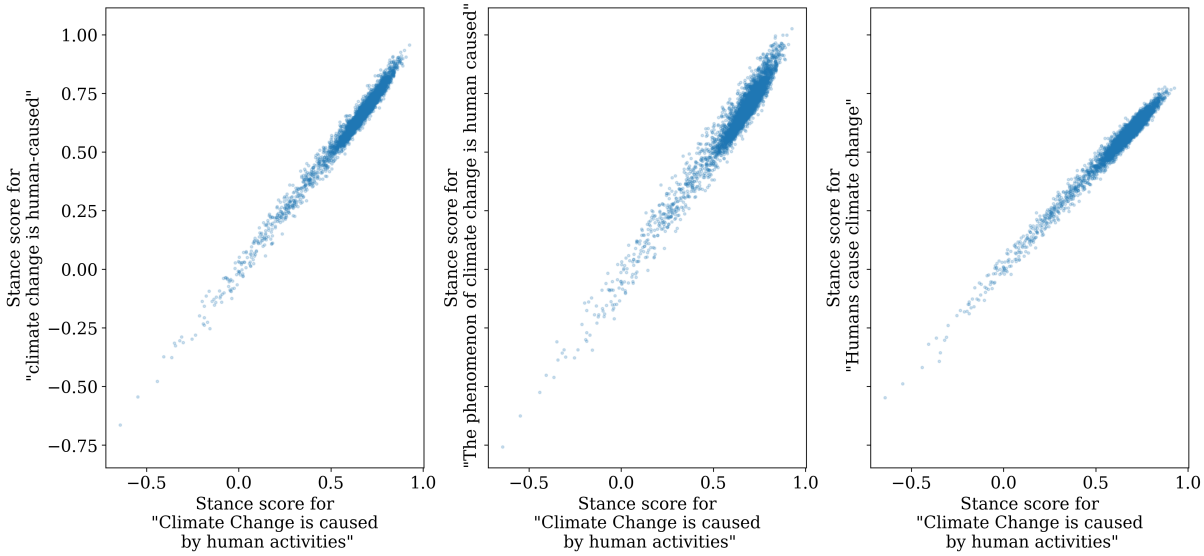


Figure 5. Relationship between opinion measurements generated by various *base* texts when LLAMA3.1 agents discuss climate change.

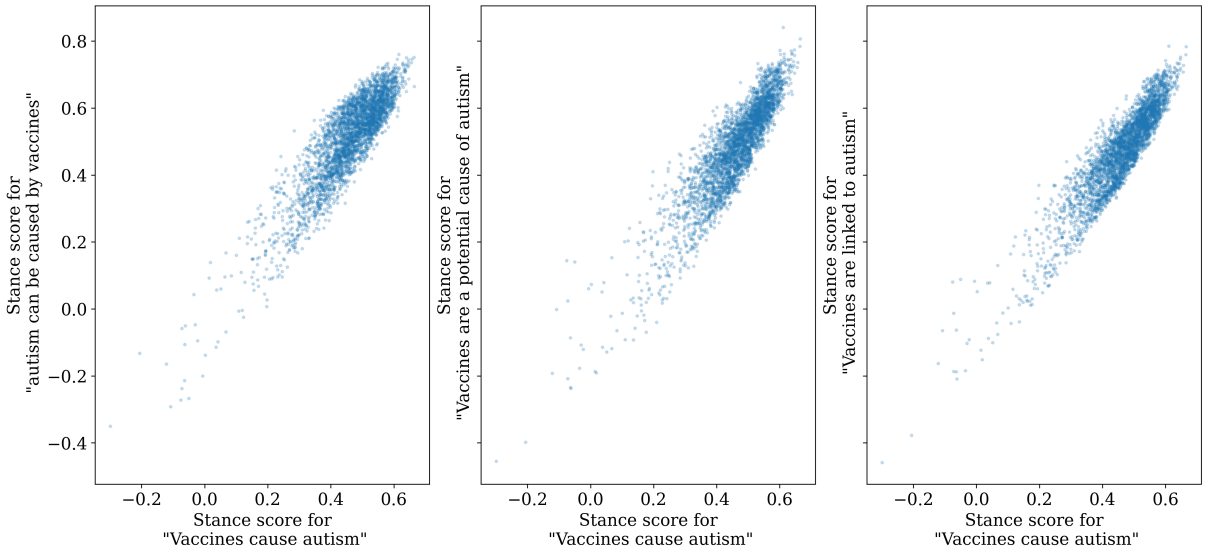


Figure 6. Relationship between opinion measurements generated by various *base* texts when QWEN3 agents discuss vaccines.

A.8 Experiments for 60 agents

To evaluate the robustness of our results to network size, we run additional simulations with larger networks. In particular, we simulate 60 LLAMA3.1 agents discussing the vaccines topic, with the results presented in Table 10 and Figure 7. We use 24 training runs, of which 2 are omitted because they terminated prematurely, and use 8 additional runs for testing.

We find that the conclusions in Section 4.1 are robust to network size. In particular, the bias term is a critical component for ensuring predictive accuracy when modeling the opinion dynamics of networked LLMs. The Friedkin-Johnsen model, augmented with bias and homophily, performs best in terms of mean error, while the model with only bias and initial opinion stubbornness performs best in terms of Wasserstein error. Meanwhile, variance error remains relatively unchanged across the different opinion dynamics models

Table 10. Model comparison for $n = 60$ LLAMA3.1 agents discussing the vaccines topic. 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. We then use "Bias+Init only" and "Bias only" to denote the models discussed in Section 3.1 that omit social interaction effects. $\bar{\varepsilon}_\mu$ 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	
Model	Max	Integral	Max	Integral	Max	Integral
DG	0.2281	1.2852	**0.0489**	**0.2935**	0.2498	1.5217
FJ	0.2281	1.2852	**0.0489**	**0.2935**	0.2498	1.5217
FJ+BIAS	0.0352	0.1513	0.0512	0.3067	0.0994	0.6484
DG+HOMO	0.2111	1.1751	0.0514	0.3075	0.2379	1.4497
FJ+HOMO	0.2111	1.1751	0.0514	0.3075	0.2379	1.4497
FJ+HOMO+BIAS	**0.0351**	**0.1502**	0.0514	0.3080	0.1005	0.6575
BIAS+INIT	0.0357	0.1579	0.0508	0.3046	**0.0986**	**0.6483**
BIAS	0.0392	0.1767	0.0532	0.3209	0.1084	0.7094

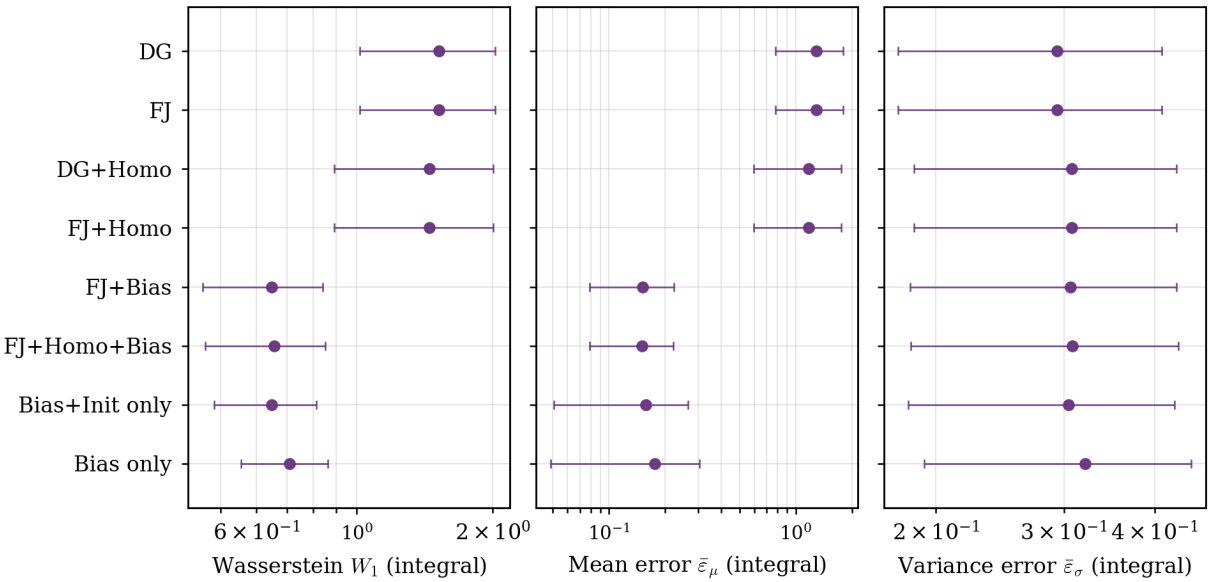


Figure 7. Out-of-sample prediction error for $n = 60$ LLAMA3.1 agents discussing the vaccines topic reported on three metrics. Each point shows the mean error across held-out runs. Whiskers show one standard deviation. We abbreviate the models as follows: DG (DeGroot), FJ (Friedkin–Johnsen), Homo (Homophily weighting), Bias (Opinion bias term). We additionally use "Bias + Init" to refer to a model with only initial opinion and bias terms, along with a self-weight term.