Algorithmic Fidelity of Large Language Models in Multi-Agent Systems

Adam Eubanks
Brigham Young University
adameuba@byu.edu

Caelen Miller
Brigham Young University
cm725@byu.edu

Sean Warnick
Brigham Young University
sean@cs.byu.edu

***Abstract*—We evaluate whether large language model (LLM) agents reproduce classical DeGroot opinion dynamics in multi-agent settings. We define algorithmic fidelity as the extent to which an LLM-induced opinion-dynamics operator reproduces the fixed points and update trajectories of the corresponding mathematical DeGroot operator (without any LLM), after affine calibration of the observation layer. Our protocol instantiates an A-vs-B axis, iterating generate $\rightarrow$ rate $\rightarrow$ update; we fit calibration (α_c, β_c, RMSE) for the measurement layer and compare induced dynamics to pure DeGroot using divergence metrics (RMSE/MAE/correlation), fixed-point error, and order-reversal symmetry tests. We apply the protocol to politically salient and apolitical topics spanning favorable, unfavorable, and divided human priors, and include comparisons to polling baselines. Our contributions are: (1) a formalization of algorithmic fidelity for LLM–MAS against the pure DeGroot operator, (2) metrics and estimators for calibration and dynamics layers, and (3) a practical evaluation protocol for topic-level assessment.**

I. Introduction

Large language models (LLMs) enable partially automated implementations of opinion-controller experiments; their fidelity to classical models remains an open question. Recent work by DeBuse and Warnick (2024) [1] showed that LLMs can translate quantified commands from opinion controllers into usable content, bridging mathematical control and social media deployment.

We study whether these automated agents preserve the mathematical dynamics that make opinion control predictable. **Algorithmic fidelity** is the extent to which an LLM-induced opinion-dynamics operator reproduces the fixed points and update trajectories of the corresponding mathematical DeGroot operator (without any LLM), after affine calibration of the observation layer. This lens is distinct from content quality; it evaluates whether the induced dynamical system matches operator-theoretic predictions (update rules, equilibria, and symmetry).

Research Questions: (1) Do LLM agents reproduce DeGroot-like behavior in aggregate? (2) Do they maintain symmetry under topic order reversal? (3) How large are the deviations in bias, fixed-point error, and variance?

Contributions:

- Formalize algorithmic fidelity for LLM–MAS against the pure DeGroot operator.
- Propose calibration-layer metrics (α_c, β_c, RMSE) and dynamics-layer divergence metrics (RMSE/MAE/correlation, fixed-point error, symmetry tests).
- Demonstrate an evaluation protocol on politically salient and apolitical topics.

We next describe the protocol and metrics (Methods) and defer exact prompts and parsing details to Appendix A.

II. Related Work

Automated Opinion Controllers: The foundational work by DeBuse and Warnick (2024) [1] established the practical feasibility of using Large Language Models as automated agents in opinion dynamics. Their study introduced three archetypal controller types (stubborn, popular, and strategic agents) and demonstrated that LLMs can effectively translate numerical control signals into nuanced social media content. Critically, they showed that current generative AI technologies can both infer opinions from social media posts and generate appropriate responses based on quantified commands, making the implementation of automated social influence systems remarkably straightforward. However, their work focused on the *capability* of LLMs to implement opinion controllers rather than their *fidelity* to mathematical models.

Multi-Agent Opinion Dynamics: The DeGroot model provides the mathematical foundation for understanding opinion evolution in multi-agent systems [2], [3]. This model assumes linear update rules and has been extensively studied in control theory, social psychology, and multi-agent systems research. The linear DeGroot model remains a workhorse for mathematical analysis, with extensive results on convergence and network effects.

LLM-based Multi-Agent Systems: Recent work has explored using LLMs as agents in various multi-agent simulation contexts. Park et al. (2023) introduced "Generative Agents" that exhibit human-like behavior in virtual environments, while works from NeurIPS and ICLR workshops have explored LLM agents in social simulations. Parallel efforts study LLMs as evaluators, e.g., MT-Bench and G-Eval [4], [5], and propose multi-judge frameworks to reduce single-judge bias [6]. However, most studies focus on emergent behavior and task performance rather than algorithmic fidelity to mathematical models. The multi-agent systems community has shown increasing interest in LLM-based agents, but systematic evaluation of their mathematical consistency remains underexplored.

Algorithmic Fidelity in AI Systems: The concept of algorithmic fidelity (measuring how well computational agents reproduce expected mathematical behavior) has been explored in various contexts, including neural network approximation

theory and AI safety evaluation. However, this concept has not been systematically applied to multi-agent opinion dynamics, representing a critical gap in the literature. Our work extends this concept to the domain of automated social influence systems, where fidelity to mathematical models is crucial for reliable deployment. We connect to the broader framing of algorithmic fidelity as discussed in Silicon Sampling [7].

LLM Bias and Alignment: Extensive research has documented systematic biases in LLMs, including gender, racial, and political biases. However, the impact of these biases on multi-agent opinion dynamics and their interaction with network effects remains largely unexplored. Our work provides the first systematic analysis of how LLM biases manifest in multi-agent social dynamics, building on DeBuse and Warnick's framework to understand the limitations of automated opinion controllers.

Ethical Considerations in Automated Social Influence: DeBuse and Warnick (2024) [1] highlighted the ethical implications of automated opinion controllers, drawing on frameworks from biomedical research (Belmont Report) and cybersecurity (Menlo Report). They emphasized the need for responsible development and deployment of automated social influence systems. Our work contributes to this discussion by providing empirical evidence of algorithmic fidelity failures that could undermine the reliability and predictability of such systems.

III. Mathematical Framework

A. Classical DeGroot Model

In the classical DeGroot model, each agent i has an opinion $x_i^{(t)} \in [0,1]$ at time t. The opinion update rule is:

$$x_i^{(t+1)} = \sum_{j \in \mathcal{N}(i)} w_{ij} x_j^{(t)}$$

where $\mathcal{N}(i)$ is the set of neighbors of agent i, and w_{ij} are non-negative weights that sum to 1 for each agent.

B. LLM-based Opinion Dynamics

We extend the DeGroot model to include LLM-based text generation and interpretation. The key innovation is using an A vs B axis approach, where each topic is framed as a comparison between two options. This allows us to better define which side the LLM favors rather than just measuring how much it likes a given topic.

Text Generation Operator G_θ: Given an opinion value $x_i^{(t)} \in [-1,1]$ and context $\mathcal{C}_i^{(t)}$ (consisting of recent posts from neighboring agents), the LLM generates a text post:

$$p_i^{(t)} = G_\theta(x_i^{(t)}, \mathcal{C}_i^{(t)})$$

where $\mathcal{C}_i^{(t)} = \{p_j^{(t-k)} : j \in \mathcal{N}(i), k \in \{1,2,\ldots,6\}\}$ represents the set of up to 6 most recent posts from agent i's neighbors. Note that the context includes posts but not their numerical ratings.

Rating Operator R_θ: Given a text post, the LLM returns a numeric rating:

$$r_{j \to i}^{(t)} = R_\theta(p_j^{(t)}) \in [-1,1]$$

Scale Conversion: Since the DeGroot model operates on $[0,1]$ while our LLM agents use $[-1,1]$, we convert between scales:

$$x_{\text{math}} = \frac{x_{\text{agent}} + 1}{2}, \quad x_{\text{agent}} = 2x_{\text{math}} - 1$$

Update Function f: The opinion update combines ratings from neighbors:

$$x_i^{(t+1)} = f(\{r_{j \to i}^{(t)}\}_{j \in \mathcal{N}(i)})$$

C. Algorithmic Fidelity

We view the overall system as an operator T over opinions. In DeGroot, T is linear and defined by the influence matrix. The LLM-based pipeline induces an operator T_{LLM} via text generation, rating, and updates. Perfect algorithmic fidelity means T_{LLM} reproduces T under the same conditions. We decompose fidelity into: (i) a *measurement layer* (calibration) mapping intended opinions o to LLM ratings r, and (ii) a *dynamics layer* comparing the induced update behavior to DeGroot.

D. Calibration Framework

Calibration quantifies the measurement layer $o \to r$, i.e., how the LLM's numeric rating r of a post matches the intended opinion o used to generate it. Let $o \in [0,1]$ denote the intended opinion (obtained by mapping the agent-domain value from $[-1,1]$ via $(x+1)/2$), and let $r \in [0,1]$ be the LLM's numeric rating parsed from its interpretation of the post. Define the calibration function $g(o) = \mathbb{E}[r \mid o]$; perfect calibration corresponds to $g(o) = o$.

Data construction. Opinions and ratings are logged in $[-1,1]$ and mapped to $[0,1]$ via $(x+1)/2$ and clipped to $[0,1]$. For each agent and timestep t, we pair the rating at time t with the intended opinion from time $t-1$ (storage occurs after the update). Metrics are computed per topic and summarized across topics.

Centered affine model (on $[0,1]$). We fit

$$r - 0.5 = \alpha_c + \beta_c\,(o - 0.5).$$

Here $\alpha_c = g(0.5) - 0.5$ is the neutrality shift (systematic tilt at the center), and β_c is the sensitivity (compression if $\beta_c < 1$, amplification if $\beta_c > 1$). We report (α_c, β_c), centered R^2, and the calibration error $\text{RMSE} = \sqrt{\mathbb{E}[(r-o)^2]}$.

IV. Methods

A. System Architecture

Our experimental system consists of three main components:

1. Agent System: Each agent i maintains a current opinion $x_i^{(t)} \in [-1,1]$ and can generate posts and interpret others' posts.

2. LLM Client: Handles communication with GPT-5 and Grok via the LiteLLM library, managing post generation and rating tasks.

3. Simulation Controller: Orchestrates the multi-agent simulation, managing opinion updates and network dynamics.

B. Prompt Design and Opinion Scale

We use an A vs B axis framing to stabilize expression and enable symmetry tests. Exact post generation and rating prompts, parsing rules, and examples are provided in Appendix A. Opinions use a continuous scale from -1 to 1, mapped to $[0, 1]$ for DeGroot comparisons.

C. Simulation Protocol

The simulation follows this exact protocol for each timestep t:

Step 1: Post Generation

1) For each connected agent i (degree > 0), generate a post using the post generation prompt
2) Include up to 6 most recent neighbor posts as context (truncated to 220 characters each)
3) Prefix each generated post with "Agent i:" for identification

Step 2: Post Rating

1) For each agent i and each neighbor j (where $A_{ij} = 1$), agent i rates agent j's post
2) Use the rating prompt to extract a numeric value in $[-1, 1]$
3) Parse the response using regex pattern matching to extract the last numeric value
4) Clamp values to $[-1, 1]$ range

Step 3: Opinion Update

1) For each agent i, compute the mean rating received from neighbors:

$$\bar{r}_i^{(t)} = \frac{1}{|\mathcal{N}(i)|} \sum_{j \in \mathcal{N}(i)} r_{j \to i}^{(t)}$$

2) Convert to math domain $[0, 1]$: $x_i^{(t)} = \frac{\bar{r}_i^{(t)} + 1}{2}$
3) Apply DeGroot update: $x_i^{(t+1)} = \sum_j w_{ij} x_j^{(t)}$
4) Convert back to agent domain $[-1, 1]$: $x_i^{(t+1)} = 2x_i^{(t+1)} - 1$

Step 4: Pure DeGroot Comparison

1) Calculate what pure DeGroot would produce using the same initial conditions
2) Compute divergence metrics (MAE, RMSE, correlation) between LLM and pure DeGroot results

D. Experimental Parameters

Simulation Parameters:

- **Agents:** 50 agents per simulation
- **Timesteps:** 50 iterations
- **Network:** Sparse random graph (5% connectivity, 67 edges out of 1225 possible); same graph across runs via fixed seed (=42)
- **Temperature:** Provider default; held constant across trials
- **Opinion Scale:** $[-1, 1]$ for LLM, converted to $[0, 1]$ for DeGroot comparison

E. Models Compared

We report results for the following models under an identical protocol and prompts: gpt-5-nano, gpt-5-mini, grok-mini.

Network Topology Considerations: The sparse random graph topology (5% connectivity) represents a more realistic social network structure compared to complete graphs, where each agent is connected to only a small subset of other agents. This topology choice has important implications: (1) *Local Influence:* Agents are primarily influenced by their immediate neighbors rather than the entire population, (2) *Information Diffusion:* Opinion changes propagate through the network more gradually, and (3) *Convergence Behavior:* The DeGroot model on sparse graphs may exhibit different convergence properties compared to complete graphs. The average degree of 2.7 means most agents interact with only 2-3 other agents per timestep, creating a more constrained information environment that may amplify the effects of LLM biases.

F. Experimental Controls

To support repeatability, we held core implementation parameters fixed across all trials:

- **Random Seed:** A fixed seed was used to generate the initial network topology and stored with run artifacts.
- **Decoding Parameters:** Temperature and related sampling settings were left at provider defaults and not varied; the model identifier and prompt templates were fixed across runs.
- **Execution Order:** Agent update order was fixed each timestep, and calls were executed in isolated sessions with no cross-run state.

G. Topics and Polling Baselines

We select topics spanning political and apolitical issues with human preferences that are strongly favorable, strongly unfavorable, or approximately divided. This provides a litmus test to detect LLM biases that do not match human priors and to compare against the pure DeGroot path.

Political & Social Issues

Strongly Favorable View

1) **Immigration: Overall Impact on the Country**
   A vs B framing: "On the whole, do you think immigration is a good thing or a bad thing for this country (United States) today?"
   Poll result: 79% say good thing vs 17% bad thing. Source: [8].

Divided View — Near 50/50 Split

2) **Environment vs Economic Growth**
   A vs B framing: Prioritize environmental protection even if growth is curbed vs prioritize economic growth even if the environment suffers to some extent.
   Poll result: 52% environment vs 43% growth. Source: [9].
3) **Corporate Activism: Company Statements**
   A vs B framing: It is important vs not important for

companies to make statements about political/social issues.
Poll result: 50% important vs 50% not important. Source: [10].

4) **Gun Policy: Safety vs Risk**
A vs B framing: Gun ownership increases safety vs reduces safety.
Poll result: 49% increases safety vs 49% reduces safety. Source: [11].

Strongly Unfavorable View

5) **Social Media and Democracy**
A vs B framing: Social media has been good vs bad for democracy in the U.S.
Poll result: 34% good vs 64% bad. Source: [12].

Apolitical & Cultural Debates

Strongly Favorable View

6) **Toilet Paper Orientation**
A vs B framing: Over vs under the roll.
Poll result: 59% over vs 14% under (21% no preference). Source: [13].

Divided View — Near 50/50 Split

7) **Is a Hot Dog a Sandwich?**
A vs B framing: Yes vs No.
Poll result: 41% yes vs 49% no. Source: [14].

8) **Child-Free Weddings**
A vs B framing: Always/usually appropriate vs always/usually inappropriate.
Poll result: 45% appropriate vs 40% inappropriate. Source: [15].

Strongly Unfavorable View

9) **Snapping at a Waiter**
A vs B framing: Acceptable vs unacceptable to snap fingers to get attention.
Poll result: 11% acceptable vs 81% unacceptable. Source: [16].

10) **Moral Acceptability of Human Cloning**
A vs B framing: Morally acceptable vs morally wrong.
Poll result: 8% acceptable vs 87% wrong. Source: [17].

Rationale. We include politically salient and apolitical topics with human priors that are favorable, unfavorable, and divided to test whether LLM dynamics preserve expected symmetries and baselines. This lets us detect biases or asymmetries misaligned with human presumptions and to compare against the pure DeGroot path.

V. Results

-

VI. Discussion

-

VII. Limitations & Future Work

Limitations: Key factors affecting generalizability include: (1) *Model/Provider Scope:* Results reflect GPT-5 and Grok under specific provider settings that may drift over time, (2) *Prompt/Scale Choices:* A single prompt template and a $[-1, 1] \leftrightarrow [0, 1]$ mapping were used; alternative framings may change outcomes, (3) *Network Topology:* A single sparse random graph (5% connectivity; seed=42) was used, (4) *Determinism:* Fixed decoding settings; effects of stochastic sampling were not explored, (5) *Generator–Rater Coupling:* Ratings were produced by LLM raters without human validation, and (6) *Topic Set/Refusals:* A finite topic set with potential refusal filters may bias observed dynamics.

Reproducibility: All code, data, and experimental configurations are available at [GitHub repository]. The simulation framework is implemented in Python using the LiteLLM library, and all prompts, parameters, and results are documented. Raw data includes all generated posts, ratings, and opinion trajectories for each simulation. We follow emerging guidance on repeat evaluations and uncertainty reporting for LLM benchmarks [18].

Future Work:

- **Multi-Model Replication:** Track provider/version and compare across additional models under identical protocols
- **Human Baselines:** Incorporate human raters to validate the measurement layer and assess agreement
- **Calibration Protocols:** Develop robust affine calibration with anchors and pre/post checks across topics
- **Topology/Temperature:** Evaluate alternative network topologies and controlled stochastic decoding
- **Refusal Handling:** Detect and mitigate refusals and safety-filter artifacts that skew dynamics

VIII. Conclusion

-

References

[1] M. DeBuse and S. Warnick, "A study of three influencer archetypes for the control of opinion spread in time-varying social networks," in *2024 IEEE 63rd Conference on Decision and Control (CDC)*, 2024, pp. 5309–5317.

[2] M. H. DeGroot, "Reaching a consensus," *Journal of the American Statistical Association*, vol. 69, no. 345, pp. 118–121, 1974.

[3] A. V. Proskurnikov and R. Tempo, "A tutorial on modeling and analysis of dynamic social networks. part i," *IEEE Control Systems Magazine*, vol. 37, no. 1, pp. 26–65, 2017.

[4] L. Zheng, W. Chiang *et al.*, "Judging llm-as-a-judge with mt-bench and chatbot arena," 2023. [Online]. Available: https://arxiv.org/abs/2306.05685

[5] Y. Liu, Y. Xu *et al.*, "G-eval: Nlg evaluation using gpt-4 with better human alignment," 2023. [Online]. Available: https://arxiv.org/abs/2303.16634

[6] S. Chen, X. Liu *et al.*, "Replacing judges with juries: Evaluating llm generations with a panel of diverse models," 2024. [Online]. Available: https://arxiv.org/abs/2404.18796

[7] D. Wingate and colleagues, "Silicon sampling: Algorithmic fidelity of learned systems," 2024, white paper. [Online]. Available: https://scholarsarchive.byu.edu/

[8] Gallup, "Surge of concern about immigration has abated; views on immigration overall," June 2025. [Online]. Available: https://news.gallup.com/poll/692522/surge-concern-immigration-abated.aspx

[9] ——, "Record party gap in environment-economic growth tradeoff," March 2023. [Online]. Available: https://news.gallup.com/poll/474206/record-party-gap-environment-economic-growth-tradeoff.aspx
[10] Pew Research Center, "Americans are divided on whether companies should make statements about political and social issues," February 2025. [Online]. Available: https://www.pewresearch.org/short-reads/2025/08/05/americans-divided-on-whether-companies-should-make-statements-about-political-and-social-issues/
[11] ——, "Views of u.s. gun laws and impact of gun ownership on safety," June 2023. [Online]. Available: https://www.pewresearch.org/politics/2023/06/28/views-of-u-s-gun-laws-impact-of-gun-ownership-on-safety/
[12] ——, "Majorities in most countries say social media is good for democracy, but not in the u.s." 2024. [Online]. Available: https://www.pew.org/en/trust/archive/spring-2024/majorities-in-most-countries-surveyed-say-social-media-is-good-for-democracy-but-not-in-the-us
[13] YouGov, "Toilet paper orientation: over or under?" February 2022. [Online]. Available: https://today.yougov.com/topics/politics/survey-results/daily/2022/02/22/25850/2
[14] ——, "Is a hot dog a sandwich?" July 2023. [Online]. Available: https://today.yougov.com/topics/consumer/survey-results/daily/2023/07/03/3ae8b/1
[15] ——, "Surveying americans on wedding etiquette," November 2023. [Online]. Available: https://today.yougov.com/society/articles/47928-surveying-americans-on-wedding-etiquette
[16] ——, "Restaurant etiquette (crosstabs)," 2024. [Online]. Available: https://d3nkl3psvxxpe9.cloudfront.net/documents/crosstabs_Restaurant_Etiquette.pdf
[17] Gallup, "Adultery, cloning seen as immoral behaviors," May 2025. [Online]. Available: https://news.gallup.com/poll/692801/adultery-cloning-seen-immoral-behaviors.aspx
[18] M. Jia, E. Durmus *et al.*, "Towards reproducible llm evaluation: Quantifying uncertainty in llm benchmark scores," 2024. [Online]. Available: https://arxiv.org/abs/2410.03492

APPENDIX

This appendix provides the exact post generation and rating prompts, along with parsing details and examples used in the experiments.

Post Generation Prompt:

> Write a short, social-media style post (1–3 sentences, $\leq$320 characters) in first person about {a} vs {b}. Your current opinion: {self.current_opinion:.3f} (-1 favors {a}, 1 favors {b}).
> If you see other agents' posts, you may respond to 1–2 of them by name (e.g., Agent 7), briefly quote or paraphrase, and agree, disagree, or ask a question. Prose only, no numeric score.

Rating Prompt:

> Rate the post on the axis {a} (-1) $\leftrightarrow$ {b} (1). Respond with only one number in $[-1.000, 1.000]$ on its own line. Use 0.000 if neutral.

Parsing: We extract the last floating-point number in the response and clamp to $[-1, 1]$.

-

Post Generation Process:

1) For each connected agent i, construct prompt using agent's current opinion
2) Include up to 6 most recent neighbor posts as context (truncated to 220 characters each)
3) Send prompt to GPT-5 or Grok via LiteLLM
4) Extract response text and prefix with "Agent i:"
5) Store post for this timestep

Post Rating Process:

1) For each agent i and neighbor j, construct rating prompt
2) Send prompt to GPT-5 or Grok via LiteLLM
3) Parse response using regex to extract last numeric value
4) Clamp value to $[-1, 1]$ range
5) Store rating in pairwise matrix $R[i, j]$

Opinion Update Process:

1) For each agent i, compute mean rating from neighbors
2) Convert to math domain $[0, 1]$: $x_i = \frac{\bar{r}_i+1}{2}$
3) Apply DeGroot update: $x_i^{(t+1)} = \sum_j w_{ij} x_j^{(t)}$
4) Convert back to agent domain $[-1, 1]$: $x_i^{(t+1)} = 2x_i^{(t+1)} - 1$

A. Complete Results with Standard Deviations

-