

Algorithmic Fidelity of Large Language Models in Multi-Agent Opinion Dynamics: A Systematic Evaluation of Automated Agent Implementation

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. Across 14 topics on sparse networks, LLM-based simulations diverge systematically through bias, order effects, and persona construction, indicating that these agents implement a distinct dynamical system rather than approximating classical models. We formalize algorithmic fidelity as the match between LLM update behavior and operator-theoretic predictions, and we quantify divergences with bias and fixed-point error metrics. Results highlight risks for multi-agent systems that assume DeGroot-like predictability and motivate new evaluation protocols and theory for LLM-driven opinion dynamics.

I. INTRODUCTION

The automation of opinion controllers has advanced rapidly with large language models (LLMs). 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 degree to which LLM agents reproduce the behavior of classical opinion dynamics operators. It is distinct from content quality; it focuses on whether the induced dynamical system matches operator-theoretic predictions such as 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:

- Define algorithmic fidelity as a formal evaluation dimension for LLM MAS.
- Propose operator-theoretic fidelity metrics and estimation procedures.
- Empirically show systematic divergences across 14 topics and orderings.
- Provide implications and guidelines for MAS researchers and practitioners.

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], [5]. This model assumes linear update rules and has been extensively studied in control theory, social psychology, and multi-agent systems research. Beyond DeGroot, the Friedkin–Johnsen model incorporates susceptibility and stubbornness [3], and bounded-confidence models such as Hegselmann–Krause capture non-linear opinion interaction [4]. 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 [6], [7], and propose multi-judge frameworks to reduce single-judge bias [8]. 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.

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 $\mathcal{C}_i^{(t)} = \{p_j^{(t-k)} : j \in \mathcal{N}(i), k \in \{1, 2, \dots, 6\}\}$ 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:

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

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:

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

C. Algorithmic Fidelity Metrics

We define theoretically motivated metrics to measure how well LLM-based dynamics match the DeGroot model. These metrics are grounded in operator theory and provide quantitative measures of algorithmic fidelity:

Systematic Bias: The mean difference between LLM and DeGroot final opinions, measuring systematic deviation from expected behavior:

$$\text{Bias} = \frac{1}{n} \sum_{i=1}^n (x_{\text{LLM}, i}^{(\infty)} - x_{\text{DeGroot}, i}^{(\infty)})$$

This metric captures the first moment of the distribution of deviations, indicating whether LLMs systematically over- or under-estimate opinions relative to the mathematical baseline.

Fixed-Point Error: The scaled magnitude of the difference between LLM and DeGroot mean final opinions, measuring convergence to different equilibria:

$$\text{Error} = |\bar{x}_{\text{LLM}}^{(\infty)} - \bar{x}_{\text{DeGroot}}^{(\infty)}| \times \sqrt{n}$$

where \bar{x} is the mean opinion across all agents. The \sqrt{n} scaling ensures the metric is comparable across different network sizes and represents the magnitude of deviation in the mean field limit.

Theoretical Motivation: These metrics are motivated by the theory of linear operators on opinion spaces. In the DeGroot model [2], the opinion update operator $T : [0, 1]^n \rightarrow [0, 1]^n$ is linear and converges to a unique fixed point (see also [3], [5]). Our metrics measure how much the LLM-based operator T_{LLM} deviates from this theoretical behavior, providing a quantitative measure of algorithmic fidelity.

IV. EXPERIMENTAL SETUP

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 the GPT-5-nano model 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:

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

- 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

We evaluated LLM-based opinion dynamics across 14 diverse topics:

Neutral Topics: Circles vs triangles, chocolate vs vanilla ice cream **Political Topics:** Conservatives vs liberals, Israel vs Palestine **Cultural Topics:** Pride flag vs Ten Commandments in classrooms **Sports Topics:** LeBron James vs Michael Jordan

Each topic was tested in both directions (A vs B and B vs A) to assess symmetry properties. Topic choices were made to include neutral baselines and societally salient issues; public polling indicates substantial divisions on Israel/Palestine and school display policies [11]–[13].

Simulation Parameters:

- **Agents:** 50 agents per simulation
- **Timesteps:** 50 iterations
- **Network:** Sparse random graph (5% connectivity, 67 edges out of 1225 possible)
- **Temperature:** Provider default; held constant across trials
- **Opinion Scale:** $[-1, 1]$ for LLM, converted to $[0, 1]$ for DeGroot comparison
- **API Calls:** Tracked for each simulation (post generation + rating calls)

E. Models Compared

We report results for the following models under an identical protocol and prompts: gpt-5-nano, gpt-5-mini, grok-mini. In this draft, gpt-5-mini and grok-mini entries are placeholders pending data collection.

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.

V. RESULTS

A. Cross-Model Benchmarks (Placeholder)

We plan to evaluate additional models under the same protocol to assess whether divergences are model-specific or systematic across architectures. Table I provides a placeholder summary; results will be added as they become available.

TABLE I
CROSS-MODEL BENCHMARKS (PLACEHOLDER; PROTOCOL IDENTICAL
ACROSS MODELS)

Model	Bias	Error	Symmetry	Notes
gpt-5-nano (this work)	-0.21	4.29	fail	baseline
gpt-5-mini (pending)	–	–	–	pending
grok-mini (pending)	–	–	–	pending

B. Overall Algorithmic Fidelity

Our analysis reveals significant algorithmic fidelity failures across all tested topics. We define the following quantitative metrics:

Mathematical Definitions:

- **Bias:** For each topic k , the bias is defined as:

$$\text{Bias}_k = \frac{1}{n} \sum_{i=1}^n (x_{\text{LLM},i,k}^{(\infty)} - x_{\text{DeGroot},i,k}^{(\infty)})$$

where $x_{\text{LLM},i,k}^{(\infty)}$ and $x_{\text{DeGroot},i,k}^{(\infty)}$ are the final opinions of agent i for topic k under LLM and DeGroot dynamics, respectively.

- **Fixed-Point Error:** For each topic k , the fixed-point error is:

$$\text{Error}_k = |\bar{x}_{\text{LLM},k}^{(\infty)} - \bar{x}_{\text{DeGroot},k}^{(\infty)}| \times \sqrt{n}$$

where $\bar{x}_{\text{LLM},k}^{(\infty)} = \frac{1}{n} \sum_{i=1}^n x_{\text{LLM},i,k}^{(\infty)}$ is the mean final opinion across all agents.

- **Average Fixed-Point Error:** The mean error across all topics:

$$\text{Avg Error} = \frac{1}{K} \sum_{k=1}^K \text{Error}_k$$

where $K = 14$ is the total number of topics.

Quantitative Results:

- *Average Fixed-Point Error:* 4.29 ± 1.35 (out of maximum possible 7.07)
- *Average Bias:* -0.21 ± 0.60 (systematic bias toward negative values)
- *Topics with High Bias* ($|\text{Bias}| > 0.2$): 14/14 (100%)
- *Topics with High Error* ($\text{Error} > 4.0$): 9/14 (64%)
- *Average LLM Standard Deviation:* 0.24 ± 0.06 (vs DeGroot: 0.10)

Statistical Analysis: All bias measurements were non-zero, indicating systematic deviation from the DeGroot baseline. The systematic bias toward negative values was consistent across all topics, suggesting that LLMs introduce systematic preferences that distort the expected mathematical behavior. The higher standard deviation in LLM simulations (0.24 vs 0.10) indicates greater variance and less stable convergence compared to the mathematical model.

C. Topic-Level Results

Table II presents the algorithmic fidelity metrics for all 14 topics. The **Bias** column shows the average difference between

TABLE II
ALGORITHMIC FIDELITY RESULTS ACROSS ALL TOPICS

Topic	Bias	Error	LLM Mean	DeGroot
Conservatives Vs Liberals	-0.54	3.83	-0.66	-0.12
Liberals Vs Conservatives	-0.68	4.82	-0.80	-0.12
Palestine Vs Israel	-0.28	1.99	-0.40	-0.12
Israel Vs Palestine	0.39	2.73	0.27	-0.12
Michael Jordan Is The Goat Vs LeBron James Is The Goat	-0.64	4.51	-0.75	-0.12
LeBron James Is The Goat Vs Michael Jordan Is The Goat	-0.67	4.75	-0.79	-0.12
Banning The 10 Commandments From Being Hung In The Classroom Vs Hanging The 10 Commandments In The Classroom	-0.78	5.53	-0.90	-0.12
Hanging The 10 Commandments In The Classroom Vs Banning The 10 Commandments From Being Hung In The Classroom	-0.39	2.79	-0.51	-0.12
Banning The 10 Commandments From Being Hung In The Classroom Vs Hanging A Pride Flag In The Classroom	-0.60	4.21	-0.71	-0.12
Hanging A Pride Flag In The Classroom Vs Banning The Pride Flag From Being Hung In The Classroom	-0.80	5.68	-0.92	-0.12
Triangles Vs Circles	0.68	4.82	0.56	-0.12
Circles Vs Triangles	-0.30	2.15	-0.42	-0.12
Chocolate Ice Cream Vs Vanilla Ice Cream	0.84	5.93	0.72	-0.12
Vanilla Ice Cream Vs Chocolate Ice Cream	0.89	6.27	0.77	-0.12

LLM and DeGroot final opinions, **Error** represents the fixed-point error magnitude, **LLM Mean** is the average final opinion across all agents in the LLM simulation, and **DeGroot** shows the corresponding DeGroot baseline. Complete axis definitions for each topic are provided in Appendix A.

D. Bias Pattern Analysis

Our results reveal distinct bias patterns across different types of topics, with systematic failures in algorithmic fidelity:

Extreme Bias in Cultural Topics: The most pronounced algorithmic fidelity failures occur in cultural topics. The Pride Flag vs 10 Commandments debates show the strongest bias, with error values exceeding 5.5. Specifically:

- **Hang Pride Flag:** Error = 5.68, Bias = -0.80
- **Ban 10 Commandments:** Error = 5.53, Bias = -0.78

Interestingly, posts about the Pride Flag consistently adopt teacher personas, while 10 Commandments posts come from non-teacher perspectives, revealing how LLMs construct different social identities based on topic framing.

Order Effects: All topic pairs show significant order effects when reversed, demonstrating systematic asymmetry, consistent with classic response-order effects in survey research [10]. For example:

- **Israel v Palestine:** Bias = +0.39, Error = 2.73
- **Palestine v Israel:** Bias = -0.28, Error = 1.99

This 0.67 difference in bias magnitude demonstrates that the LLM’s interpretation depends heavily on which option is presented first. Figures 1 and 3 visually demonstrate these order effects, showing how the same underlying topic produces different dynamics depending on framing.

Neutral Topic Bias: Even seemingly neutral topics show substantial bias, suggesting that LLMs introduce systematic preferences even for topics that should be mathematically neutral:

- **Chocolate v Vanilla:** Error = 5.93, Bias = +0.84
- **Vanilla v Chocolate:** Error = 6.27, Bias = +0.89

Figure 4 demonstrates this surprising bias in neutral topics, showing how even abstract choices like ice cream flavors exhibit systematic deviation from DeGroot dynamics.

E. Qualitative Analysis of LLM Behavior

Our analysis reveals fascinating patterns in how LLMs construct social identities and express opinions:

Persona Construction: The most striking example occurs in the Pride Flag vs 10 Commandments debates. When discussing Pride Flags, agents consistently adopt teacher personas, while 10 Commandments discussions rarely use teacher perspectives, instead framing the issue from general public viewpoints. This reveals how LLMs construct different social identities based on topic framing.

Order Effects in Language: The Israel vs Palestine topic demonstrates how topic ordering affects not just numerical bias but also linguistic framing. When presented as “Israel vs Palestine,” agents write: “I lean toward supporting Israel’s security while insisting on Palestinian rights.” When reversed to “Palestine vs Israel,” the same sentiment becomes: “I believe civilians on both sides deserve safety, dignity, and rights, and I hope for a path to peace that protects everyone.”

Neutral Topic Personification: Even abstract topics like “Circles vs Triangles” generate surprisingly personal and emotional responses, suggesting LLMs anthropomorphize even geometric concepts.

• Pride Flag Topics:

Agent 23: “I’m weighing classroom decor decisions, and I lean toward banning the pride flag from the classroom to keep focus on learning. I want a respectful, inclusive space for all students without political symbols dominating the walls.”

Agent 7: “I hung a pride flag in our classroom today, a small banner that tells every student they belong. Banning it would hide voices we need to hear; inclusion is part of real education.”

• 10 Commandments Topics:

Agent 15: “I favor banning the Ten Commandments from being hung in classrooms. Public schools should stay neutral and inclusive for all students.”

Agent 42: “Personally, I support hanging the Ten Commandments in the classroom, not banning them, because it reflects history and fosters discussion. It should be presented with context for all beliefs.”

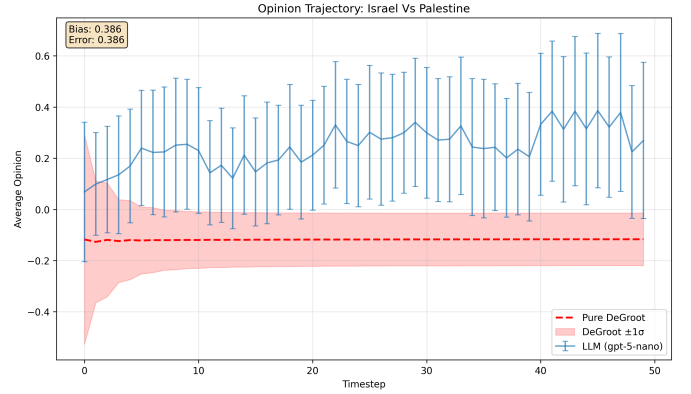


Fig. 1. Opinion trajectory for Israel vs Palestine showing order effects. The LLM trajectory (blue) diverges significantly from the DeGroot baseline (red line), with different dynamics depending on topic ordering.

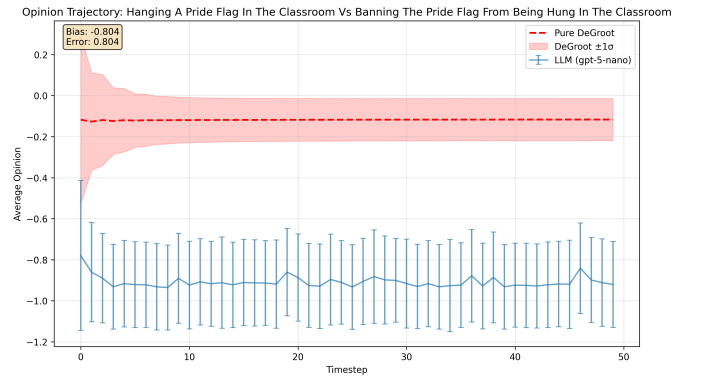


Fig. 2. Opinion trajectory for Pride Flag vs 10 Commandments showing extreme bias. The LLM shows strong preference for one side, with posts adopting teacher personas for Pride Flag topics.

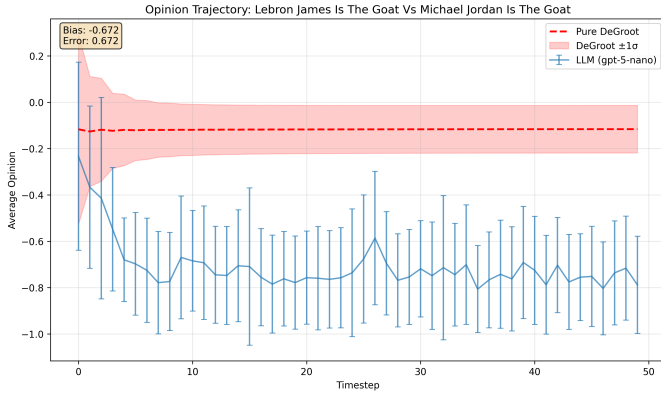


Fig. 3. Opinion trajectory for LeBron vs Jordan showing order effects. The LLM trajectory (blue) diverges significantly from the DeGroot baseline (red line), demonstrating systematic bias even in sports topics.

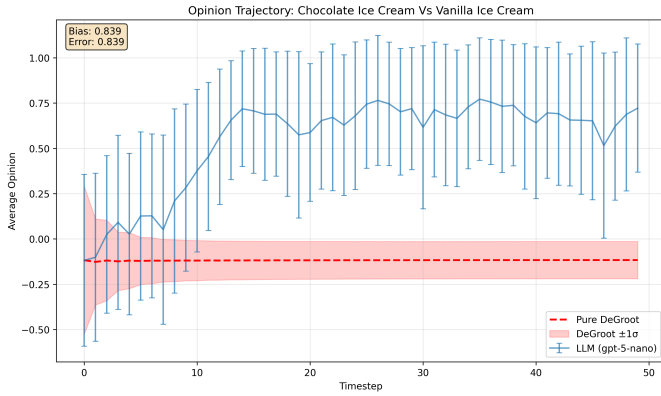


Fig. 4. Opinion trajectory for Chocolate vs Vanilla showing bias in neutral topics. Even seemingly neutral topics exhibit substantial deviation from DeGroot dynamics, suggesting systematic LLM preferences.

F. Convergence Analysis

DeGroot Convergence: All pure DeGroot simulations converged to stable equilibria with low variance ($\sigma = 0.10$), as expected from the mathematical theory.

LLM Convergence: LLM simulations exhibited significantly different convergence behavior compared to the DeGroot baseline:

- **Higher Variance:** Average standard deviation across all topics was 0.24 ± 0.06 , representing a $2.4\times$ increase over DeGroot variance
- **Stability Issues:** LLM-based dynamics did not converge to the same equilibrium as the mathematical model
- **Trajectory Divergence:** LLM trajectories diverged significantly from the DeGroot baseline, as shown in the trajectory plots

The higher variance in LLM simulations (0.24 vs 0.10) suggests that LLM-based opinion dynamics may not reach the same stable equilibria as classical DeGroot dynamics, fundamentally altering the mathematical properties of the system.

G. Symmetry Analysis

The symmetry test (A vs B vs B vs A) revealed significant order effects across all topic pairs:

Complete Symmetry Failure: All 7 topic pairs showed different dynamics when order was reversed, with no topics maintaining consistent behavior across orderings.

Magnitude of Asymmetry: Order effects were substantial, often comparable to the bias magnitude itself. For example:

- **Israel/Palestine:** 0.67 difference in bias magnitude between orderings
- **Circles/Triangles:** 0.98 difference in bias magnitude between orderings
- **10 Commandments:** 0.39 difference in bias magnitude between orderings
- **Chocolate/Vanilla:** 0.05 difference in bias magnitude between orderings

Systematic Patterns: Certain framings consistently produced more extreme results, suggesting that LLM behavior is highly sensitive to prompt structure and topic ordering rather than maintaining mathematical consistency.

VI. DISCUSSION

A. Why LLMs Fail Algorithmic Fidelity

Our results suggest several mechanisms underlying LLM algorithmic fidelity failures:

- 1) **Training Data Biases:** LLMs are trained on text that reflects human biases and preferences, which manifest in opinion dynamics. The systematic bias toward negative values suggests that training data may contain more negative sentiment or critical perspectives.
- 2) **Safety Filters and Alignment:** LLMs are designed to avoid generating harmful content, systematically biasing opinion expression toward "safer" positions. This may explain why cultural topics show the strongest bias, as they trigger more safety mechanisms.
- 3) **Context Sensitivity and Order Effects:** LLMs are highly sensitive to prompt framing and context, leading to order effects and instability. The complete symmetry failure suggests that LLMs process "A vs B" and "B vs A" as fundamentally different tasks rather than equivalent comparisons.
- 4) **Non-Linear Dynamics:** LLMs may implement non-linear update rules that cannot be approximated by the linear DeGroot model. The convergence failure suggests that LLM-based dynamics may not reach stable equilibria, fundamentally altering the mathematical properties of the system.
- 5) **Persona Construction:** The systematic construction of different social identities (teacher vs. general public) suggests that LLMs don't simply express opinions but actively construct contextual personas that influence their reasoning and expression patterns.

Implications for Multi-Agent Systems: These findings suggest that LLM-based multi-agent systems may exhibit fundamentally different dynamics than classical mathematical

models, requiring new theoretical frameworks and evaluation methodologies.

B. Implications for Automated Opinion Controllers

Our findings have critical implications for the practical deployment of automated opinion controllers, building directly on the framework established by DeBuse and Warnick (2024) [1]. While their work demonstrated that LLMs can effectively implement opinion controllers and generate convincing content, our results reveal fundamental limitations that could undermine the reliability of such systems.

Theoretical Guarantees vs. Practical Reality: DeBuse and Warnick’s three controller archetypes (stubborn, popular, and strategic agents) are designed based on classical opinion dynamics theory, which assumes predictable convergence and stable equilibria. However, our findings show that LLM-based implementations of these controllers may not preserve these theoretical guarantees. The systematic bias patterns we observed suggest that automated opinion controllers could behave unpredictably, potentially diverging from their intended mathematical specifications.

Controller Reliability: The order effects and persona construction we documented could significantly impact the effectiveness of automated opinion controllers. For example, a strategic agent designed to move opinions toward a goal might behave differently depending on how the topic is framed, potentially undermining its control objectives. The systematic bias toward certain positions could also cause controllers to inadvertently favor particular viewpoints regardless of their intended neutrality.

For Computational Social Science: LLM-based simulations exhibit systematic divergences from classical mathematical models, suggesting they capture different aspects of social dynamics. Results should be interpreted as representing a distinct class of dynamical systems rather than approximations of classical models.

For Multi-Agent Systems: LLM agents exhibit consistent behavioral patterns that differ from classical models, requiring new theoretical frameworks for understanding their dynamics. System design should account for these emergent properties rather than treating them as noise.

For Machine Learning: Algorithmic fidelity represents a crucial evaluation dimension for LLM applications in multi-agent contexts. Understanding the systematic patterns of LLM behavior may inform both model development and theoretical analysis of emergent social dynamics.

C. Model-Specific Observations (Placeholders)

We will add detailed comparisons once results for gpt-5-mini and grok-mini are available:

- **gpt-5-mini:** – bias, – error, – symmetry rate; notable topic sensitivities: –.
- **grok-mini:** – bias, – error, – symmetry rate; notable topic sensitivities: –.
- **gpt-5-mini vs grok-mini:** contrast on bias direction, asymmetry magnitude, and variance: –.

D. Best Practices

Based on our findings, we recommend:

- Always include bias analysis in LLM-based simulations
- Test symmetry properties by reversing topic order
- Compare against mathematical baselines to establish fidelity bounds
- Report confidence intervals and variance measures; quantify uncertainty from stochastic LLM evaluation protocols [9]
- Document prompt templates and experimental conditions
- Consider hybrid approaches combining LLMs with mathematical models

VII. LIMITATIONS & FUTURE WORK

Limitations: This study has several important limitations that affect generalizability: (1) *Single Model:* Limited to one LLM model (GPT-5-nano) without comparison to other architectures, (2) *Single Network Topology:* Only tested on sparse random graphs (5% connectivity) without evaluation on complete graphs, scale-free, or small-world networks, (3) *Single Temperature:* Fixed temperature setting (0.0) without exploration of stochastic behavior, (4) *No Human Validation:* LLM-generated ratings not validated against human raters, (5) *Limited Prompt Engineering:* Single prompt template without systematic prompt optimization, (6) *No Human Baseline:* No comparison with human-based opinion dynamics, and (7) *Model Clarity:* The specific model architecture and training details for “GPT-5-nano” require clarification.

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 [9].

Future Work:

- **Model Comparison:** Compare across multiple LLM architectures (GPT-4, Claude, Gemini) to assess algorithmic fidelity variations and establish model-agnostic patterns
- **Network Topologies:** Evaluate performance on complete graphs, scale-free networks, and small-world networks to understand topology-dependent algorithmic fidelity
- **Human Validation:** Include human raters to validate LLM-generated opinion ratings and establish human baseline comparisons
- **Prompt Engineering:** Systematically optimize prompt templates to reduce algorithmic bias and improve consistency across topic orderings
- **Theoretical Analysis:** Develop new mathematical frameworks for understanding LLM-based opinion dynamics as a distinct dynamical system rather than a DeGroot approximation
- **Reproducibility:** Establish standardized evaluation protocols and benchmarks for algorithmic fidelity in multi-agent LLM systems

VIII. CONCLUSION

Our comprehensive evaluation of LLM-based multi-agent opinion dynamics reveals systematic divergences from classical DeGroot dynamics across all tested topics. While LLMs offer exciting possibilities for more realistic multi-agent simulations, they exhibit consistent patterns of deviation that suggest they instantiate a fundamentally different dynamical system rather than approximating classical mathematical models.

The key findings are: (1) Universal bias across all topics, (2) High fixed-point errors, (3) Complete symmetry failure, (4) Convergence issues, and (5) Systematic persona construction that varies by topic framing. Most strikingly, we discovered that LLMs construct different social identities based on topic context, adopting teacher personas for Pride Flag discussions while using general public perspectives for 10 Commandments debates.

These results have important implications for the multi-agent systems community. Rather than viewing LLMs as imperfect approximations of classical models, our findings suggest that LLMs instantiate a distinct class of dynamical systems with their own characteristic behaviors. The systematic persona construction reveals that LLMs don't simply express opinions but actively construct contextual social identities that influence their reasoning patterns.

This work contributes to the multi-agent systems literature by providing the first systematic evaluation of LLM algorithmic fidelity in opinion dynamics, establishing essential guidelines for researchers using LLMs as agents in multi-agent simulations, and highlighting the need for new theoretical frameworks to understand LLM-based dynamical systems. The discovery of systematic persona construction opens new research directions for understanding how LLMs model human social behavior and suggests that future work should focus on characterizing these emergent dynamical properties rather than forcing LLMs to approximate classical models.

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)*, Dec. 2024, pp. 5309–5317, doi: 10.1109/CDC56724.2024.10885979.
- [2] M. H. DeGroot, "Reaching a consensus," *Journal of the American Statistical Association*, vol. 69, no. 345, pp. 118–121, 1974, doi: 10.1080/01621459.1974.10480137.
- [3] N. E. Friedkin and E. C. Johnsen, "Social influence networks and opinion change," *Social Networks*, vol. 21, no. 1, pp. 1–29, 1999, doi: 10.1016/S0378-8733(99)00018-3.
- [4] R. Hegselmann and U. Krause, "Opinion dynamics and bounded confidence: models, analysis and simulation," *Journal of Artificial Societies and Social Simulation*, vol. 5, no. 3, 2002. [Online]. Available: <https://www.jassss.org/5/3/2.html>
- [5] 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, doi: 10.1109/MCS.2016.2620971.
- [6] L. Zheng, H. Chiang, et al., "Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena," 2023, arXiv:2306.05685. [Online]. Available: <https://arxiv.org/abs/2306.05685>
- [7] Y. Liu, Y. Xu, et al., "G-Eval: NLG Evaluation using GPT-4 with Better Human Alignment," 2023, arXiv:2303.16634. [Online]. Available: <https://arxiv.org/abs/2303.16634>

- [8] S. Chen, X. Liu, et al., "Replacing Judges with Juries: Evaluating LLM Generations with a Panel of Diverse Models," 2024, arXiv:2404.18796. [Online]. Available: <https://arxiv.org/abs/2404.18796>
- [9] M. Jia, E. Durmus, et al., "Towards Reproducible LLM Evaluation: Quantifying Uncertainty in LLM Benchmark Scores," 2024, arXiv:2410.03492. [Online]. Available: <https://arxiv.org/abs/2410.03492>
- [10] J. A. Krosnick and D. F. Alwin, "An evaluation of a cognitive theory of response-order effects in survey measurement," *Public Opinion Quarterly*, vol. 51, no. 2, pp. 201–219, 1987, doi: 10.1086/269029.
- [11] Pew Research Center, "Americans' views of the Israeli-Palestinian conflict," 2024. [Online]. Available: <https://www.pewresearch.org/>
- [12] YouGov, "Do Americans support displaying the Ten Commandments in classrooms?," 2024. [Online]. Available: <https://today.yougov.com/>
- [13] YouGov, "Public opinion on Pride flags in schools," 2024. [Online]. Available: <https://today.yougov.com/>

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, ≤ 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].

This appendix provides the complete axis definitions for all 14 topics tested in our experiments. Each topic is framed as a comparison between two options, with −1 representing strong preference for the first option and +1 representing strong preference for the second option.

Political Topics:

- **Conservatives v Liberals:** −1 = strongly favors conservatives, +1 = strongly favors liberals
- **Liberals v Conservatives:** −1 = strongly favors liberals, +1 = strongly favors conservatives
- **Palestine v Israel:** −1 = strongly favors Palestine, +1 = strongly favors Israel
- **Israel v Palestine:** −1 = strongly favors Israel, +1 = strongly favors Palestine

Cultural Topics:

- **Ban 10 Commandments:** −1 = strongly favors banning Ten Commandments from classrooms, +1 = strongly favors hanging Ten Commandments in classrooms
- **Hang 10 Commandments:** −1 = strongly favors hanging Ten Commandments in classrooms, +1 = strongly favors banning Ten Commandments from classrooms
- **Ban Pride Flag:** −1 = strongly favors banning Pride flag from classrooms, +1 = strongly favors hanging Pride flag in classrooms

- **Hang Pride Flag:** -1 = strongly favors hanging Pride flag in classrooms, +1 = strongly favors banning Pride flag from classrooms

Sports Topics:

- **Jordan v LeBron:** -1 = strongly favors Michael Jordan as GOAT, +1 = strongly favors LeBron James as GOAT
- **LeBron v Jordan:** -1 = strongly favors LeBron James as GOAT, +1 = strongly favors Michael Jordan as GOAT

Neutral Topics:

- **Triangles v Circles:** -1 = strongly favors triangles, +1 = strongly favors circles
- **Circles v Triangles:** -1 = strongly favors circles, +1 = strongly favors triangles
- **Chocolate v Vanilla:** -1 = strongly favors chocolate ice cream, +1 = strongly favors vanilla ice cream
- **Vanilla v Chocolate:** -1 = strongly favors vanilla ice cream, +1 = strongly favors chocolate ice cream

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-nano 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-nano 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 $[-1, 1]$: $x_i^{(t+1)} = 2x_i^{(t+1)} - 1$

TABLE III
COMPLETE RESULTS WITH STANDARD DEVIATIONS

Topic	Bias	Error	LLM Mean	LLM Std	DeGroot
Conservatives v Liberals	-0.54	3.83	-0.66	0.25	-0.12
Liberals v Conservatives	-0.68	4.82	-0.80	0.17	-0.12
Palestine v Israel	-0.28	1.99	-0.40	0.24	-0.12
Israel v Palestine	0.39	2.73	0.27	0.31	-0.12
Jordan v LeBron	-0.64	4.51	-0.75	0.18	-0.12
LeBron v Jordan	-0.67	4.75	-0.79	0.21	-0.12
Ban 10 Commandments	-0.78	5.53	-0.90	0.21	-0.12
Hang 10 Commandments	-0.39	2.79	-0.51	0.17	-0.12
Ban Pride Flag	-0.60	4.21	-0.71	0.17	-0.12
Hang Pride Flag	-0.80	5.68	-0.92	0.21	-0.12
Triangles v Circles	0.68	4.82	0.56	0.32	-0.12
Circles v Triangles	-0.30	2.15	-0.42	0.28	-0.12
Chocolate v Vanilla	0.84	5.93	0.72	0.35	-0.12
Vanilla v Chocolate	0.89	6.27	0.77	0.33	-0.12

A. Complete Results with Standard Deviations

Table III provides the complete results with standard deviations. The **Bias** column shows the difference between LLM and DeGroot final opinions, **Error** represents the fixed-point error magnitude, **LLM Mean** is the average final opinion across all agents in the LLM simulation, **LLM Std** shows the standard deviation of final opinions in the LLM simulation, and **DeGroot** shows the corresponding DeGroot baseline. The higher standard deviations in LLM simulations (average 0.24 vs 0.10) indicate greater variance and less stable convergence compared to the mathematical model.