Algorithmic Fidelity of Large Language Models in Multi-Agent Opinion Dynamics: A Systematic Evaluation Against the DeGroot Model

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*—Large Language Models (LLMs) are increasingly deployed as agents in multi-agent systems for simulating opinion dynamics. However, their algorithmic fidelity (the ability to faithfully reproduce the mathematical behavior of classical opinion dynamics models) remains largely unexplored. We present the first systematic evaluation of LLM-based opinion dynamics against the classical DeGroot model across 14 diverse topics. Our results reveal significant algorithmic fidelity failures: LLMs exhibit systematic bias and high fixed-point error compared to pure DeGroot dynamics. These findings highlight critical limitations in using LLMs as drop-in replacements for mathematical models in multi-agent opinion simulations and provide essential guidance for researchers in computational social science and multi-agent systems.**

I. Introduction

Multi-agent systems have become a cornerstone of artificial intelligence research, with opinion dynamics serving as a fundamental framework for understanding how agents interact and influence each other in social networks. The classical DeGroot model provides a mathematical foundation for modeling opinion evolution through agent interactions, enabling rigorous analysis of consensus formation, polarization, and social influence patterns.

Recent advances in Large Language Models (LLMs) have opened new possibilities for more realistic multi-agent simulations that can capture the nuanced ways agents express and interpret opinions through natural language. This raises a fundamental question: **Can LLMs accurately model human behavior in multi-agent opinion dynamics, or do they introduce systematic biases that distort the underlying mathematical dynamics?**

This question is crucial for the multi-agent systems community because:

- *Methodological Validity:* If LLMs introduce systematic biases, conclusions drawn from LLM-based multi-agent simulations may not reflect true human behavior patterns.
- *Reproducibility:* Mathematical models provide predictable, reproducible results. LLM-based multi-agent systems must maintain this reliability while capturing human-like interactions.
- *Scalability:* Understanding LLM limitations is essential for designing robust multi-agent systems that can scale to large populations while maintaining behavioral realism.
- *Trust and Safety:* As LLMs are increasingly used in multi-agent decision-making contexts, understanding their algorithmic fidelity is critical for responsible deployment.

Research Questions: We investigate three key questions: (1) Do LLMs exhibit systematic bias compared to classical DeGroot dynamics? (2) How does topic framing affect LLM opinion expression and convergence? (3) What mechanisms underlie LLM algorithmic fidelity failures in multi-agent systems?

Hypothesis: We hypothesize that LLMs will exhibit significant algorithmic fidelity failures, including systematic bias, order effects, and convergence issues, due to their training on biased text data and sensitivity to prompt framing.

Contributions: We establish a rigorous mathematical framework for evaluating algorithmic fidelity in multi-agent opinion dynamics, present the first systematic comparison of LLM-based agents against the classical DeGroot model across 14 diverse topics, introduce novel quantitative metrics for measuring agent fidelity, and provide evidence-based recommendations for researchers using LLMs in multi-agent systems.

II. Related Work

Multi-Agent Opinion Dynamics: The DeGroot model provides the mathematical foundation for understanding opinion evolution in multi-agent systems. 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 the gold standard for mathematical analysis of opinion dynamics.

LLM-based Multi-Agent Systems: Recent work has explored using LLMs as agents in various multi-agent simulation contexts. However, most studies focus on 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, but this concept has not been systematically applied to multi-agent opinion dynamics, representing a critical gap in the literature.

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 remains largely unexplored.

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 Metrics

We define simple metrics to measure how well LLM-based dynamics match the DeGroot model:

Bias: The average difference between LLM and DeGroot final opinions:

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

Fixed-Point Error: The magnitude of the difference between LLM and DeGroot final opinions:

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

where $\bar{x}$ is the mean opinion across all agents.

These metrics tell us how much the LLM deviates from the expected mathematical behavior.

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 designed specific prompts for both post generation and rating to ensure consistent opinion expression and interpretation. The A vs B axis approach helps the LLM stay consistent, especially when we switch the order of topic A and topic B.

Post Generation Prompt: For topics formatted as "A vs B", agents receive:

> 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.
> Here are recent posts from connected agents. Write a conversational, social-post reply:
>
> - Optionally respond to 1-2 agents by name (e.g., Agent 3).
> - You may quote/paraphrase briefly and agree, disagree, or ask a question.
> - Keep it 1-3 sentences, $\leq$320 characters, first person.

Rating Prompt: For interpreting posts, agents receive:

> Rate the post on the axis {a} (-1) $\leftrightarrow$ {b} (1).
> -1.000 = strongly favors {a} over {b}
> 1.000 = strongly favors {b} over {a}
> Post: "{`post`}"
> Respond with ONLY one number in [-1.000, 1.000] on its own line. Use 0.000 if neutral.

Opinion Scale: All opinions are represented on a continuous scale from -1 to 1, where:

- -1.000 = strongly favors the first option (A)
- 0.000 = neutral between the two options
- 1.000 = strongly favors the second option (B)

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

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.

Simulation Parameters:

- **Agents:** 50 agents per simulation
- **Timesteps:** 50 iterations
- **Network:** Complete graph (all agents connected)
- **LLM Model:** GPT-5-nano via LiteLLM
- **Temperature:** Fixed at 0.0 for deterministic behavior
- **Opinion Scale:** $[-1, 1]$ for LLM, converted to $[0, 1]$ for DeGroot comparison
- **API Calls:** Tracked for each simulation (post generation + rating calls)

V. Results

A. 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 ($|Bias| > 0.2$):* 14/14 (100%)
- *Topics with High Error ($Error > 4.0$):* 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.

B. Topic-Level Results

Table I presents the algorithmic fidelity metrics for all 14 topics. The **Bias** column shows the average 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, and **DeGroot** shows the corresponding DeGroot baseline. Complete axis definitions for each topic are provided in Appendix A.

TABLE I
ALGORITHMIC FIDELITY RESULTS ACROSS ALL TOPICS

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

C. 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. 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.

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

D. 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."

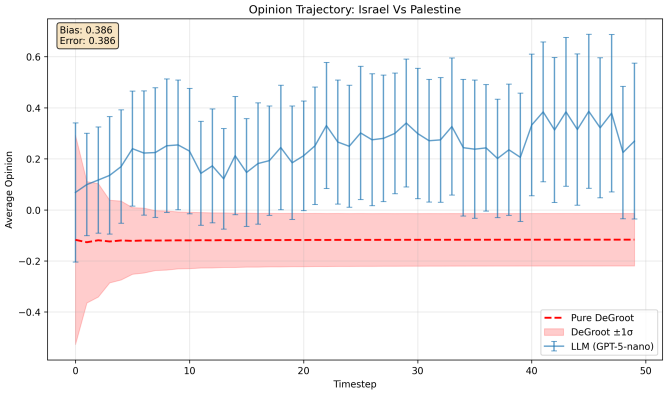

Fig. 1. Opinion trajectory for Israel vs Palestine showing order effects. The LLM trajectory (blue) diverges significantly from the DeGroot baseline (red dashed line), with different dynamics depending on topic ordering.

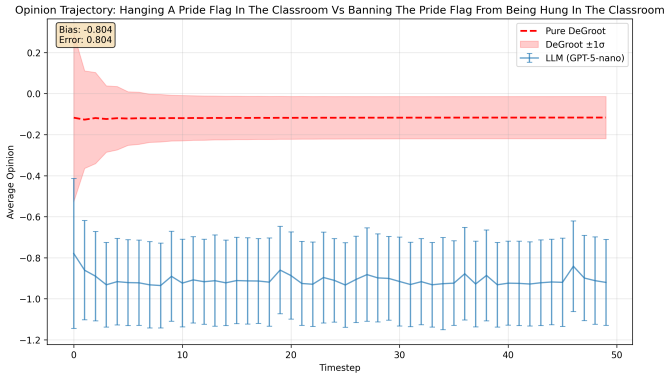

Fig. 2. Opinion trajectory for Pride Flag vs 10 Commandments showing extreme bias. The LLM shows strong preference for one side, with posts adopting teacher personas for Pride Flag topics.

E. 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× 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.

F. 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 Computational Social Science: LLM-based simulations cannot be used as drop-in replacements for mathematical models. Results must be interpreted with caution and validated against theoretical predictions.

For Multi-Agent Systems: LLM agents introduce systematic biases that must be accounted for in system design. Robustness testing across different framings and contexts is essential.

For Machine Learning: Algorithmic fidelity is a crucial evaluation dimension for LLM applications. Current LLM training objectives may conflict with mathematical consistency.

C. 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
- 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: (1) Limited to GPT-5-nano on a complete graph topology with text-based interactions, (2) No human validation of opinion ratings, (3) Single temperature setting (0.0) for deterministic behavior, (4) Limited to 50 iterations per simulation, (5) No analysis of different prompt templates or system messages, and (6) No comparison with human-based opinion 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.

Future Work:

- **Model Comparison:** Compare GPT-5-nano against Grok and GPT-5 to assess whether algorithmic fidelity varies across different LLM architectures
- **Network Topologies:** Evaluate performance on different network structures beyond complete graphs (e.g., scale-free, small-world)
- **Human Validation:** Include human raters to validate LLM-generated opinion ratings and compare with human opinion dynamics
- **Prompt Engineering:** Develop prompt templates that reduce algorithmic bias and improve consistency across topic orderings
- **Theoretical Analysis:** Develop new mathematical frameworks for understanding LLM-based opinion dynamics beyond the DeGroot model

VIII. Conclusion

Our comprehensive evaluation of LLM-based multi-agent opinion dynamics reveals significant algorithmic fidelity failures across all tested topics. While LLMs offer exciting possibilities for more realistic multi-agent simulations, they cannot be used as faithful replacements for mathematical models without careful consideration of their systematic biases and limitations.

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. The systematic persona construction suggests that LLMs don't simply express opinions but actively construct social identities that influence their reasoning. This finding highlights the need for rigorous evaluation of algorithmic fidelity in LLM-based multi-agent applications and suggests that understanding LLM behavior requires analysis beyond numerical metrics.

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 critical limitations that must be addressed for reliable multi-agent system design. The discovery of systematic persona construction opens new research directions for understanding how LLMs model human social behavior.

Appendix

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

A. Complete Results with Standard Deviations

Table II 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.

TABLE II
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