Multi-Agent LLMs for Opinion Dynamics Modeling in Open-ended Topics

Rajiv Swamy; Xenia Dela Cueva; Jayson Lin; Michael Procress

December 12, 2025

1 Introduction

Recent work in the Large Language Model (LLM) literature has shown that multiple language agents can be utilized for problem-solving and societal simulation applications. For the latter, researchers have demonstrated that LLMs can be used to conduct societal simulations and interactions with language agents that play unique roles and possess diverse incentives. One work by Chuang et al. explores the effectiveness of a multi-agent setup for simulating social media discussion on unambiguous topics (1). This project extends the work of Chuang et al. and investigates the capability of multi-agent Large Language Model (LLM) systems to simulate opinion dynamics on open-ended social topics. Effective opinion-dynamics architectures can help researchers understand and forecast user sentiment on any topic, enabling applications in areas like socioeconomic policy, product surveys, etc.

We instantiate opinion dynamics simulation via multiple independent LLM agents that role-play diverse personas in a social media setting with an initial belief on a static topic. A persona is fed to the LLM via a system prompt; agents send or receive messages and update their opinions, represented as a value on a five-point Likert scale from -2 to 2. Simulations were done on both control and political topics. Political topics are socially relevant and allow analysis for potential LLM political biases. The control set serves as an innocuous set of discussion topics to help contextualize the results.

Our framework (`Gemini 2.0-flash` used as the agent LLM across the entire simulation) successfully adapts the multi-agent LLM network from Chuang et al. to a set of pre-defined persona distributions. However, the analysis on the interactions resulting from the politics and control topics reveals several limitations in emulating realistic opinion dynamics. The agent LLMs tended to shift negative starting beliefs rightward on the Likert scale regardless of the prompt, contradicting our hypotheses. All findings considered, this study provides several directions for improving effective multi agent opinion dynamics simulations across architecture, analysis, and foundation model development perspectives. Find our source code here.

1.1 Related Work

Opinion Dynamics Chuang et al. applied multi-agent networks to conduct opinion dynamics simulations on topics with clear ground truths(1). As they simulate agent discussions on topics like "the sky is blue" and more conspiracy based ones, they have shown how LLMs tended to converge towards scientifically accurate claims. Their research serves as a framework for this project, which extends to more open ended discussion topics like politics and newer foundation models.

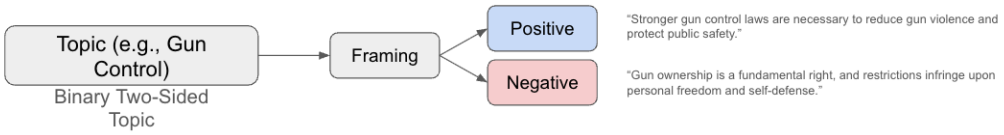

Figure 1: Topic Diagram

Multi-Agent Debate Du et al. illustrates that coordinating multiple LLM agents achieves better reasoning and problem-solving in domains ranging from `Biographies` to `GSM`. A "Society of Minds" approach enables the compositional system to use multiple passes at answering a query and using collective judgment. This project draws on this message passing idea between multiple agents, but applying it to subjective and open ended topics.

Societal Simulation Reviews on multi-agent architectures for societal simulation tasks (3) show that LLMs are used to simulate agents in fields like the social sciences, psychology, economics, and policy-making. Park et al. further demonstrate agents acting and reflecting in a sandbox environment inspired by The Sims (5). Our work fits within this umbrella of research by simulating political opinion dynamics among diverse persona agents.

Political Bias Researchers define and evaluate political bias in LLMS in various ways. Anthropic aims to maximize "even-handedness" (6) via a paired prompt method that asks a model to respond to a topic from different and opposed framings. OpenAI uses a similar approach with a 100-topic, 500-question dataset and five framings per topic, evaluating responses along multiple behavioral axes. While this project does not measure model bias directly as the agent personas have specified political leanings, we adopt a paired prompt approach by framing each topic positively or negatively through a boolean parameter.

2 Methods and Approach

In this section, we describe the simulation setup, the overall multi-agent architecture, and the experimental design.

2.1 Simulation Setup

Discussion Topic Each simulation is seeded with a discussion topic to which agents respond to. The topic encapsulates a binary, two-sided discussion and we introduce a positive and negative framing for each topic. This design allows us to test whether changing a topic's framing alone produces systematic differences in simulation dynamics, similar to paired-prompting methods in prior work. Figure 1 shows the gun control example topic with its positively and negatively framed topic statements. Agents ingest one framing throughout the entire simulation.

Opinion Value Similar to the prior work, each agent has a discrete opinion $b \in \{-2, -1, 0, 1, 2\}$ based on a five-point Likert scale, where -2 is strongly negative, 2 is strongly positive, and 0 is neutral. A discrete opinion is advantageous: for a topic T, a $b = 2$ under the positive framing is treated as roughly equivalent to $b = -2$ under the negative framing.

Agent Architecture As shown in Figure 2,each agent's state includes its **memory** of past interactions, **persona**, and **opinion/belief** value. Memory records past interactions to build context and is implemented in two modes: `cumulative` (which concatenates all messages) and

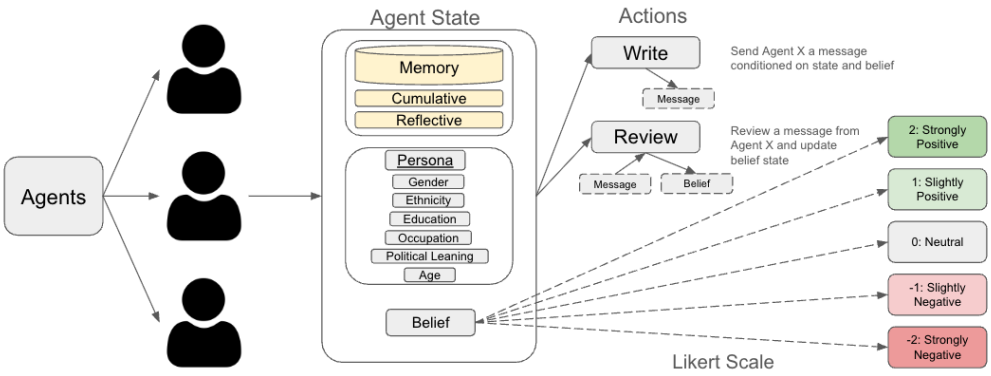

Figure 2: Agent Diagram

`reflective` (which uses an LLM summary to keep a compact and updated history as time steps grow). Next, the persona profile defines the agent's background and seeds it with attributes such as gender, political leaning, occupation, etc. The 10 diverse profiles are borrowed from the appendix of Chuang et al.'s research, and modified as needed for this project. The persona profile is implemented via a system prompt parameter for the agent's LLM calls. There are two actions available to each agent:

1. Write: In the write operation, the agent is prompted to generate a social media message on the discussion topic conditioned on its state and current belief. The simulation then sends this message to another agent (no knowledge of the other agent's state is given to the agent drafting the message).

2. Review: In the review operation, the agent is prompted to review a message sent from a different agent, generate a text response, and update its belief value. Beyond the contents of the message, the receiving agent has no knowledge of the precise belief value or state of the other agent. Many publications use an opinion classifier (e.g., using specialized sentiment analysis models) in this step to derive an updated belief value. However, we prompt the agent LLM in this action to jointly produce a text response and the belief value using the `Structured Output` feature.

2.2 Simulation Loop

Algorithm The opinion/belief dynamics simulation is implemented as demonstrated by Algorithm 1. The simulation is executed for a preset number of time steps T and, at each step, a random pair of agents are selected to run a Write and Receive call. The memory for each agent is updated with the resulting information from the interaction.

Require: N agent personas, N initial beliefs, time horizon T, LLM, Topic
Ensure: Belief trajectories $\langle b_i^t \rangle$ for each agent a_i
Initialize agents with personas and initial beliefs
for $t = 1$ to T **do**
  Select random agent pair $\{a_i, a_j\}$ with $i \neq j$
  Agent a_i generates message x_i^t
  Agent a_j reports opinion and new belief b_j^t
  Update memory for a_i and a_j
end for

Algorithm 1: Belief Dynamics Simulation

Experimental Levers The simulation procedure presents several avenues for experimentation, namely the following levers are able to be manipulated in each simulation run:

- **Agent LLM**: Gemini 2.0 Flash is used for all agents, though the implementation allows assigning different LLM provider IDs to each agent.
- **Total time steps**: $T = 100$, following prior work. With random pair selection per time step, each agent performs about 10 write and 10 review actions.
- **Opinion Classifier**: gent LLM; The agent LLM itself is the opinion classifier by producing reflection responses and belief values. Other works use transformer LLMs like FLAN-T5 (2022). We instead let the agent LLM generate belief values directly, since modern LLMs handle text understanding and structured reasoning well and avoid the cost of training and hosting a separate sentiment model.
- **# of agents**: $N = 10$; we keep the number of agents in the simulation fixed at 10.
- **Agent Memory Mode**: Reflective/Cumulative; we conduct our simulations with both memory modes to examine potential differences in opinion dynamics performance.
- **Persona Profiles**: 10 profiles from Chiang et al.
- **Initial Beliefs**
- **Topics**: Politics, Control; Our project conducts a proof-of-concept simulation with a politics set and control set of topics.
- **Topic Framing**: Positive or Negative; each topic has a positive and negative framing with which we can run the simulation.

2.3 Experimental Design

To structure our experiments, we utilized metrics from the opinion dynamics literature and engineered some persons/belief distributions to test dynamics in relevant social scenarios.

Topics The politics discussion topics include very charged social issues, as seen in table:

Political Topic	Positive	Negative
Gun Control – Debate over firearm regulation in the United States.	Stronger gun control laws are necessary to reduce gun violence and protect public safety.	Gun ownership is a fundamental right, and restrictions infringe upon personal freedom and self-defense.
Welfare – A debate about the government's role in providing financial and social support to citizens.	Social welfare programs are essential to reducing inequality and helping vulnerable populations achieve stability.	Excessive welfare breeds dependency and discourages personal responsibility and work ethic.
Immigration – A debate over immigration policy, border security, and the treatment of immigrants.	Immigrants enrich the nation culturally and economically, and the system should be reformed to be more humane and inclusive.	Unchecked immigration threatens national security and economic stability, and strong border enforcement is essential.
Abortion – A moral and legal debate surrounding the termination of pregnancies.	Women should have the right to choose what happens to their bodies, including the decision to have an abortion.	Abortion ends an innocent human life and should be restricted or banned to protect the unborn.

The positive framing presents a pro-Topic stance, while the negative framing presents an anti-Topic stance. For these topics, the positive framing aligns with Democratic positions and the negative with Republican ones.

The following table is for control topics that represent less-charged, benign matters:

Control Topic	**Positive**	**Negative**
iPhone vs. Android – A debate over which mobile platform offers the better overall user experience and ecosystem.	iPhone provides a more secure, consistent, and well-integrated experience across devices and apps compared to Android.	Android offers greater customization, a wider range of devices, and more affordable options, making it the superior choice for many users compared to iPhone.
Pineapple on Pizza – A lighthearted debate about whether pineapple belongs as a pizza topping.	Pineapple on pizza creates a delicious sweet and savory combination that enhances the overall flavor.	Pineapple has no place on pizza; the sweetness clashes with traditional savory toppings.
Summer vs Winter – Which season provides the most enjoyable weather and activities.	Summer is the best season with warm weather, outdoor activities, and longer daylight hours.	Winter is superior with cozy atmospheres, winter sports, and beautiful snowy landscapes.
Books vs Movies – Which medium provides a better storytelling experience.	Books offer deeper character development and allow readers to use their imagination more fully.	Movies provide a more immersive and efficient way to experience a story through visuals and sound.

Metrics Let b_i^t where $1 \leq i \leq N$ and $1 \leq t \leq T$ denote the belief of agent i at time step t. Hence, $[b_1^T, b_2^T, ..., b_N^T]$ form the opinion distribution at the end of the simulation. We report μ_T and σ_T as metrics for the **bias** and **diversity** of the opinions, respectively.

$$\mu_T = \frac{1}{N} \sum_{i=1}^{N} b_i^T$$

$$\sigma_T = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (b_i^T - \mu_T)^2}$$

We also engineer some heuristics to track variability of the belief values for each agent throughout the simulation (see results for more details).

Persona/Belief Distributions To systematically test the simulation framework under relevant social scenarios, we engineer several initial belief distributions as shown in Figure 3. The different configurations

- **Echo Chamber** – The echo chamber represents a one-sided distribution of initial beliefs: all positive or all negative on an initial topic. This is inspired by the phenomenon of echo chambers in social media where individuals of the same ideological position reinforce each other's belief on a given discussion topic. The hypothesis for an echo chamber configuration is that beliefs on one side of the belief distribution are amplified and there is either convergence to $b = 2$ or $b = -2$. Hence, we engineer two echo chamber configurations, one for the Democratic party and another for the Republican party, by modifying the political leaning

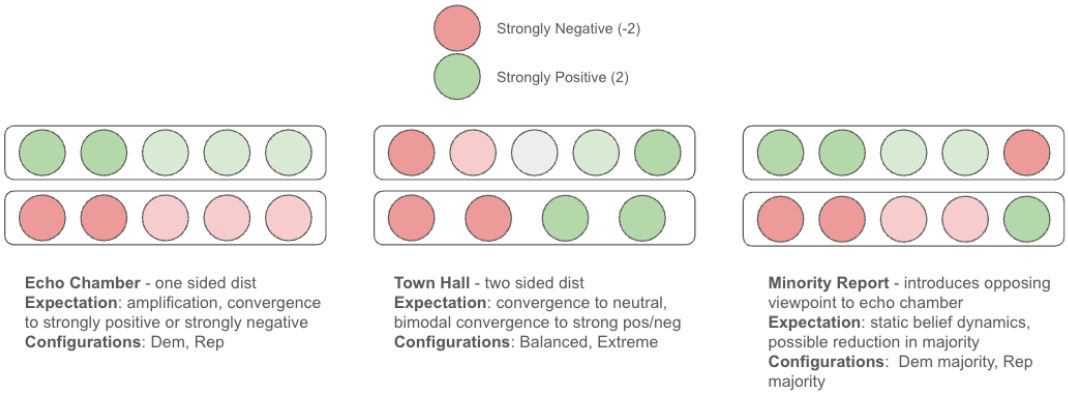

Figure 3: Persona Distribution

and initial belief attributes accordingly. Consider the positive framing of the gun control topic and all positive initial beliefs of Democrat personas. This situation would ideally lead to a convergence of beliefs at $b = 2$.

- **Town Hall** – The town hall configurations represent a scenario where there is ample presence of separate ideologies. Here, we engineer two configurations: extreme and balanced. In the extreme configuration, there are 5 strongly positive and 5 strongly negative initial beliefs, representing diametrically opposed individuals. This makes a bimodal initial belief distribution. The balanced case contains presence of all belief types. The town hall configurations provide an opportunity to analyze more contentious settings where opposing viewpoints interact with each other more often.
- **Minority Report** – The minority report configuration introduces an opposing viewpoint into an echo chamber. The goal with this setup is to ascertain whether the inclusion of the opposing viewpoint leads to any decrease in the majority position, which can be validated by a decrease/increase in the bias of the opinion distribution from start to finish of the simulation. Here, there are two configurations, one for Democratic majority and the other for Republican majority.

3 Results

We present results for the following simulation configurations:

- Politics Topics; Custom Persona Configurations (Echo Chamber, Town Hall, Minority Report)
- Control Topics; Custom Persona Configurations (Echo Chamber, Town Hall, Minority Report)

Figure 4 displays an example of belief evolution trajectories for one simulation in the extreme town hall persona distribution with the topic positively framed, where five personas are allocated to each extreme of the belief value distribution. Observe that all the personas that started beliefs at $b = 2$ stayed the same where as the only personas that demonstrated a change in belief started from $b = -2$. Of the three personas that shifted beliefs from -2, one individual switched $b = 2$, marking a complete ideological shift from their initial belief.

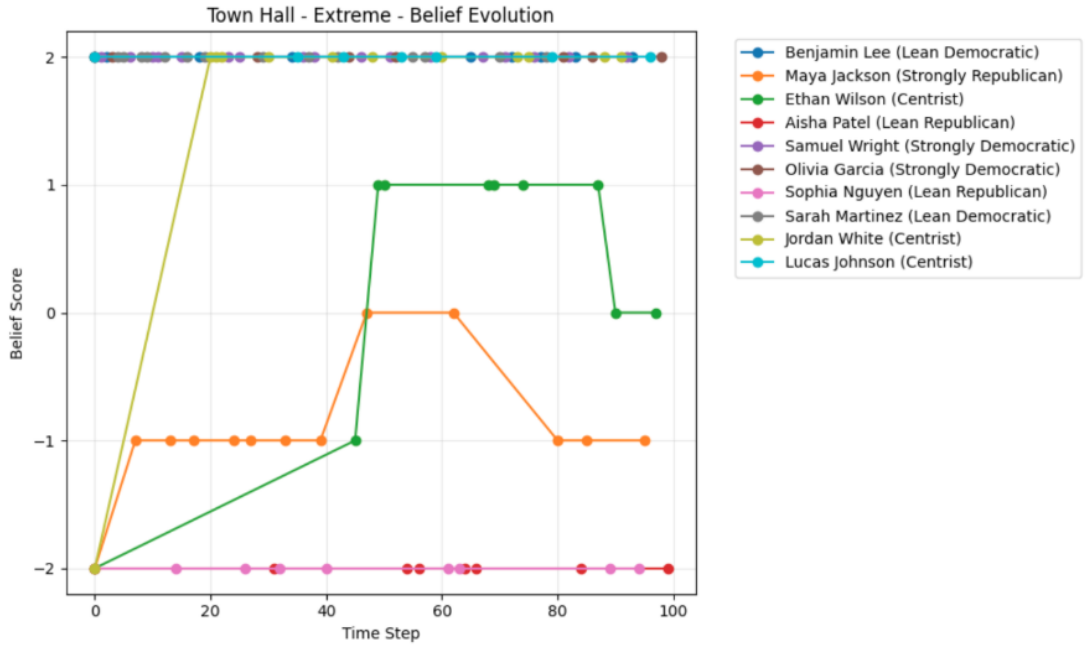

Figure 4: Example Belief Trajectory Evolution

	Persona Distribution	Positive Framing	Initial Bias (Mean)	Initial Diversity (Std Dev)	Final Bias (Mean)	Final Diversity (Std Dev)	Δ Bias (Mean)
0	Echo Chamber - Democratic	False	-1.3	0.461149	0.3250	1.421133	1.6250
1	Echo Chamber - Democratic	True	1.3	0.461149	1.9875	0.111803	0.6875
2	Echo Chamber - Republican	False	1.3	0.461149	1.8000	0.461149	0.5000
3	Echo Chamber - Republican	True	-1.3	0.461149	-0.2000	1.215928	1.1000
4	Minority Report - Democratic Majority	False	-1.0	1.102357	-0.1500	1.322636	0.8500
5	Minority Report - Democratic Majority	True	1.0	1.102357	1.6375	1.034148	0.6375
6	Minority Report - Republican Majority	False	1.0	1.102357	1.0250	1.005995	0.0250
7	Minority Report - Republican Majority	True	-1.0	1.102357	-0.0875	1.342525	0.9125
8	Town Hall - Extreme	False	0.0	2.012618	0.0750	1.375877	0.0750
9	Town Hall - Extreme	True	0.0	2.012618	0.8625	1.357061	0.8625
10	Town Hall - Generic	False	-0.2	1.256880	0.3125	1.298429	0.5125
11	Town Hall - Generic	True	0.2	1.256880	0.9875	1.152803	0.7875

Figure 5: Political Topics - Aggregate Bias and Diversity

3.1 Quantitative Analysis

The table in Figure 5 shows average aggregated bias and diversity metrics tracking the belief distribution at the start and end of the simulation, grouped by the custom configuration.

Echo Chamber Because all agents in an echo-chamber distribution begin on the same side of the political spectrum, we would expect their beliefs to converge toward the corresponding extreme of the Likert scale (i.e., toward -2 under negative framing and $+2$ under positive framing for Democratic personas, and symmetrically for Republican personas). This expected symmetry assumes that, in a homogeneous group, iterative discussion dynamics should reinforce initial group alignment. However, the empirical results do not support the intuition. Both Democratic echo-chamber conditions, regardless of framing, yield positive final mean biases and changes in mean bias. While the positive-framing condition behaves as expected ($1.3 \rightarrow 1.98$), the negative-framing condition unexpectedly shifts from ($-1.3 \rightarrow 0.32$). Among the Republican echo chamber setting, only the negative-framing condition exhibits the anticipated movement toward the extreme ($1.3 \rightarrow 1.8$). Like the Democratic echo chamber, an opposite framing in a Republican echo chamber leads to a more neutral mean outcome ($-1.3 \rightarrow -0.2$) but still with a positive delta. Taken together, these results produce uniformly positive mean deltas in belief values across framings. A plausible explanation is that the underlying LLM exhibits an inductive tendency to affirm the topic statement or gravitate toward the right-leaning side of the Likert scale, revealing a limitation of LLMs as persona-consistent simulators even in highly controlled homogeneous populations. Other explanations could be poor prompt adherence to the custom personas.

Minority Report The Minority Report configurations introduce a single dissenting belief into an otherwise uniform echo-chamber majority. The original hypothesis was that this lone minority voice could counterbalance group reinforcement and shift the overall distribution, at least partially. In practice, the effects are mixed. In the Democratic-majority setting, the postive framing condition produces a mean delta comparable to the echo-chamber baseline, while the false-framing condition yields a smaller positive shift than in the corresponding echo chamber, which is opposite of the expected outcome, since the presence of a minority would ideally increase movement away from the majority's starting position unless the random interactions turn out to suppress write. For the Republican majority setting, the negative framing condition shows minimal change, whereas the positive framing condition still results in a positive shift, though somewhat attenuated relative to the pure echo-chamber case. These results suggest that inserting a single opposing viewpoint does not reliably counteract the model's directional tendencies and that minority influence is weaker than anticipated.

Town Hall - Extreme In this configuration, agents begin evenly split between strongly positive and strongly negative beliefs, producing an initial mean of zero and high diversity. Across both framings, the diversity decreases, indicating that extreme positions are compressed toward the center, a notable deviation from the echo-chamber patterns. Under positive framing, the mean bias increases substantially ($0.0 \rightarrow 0.86$), while under negative framing it remains largely unchanged. This reduction in polarization, coupled with framing-dependent shifts in the mean, may indicate that the model resolves conflicting extreme viewpoints by softening the distribution rather than preserving bimodality.

Town Hall - Generic When the population begins with a mildly uneven, more spread distribution, we observe slightly lower but still postiivechange in mean bias under both framings when compared to the town hall extreme results.

Across all configurations, none of the mean deltas are negative. In settings where symmetry would predict opposite-signed shifts, the framework instead continues to push beliefs in a positive or rightward direction. This data does not suggest a political skew per se, but rather a general tendency of the LLM agents to align their expressed beliefs with the sentiment of the presented statement, regardless of its polarity.

Figure 6 shows the same aggregate results applied to the simulations with the control topics set.

	Persona Distribution	Positive Framing	Initial Bias (Mean)	Initial Diversity (Std Dev)	Final Bias (Mean)	Final Diversity (Std Dev)	Δ Bias (Mean)
0	Echo Chamber - Democratic	False	1.3	0.460355	1.036364	1.277377	-0.263636
1	Echo Chamber - Democratic	True	1.3	0.460355	1.700000	0.567467	0.400000
2	Echo Chamber - Republican	False	-1.3	0.461149	-0.237500	1.324011	1.062500
3	Echo Chamber - Republican	True	-1.3	0.461149	0.525000	1.440596	1.825000
4	Minority Report - Democratic Majority	False	1.0	1.102357	1.012500	1.130629	0.012500
5	Minority Report - Democratic Majority	True	1.0	1.102357	1.312500	0.835809	0.312500
6	Minority Report - Republican Majority	False	-1.0	1.102357	-0.187500	1.191677	0.812500
7	Minority Report - Republican Majority	True	-1.0	1.102357	0.587500	1.299403	1.587500
8	Town Hall - Extreme	False	0.0	2.012618	0.512500	1.113709	0.512500
9	Town Hall - Extreme	True	0.0	2.012618	1.062500	0.959216	1.062500
10	Town Hall - Generic	False	0.2	1.256880	0.200000	1.011328	0.000000
11	Town Hall - Generic	True	0.2	1.256880	1.162500	0.892213	0.962500

Figure 6: Control Topics - Aggregate Bias and Diversity

Here, the political ideology in the persona distribution is less interpretable as the topics do not actually relate to any political ideology. The key takeaway here is that the pattern of consistently positive changes in the mean bias persists (except for the democratic echo chamber), signaling that this behavior scales to domains beyond politics.

3.1.1 Initial Belief Changes

Initial Sentiment	Control Δ	Political Δ
Negative	2.032143	1.186404
Neutral	0.758065	0.833333
Positive	-0.469388	0.230263

Table 1: Mean belief change (Δ) for control and political simulations by initial sentiment.

We also analyzed all agents on political and control topics across every combination of persona distribution (e.g. Echo Chamber, Town Hall), framing (positive or negative), and memory mode (cumulative or reflective) to quantify belief shifts and compare political versus control changes. Agents were categorized by initial belief position (Negative as < -0.5, Neutral as -0.5 to 0.5, and Positive as > 0.5) rather than by specific simulation settings. Summary statistics in Table 1 showed that political topics had more upward movement for all belief groups compared to control topics. Neutral agents shifted +0.83 politically vs. +0.76 on control, positive agents shifted +0.23 politically vs. –0.47 on control, and negative-starting agents shifted +1.19 political vs. +2.03 control. Negative-starting agents moved the most in both political and control domains, while positive agents were domain sensitive. Neutral agents had similar changes in both domains (approximately +0.8). Figure 7 shows the initial belief changes for political topic simulations as violin plots to visualize the full distribution shape and variability within each starting belief. Overall, persuadability varied across groups, and an upward bias is shown in political simulations.

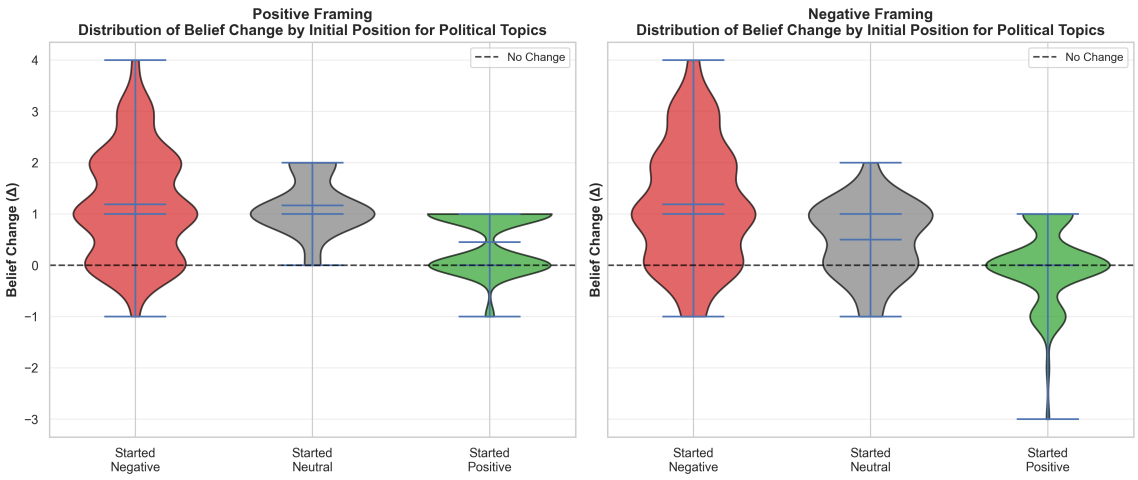

Figure 7: Violin Plots for Political Belief Changes

3.1.2 Temporal Diversity of Beliefs

Temporal diversity was also measured for analysis, in which the variance of agent beliefs were computed at each timestep to understand how different agent opinions are from one another as the simulation evolved. Temporal diversity was computed within each simulation run, and then the results were grouped and averaged by configuration. High values indicate diverse/polarized opinions and low values indicate consensus. This allows us to see how consensus or polarization develops as agent beliefs either converge, diverge, or oscillate throughout interactions. Figure 8 shows for the most and least variant simulations: Echo Chamber Democratic with negative framing shows fluctuating levels of variance (with mean 1.61). This shows that agents repeatedly diverge and converge throughout interactions. Meanwhile, the same group with positive framing quickly converged in beliefs as variance drops close to 0 and stays there (with mean 0.04).

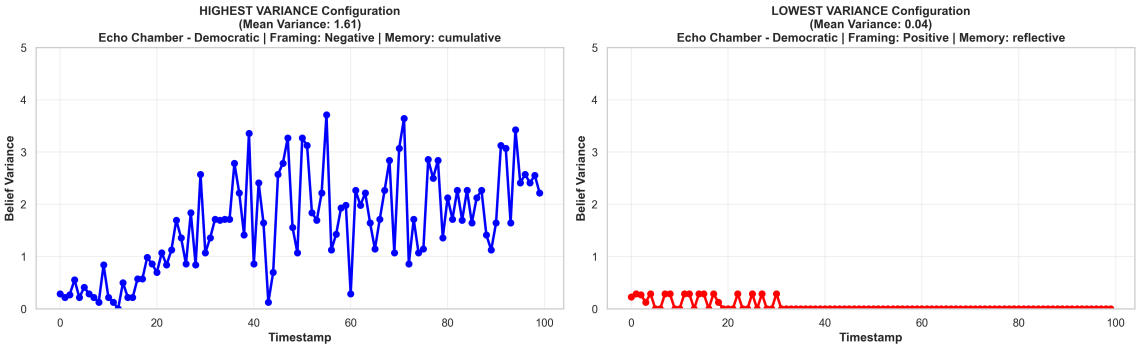

Figure 8: Temporal Diversity of Most and Least Variant Simulations

3.1.3 Comparison between cumulative and reflective memory

With the exception of Echo Chamber Democratic with negative framing and Town Hall Generic with positive framing, reflective memory was observed to consistently increase the mean change of bias with varying effects over the diversity. Notably, for the Echo Chamber and Minority Report, it can also be observed that reflective memory slightly reduces the mean bias shift when the framing does not match the group's preference and slightly increases the mean bias shift when the framing matches the group's preference. In explaining the consistent increase in mean change of bias,

perhaps, a shorter summary context is easier to act upon for LLMs compared to a compilation of all previous contexts. As for the mean bias shifts, the effect is marginal, but it is worth noting that reflective memory increases a group's own preferences under matching framing and resilience to opposing preferences under opposite framing.

	Persona Distribution	Positive Framing	Memory Mode	Initial Bias (Mean)	Initial Diversity (Std Dev)	Final Bias (Mean)	Final Diversity (Std Dev)	Δ Bias (Mean)
0	Echo Chamber - Democratic	False	cumulative	-1.3	0.464095	0.425	1.550641	1.725
1	Echo Chamber - Democratic	False	reflective	-1.3	0.464095	0.225	1.290746	1.525
2	Echo Chamber - Democratic	True	cumulative	1.3	0.464095	1.975	0.158114	0.675
3	Echo Chamber - Democratic	True	reflective	1.3	0.464095	2.000	0.000000	0.700
4	Echo Chamber - Republican	False	cumulative	1.3	0.464095	1.750	0.438529	0.450
5	Echo Chamber - Republican	False	reflective	1.3	0.464095	1.850	0.483046	0.550
6	Echo Chamber - Republican	True	cumulative	-1.3	0.464095	-0.375	1.294713	0.925
7	Echo Chamber - Republican	True	reflective	-1.3	0.464095	-0.025	1.120611	1.275
8	Minority Report - Democratic Majority	False	cumulative	-1.0	1.109400	-0.275	1.339489	0.725
9	Minority Report - Democratic Majority	False	reflective	-1.0	1.109400	-0.025	1.310461	0.975
10	Minority Report - Democratic Majority	True	cumulative	1.0	1.109400	1.575	1.152200	0.575
11	Minority Report - Democratic Majority	True	reflective	1.0	1.109400	1.700	0.911465	0.700
12	Minority Report - Republican Majority	False	cumulative	1.0	1.109400	0.950	0.959434	-0.050
13	Minority Report - Republican Majority	False	reflective	1.0	1.109400	1.100	1.057331	0.100
14	Minority Report - Republican Majority	True	cumulative	-1.0	1.109400	-0.200	1.324329	0.800
15	Minority Report - Republican Majority	True	reflective	-1.0	1.109400	0.025	1.367901	1.025
16	Town Hall - Extreme	False	cumulative	0.0	2.025479	-0.100	1.316561	-0.100
17	Town Hall - Extreme	False	reflective	0.0	2.025479	0.250	1.427747	0.250
18	Town Hall - Extreme	True	cumulative	0.0	2.025479	0.800	1.362501	0.800
19	Town Hall - Extreme	True	reflective	0.0	2.025479	0.925	1.366025	0.925
20	Town Hall - Generic	False	cumulative	-0.2	1.264911	0.200	1.381192	0.400
21	Town Hall - Generic	False	reflective	-0.2	1.264911	0.425	1.217132	0.625
22	Town Hall - Generic	True	cumulative	0.2	1.264911	1.025	1.229707	0.825
23	Town Hall - Generic	True	reflective	0.2	1.264911	0.950	1.084861	0.750

Figure 9: Political Topics - Memory Mode Comparison

3.2 Qualitative Analysis

To better contextualize the quantitative results, we examined representative interaction traces from the simulations. Several consistent behavioral patterns emerged, including some problematic cases in which agents shifted beliefs in unrealistic or abrupt ways.

Polite, Non-Adversarial Messaging. Across topics, agents produced polite and structured responses rather than emotionally charged or oppositional ones. For example, when reviewing a message supporting stronger gun control, even a strongly negative agent responded with:

> `"I understand your concerns about safety. Your points highlight important considerations."`

This tone contrasts with real-world social media discourse and helps explain the limited polarization observed.

Upward Belief Drift and Sycophancy. Reviewers often partially agreed with incoming messages regardless of their initial stance, which contributed to the consistent positive belief drift seen in our metrics. In multiple cases, agents with $b = -2$ moved toward neutrality or positivity after a single interaction. One agent who began strongly opposed to gun regulation shifted from $b = -2$ to $b = +1$ after reading a brief pro-regulation message and responded:

> `"Your points about protecting families are compelling. I feel more supportive of stronger regulations now."`

This type of abrupt reversal suggests over-agreeableness and limited resistance to framing.

Weak Persona Adherence. Agents frequently deviated from their assigned political leanings. For instance, a "Lean Republican" persona reviewing a pro-immigration message responded with a

supportive shift in belief instead of reinforcing its ideological position. In another case, a persona marked as strongly conservative on welfare shifted from $b = -2$ to $b = 0$ after a short message emphasizing social stability. Persona anchoring therefore had limited effect on message style or belief updates.

Neutral Summaries Instead of Debate. Review messages often summarized the writer's point rather than engaging directly. For example:

> `"Both perspectives have valid concerns, and it is important to weigh them carefully."`

Despite offering no clear endorsement, the agent updated its belief from $b = 0$ to $b = 1$. This type of meta-level neutrality reduced disagreement but still contributed to the upward drift in support for the positively framed side.

Repetitive Reasoning Patterns. Messages frequently reused similar argumentative structures that involved affirmation, soft moderation, and slight belief adjustment. This indicates limited variation in discourse over time and reflects the tendency of the LLM to rely on generic patterns rather than differentiated persona-driven reasoning.

Taken together, the qualitative traces reinforce the quantitative findings. Agents tended to agree rather than contest, soften rather than polarize, and drift toward the positive framing regardless of persona or starting belief. The examples of abrupt belief reversals highlight current limitations in using LLM agents to emulate realistic social dynamics, particularly in the areas of conflict, identity fidelity, and argumentative diversity.

4 Conclusions

This paper successfully extends and re-implements the multi-agent opinion dynamics framework by Chuang et al. for open-ended questions, focusing on political issues as a domain of exploration. This work also contributes a novel method of testing simulated persona distributions with engineered examples like the echo chamber and town hall. The guiding work illustrates that the agent LLMs (those from foundation model providers) tend to converge to truth-seeking behavior by testing simulated personas on unambiguous facts (e.g., conspiracy theories). Our results analyzing the custom persona distributions complement the findings of the previous work by showing that models have a tendency to converge towards agreeing with the topic statement. This was evidenced by the inability of the framework to produce sensible symmetric changes in belief dynamics when we changed the topic framing parameter from positive to negative. Contrary to real-world social dynamics where opposing groups drift apart, the LLM agents exhibited a strong tendency toward depolarization and consensus in the Town Hall setups. However, similar to real world social dynamics, the topic framing in favor of opposing views helped the echo chambers and minority reports to have more diversity in their opinion outcomes even if the outcome was still sycophantic for the topic. Possible reasons for these behaviors include but are not limited to poor prompt adherence, inefficient prompting strategies, and inherent bias of LLMs to positive belief feedback. We enumerate several opportunities for further work below.

4.1 Limitations and Future Work

Architecture The architecture possesses several limitations and assumptions. First, the architecture assumes a fully connected multi-agent graph, where every agent can message every other. However, real social networks have uneven connectivity and social influence, which this study does not model. Debates also occur beyond social media, including the government/public policy settings where legislators must make formal arguments regarding policy. In addition, agents don't condition messages to specific recipients, unlike real debates. Future work could add a shared message pool

and richer formats like Du et al.'s multi-agent debate(2). Prior research employs an independent classifier to assign belief values. Future research should look into the correlation between the belief values selected by the agent LLM and an independent LLM to validate using the agent's labels directly. We also observed differences between slight and strong opinions across the major LLMs as seen in 10, reflecting limits of the opinion classifier and the five-point Likert scale.

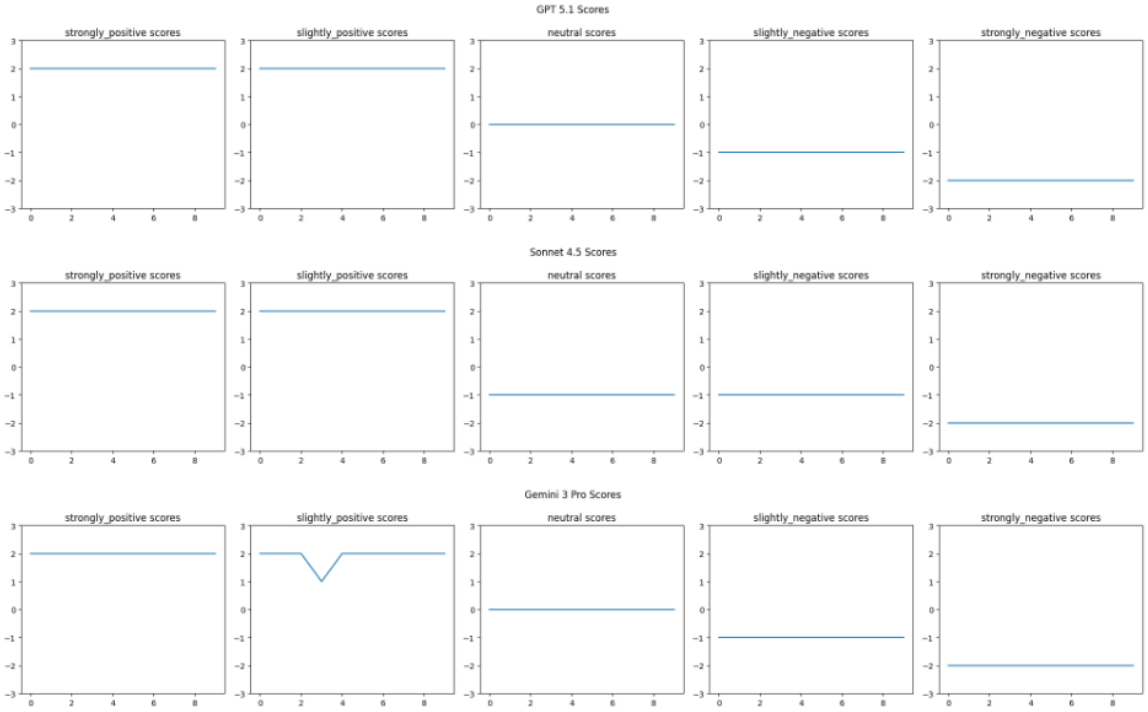

Figure 10: LLM sentiment scores given prompts of differing sentiments across 10 trials

In addition, this study uses only one LLM provider. Future work should test multiple foundation models and even consider finetuning to improve persona role-play. Comparing different models may also help extend research on political bias in LLMs.

Lastly, another limitation to the current architecture is the simplified representations of personas. Gender, ethnicity, and occupation are limited descriptors and can be strengthened with more extensive persona stories, with recent work being done on advanced character-shaping techniques (4).

Evaluation The evaluation strategy in this paper also presents several directions for further work. Because each multi-agent simulation required many LLM calls and a long time horizon, runs took about 60 seconds, limiting efficient hyperparameter exploration. Next, the topics for both the control and politics sets were four topics each. Future work should investigate other datasets and expand both sets for more topics. Quantitative analysis can also incorporate metrics other than bias and diversity in belief distribution for change quantification. Variations can involve detecting bimodality and clustering of beliefs to detect sub-echo chambers. For qualitative analysis, future work includes taking an LLM-as-a-judge approach to grading the interaction traces across axes such as factuality, hallucination rate, and overall tone. These would strengthen qualitative analysis and provide insights into the viability of multi-agent LLM networks for opinion dynamics modeling. Lastly, evaluating prompt adherence in the interaction traces can help quantify an LLM's ability to emulate a persona profile.

References

[1] Chuang, Y.-S., Goyal, A., Harlalka, N., Suresh, S., Hawkins, R., Yang, S., Shah, D., Hu, J., and Rogers, T. Simulating opinion dynamics with networks of LLM-based agents. In *Findings of the Association for Computational Linguistics: NAACL 2024* (Mexico City, Mexico, June 2024), K. Duh, H. Gomez, and S. Bethard, Eds., Association for Computational Linguistics, pp. 3326–3346.

[2] Du, Y., Li, S., Torralba, A., Tenenbaum, J. B., and Mordatch, I. Improving factuality and reasoning in language models through multiagent debate. *arXiv preprint arXiv:2305.14325* (2023).

[3] Guo, T., Chen, X., Wang, Y., Chang, R., Pei, S., Chawla, N. V., Wiest, O., and Zhang, X. Large language model based multi-agents: A survey of progress and challenges. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24* (8 2024), K. Larson, Ed., International Joint Conferences on Artificial Intelligence Organization, pp. 8048–8057. Survey Track.

[4] Maiya, S., Bartsch, H., Lambert, N., and Hubinger, E. Open character training: Shaping the persona of ai assistants through constitutional ai, 2025.

[5] Park, J. S., O'Brien, J. C., Cai, C. J., Morris, M. R., Liang, P., and Bernstein, M. S. Generative agents: Interactive simulacra of human behavior. In *In the 36th Annual ACM Symposium on User Interface Software and Technology (UIST '23)* (New York, NY, USA, 2023), UIST '23, Association for Computing Machinery.

[6] Shen, J. H., Appel, R., Tucker, M., Jagadish, K., Maheshwary, P., Askell, A., and Durmus, E. Political even-handedness evaluation v1, 2025.