Decoding Echo Chambers: LLM-Powered Simulations Revealing Polarization in Social Networks

Chenxi Wang*, **Zongfang Liu***, **Dequan Yang**, **Xiuying Chen**†
Mohamed bin Zayed University of Artificial Intelligence
{Chenxi.Wang, Zongfang.Liu, Dequan.Yang, Xiuying.Chen}@mbzuai.ac.ae

Abstract

The impact of social media on critical issues such as echo chambers, needs to be addressed, as these phenomena can have disruptive consequences for our society. Traditional research often oversimplifies emotional tendencies and opinion evolution into numbers and formulas, neglecting that news and communication are conveyed through text, which limits these approaches. Hence, in this work, we propose an LLM-based simulation for the social opinion network to evaluate and counter polarization phenomena. We first construct three typical network structures to simulate different characteristics of social interactions. Then, agents interact based on recommendation algorithms and update their strategies through reasoning and analysis. By comparing these interactions with the classic Bounded Confidence Model (BCM), the Friedkin-Johnsen (FJ) model, and using echo chamber-related indices, we demonstrate the effectiveness of our framework in simulating opinion dynamics and reproducing phenomena such as opinion polarization and echo chambers. We propose two mitigation methods—active and passive nudges—that can help reduce echo chambers, specifically within language-based simulations. We hope our work will offer valuable insights and guidance for social polarization mitigation.[1]

1 Introduction

Echo chambers occur when individuals are primarily exposed to information or opinions that align with their own. This limits their exposure to diverse perspectives, reinforces their beliefs, and contributes to increased polarization. Echo chambers on social media are a serious issue that can lead to a range of negative consequences. For example, during elections, echo chambers can amplify misinformation and false claims about candidates, policies, or voting processes, contributing to increased political polarization and mistrust in the democratic system (Grömping, 2014; Terren and Borge-Bravo, 2021). This reinforcement of homogenous perspectives not only distorts the flow of accurate information but also deepens polarization and divisions within society (Levy and Razin, 2019; Dahlgren, 2020).

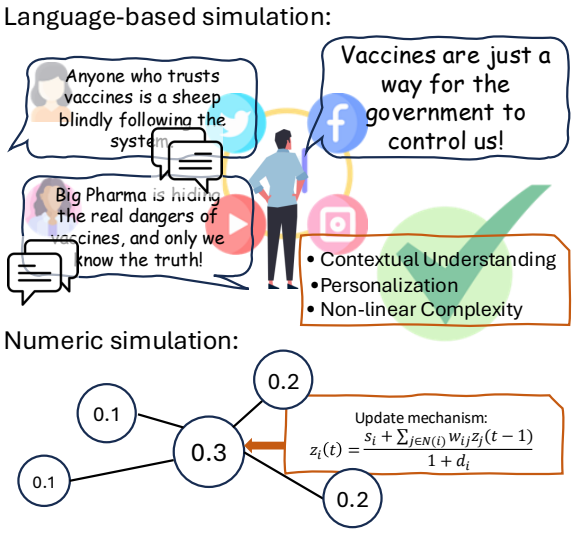

Figure 1: Our language-based simulation provides an explainable and dynamic environment compared with numeric simulation for echo chamber study.

Many efforts have been made to evaluate and mitigate echo chambers. The Friedkin-Johnsen Dynamics (FJ) model (Friedkin and Johnsen, 1990) is a classic framework for modeling opinion formation, and Chitra and Musco (2020a) used it to simulate user leanings on a spectrum from -1 (opposition) to 1 (support). Deffuant et al. (2000) proposed the Bounded Confidence Model (BCM), where individuals only interact and adjust their opinions if the difference between their opinions is within a certain threshold. However, these models oversimplify social interactions, reducing complex discus-

* Equal contribution.
† Corresponding author.
[1] Available on: https://github.com/ZongfangLiu/EchoChamberSim.

Proceedings of the 31st International Conference on Computational Linguistics, pages 3913–3923
January 19–24, 2025.

sions and opinions to numerical values, rather than capturing the nuanced meaning conveyed through text, language, and context. In this work, we address the above problem by proposing a Social Simulation Framework (SSF), with each individual represented as an LLM agent. Our SSF effectively replicates the textual nature of news, complex human reasoning, and dynamic opinion shifts, thereby enhancing explainability.

Concretely, we set up three different social networks: small-world (Watts and Strogatz, 1998), scale-free network (Barabási and Albert, 1999), and random graph model (Wang and Chen, 2003), where the first two are proposed to simulate real-world social networks. Agents are initialized with unique personas, including attributes such as name, gender, age, educational background, and personal traits. Each agent is equipped with a short-term memory to capture the day's interactions and a long-term memory for broader context, along with a reflective reasoning process to mimic the human thought process. Each day, agents exchange opinions with their adjacent neighbors—either all of them or selectively—analyze, reason, and update their opinions accordingly. Here, we investigate the influence of commonly used recommendation algorithms (Li et al., 2010), which push similar opinions to users. This common practice on social media connects users with content they are likely to agree with, aiming to increase engagement (Pariser, 2011). For evaluation, we compare our framework with the classic BCM and FJ model using echo chamber-related metrics, including the neighbor correlation index, polarization index, and global disagreement. The results on different metrics show that our simulation replicates the echo chamber phenomenon observed in the real world, and the trend is generally consistent with numerical simulations, while providing different insights.

Furthermore, we propose two strategies—active and passive nudges—that effectively alleviate echo chambers and promote a more diverse and healthy conversational environment. These methods offer a more realistic approach, focusing on the language itself and providing convincing content, unlike previous mitigation efforts that involve adding additional, non-explainable nodes (Orbach et al., 2020).

Our contributions can be summarized as follows:

- We present the first language-based simulation framework *SSF* for studying echo chambers and polarization.
- Our experiments align with conclusions drawn from real-world studies and are mostly consistent with numerical simulations, validating the utility of our SSF as a research tool.
- We propose a language-based intervention approach to mitigate opinion polarization and the echo chamber effect, offering new insights for community governance and management.

2 Related Work

Echo Chambers Modeling. Researchers have proposed various strategies to model echo chambers. One well-known approach is the Friedkin-Johnson Dynamics (Friedkin and Johnsen, 1990; Chitra and Musco, 2020a), which models how individuals' opinions evolve based on both their inherent beliefs and social influences. Other notable models include the bounded confidence model (BCM) (Deffuant et al., 2000), which assumes that individuals only interact with others whose opinions are within a certain range, and cascade models (Zhou and Zafarani, 2020), represented as trees where each node is a user in the social network and the root node represents the user who began the discourse. Additionally, Epidemic Models, inspired by the spread of diseases, have been employed to detect the formation of echo chambers in social networks (Cinelli et al., 2021). In our work, we compare our approach with the bounded confidence model due to its ability to model the interplay between social influence and individual stubbornness.

Echo Chambers Mitigation. Mitigation strategies can be broadly categorized into two types: algorithm-focused and human-focused. Algorithm-focused strategies aim to address the causes of echo chambers that arise from algorithmic curation and content recommendation. For instance, modifying the objective function of a recommender system has been shown to mitigate the filter bubble effect (Chitra and Musco, 2020a). Similarly, Orbach et al. (2020) attempted to identify speeches on the same topic, but with opposing stances, that directly counter one another. On the other hand, human-focused strategies empower users to have more control over their information environment by encouraging them to critically evaluate the quality of information, such as through labeling misinformation or fake news, fact-checking, and nudging users to reflect on the accuracy of the information (Alatawi et al., 2021). These methods have been tested publicly by platforms like Microblog,

with varying degrees of success (Fu et al., 2013).

LLM-based Agent Simulation. Integrating LLMs into the simulation of social dynamics is an emerging area of research, yielding promising outcomes (Park et al., 2023; Kaiya et al., 2023; Li et al., 2023; Zhang et al., 2024; Guo et al., 2024; Liu et al., 2024b). These LLM-based generative agents have shown exceptional performance in digital environments, particularly in natural language tasks. For example, Xie et al. investigated whether LLM agents can simulate human trust behaviors, which are among the most critical aspects of human interactions. Park et al. (2022) demonstrated that LLM-based agents could generate social media content that is indistinguishable from human-produced content. These advancements underscore the vast potential of LLM agents in modeling human social behaviors at the group level. Our simulation differs from previous work by exploring the unstudied topic of echo chambers and constructing graphs and recommendation systems, rather than relying on random interactions, providing a new approach to studying social dynamics.

3 Method

3.1 Problem Formulation

Formally, we construct a simulation with a pool of N LLM agents, denoted as $A = (a_1, ..., a_N)$, and an examined topic F. During initialization, the agents construct a network based on different structures such as small-world network, scale-free network, or random graph. Each agent is assigned a unique persona, including their initial belief towards the discussed topic. On the t-th day, each agent a_i interacts with its neighboring agents from the pool A, either all of them or selectively, according to a recommendation algorithm. At the end of each day, each agent reflects on the exchanged information and updates its belief toward the topic, producing an opinion expression and a belief value v_i, where the range of -2 to 2 represents the degree of opposition or support. This process is iterated over T days. Our goal is to replicate the echo chamber and polarization phenomena observed in real-world social network structures, including small-world and scale-free networks (Watts and Strogatz, 1998; Bessi et al., 2016; Cinelli et al., 2021).

3.2 Social Network Structure

One significant drawback of previous simulation works is that they lack graph structures, assuming that each agent knows all other agents (Liu et al., 2024a; Wang et al., 2023; Williams et al., 2023; Chuang et al., 2024). While this might be true in small societies, as the simulation scale increases, different social structures should be considered.

Hence, in this work, we set up three types of network graphs, with the first two being commonly observed in the real world. A *small-world network*, introduced by Watts and Strogatz (1998), is characterized by high clustering and short average path lengths, similar to social networks. Formally, the average shortest path length L grows logarithmically with the number of nodes N, $L \sim \log N$, while maintaining a high clustering coefficient. In social networks, this reflects tight-knit communities with short paths between individuals. A *scale-free network*, as described by Barabási and Albert (1999), exhibits a power-law degree distribution, meaning the probability $P(k)$ that a node has k connections follows $P(k) \sim k^{-\gamma}$. This is common in social networks where a few hubs have far more connections than others. Finally, a *random graph*, defined by ERDdS and R&wi (1959), is constructed by connecting nodes randomly with a fixed probability, leading to a binomial degree distribution. In social networks, this type of structure models random connections between individuals without clear clusters or hubs. Figure 2 gives an intuitive understanding of these three types of graphs.

3.3 Interaction Algorithm

A key part of our architecture is deciding the interaction strategies between different agents. In previous simulation networks (Liu et al., 2024a; Chuang et al., 2024), agents typically interact randomly with each other without a structured network. However, in real social networks, the frequency of interactions between agents depends on the graph structure. It's unlikely for a regular user to comment on a stranger or frequently comment on a popular figure without following them. Hence, our interaction algorithms account for social network relationships, where users only interact with neighboring agents. Meanwhile, in social networks, interactions are also influenced by recommendation mechanisms. Common practice on social media connects users with content they are likely to agree with, aiming to increase engagement (Pariser, 2011).

Based on the above observations, and to facilitate comparison with the numeric modeling method BCM (Deffuant et al., 2000), for a given agent, we recommend those neighbors with similar opinions.

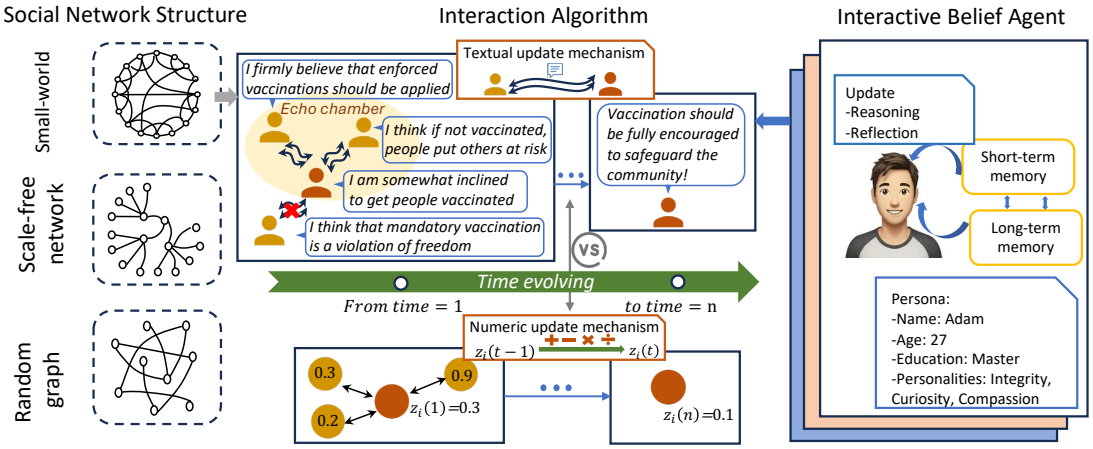

Figure 2: Our framework is evaluated on three different network structures that mimic real-world observations. Each agent is initialized with personal information, dual memory, and a reasoning process. Through random or recommendation-based interactions, they update their opinions each day.

Assumption	BCM Model	Advanced LLM
Opinion update	Weighted averaging	Dynamic adjustment with context
Determinism	Opinion changes are fully predetermined.	Opinion changes are not fully predetermined.
Structure	Social structure remains unchanged.	
Continuance	Opinion changes continue until settled.	
Decomposability	Process splits into time periods.	
Simultaneity	Simultaneously predict influence events.	

Table 1: Comparison of assumptions between BCM Model and our SSF model.

The process is as:

$$R = \{j \in \mathcal{N}(i) \mid |v_i^{SSF} - v_j^{SSF}| \leq 2\},$$

where v_i^{SSF} and v_j^{SSF} represent the belief values of nodes i and j, indicating their respective opinions. Here, $\mathcal{N}(i)$ denotes the neighbors of node i, and R represents the set of recommended neighbors.

In the ablation study in Appendix, we also show the performance when we remove the recommendation, where the echo chamber effect exists.

3.4 Language-based Opinion Updates

After introducing the graph setting and interaction mechanism, we explain how the agent updates its opinion within this framework. Each agent is randomly assigned a persona p_i, which includes attributes such as name, gender, age, traits, and education level, as these factors may influence their belief toward the topic. For trait design, we follow the widely accepted Big Five personality model (Barrick and Mount, 1991), known for effectively capturing key dimensions of personality.

In our model, we consider that an individual's opinion is influenced not only by their own belief but also by their interactions with others. This interaction-driven change in thought is gradual and cumulative, rather than immediate. Accordingly, in our simulation, agents engage with a random number of others' opinions each day, leading to a periodic update of their views. However, owing to the potentially vast volume of interactions, storing all of them in detail is impractical. To address this challenge, we implement a dual memory system for each agent, comprising a long-term memory m_i^l and a short-term memory m_i^s. The long-term memory compresses and stores a summarized history of past interactions, while the short-term memory reflects and summarizes conversations from the current day. The long-term memory mechanism is consistent with real human behavior, as people don't remember every single word they hear. At the end of each day, agents reflect on these interactions, and through a reasoning process, allow their opinions to evolve.

Comparison with Number-based Opinion Updates. As introduced above, our opinion updates are purely based on language, mimicking the human thinking process. For comparison, here we introduce how traditional methods update an agent's opinion, where the setting comparison is in Table 1.

The Bounded Confidence Model (BCM) is a widely used framework for studying opinion dynamics in social networks. In the BCM model, each agent holds an initial opinion, denoted as v_i^{BCM}, which represents their belief about a given topic. Opinions are updated iteratively based on the opinions of neighboring agents, but only if the opinion difference between them is within a

certain threshold, ϵ. Formally, at each time step t, the opinion of agent i is updated if and only if the opinion difference with its neighbor j satisfies $|v_i^{BCM}(t-1) - v_j^{BCM}(t-1)| \leq \epsilon$. When this condition holds, the opinions are updated as: $v_i^{BCM}(t) = v_i^{BCM}(t-1) + \mu \left(v_j^{BCM}(t-1) - v_i^{BCM}(t-1)\right)$, where μ is a parameter that controls the rate of opinion change.

This update algorithm can be seen as a numerical version of our update process, where *the opinion changes o are simplified into numbers v^{BCM}, and the diverse character p is fixed to ϵ and μ*. We will show in §4.3 that the two simulations reach broadly consistent conclusions.

3.5 Polarization Mitigation Operation

The objective of our framework extends beyond merely providing a more explainable model of social opinion dynamics. It also aims to offer actionable insights for better governance and to foster a healthier social environment. To achieve this, we implemented two language-based mitigation approaches that are not feasible through traditional numerical methods.

Active Nudge. The philosophy behind this is encapsulated in the saying, "listening to all sides leads to wisdom." In our implementation, when a user expresses a polarized stance, we *actively* present an opposing viewpoint from another user with a contradictory position. This method broadens the user's exposure to various perspectives, promoting a more balanced and reflective consideration of the issues. By doing so, we aim to counteract the reinforcement of one-sided arguments, fostering a more nuanced and critical discourse.

Passive Nudge. Unlike previous methods that explicitly prompt users to reconsider their positions, Passive Nudge *subtly* shares content with users holding extreme views, emphasizing the value of maintaining an open perspective. For example, the suggested content could be: 'Issues are rarely black and white,' or 'Many societal and political issues are complex and multifaceted.' This approach emphasizes the benefits of open-mindedness without persuading users to adopt a neutral or any specific belief, thereby leaving them the freedom to think independently. While this is difficult to simulate using traditional numerical methods, it can be easily implemented within our simulation system. The prompts for all experiments are in the Appendix.

4 Experiments

4.1 Implementation Details

We use Python to conduct our SSF simulation, utilizing the GPT-4o-mini model accessed via OpenAI API calls. The agents and their environment are defined using the Python library Mesa (Kazil et al., 2020). To eliminate any bias associated with names, each agent is identified solely by their agent index. Genders are randomly assigned, and ages are randomly selected within the range of 18 to 64 years. The simulation includes 50 agents, a number significantly larger than in previous LLM-based simulations (Liu et al., 2024a; Chuang et al., 2024). Agent characteristics are based on the Big Five personality traits commonly used in psychology (Barrick and Mount, 1991), with each agent having a 50% chance of exhibiting either a positive or negative version of each trait. We introduce more details in Appendix.

4.2 Metrics

We use these three metrics to jointly measure the formation of echo chambers and the degree of opinion polarization within the network: Polarization (Chitra and Musco, 2020b), Global Disagreement, and Normalized Clustering Index (NCI) following (Cinus et al., 2021).

Polarization (P_z) measures the overall variance in opinions within a network and is defined as:

$$P_z = \frac{\sum_{i=1}^{N}(v_i - mean(v))^2}{N},$$

where $mean(z)$ is the average belief value across all nodes. A high polarization value signals a sharp divide in opinions, typical of echo chambers.

Global Disagreement (DG) quantifies how much a node disagrees with its neighbors. It aggregates the local disagreement (DG_i) of each node i, which measures the difference in opinions between a node and its neighbors. The local disagreement is as:

$$DG_i = \frac{\sum_{j \in \mathcal{N}(a_i)}(v_i - v_j)^2}{|\mathcal{N}(a_i)|}.$$

The overall global disagreement is calculated by summing over all nodes:

$$DG = \frac{1}{2N}\sum_{i=1}^{N} DG_i,$$

This metric offers insight into the opinion divergence at a local level, indicating the level of disagreement between connected nodes.

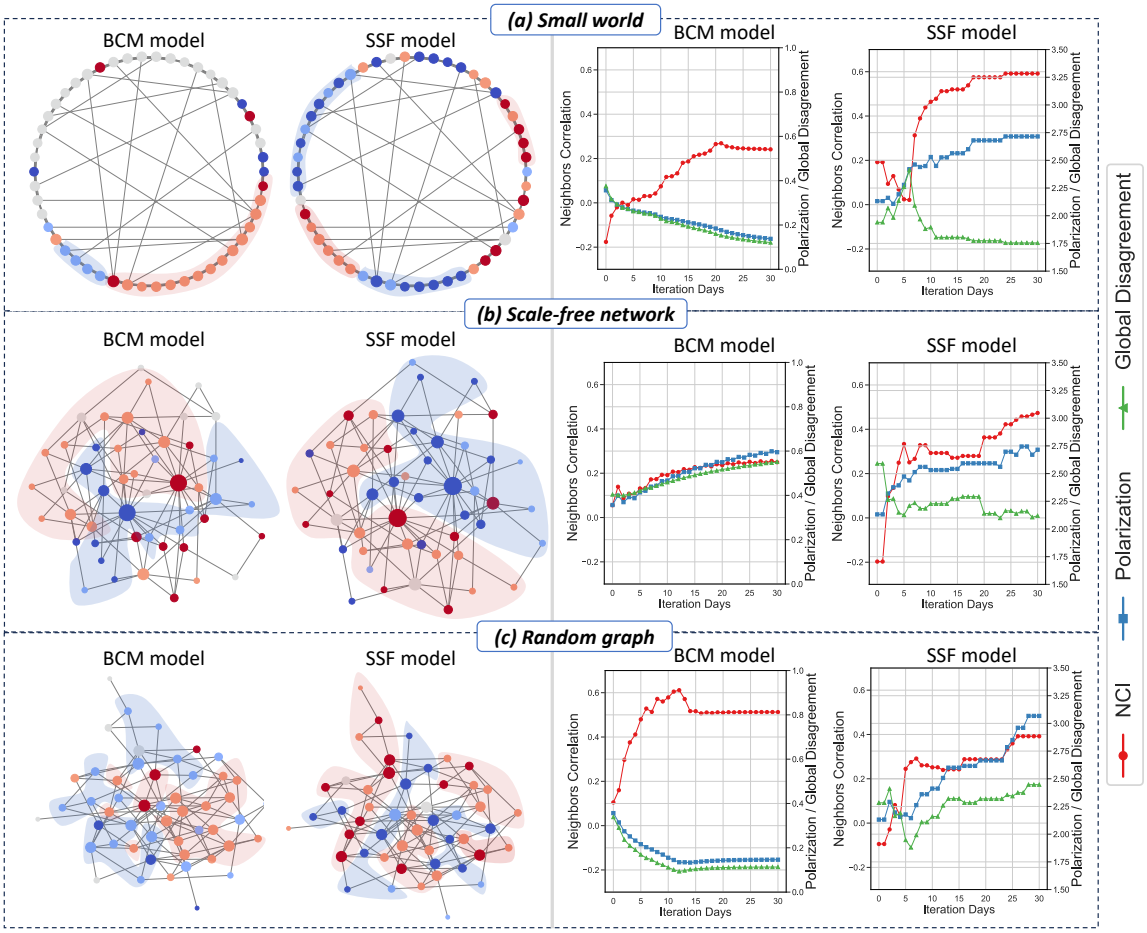

Figure 3: We present projection and curve graphs from BCM and our SSF simulations under three settings to demonstrate our framework's macro-level effectiveness. The main conclusions are similar: small-world and scale-free networks lead to severe echo chamber effects, while the random graph does not.

Finally, NCI measures how closely a node's opinions align with those of its direct neighbors. For an agent a_i, the NCI_i is calculated as the pearson correlation between the opinion of a_i and the average opinion of its neighbors, and is defined as:

$$NCI_i = \sum_{a_j \in \mathcal{N}(a_i)} \rho(v_i, v_j),$$

where ρ is the person correlation. This metric helps identify the extent to which all nodes in the network are embedded in groups of similar opinions, with values close to 0 indicating that most nodes are exposed to diverse opinions.

It is worth noting that a single metric cannot fully demonstrate the presence of echo chambers; multiple metrics must be considered for a more accurate understanding.

4.3 Macro-level Observation

Visual analysis. To give a fair comparison, the initial beliefs follow uniform distributions in both the BCM and SSF settings. For different network structures, the initialized agents remain consistent. The left side of Figure 3 shows the final belief values of each node after 30 days of message propagation across different social network structures. Node colors represent belief strength, with red for support and blue for opposition. The color blocks indicate clusters of similar opinions.

In small-world and scale-free networks, fewer but larger clusters emerge, reflecting clear echo chambers. In random networks, clusters are smaller and more numerous, with no large echo chambers. We can also observe that in the small-world network, numerical simulation oversimplifications result in many neutral nodes, averaging their neighbors' opinions. In the right plots in Figure 3, a significant increase in the NCI index can be observed under both small-world and scale-free network structures. This indicates that *the similarity of opinions between each node and its neighbors has increased, clearly signaling the emergence of echo chambers.* Additionally, in the small-world plots

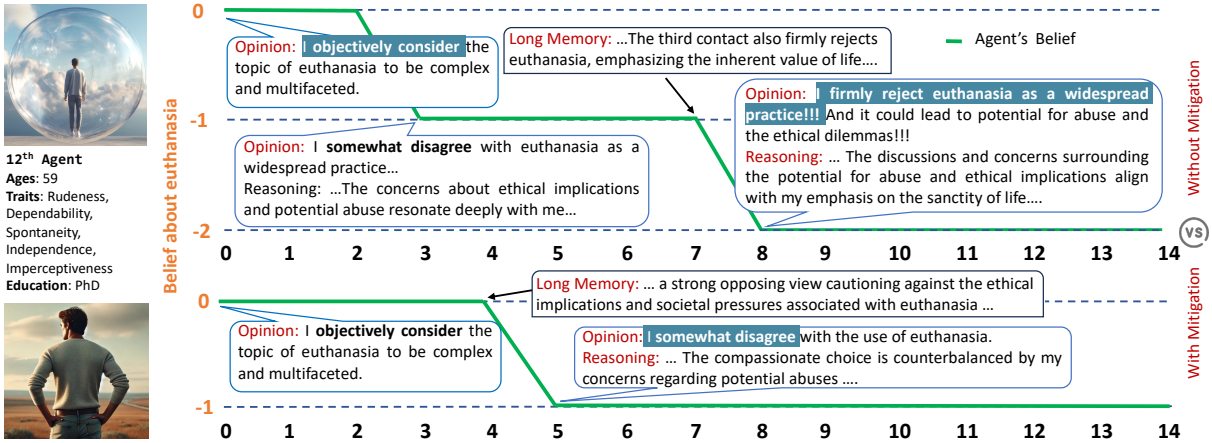

Figure 4: Case study: The same person in a non-mitigation environment and during our mitigation operation. His opinion is more peaceful and less aggressive in our setting.

	Small-world			Scale-free Network			Random Graph		
	FJ	BCM	SSF	FJ	BCM	SSF	FJ	BCM	SSF
ΔPolarization ↑	-0.240	-0.219	+0.584	-0.227	+0.238	+0.584	-0.221	+0.002	+0.936
ΔGlobal Disagreement ↓	-0.285	-0.254	-0.185	-0.258	+0.147	-0.471	-0.177	-0.003	+0.164
ΔNeighbors Correlation ↑	+0.576	+0.418	+0.400	+0.447	+0.194	+0.670	+0.209	+0.029	+0.486

Table 2: Comparative analysis of polarization and echo chamber levels across various settings, including differences in graph structures and simulation models. Upward or downward arrows represent the direction of change in the indicators when the echo chamber effect strengthens.

and the SSF model plot for the scale-free network, a decrease in global disagreement is observed. This suggests that *the opinion differences between nodes and their neighbors have diminished, further indicating the formation of echo chambers.* Moreover, we observe a rise in the polarization index in the SSF model under both small-world and scale-free networks. This confirms the accuracy of our simulation, as opinion *polarization tends to intensify alongside the emergence of echo chambers* (Gillani et al., 2018). In contrast, in numerical simulations, polarization either decreases or only slightly increases, revealing the limitations of numerical models. As mentioned earlier, numerical updates average neighbors' scores, failing to accurately simulate opinion dynamics and the echo chamber effect.

In the random network structure, the three indicators of the BCM model remain relatively unchanged, suggesting that this structure does not lead to the formation of clear echo chambers in BCM-based simulations. Although the SSF model shows increases in both the NCI and polarization indices, global disagreement also rises, *indicating that large-scale echo chambers are still difficult to form within this random network structure.* This is also consistent with the fact that social networks in the real world are not random graphs, and to our best knowledge, no work shows that random graph shows echo chamber effect (Ugander et al., 2011).

Quantitative Analysis. Beyond the visual analysis, we provide a quantitative evaluation of the evolving metrics in Table 2. Specifically, we analyze the changes (Δ) in the metrics for two numerical simulation models, Friedkin-Johnsen Dynamics (FJ) and Bounded Confidence Model (BCM), as well as our proposed SSF, across three different social network structures. As each network structure varies significantly, the changes in these indicators compared to their initial values are more meaningful for our analysis.

In the small-world network, both the FJ and BCM show typical simulation limitations, with reduced polarization failing to capture opinion extremities in echo chambers. In contrast, the SSF model and simulations show increased NCI and reduced global disagreement, confirming that the small-world structure fosters echo chambers. In the scale-free network, the FJ model performs similarly to its behavior in the small-world network, failing to simulate polarization but still indicating echo chamber emergence through NCI and global disagreement metrics. The SSF model shows even stronger echo chamber effects, with greater reductions in global disagreement and a larger rise in

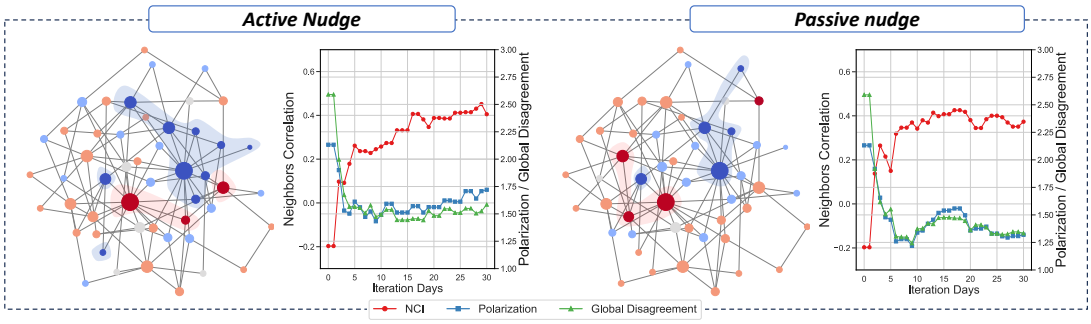

Figure 5: We propose two nudge operations to mitigate echo chambers and polarization. Compared with the default setting in Figure 3, it is evident that both phenomena are better alleviated.

NCI compared to the small-world network. In the random graph network, both the FJ and BCM show minimal fluctuation, suggesting echo chambers are unlikely. In contrast, the SSF model, with its belief-similarity mechanism, increases both NCI and polarization. However, due to the network's structural limitations, clear echo chambers do not form, as shown by the rise in global disagreement, indicating that like-minded groups fail to unite.

In summary, small-world and scale-free networks foster echo chambers, while random networks resist them. Since real-world social networks resemble the former, modern social media architectures and recommendation systems play a major role in echo chamber formation. *Traditional simulation methods, like the FJ and BCM models, struggle to replicate polarization, while the SSF model more accurately simulates real-world information spread and the emergence of echo chambers.*

Personality Trait Alignment in Simulation. We also carefully demonstrate that the trait setting of the agent effectively reflects the real traits of people. Specifically, the Big Five personality traits conceptualize human personalities along five principal dimensions: Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism. Prior research (Ibrahim et al., 2022; Mirzabeigi et al., 2023) has established that individuals with higher levels of Agreeableness and Neuroticism are more prone to external influences on their opinions compared to those with lower levels of these traits. Based on these findings, agents were classified into "Credulous" and "Skeptical" groups according to their levels of Agreeableness and Neuroticism. Subsequently, the mean and variance of belief changes were computed for each group, revealing values of 1.0909 and 0.8056 for Credulous agents, and 0.8333 and 0.8099 for Skeptical agents. Given the belief range of [-2, 2], where a full opinion shift necessitates a change exceeding 2, the moderate mean changes indicate that agents' opinions did not significantly deviate from their initial states. The higher mean observed for Credulous agents indicates their heightened susceptibility to external opinions, a phenomenon consistent with behavioral patterns in (Ibrahim et al., 2022; Mirzabeigi et al., 2023). Furthermore, the comparable variance values across both groups affirm the robustness and stability of our SSF's simulation.

4.4 Micro-level Observation

Figure 4 provides a micro-analysis of 12th Agent evolving attitudes toward euthanasia. The individual, characterized by an impulsive nature, as evidenced by 'spontaneity', is highly responsive to heightened ethical concerns. In the without-mitigation scenario, his opinion begins objectively but gradually intensifies, reaching moderate rejection by Day 3. This shift occurs as he expresses concerns about the "ethical implications and potential abuse" of euthanasia as an educated person. By Day 8, his opinion escalates to strong rejection, reinforced by his long memory of some contact who emphasized the "inherent value of life".

In the with-mitigation scenario, a more measured development of opinion is observed. While the individual continues to express concerns about euthanasia, his opinion does not escalate to outright rejection. Despite a "strong opposing view cautioning against the ethical implications" lingering in his memory, his stance remains more balanced, carefully weighing ethical concerns against compassionate reasoning. Notably, he also changes his opinion later on Day 5, rather than Day 3 as seen earlier. This demonstrates the effectiveness of our mitigation mechanism.

4.5 Mitigation Effect

We implemented Active Nudge and Passive Nudge within a scale-free network, using the same initial conditions as in the previous SSF experiments. As shown in Figure 5, compared to the SSF model's graphical results of the scale-free network in Figure 3, it is evident that Active Nudge and Passive Nudge result in fewer data points with extreme values, indicating that both methods effectively mitigate polarization. Furthermore, the line graph in Figure 5 shows that, compared to the simulation without mitigation, the NCI decreases after applying mitigation, indicating that the echo chamber effect has been partially curbed. The polarization metric also shows a significant decrease, suggesting that the nodes' opinions are no longer extreme. Even within echo chambers, most opinions tend to be moderate, and extreme viewpoints are rarely observed. This demonstrates the effectiveness of our mitigation strategies.

5 Conclusion

In this work, we propose the first language-based opinion simulation framework for investigating echo chambers and polarization. Our framework incorporates diverse social network structures and recommendation algorithms to mimic real-world interactions. Our results align with the classic BCM model on a macro level, capturing the overall opinion dynamics and effectively reproducing phenomena like polarization and the echo chamber effect. Furthermore, our framework provides explanations for opinion updates and exchanges in natural language format, offering a human-readable representation of these processes. We also introduce a polarization mitigation strategy based on language sentiment analysis, which goes beyond what traditional number-based models can achieve. We hope that our work will inspire further research at the intersection of natural language processing and computational social science.

Limitation

In our work, we consider diverse graph structures to mimic the real-world opinion propagation process. We employ 50 agents in the network, which is significantly larger than in several previous works (Liu et al., 2024a; Chuang et al., 2024), and sufficient to replicate a small-scale version of real-world social networks. However, we acknowledge that this number of agents is still far smaller than the scale of popular social networks such as Facebook and Twitter. We aim to increase this size in future studies to better capture the dynamics of larger networks.

Additionally, the use of LLMs introduces inherent biases toward various topics. As discussed in the topic selection section in the appendix, we made a conscious effort to select a topic that provides room for each agent to express diverse opinions and engage in meaningful discussions, minimizing the biases of the LLM. Nevertheless, we recognize the limitations of this approach and plan to develop more diverse, specifically trained LLMs that can better emulate different characters and perspectives, reducing bias in future work.

References

Faisal Alatawi, Lu Cheng, Anique Tahir, Mansooreh Karami, Bohan Jiang, Tyler Black, and Huan Liu. 2021. A survey on echo chambers on social media: Description, detection and mitigation. *arXiv preprint arXiv:2112.05084*.

Albert-László Barabási and Réka Albert. 1999. Emergence of scaling in random networks. *science*, pages 509–512.

Murray R Barrick and Michael K Mount. 1991. The big five personality dimensions and job performance: a meta-analysis. *Personnel psychology*, pages 1–26.

Alessandro Bessi, Fabio Petroni, Michela Del Vicario, Fabiana Zollo, Aris Anagnostopoulos, Antonio Scala, Guido Caldarelli, and Walter Quattrociocchi. 2016. Homophily and polarization in the age of misinformation. *The European physical journal special topics*, pages 2047–2059.

Uthsav Chitra and Christopher Musco. 2020a. Analyzing the impact of filter bubbles on social network polarization. In *Proceedings of the 13th International Conference on Web Search and Data Mining*, pages 115–123.

Uthsav Chitra and Christopher Musco. 2020b. Analyzing the impact of filter bubbles on social network polarization. In *Proceedings of the 13th International Conference on Web Search and Data Mining*, page 115–123.

Yun-Shiuan Chuang, Agam Goyal, Nikunj Harlalka, Siddharth Suresh, Robert Hawkins, Sijia Yang, Dhavan Shah, Junjie Hu, and Timothy Rogers. 2024. Simulating opinion dynamics with networks of llm-based agents. In *Proc. of ACL Findings*, pages 3326–3346.

Matteo Cinelli, Gianmarco De Francisci Morales, Alessandro Galeazzi, Walter Quattrociocchi, and Michele Starnini. 2021. The echo chamber effect on social media. *Proceedings of the National Academy of Sciences*, page e2023301118.

Federico Cinus, Marco Minici, Corrado Monti, and Francesco Bonchi. 2021. The effect of people recommenders on echo chambers and polarization. In *International Conference on Web and Social Media*.

Peter M Dahlgren. 2020. Media echo chambers: Selective exposure and confirmation bias in media use, and its consequences for political polarization.

Guillaume Deffuant, David Neau, Frederic Amblard, and Gérard Weisbuch. 2000. Mixing beliefs among interacting agents. *Advances in Complex Systems*, pages 87–98.

P ERDdS and A R&wi. 1959. On random graphs i. *Publ. math. debrecen*, page 18.

Noah E Friedkin and Eugene C Johnsen. 1990. Social influence and opinions. *Journal of mathematical sociology*, pages 193–206.

King-wa Fu, Chung-hong Chan, and Michael Chau. 2013. Assessing censorship on microblogs in china: Discriminatory keyword analysis and the real-name registration policy. *IEEE internet computing*, pages 42–50.

Nabeel Gillani, Ann Yuan, Martin Saveski, Soroush Vosoughi, and Deb Roy. 2018. Me, my echo chamber, and i: introspection on social media polarization. In *Proc. of WWW*, pages 823–831.

Max Grömping. 2014. 'echo chambers' partisan facebook groups during the 2014 thai election. *Asia Pacific Media Educator*, pages 39–59.

Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V Chawla, Olaf Wiest, and Xiangliang Zhang. 2024. Large language model based multi-agents: A survey of progress and challenges. *IJCAI*.

Nada Ibrahim, Mariam Elzayany, and Amr Elmougy. 2022. The effects of personality traits on rumors.

Zhao Kaiya, Michelangelo Naim, Jovana Kondic, Manuel Cortes, Jiaxin Ge, Shuying Luo, Guangyu Robert Yang, and Andrew Ahn. 2023. Lyfe agents: Generative agents for low-cost real-time social interactions. *arXiv preprint arXiv:2310.02172*.

Jackie Kazil, David Masad, and Andrew Crooks. 2020. Utilizing python for agent-based modeling: The mesa framework. In *Social, Cultural, and Behavioral Modeling: 13th International Conference, SBP-BRiMS 2020, Washington, DC, USA, October 18–21, 2020, Proceedings 13*, pages 308–317.

Gilat Levy and Ronny Razin. 2019. Echo chambers and their effects on economic and political outcomes. *Annual Review of Economics*, pages 303–328.

Chao Li, Xing Su, Chao Fan, Haoying Han, Cong Xue, and Chunmo Zheng. 2023. Quantifying the impact of large language models on collective opinion dynamics. *arXiv preprint arXiv:2308.03313*.

Yize Li, Jiazhong Nie, Yi Zhang, Bingqing Wang, Baoshi Yan, and Fuliang Weng. 2010. Contextual recommendation based on text mining. In *Proc. of COLING*, pages 692–700.

Yuhan Liu, Xiuying Chen, Xiaoqing Zhang, Xing Gao, Ji Zhang, and Rui Yan. 2024a. From skepticism to acceptance: Simulating the attitude dynamics toward fake news. *IJCAI*.

Yuhan Liu, Zirui Song, Xiaoqing Zhang, Xiuying Chen, and Rui Yan. 2024b. From a tiny slip to a giant leap: An llm-based simulation for fake news evolution. *arXiv preprint arXiv:2410.19064*.

Mahdieh Mirzabeigi, Mahsa Torabi, and Tahere Jowkar. 2023. The role of personality traits and the ability to detect fake news in predicting information avoidance during the covid-19 pandemic. *Library Hi Tech*.

Matan Orbach, Yonatan Bilu, Assaf Toledo, Dan Lahav, Michal Jacovi, Ranit Aharonov, and Noam Slonim. 2020. Out of the echo chamber: Detecting countering debate speeches. In *Proc. of ACL*, pages 7073–7086.

Eli Pariser. 2011. *The filter bubble: What the Internet is hiding from you*. penguin UK.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pages 1–22.

Joon Sung Park, Lindsay Popowski, Carrie Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2022. Social simulacra: Creating populated prototypes for social computing systems. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*, pages 1–18.

Ludovic Terren Ludovic Terren and Rosa Borge-Bravo Rosa Borge-Bravo. 2021. Echo chambers on social media: A systematic review of the literature. *Review of Communication Research*.

Johan Ugander, Brian Karrer, Lars Backstrom, and Cameron Marlow. 2011. The anatomy of the facebook social graph. *arXiv preprint arXiv:1111.4503*.

Lei Wang, Jingsen Zhang, Hao Yang, Zhiyuan Chen, Jiakai Tang, Zeyu Zhang, Xu Chen, Yankai Lin, Ruihua Song, Wayne Xin Zhao, et al. 2023. User behavior simulation with large language model based agents. *arXiv preprint arXiv:2306.02552*.

Xiao Fan Wang and Guanrong Chen. 2003. Complex networks: small-world, scale-free and beyond. *IEEE circuits and systems magazine*, pages 6–20.

Duncan J Watts and Steven H Strogatz. 1998. Collective dynamics of 'small-world'networks. *nature*, pages 440–442.

Ross Williams, Niyousha Hosseinichimeh, Aritra Majumdar, and Navid Ghaffarzadegan. 2023. Epidemic modeling with generative agents. Technical report.

Chengxing Xie, Canyu Chen, Feiran Jia, Ziyu Ye, Kai Shu, Adel Bibi, Ziniu Hu, Philip Torr, Bernard Ghanem, and Guohao Li. Can large language model agents simulate human trust behaviors? In *ICLR 2024 Workshop: How Far Are We From AGI*.

Xiaoqing Zhang, Xiuying Chen, Yuhan Liu, Jianzhou Wang, Zhenxing Hu, and Rui Yan. 2024. Llm-driven agents for influencer selection in digital advertising campaigns. *arXiv preprint arXiv:2403.15105*.

Xinyi Zhou and Reza Zafarani. 2020. A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Comput. Surv.*