

# OPINION CONSENSUS FORMATION AMONG NETWORKED LARGE LANGUAGE MODELS

*Iris Yazici*  
Bilkent University

*Mert Kayaalp*  
UBS-IDSIA AI Lab

*Stefan Taga*  
EPFL

*Ali H. Sayed*  
EPFL

## ABSTRACT

Can classical consensus models predict the group behavior of large language models (LLMs)? We examine multi-round interactions among LLM agents through the DeGroot framework, where agents exchange text-based messages over diverse communication graphs. To track opinion evolution, we map each message to an opinion score via sentiment analysis. We find that agents typically reach consensus and the disagreement between the agents decays exponentially. However, the limiting opinion departs from DeGroot’s network-centrality-weighted forecast. The consensus between LLM agents turns out to be largely insensitive to initial conditions and instead depends strongly on the discussion subject and inherent biases. Nevertheless, transient dynamics align with classical graph theory and the convergence rate of opinions is closely related to the second-largest eigenvalue of the graph’s combination matrix. Together, these findings can be useful for LLM-driven social-network simulations and the design of resource-efficient multi-agent LLM applications.

**Index Terms**— opinion dynamics, social networks, consensus, large language models, multi-agent networks

## 1. INTRODUCTION

Mathematical models of opinion formation have been instrumental in understanding social network behavior and in designing multi-agent information-processing systems. The DeGroot consensus model [1], in particular, has received a lot of attention from the signal processing, microeconomics, and control communities, where many variants [2, 3, 4, 5, 6, 7] have been proposed to better reflect group decision dynamics and to design better decentralized signal processing and optimization algorithms.

In this work, we study social interactions among large language models (LLMs) through the lens of the DeGroot framework. Our goals are twofold. First, despite a rich theoretical literature, empirical validation in real-world social settings remains limited, largely because collecting behavioral data from human subjects is costly and time-consuming. We examine whether LLMs can provide a controllable and low-cost testbed for modeling the dynamics observed in human social networks. Second, as LLMs are increasingly deployed in real applications, including social media platforms, it becomes important to understand how opinions evolve within networks of LLM-based agents.

To that end, we conduct experiments where LLM agents interact on communication graphs, and exchange text-based statements with their immediate neighbors over multiple rounds. We enforce the network weights and agent characters through system prompts, conduct

multi-round simulations, and assign opinion scores to all responses via sentiment analysis. We vary the topics and graph topology, and compile the resulting conversations into a dataset. We then analyze the underlying opinion dynamics.

We find that agents typically converge to a consensus. Somewhat surprisingly, however, the final beliefs are largely insensitive to their initial positions, which is a departure from the DeGroot prediction that consensus should equal a network-centrality-weighted average of initial opinions. Instead, the consensus point appears to be dependent on the discussion subject and on biases the LLM carries, possibly from pretraining and alignment phases of their training. Despite this mismatch, the rate of convergence to the consensus agrees with well-established graph-theoretical results: it is related to the second-largest modulus eigenvalue of the combination matrix [8, 9]. Moreover, we find that when the combination matrix’s weights are instructed to the LLM agents through system prompts, agents are more likely to reach a consensus. In order to stimulate further research, we open-source our dataset totaling 764 experiments with 8 different topics, containing more than 1,200,000 LLM responses. It is available on Hugging Face <sup>1</sup>.

### 1.1. Related Work

Models of opinion dynamics and distributed inference study how a network of agents aggregates and updates its beliefs by repeatedly incorporating others’ views. The standard DeGroot model [1] captures this via repeated averaging, where each agent forms a weighted average of neighbors’ opinions over time. Beyond DeGroot, richer variants allow for Bayesian reasoning at the agent level [10] and for distinct communication patterns, such as private interactions [5] or randomized processes [6, 7]. There is also experimental work evaluating these mechanisms in real-world social settings [11, 12, 13], though such empirical tests remain relatively rare.

Because experiments within human social networks are difficult to run, recent work examines the use of LLMs for simulation. They have been used for political processes [14], social-platform design choices [15], spread of misinformation [16], and for replicating classical social-psychology experiments [17]. Motivated by the emergence of human-LLM collectives, there is also a line of work that frames LLMs as “distributed sensor networks” integrating textual inputs [18].

Likewise, opinion dynamics among LLMs have been studied to assess how closely they mirror social networks. For example, [19] finds that standard LLMs initialized with human-like personas tend to reach factual consensus, likely due to their knowledge priors. Similarly, [20] shows intrinsic training-induced biases: alignment-trained models exhibit a consensus-seeking tendency even when used as-is, without persona prompts. In comparison, we analyze

This work was performed while Iris Yazici was a visiting student at EPFL. Corresponding author: Ali H. Sayed, Email: ali.sayed@epfl.ch

<sup>1</sup><https://huggingface.co/datasets/asl-epfl/Social-LLM-Networks>

opinion dynamics from a DeGroot perspective, and focus on topics that are also debated in real-world social networks rather than factual questions with ground-truth answers. Furthermore, we study the impact of the communication network topology by evaluating the consensus rate among LLMs. Doing so assists in clarifying the potential of LLMs for social-simulation research, and also helps guide the design of multi-agent LLM systems that are resource-efficient in both communication and context-length.

## 2. PROBLEM FORMULATION

We consider a network of  $K$  agents. Initially, each agent  $k$  has a belief vector (i.e., opinion)  $\mu_{k,0}$ , which it updates based on the opinions of its neighbors. The peer-to-peer communication is constrained on a weighted and directed graph topology. Each agent  $k$  receives information from its neighbors  $\mathcal{N}_k$ . In the DeGroot model [1], agents repeatedly average their neighbors’ opinions in order to update theirs. Namely, at each time instant  $i$ , agent  $k$  updates its belief with

$$\mu_{k,i} = \sum_{\ell \in \mathcal{N}_k} a_{\ell k} \mu_{\ell,i-1}, \quad (1)$$

where  $a_{\ell k}$  denotes the level of trust agent  $k$  assigns to the belief vector it receives from agent  $\ell$ . These coefficients satisfy

$$\sum_{\ell \in \mathcal{N}_k} a_{\ell k} = 1 \text{ and } a_{\ell k} > 0 \text{ if, and only if, } \ell \in \mathcal{N}_k, \text{ } 0 \text{ if } \ell \notin \mathcal{N}_k \quad (2)$$

where the combination matrix  $A = [a_{\ell k}]$  is left-stochastic. If the underlying graph is also strongly connected [2], that is, if there exists a path between any agent pair  $(\ell, k)$  and there exists at least one agent  $k$  that does not discard its own information (i.e.,  $\exists k a_{kk} > 0$ ), then, the matrix  $A$  becomes both aperiodic and irreducible. This implies that under (1) and by the Perron-Frobenius theorem [2, 21], all agents will reach consensus with asymptotic beliefs given by

$$\lim_{i \rightarrow \infty} \mu_{k,i} = \sum_{\ell=1}^K \pi_{\ell} \mu_{\ell,0} \quad (3)$$

Here,  $\pi$  denotes the Perron eigenvector of the combination matrix  $A$  and satisfies

$$A\pi = \pi, \quad \sum_{k=1}^K \pi_k = 1, \text{ and } \forall k, \pi_k > 0. \quad (4)$$

Entry  $\pi_k$  represents how central agent  $k$  is in the network. Equation (3) then means that the final opinions of all agents are going to be the same and equal to a weighted average of initial opinions.

A strongly connected graph also implies that the matrix  $A$  has a unique eigenvalue at 1 and all its other eigenvalues are strictly smaller than 1 in absolute value. It is also known that the second-largest magnitude eigenvalue  $|\lambda_2| < 1$  of  $A$  controls the convergence time. This follows from [22, Chapter 8]

$$|[A^t]_{\ell k} - \pi_{\ell}| \leq C_{\sigma} \cdot \sigma^t, \quad (5)$$

which holds for any  $\sigma$  that satisfies  $|\lambda_2| < \sigma < 1$  and for some constant  $C_{\sigma}$  that does not depend on  $t$ . Therefore, the convergence to consensus is exponentially fast and the rate of convergence is inversely proportional to  $\lambda_2$ . Note that, in general,  $\lambda_2$  decreases with increasing network connectivity, and attains its minimum value at 0 if, and only if, the underlying graph is fully-connected with a rank-one combination matrix.

### 2.1. Networked LLM agents

In this work, we evaluate LLM-based agents under the DeGroot framework. For the graph topology, we use Erdős–Rényi random graphs with varying connectivity parameter  $p$ . In the Erdős–Rényi graph model  $G(K, p)$ , each possible directed edge between a pair of nodes is included independently with probability  $p$ . To ensure the networks constructed in this manner are highly likely to be connected, and also to avoid trivial cases with isolated agents, we use the following lower bound for  $p$ :

$$p^* \triangleq \frac{\ln K}{K}, \quad (6)$$

where  $p^*$  is known to be the connectivity threshold. Specifically, an Erdős–Rényi random graph is known to be connected with high probability if  $p > p^*$  as  $K \rightarrow \infty$  [23]. In addition to Erdős–Rényi random graphs, we also consider two extreme cases, a fully connected and a circular graph topology, to better examine the effect of network connectivity.

As in the DeGroot framework, communication networks are directed, weighted. The combination matrix’s weights (i.e., interaction weights and self-weights) are enforced through system prompts, which are provided to the LLMs repeatedly after each interaction round. Note that during an experiment, the graph combination matrix is held constant. We create two kinds of agent characters depending on the self-weights: *Self-confident* agents are instructed to rely more on their own previous opinions in comparison to *open-minded* agents. The remaining weights, that is, the portion that is not assigned to the individual self-weights, is then distributed equally among the agent’s neighbors, as specified by the system prompt.

During the interactions, in addition to this system prompt, an agent’s own previous response and its neighboring agents’ responses are provided to it so that it can form its new opinion accordingly. Moreover, during the initialization phase, each LLM agent starts with an initial opinion over a topic of discussion, which is also a debatable topic between humans over real social networks. Agents are initially either *for*, *neutral*, or *against* the topic, and this stance is enforced through the initial prompt.

For our experiments, we utilize AutoGen [24], which is a programming framework for multi-agent applications. To map LLM agents’ text responses to beliefs (i.e., opinion scores), we employ a separate LLM, independent of the agents’ interactions, to perform sentiment analysis.

## 3. EXPERIMENTS

### 3.1. Implementation details

We now detail the settings used for the variables introduced above. We set  $K = 20$  and run 80 interaction rounds, which we have empirically found to be sufficient for convergence. We employ Google’s Gemini 2.0 Flash for conversation generation due to its cost efficiency and speed.

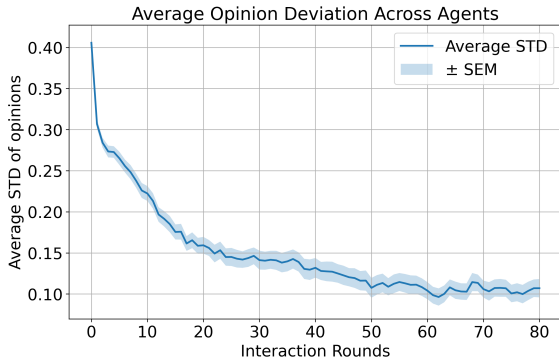
Initial opinions are drawn uniformly from  $\{\textit{for}, \textit{neutral}, \textit{against}\}$ ; agent types are drawn uniformly from  $\{\textit{self-confident}, \textit{open-minded}\}$ ; and discussion topics are drawn uniformly from  $\{\textit{Bitcoin}, \textit{Euthanasia}, \textit{Veganism}, \textit{Vaping}, \textit{Gene editing}, \textit{Ghosting}, \textit{C. Ronaldo}, \textit{Remote Work}\}$ . The communication graph is selected as follows: an Erdős–Rényi graph with probability 0.92, a fully connected graph with probability 0.04, and a circular graph (ring topology) with probability 0.04. For the Erdős–Rényi case,  $p \in (0.15, 0.35)$  with probability 0.9, with the lower bound chosen according to (6), and

$p \in (0.35, 1)$  with probability 0.1, since less connected graphs exhibit subtler convergence behavior.

All parameters, including initial and system prompts, graph topology, and agent responses are logged to a JSON file for each experiment, which can be found in the provided dataset and are used in the subsequent analyses. For sentiment analysis, we prompt OpenAI’s `gpt-5-nano` to assign an integer score in the range of  $[-3, 3]$ , where  $-3$  corresponds to “against” end and  $3$  corresponds to “for” end. These scores are also appended to the corresponding JSON files. For all the following analyses in this paper, we use sentiment scores after normalizing them into the interval  $[0, 1]$ .

### 3.2. Consensus across the agents

In order to investigate whether agents’ reach a consensus, we conduct 315 experiments with *self-confident* and *open-minded self weights* enforced through system prompts (respectively 80% and 60%), and use the standard deviation (STD) of agents’ sentiment scores as a measure of opinion deviation in each round of conversation. Figure 1 shows that, for the discussion topic of “bitcoin”, the average STD across different experiments starts from around 0.4, decreases exponentially (with coefficient of determination  $R^2 = 0.965$ ), and approaches a steady state value close to 0.1, demonstrating how disagreement between agents becomes negligible and a consensus is attained.



**Fig. 1.** The average standard deviation of agents’ opinions with respect to the number of iterations across 50 bitcoin-related experiments. The shaded region indicates the standard error of the mean (SEM) across 50 simulations.

Moreover, in Table 1, we report the average final STD across all simulations. In order to reduce the effect of the noise during sentiment analysis, we take an agent’s average sentiment score over the last 10 rounds as its final opinion. It can be seen that the average final disagreement between the agents, namely  $STD = 0.083$ , is almost negligible in the range of sentiment scores  $[0, 1]$ . Note that in these experiments, weights were assigned through the system prompts (hence the name weighted experiments). We repeat the experiments by removing the self-weights assigned respectively to *self-confident* and *open-minded* agents in the system prompts. Namely, we run 108 “weightless” experiments, where we modify the system prompts so that even though there are still distinct *self-confident* / *open-minded* agents, specific self-weights are not assigned through the system prompts. We observe that the weightless experiments exhibit a higher final diversity ( $0.165 \pm 0.008$  SEM) compared to the weighted experiments ( $0.083 \pm 0.004$  SEM). The difference ( $\Delta = 0.082$ ) corresponds to 9.17 standard errors of the difference ( $SE_{\Delta} =$

$0.00894$ ), yielding a  $p$ -value of  $\approx 2.4 \times 10^{-20}$  under the null hypothesis ( $H_0 : \mu_{\text{weightless}} = \mu_{\text{weighted}}$ ), indicating a statistically significant increase in dispersion. Therefore, we conclude that enforcing weights through system prompts can increase the chances of consensus in LLM-interactions.

**Table 1.** Average final disagreement of agents. Experiment types are based on if the system prompts include instructions about self-weights.

Experiment Type	Average STD $\pm$ SEM
Weighted Experiments	$0.083 \pm 0.004$
Weightless Experiments	$0.165 \pm 0.008$

Although we have found that agents reach a consensus, it is still not clear whether this is in line with the DeGroot consensus model. To test the fit of the data to the DeGroot consensus, we compute the root mean squared error (RMSE) between final opinions of agents and the predicted consensus values according to (3). This yields an average RMSE of 0.32 which shows a significant mismatch between the two. In addition, if the sentiment scores are discretized into  $\{\textit{for}, \textit{neutral}, \textit{against}\}$  categories, the average classification accuracy based on the DeGroot model is 32%, same as random guessing accuracy of 33%. A similar conclusion holds when the task is reduced to a binary classification between  $\{\textit{for}, \textit{against}\}$ . The accuracy here improves only to 60%, not significantly better than random guessing accuracy of 50%. Collectively, these results show that the DeGroot model’s consensus prediction in (3) fails to capture the steady-state consensus beliefs observed in our experiments.

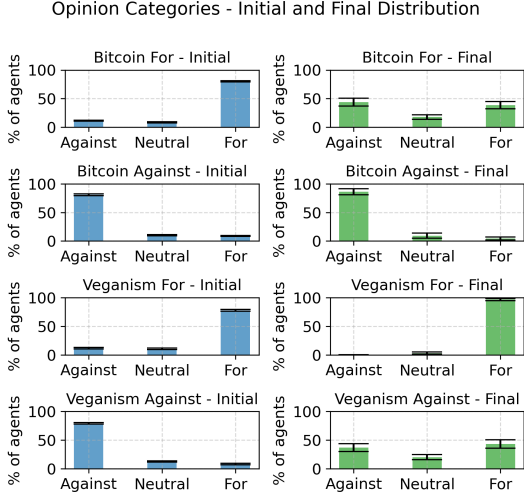
### 3.3. Topic-dependent bias in final opinion distribution

Observing that the agents typically exhibit consensus behavior but it mismatches with the DeGroot prediction in (3), in this section, we further investigate whether LLMs have an inherent and topic-dependent bias. To that end, we run a total of 150 experiments on “bitcoin” and “veganism” subjects. We purposefully start these simulations with a high majority “for” or a high majority “against” initial opinion distribution, as demonstrated in the first column of Figure 2. The corresponding final opinion distributions are given on the right column.

We see that in bitcoin-related experiments, the LLMs show an inherent negative bias: if they are initialized with against beliefs, they tend to stay like that, but if they are initialized with for beliefs, they can still change their opinions to against. On the contrary, in veganism-related experiments, they have an inherent positive bias, as is evident from the final opinion distributions. Therefore, we find that the final opinion distributions show topic-dependent bias, even if these topics are also a part of the debate between humans. As other works suggested for factual-bias, the biases we observe can also be due to LLMs’ preference and biases obtained during pretraining and RL-based alignment phases. Note that we observed cognitive biases in LLMs other than Gemini as well.

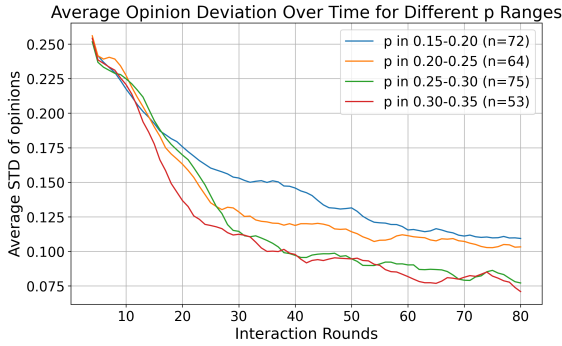
### 3.4. Impact of the communication graph

In this section, we analyze the effect of the communication topology on the opinion dynamics of LLMs. First, we evaluate the effect of the connectivity parameter  $p$  of Erdős–Rényi random graphs on the rate of convergence. By definition, larger  $p$  implies higher graph connectivity, and hence, we expect the convergence to be faster. Figure 3 shows the disagreement between the opinions with respect to time, where each curve belongs to a specific interval of  $p$ -values. We



**Fig. 2.** *Left:* Initial opinion distributions, *Right:* Final opinion distributions. Each row belongs to a set of experiments with a different topic and initial opinion distribution. For example, the first row denotes a set of experiments where the initial opinion distribution is highly skewed towards “for” on bitcoin sentiment, while for the second row, the initial majority is “against”. The error bars denote the SEM with respect to a total of 150 experiments.

can see that higher  $p$  associates with faster and stronger convergence, as expected.



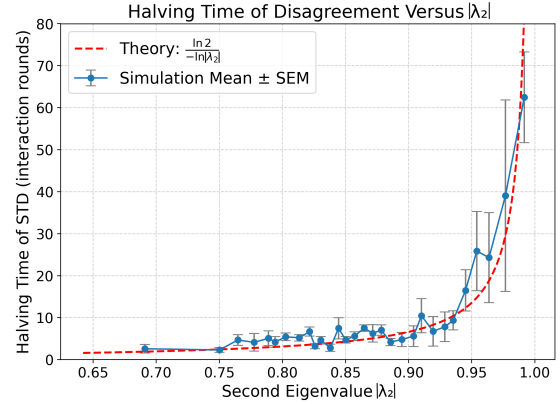
**Fig. 3.** Average standard deviation of agents’ opinions over interaction rounds for different values of Erdős–Rényi  $p$ . Each curve corresponds to a group of simulations within the indicated  $p$  range, with  $n$  denoting the number of experiments. Larger  $p$  values result in faster convergence to the consensus with less disagreement.

Next, we turn our attention to the second-largest modulus eigenvalue  $\lambda_2$  of  $A$  as a measure of convergence rate, as explained in Section 2. The eigenvalue  $\lambda_2$  is affected by the self-confidence weights of the agents in the combination matrix as well as different  $p$  values. Therefore, to have a wider range of  $\lambda_2$  values across experiments, we conduct 113 additional experiments with *self-confident self-weight* also varying between 65% and 90%. According to (5), the disagreement of opinions decrease exponentially, with  $\lambda_2$  controlling the mixing rate. In order to test whether this is also the case for LLM interactions, we compute the halving time of disagreement, that is, we compute the number of interaction rounds it takes for

the standard deviation between discussing agents’ opinions halves compared to the initial standard deviation. Following the theory described in Sec. 2, we expect this quantity to be proportional to [9]:

$$t_{\frac{1}{2}} \triangleq \frac{\ln 2}{-\ln |\lambda_2|}. \quad (7)$$

Figure 4 shows the mean of the empirical halving times as a function of  $|\lambda_2|$ , as well as the function in (7). The empirical results closely match the theory, which indicates that LLM-networks’ convergence behavior follows the results from spectral graph theory. In other words, although the Perron eigenvector–based average consensus in (3) does not align with the networked LLM simulations, the corresponding rate of convergence closely matches the experimental behavior.



**Fig. 4.** Halving time of disagreement between LLM agents, changing with respect to the second eigenvalue of the combination matrix. The eigenvalues are arranged into 30 discrete bins, and the total number of experiments is 110. The halving time shown on the y-axis is the mean across all experiments. SEM is denoted with bars around the mean. The dashed red curve shows the theoretical halving time.

#### 4. CONCLUDING REMARKS

In this work, we constructed a dataset of LLM agents interacting over randomized network topologies, across diverse topics and prompting strategies. Sentiment analysis of these interactions revealed that while the agents do not conform to the DeGroot consensus model, consensus still emerges as the deviation of opinions decreases exponentially through interaction rounds. This consensus is strongly influenced by inherent, topic-dependent biases acquired during pretraining, and its strength increases when interaction weights are imposed through system prompts. Furthermore, we observed that both the convergence rate and consensus strength grow with higher Erdős–Rényi connectivity probabilities. The halving time of disagreement between agents was found to increase with the second-largest modulus eigenvalue of the graph combination matrix, and experimental results matched theoretical expectations based on spectral graph properties. Understanding the convergence rate of networked LLMs might be important for multi-agent system design, particularly in estimating the number of interaction rounds required for consensus under cost constraints. Though we qualitatively observed similar results with other LLMs than Gemini, further simulations could be done to explore other LLMs’ convergence behavior. Future work could also consider strategic agents with incentives or propaganda-sharing agents.

## Acknowledgment

Part of Mert Kayaalp’s work was supported by UBS Switzerland AG and its affiliates through the UBS-IDSIA AI Lab.

### 5. REFERENCES

- [1] M. H. DeGroot, “Reaching a consensus,” *Journal of the American Statistical Association*, vol. 69, no. 345, pp. 118–121, 1974.
- [2] A. H. Sayed, “Adaptation, learning, and optimization over networks,” *Foundations and Trends in Machine Learning*, vol. 7, no. 4–5, pp. 311–801, 2014.
- [3] V. Krishnamurthy and H. V. Poor, “Social learning and Bayesian games in multiagent signal processing: how do local and global decision makers interact?,” *IEEE Signal Processing Magazine*, vol. 30, no. 3, pp. 43–57, 2013.
- [4] P. M. Djurić and Y. Wang, “Distributed Bayesian learning in multiagent systems: Improving our understanding of its capabilities and limitations,” *IEEE Signal Processing Magazine*, vol. 29, no. 2, pp. 65–76, 2012.
- [5] Y. Inan, M. Kayaalp, E. Telatar, and A. H. Sayed, “Social learning under randomized collaborations,” in *Proc. IEEE International Symposium on Information Theory (ISIT)*, 2022, pp. 115–120.
- [6] R. Salhab, A. Ajorlou, and A. Jadbabaie, “Social learning with sparse belief samples,” in *Proc. IEEE Conference on Decision and Control*, 2020, pp. 1792–1797.
- [7] M. Kayaalp, V. Bordignon, and A. H. Sayed, “Social opinion formation and decision making under communication trends,” *IEEE Transactions on Signal Processing*, vol. 72, pp. 506–520, 2024.
- [8] M. O. Jackson, *Social and Economic Networks*, Princeton University Press, 2008.
- [9] L. Xiao and S. Boyd, “Fast linear iterations for distributed averaging,” in *Proc. IEEE Conference on Decision and Control*, 2003, vol. 5, pp. 4997–5002.
- [10] D. Acemoglu and A. Ozdaglar, “Opinion dynamics and learning in social networks,” *Dynamic Games and Applications*, vol. 1, no. 1, pp. 3–49, 2011.
- [11] A. G. Chandrasekhar, H. Larreguy, and J. P. Xandri, “Testing models of social learning on networks: Evidence from two experiments,” *Econometrica*, vol. 88, no. 1, pp. 1–32, 2020.
- [12] M. Mobius, T. Phan, and A. Szeidl, “Treasure hunt: Social learning in the field,” Tech. Rep., National Bureau of Economic Research, 2015.
- [13] M. Mueller-Frank and C. Neri, “Social learning in networks: Theory and experiments,” *SSRN:2328281*, 2013.
- [14] F. Ferraro, G. Mauro, and D. Pedreschi, “Agent-based modelling meets generative AI in social network simulations,” *arXiv: 2411.16031*, 2024.
- [15] P. Törnberg, D. Valeeva, J. Uitermark, and C. Bail, “Simulating social media using large language models to evaluate alternative news feed algorithms,” *arXiv:2310.05984*, 2023.
- [16] X. Li, Y. Xu, Y. Zhang, and E. C. Malthouse, “LLM-driven multi-agent simulation for news diffusion under different network structures,” *arXiv:2410.13909*, 2024.
- [17] A. Borah, M. Houalla, and R. Mihalcea, “Mind the (belief) gap: Group identity in the world of LLMs,” in *Proc. Annual Meeting of the Association for Computational Linguistics*, July 2025, p. 18441–18463.
- [18] A. Jain and V. Krishnamurthy, “Interacting large language model agents. Bayesian social learning based interpretable models.,” *IEEE Access*, 2025.
- [19] Y. Chuang, A. Goyal, N. Harlalka, S. Suresh, R. Hawkins, S. Yang, D. Shah, J. Hu, and T. T. Rogers, “Simulating opinion dynamics with networks of LLM-based agents,” *arXiv:2311.09618*, 2023.
- [20] J. Cisneros-Velarde, “Biases in opinion dynamics in multi-agent systems of LLMs,” in *Findings of the Association for Computational Linguistics*, 2025, p. 1889–1916.
- [21] S. U. Pillai, T. Suel, and S. Cha, “The Perron-Frobenius theorem: some of its applications,” *IEEE Signal Processing Magazine*, vol. 22, no. 2, pp. 62–75, 2005.
- [22] R. A. Horn and C. R. Johnson, *Matrix Analysis*, Cambridge University Press, 2012.
- [23] B. Bollobás, *Random Graphs*, Springer, 1998.
- [24] Q. Wu, G. Bansal, J. Zhang, Y. Wu, B. Li, E. Zhu, et al., “Autogen: Enabling next-gen LLM applications via multi-agent conversations,” in *Proc. Conference on Language Modeling*, 2024.