Bots of a Feather: Mixing Biases in LLMs' Opinion Dynamics

Erica Cau[1,2], Andrea Failla[1,2], and Giulio Rossetti[2]

[1] Department of Computer Science, University of Pisa
56127 Pisa, Italy
[name].[surname]@phd.unipi.it,
[2] ISTI-CNR, National Research Council
56124 Pisa, Italy
giulio.rossetti@isti.cnr.it

Abstract. The rapid integration of Large Language Models (LLMs) into everyday applications raises critical questions about their group interactions, consensus formation, and potential to mimic human-like behavior. Although initial research has explored the evolution of opinions within LLM populations, these efforts often rely on simplistic network assumptions, such as uniform connections among agents, thereby overlooking the influence of more realistic network topologies. This paper introduces a framework for examining opinion dynamics among LLM agents within various network structures. We perform several multi-model simulations on network topologies with known locally assortative/disassortative mixing patterns. We find that convergence is quicker in mostly-disassortative networks compared to networks with no mixing biases. However, the joint effect of assortative and disassortative patterns leads to slower/no convergence.

Keywords: Large Language Models, Opinion Dynamics, Homophily, Social Networks

1 Introduction

Human societies are gradually integrating Artificial Intelligence (AI) models in their daily lives, to solve a variety of both trivial and complex tasks. Large Language Models (LLMs) are instances of these models that are able to simulate coherent human-like conversations [3], mimic behaviors of specific personae [20], and even achieve remarkable accuracy in complex natural language processing tasks such as translation [3], sentiment analysis [1], and stance detection [23]. As the presence of these models becomes more widespread, it is essential to understand how they interact in large groups, reach consensus and/or disagreement, and whether they can replicate human-like behaviors without any specific training but as an emergent ability [41]. In this direction, recent work in psychology suggests that LLMs might have developed a *Theory of Mind* [11, 21, 4], i.e., the ability to impute mental states to themselves and others. If they do have a Theory of Mind, then LLMs are able to manifest emotions, beliefs and desires in

a coherent way; most importantly, they are able to understand/predict these mental states when interacting with others. Consequently, LLMs' beliefs could be studied/modeled in the same ways we currently study/model human ones [11, 36]. Some works have already investigated how the opinions of LLM populations evolve by leveraging statistical physics and network analysis tools [6, 7, 39]. Nevertheless, these studies often assume mean-field topologies (i.e., settings where each agent is connected with/affected by all other agents), and the effects of network topology remain largely unexplored. In this work, we address this research gap by introducing a general framework for opinion dynamics with LLM agents, describing an algorithm for LLM debates, and performing various simulations on different network configurations. We aim to understand whether locally assortative/disassortative initial conditions affect the evolution of LLMs' opinions.

The rest of this work is organized as follows. In section 2, we provide an overview of the main topics related to our study; section 3 formalizes a framework for studying LLMs' Opinion Dynamics; section 4 introduces the experimental campaign and discusses the results. Finally, in section 5, we conclude the work and suggest future research directions.

2 Related Work

In this section, we provide the reader with background on the main themes related to our study. First, we outline classical opinion dynamic models; then, we move to more complex LLM-based simulations.

Opinion Dynamics. Opinion dynamics (OD) can be understood as a process simulating how individuals' opinions change over time due to empirically observed social phenomena [5, 37]. OD simulations typically involve a set of connected agents (the *population*) holding opinions encoded as real numbers. Social phenomena are implemented as rules that govern how a population's individual opinions are updated over time, leveraging statistical physics tools, such as graph and probability theories. For instance, an individual's opinion could be updated by randomly copying a neighbor [19] (which captures the tendency to mimic others) by averaging the opinions in its neighborhood [13] (which captures homophilic tendencies [25]), by computing the most frequent opinion therein [22] (which captures peer pressure and mutual awareness effects [10]). The goal of OD is generally to understand whether the interplay of the considered social mechanisms eventually leads to one of the following mutually exclusive scenarios: (i) consensus, when all agents hold the same opinion; (ii) polarization, when opinions are clustered around two distinct poles; (iii) fragmentation, when opinions are clustered around more than two poles [28, 38].
Moreover, OD models are often categorized based on how opinion variables are defined. In *discrete* opinion models, opinions can take on a limited set of distinct values, and rules dictate changes between them. The Voter Model [9], for instance, focuses on scenarios where individuals can adopt one of two opposing

stances (e.g., agree/disagree). In contrast, *continuous* opinion models represent opinions as points within a continuous spectrum. Classical continuous models, such as the DeGroot Model [14], assume that any opinion within the range [-1, 1] is possible, and interactions between individuals can shift opinions gradually.

LLM-based social simulations Recent research has focused on analyzing the dynamics of interacting LLMs. Generally, LLM agents show believable behaviors both at the individual and collective levels [30], and human-like behaviors naturally emerge from their interactions, e.g., scale-free networks [12], information spreading [18], and trust [42]. When prompted with demographic information and personality traits, they more closely reproduce human behavior [31, 30], including human biases and political leaning [6]. The combination of LLMs and traditional Agent-Based modeling [24] as in [40] demonstrated the impact of recommender systems over the quality of the discussion, showing that the exposure of agents to popular contents foster toxicity and interpartisan interaction, reaching levels of toxicity similar to the ones found on tweets by people living in the U.S. in 2019. Additionally, simulations have shown how confirmation bias among agents can lead to fragmentation [6], in line with research on opinion dynamics. LLMs can replicate humans' persuasion dynamics [2, 26], producing well-reasoned arguments incorporating psycho-linguistic opinion change theories and can be leveraged for completed social media simulations [34]. However, LLMs have an inherent bias towards accurate information and, unless prompted accurately, avoid producing utterances that conflict with scientific fact [6]. This can cause agents to ignore their assigned personality or role [39], leading to less believable behaviors. LLM agents also tend to avoid harsh conflicts [8, 40] and are sensible to several social/political topics [35, 16].

3 LLM Opinion Dynamics: A General Framework

In the following, we describe the methodological framework used to simulate LLM discussions and track their opinions. We define our framework for LLM Opinion Dynamics as the following tuple:

$$\mathfrak{M} = \langle \mathcal{G}, \mathcal{S}, \mathcal{O}, f, T, \mathcal{D} \rangle, \tag{1}$$

where:

- $\mathcal{G} = \langle V, E \rangle$ is a graph structure, such that V denotes the set of LLM agents and $E \subseteq V \times V$ denotes the set of unordered agent pairs (i, j) that can engage in a discussion. $\mathcal{G}$ may be a complete graph (i.e., all possible edges exist, a scenario referred to as *mean field*), as well as any real or simulated topology ;
- $\mathcal{S}$ is a statement, namely a claim that is used to initiate the discussion. The statement sets the topic of discussion for the simulation and will serve as a reference for interpreting opinion values.

- $\mathcal{O}$ represents the opinion space, namely the set of possible opinions. Depending on the context, this can be either discrete (e.g., $\mathcal{O} = \{0, 1\}$ for binary opinions) or continuous (e.g., $\mathcal{O} = [0, 1]$ for opinions ranging between 0 and 1). Opinions depend on the topic set by the statement $\mathcal{S}$, and range from perfect disagreement ($min(\mathcal{O})$) to perfect agreement ($max(\mathcal{O})$);
- $f : \mathcal{O}^n \times \mathbb{R}^m \to \mathcal{O}$ is a function that describes how an agent updates its opinion based on its current opinion, the opinions of its neighbors, and possibly other parameters or external influences. Here, n is the number of neighbors and m represents other influencing factors. Note that in an LLM-powered model, the full form of f is unknown, as it also depends on the LLM's (unknown) parameters.
- T defines the temporal evolution of the system, namely an ordered set of adjacent positive integers $t \in \mathbb{Z}^+$.
- $\mathcal{D}(t) = \{o_i(t) \mid i \in V\}$ is the distribution of opinions over all agents at time t, where $o_i(t) \in \mathcal{O}$ represents the opinion of agent i at time t.

This framework generalizes to all social topologies, interaction scenarios, and discussion topics. In the following, we introduce a specialization of the framework used in the experimental campaign. A general overview of its rationale is provided in Algorithm 1. In detail, we start by defining the graph topology, the

Algorithm 1 LLM Opinion Dynamics

Require: Graph: $\mathcal{G}$, Statement: $\mathcal{S}$, Initial opinions: $\mathcal{D}(0) = \{o_i(0) \mid i \in V\}$, Time steps: $T_{\max}$,
Ensure: Opinion Distribution Evolution $\mathcal{D}_{all}$
Init $\mathcal{D}_{all}$ as an empty list
$\mathcal{D}_{all}$.append($\mathcal{D}(0)$)
for $t = 0$ to $T_{\max} - 1$ **do**
  for each agent $i \in V$ **do**
    Select random neighbor $v_j \in Neighbors(i)$
    $\Delta o_i(t) = debate(o_i(t), o_j(t), \mathcal{S})$
    Update opinion: $o_i(t+1) = o_i(t) + \Delta o_i(t)$
  end for
  Update opinion distribution: $\mathcal{D}(t+1) = \{o_i(t+1) \mid i \in V\}$
  $\mathcal{D}_{all}$.append($\mathcal{D}(t+1)$)
end for
return $\mathcal{D}_{all}$

initial statement, the opinion distribution, and the number of iterations. Each agent is initialized with a discrete opinion on a 7-point Likert scale, such that 0 corresponds with strong disagreement with S, 3 corresponds with neutrality, and 6 denotes strong agreement. At each time step, for each agent $i \in V$ a neighbor j is randomly selected to engage in a debate on $\mathcal{S}$. As a result of the debate, i's opinion is either (i) brought closer to j's if they agree, (ii) brought farther from j's if they disagree, or (iii) not modified if j's arguments are ignored. Opinion updates are stored at each iteration, and returned at the end.

The debate unfolds according to Algorithm 2. The selected agents are first assigned one of two mutually exclusive roles, namely discussant (agent i) or

opponent (agent j). The discussant is the agent who starts the conversation. It does so by expressing and justifying its opinion on S. The discussant is the only agent allowed to change its opinion (after engaging with the opponent). The opponent is tasked to produce a persuasive decision with the goal of convincing the discussant. It is a stubborn agent, i.e., it is not allowed to change opinion. The debate proceeds in rounds, where each round consists of one inter-

Algorithm 2 Debate

Require: Initial opinions: o_i, o_j, Statement: $\mathcal{S}$, Maximum number of interaction rounds: `max_rounds`
Ensure: Final opinion delta of the Discussant: $o_i^{\Delta+}$

Init `Discussant`'s Opinion Delta $o_i^{\Delta+} \leftarrow 0$
`Discussant` states its opinion o_i on $\mathcal{S}$
for `round` = 1 to `max_rounds` = 3 **do**
    Opponent's Turn:
    `Opponent` evaluates `Discussant`'s argument
    `Opponent` decides and justifies: `ACCEPT`, `REJECT`, `IGNORE`
    if decision is `IGNORE` or `ACCEPT` **then**
        break
    end if
    Discussant's Turn:
    `Discussant` evaluates `Opponent`'s argument
    `Discussant` decides and justifies: `ACCEPT`, `REJECT`, `IGNORE`
    if decision is `ACCEPT` **then**
        $o_i^{\Delta+} \leftarrow 1$ if $o_j > o_i$ else -1
        break
    else if decision is `REJECT` **then**
        $o_i^{\Delta+} \leftarrow$ -1 if $o_j > o_i$ else 1
        break
    end if
end for
return $o_i^{\Delta+}$

action from each agent. The total number of rounds is capped by a maximum limit, `max_rounds`, to ensure the debate does not continue indefinitely. Initially, the Discussant is prompted to express its opinion on $\mathcal{S}$. During each round, the Opponent first evaluates the Discussant's message and then decides whether to `ACCEPT`, `REJECT`, or `IGNORE` the Discussant's viewpoint. If the Opponent chooses to `IGNORE` or `ACCEPT`, the debate concludes. Otherwise, the Discussant evaluates the Opponent's argument, and the debate continues. The final opinion shift of the Discussant, $o_i^{\Delta+}$, is determined based on its final choice. If the Discussant `ACCEPT`s, its opinion is brought closer to the Opponent's, and if it `REJECT`s, it is brought further (e.g., backfire effect [29]). the Opponent's argument. If the Discussant does neither, or if the debate terminates earlier, its opinion remains unchanged.

4 Opinion Dynamics Analysis

Analytical Setting. To study the effects of different node mixing patterns on LLM opinion evolution, we perform simulations on Peel's Quintet [32]. The Quintet is a collection of five graphs characterized by the same number of nodes

(40) and edges (160), the same Nominal Assortativity Coefficient [27], but different local mixing patterns (i.e., assortative/disassortative areas in each graph). Each node is assigned one of two labels (classes), and edges are placed so that the number of connections between and within classes is the same across the quintet. We provide a visual representation in Fig. 1.

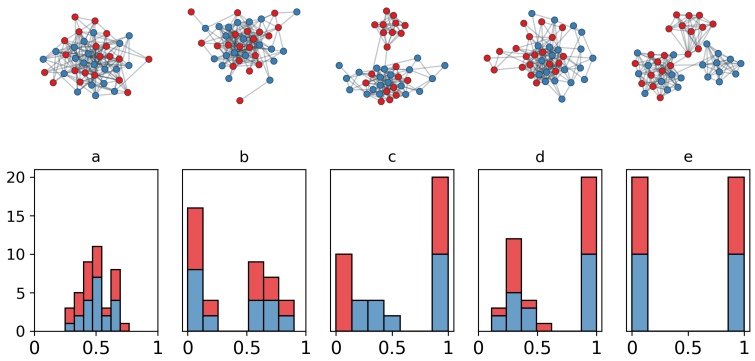

Fig. 1: Peel's Quintet. The top row displays the quintet using the Fruchterman-Reingold layout algorithm [17]. Colors identify node classes. The bottom row displays the distribution of the ratios of same-class neighbors for each node in the corresponding graph (later called *Homogeneity* (Eq. 2.))

The chosen discussion statement refers to the paradox of Theseus' Ship, first recorded in Plutarch's Parallel Lives [15]. In essence, the paradox debates whether the ship did or did not keep its identity after all of its parts were replaced with new ones. The rationale behind this choice is to mitigate LLM bias toward accurate scientific truth, as this causes a faster convergence and reduces the variety of persuasive arguments to support their claims [6]. The input statement in our experiments declares that the ship remains the same. We set red nodes in Fig. 1 to *strongly disagree* with the statement (opinion 0), and blue nodes to *strongly agree* (opinion 6). Since we are also interested in understanding potential behavioral differences among different language models, we employ two LLMs: `Mistral-7B Instruct` (Mistral, henceforth) and `Llama-3-8B` (Llama3) and perform two sets of simulations. In the first set, we set Mistral agents for all red nodes in Fig. 1, and Llama agents for blue ones. In the second set, we set Llama agents for red nodes, and Mistral agents for blue ones. Each simulation is observed for 100 iterations.

In the following, we first analyze opinion trends for both sets of simulations.

Then, we study the extent to which a node is surrounded by peers with the same opinion over time. To measure this, we define ego network *homogeneity* as follows:

$$Homogeneity(u) = \frac{|\{v : v \in neighbors(u) \wedge o_u = o_v\}|}{|neighbors(u)|}, \quad (2)$$

where u is the target node with opinion o_u, and v is any of u's neighbors such that they have the same opinion value. All code and data produced in this study are available in a dedicated GitHub repository. [3]

Opinion Trends. Fig. 2 displays opinion trends for both sets of experiments. Panels in the top row refer to the first set, and panels in the bottom one refer to the second. In all configurations, agents tend to accept opponents' opinions and ultimately stabilize on *agree* (5). This effect could be explained by the excessive politeness of LLMs, which are often assertive and tend to avoid conflicts [33].
The most noticeable difference relates to the rate at which negative opinions disappear. In the absence of biases (Figs. 2a and 2f), for instance, opinions take 70-80 iterations to converge. For mostly disassortative networks (Figs. 2b and 2g), convergence is quicker (20-60 iterations). In configurations where nodes show high assortativity, behaviors vary widely. Indeed, the systems in Figs. 2c and 2h, where nodes show high assortativity coupled with some disassortativity, do not converge (in the first simulation) or converge at the very end (in the second experiment). As mixing biases (i.e., both dis/assortative behaviors) become more prominent, systems take 75 iterations or more to converge in the first set of simulations (Figs. 2d and 2e). In the second set, similar behaviors can be noted when assortative and disassortative mixing are equally present (Fig. 2j). However, faster convergence can be observed when assortativity is high, and disassortativity is moderate. In general, in the second set of simulations (where we assign *strongly disagree* opinions to Llama agents and *strongly agree* opinions to Llama ones), we observe relatively faster convergence compared to the first set. Most notably, in Fig. 2g, the bias toward positive opinions leads to the unique emergence (and fast convergence) of *strongly agree*ing agents. We attribute this difference to the inherent biases of LLM models.
Neighborhood Homogeneity. Fig. 3 displays Homogeneity distributions over time for both sets of experiments. Figs. 3a-3e show results for the first set of simulations, and Figs. 3f-3j show results for the second. Note that the upper curve in each panel (labeled as `0`) displays the initial homogeneity values, that is the same observed in Fig. 1. Moving downwards, the curves outline the homogeneity distributions every 25 iterations. We observe the tendency of nodes toward local homogeneity, where most nodes show full homogeneity within 100 iterations in all scenarios. What is more, agents show mostly-homogeneous neighborhoods early on. While results in the previous section highlight the tendency of accepting the initial statement, here we also observe the tendency to gradually align with local opinions. When no mixing biases are present (Figs. 3a and 3f) LLM neighborhoods show both strong heterogeneity and homogeneity before converg-

[3] https://github.com/GiulioRossetti/LLM_Opinion_multi_model

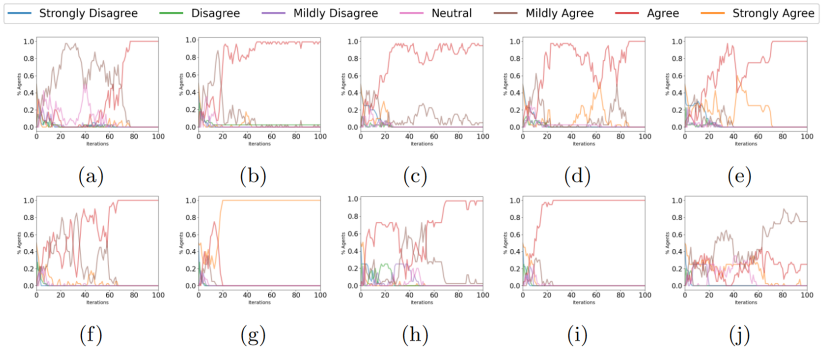

Fig. 2: Percentage of agents holding opinion o as a function of time. Opinions are mapped to the following labels: strongly disagree, disagree, mildly disagree, neutral, mildly agree, agree, and strongly agree. The top panels refer to simulations where Mistral agents are assigned the red class, and Llama agents are assigned the blue class. The bottom panels refer to the opposite scenario.

ing. Similarly, when both assortative and disassortative behaviors are present (Figs. 3e and 3j), neighborhoods maintain this duality, albeit with less prevalent heterogeneous patterns. In high-assortativity networks (Figs. 3c, 3d, 3h, 3i), instead, homogeneous neighborhoods dominate the social landscape, and only a negligible fraction show heterogeneous patterns before disappearing.

5 Discussion and Conclusions

In this paper, we formalized a framework for studying Opinion Dynamics of Large Language Models and evaluated how different local node mixing patterns impact the evolution of opinions. We observe that, in most cases, most agents tend to converge towards agreeing with the initial statement within the first 100 iterations, regardless of the extreme views in the initial state. Furthermore, upon reaching convergence, the agents do not adopt the most extreme opinion (6) but maintain a moderate one (5). We also find that mostly disassortative LLM societies show faster convergence with respect to cases where no mixing bias is present. However, the joint effect of assortative and disassortative patterns seems to lead to even slower/no convergence. We also highlight that LLM opinion dynamics are not oblivious to initial LLM configurations: in our simulations, assigning *strongly agree* opinions to Llama and *strongly disagree* opinions to Mistral agents leads to comparatively faster convergence than the opposite scenario. Finally, the homogeneity analysis unveiled LLMs' tendency to gradually align with local opinions, especially when the initial conditions are already assortative. Future work could replicate our study on larger populations or in more realistic simulation environments [34], investigating the role of other network properties in shaping LLMs' opinions, such as network density or different

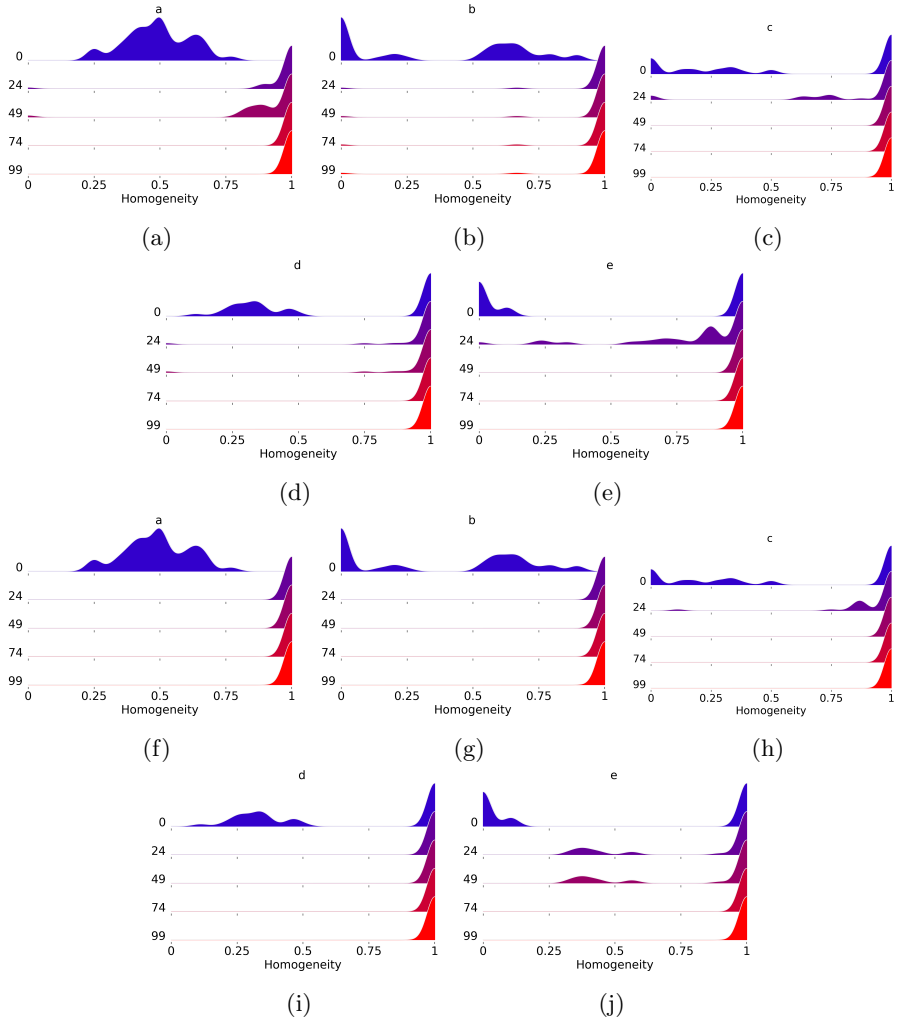

Fig. 3: Kernel Density Estimation of Homogeneity values over time for the first (a-e) and second (f-j) sets of simulations. In each panel, values are shown every 25 iterations, with the top curve being the starting condition (i.e., the same setting as Fig. 1, bottom row).

degree distributions. Moreover, by leveraging LLMs' textual productions, future works could also study language dynamics and whether/how they are affected by different network configurations.

Acknowledgements

This work is supported by (i) the European Union – Horizon 2020 Program under the scheme "INFRAIA-01-2018-2019 – Integrating Activities for Advanced Communities", Grant Agreement n.871042, "SoBigData++: European Integrated Infrastructure for Social Mining and Big Data Analytics" (http://www.sobigdata.eu); (ii) SoBigData.it which receives funding from the European Union – NextGenerationEU – National Recovery and Resilience Plan (Piano Nazionale di Ripresa e Resilienza, PNRR) – Project: "SoBigData.it – Strengthening the Italian RI for Social Mining and Big Data Analytics" – Prot. IR0000013 – Avviso n. 3264 del 28/12/2021; (iii) EU NextGenerationEU programme under the funding schemes PNRR-PE-AI FAIR (Future Artificial Intelligence Research).

References

1. Aroyehun, S.T., Malik, L., Metzler, H., Haimerl, N., Di Natale, A., Garcia, D.: Leia: Linguistic embeddings for the identification of affect. EPJ Data Science **12**(1), 52 (2023)
2. Breum, S.M., Egdal, D.V., Mortensen, V.G., Møller, A.G., Aiello, L.M.: The persuasive power of large language models. In: Proceedings of the International AAAI Conference on Web and Social Media, vol. 18, pp. 152–163 (2024)
3. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. Advances in neural information processing systems **33**, 1877–1901 (2020)
4. Brunet-Gouet, E., Vidal, N., Roux, P.: Do conversational agents have a theory of mind? a single case study of chatgpt with the hinting, false beliefs and false photographs, and strange stories paradigms (2023)
5. Castellano, C., Fortunato, S., Loreto, V.: Statistical physics of social dynamics. Reviews of modern physics **81**(2), 591–646 (2009)
6. Chuang, Y.S., Goyal, A., Harlalka, N., Suresh, S., Hawkins, R., Yang, S., Shah, D., Hu, J., Rogers, T.T.: Simulating opinion dynamics with networks of llm-based agents. arXiv preprint arXiv:2311.09618 (2023)
7. Chuang, Y.S., Harlalka, N., Suresh, S., Goyal, A., Hawkins, R., Yang, S., Shah, D., Hu, J., Rogers, T.T.: The wisdom of partisan crowds: Comparing collective intelligence in humans and llm-based agents. In: Proceedings of the Annual Meeting of the Cognitive Science Society, vol. 46 (2024)
8. Cisneros-Velarde, P.: On the principles behind opinion dynamics in multi-agent systems of large language models (2024). URL https://arxiv.org/abs/2406.15492
9. CLIFFORD, P., SUDBURY, A.: A model for spatial conflict. Biometrika **60**(3), 581–588 (1973). DOI 10.1093/biomet/60.3.581. URL http://dx.doi.org/10.1093/biomet/60.3.581
10. Crespi, I.: The public opinion process: How the people speak. Routledge (2013)
11. Cuzzolin, F., Morelli, A., Cîrstea, B., Sahakian, B.J.: Knowing me, knowing you: theory of mind in ai. Psychological Medicine **50**(7), 1057–1061 (2020). DOI 10.1017/s0033291720000835. URL http://dx.doi.org/10.1017/S0033291720000835

12. De Marzo, G., Pietronero, L., Garcia, D.: Emergence of scale-free networks in social interactions among large language models. arXiv preprint arXiv:2312.06619 (2023)
13. Deffuant, G., Neau, D., Amblard, F., Weisbuch, G.: Mixing beliefs among interacting agents. Advances in Complex Systems **3**(01n04), 87–98 (2000)
14. Degroot, M.H.: Reaching a consensus. Journal of the American Statistical Association **69**(345), 118–121 (1974). DOI 10.1080/01621459.1974.10480137. URL http://dx.doi.org/10.1080/01621459.1974.10480137
15. Dryden, J., Clough, A.: Parallel Lives. e-artnow (2018). URL https://books.google.it/books?id=-amSDwAAQBAJ
16. Farina, M., Lavazza, A.: Chatgpt in society: emerging issues. Frontiers in Artificial Intelligence **6** (2023). URL https://api.semanticscholar.org/CorpusID:259168081
17. Fruchterman, T.M., Reingold, E.M.: Graph drawing by force-directed placement. Software: Practice and experience **21**(11), 1129–1164 (1991)
18. Gao, C., Lan, X., Lu, Z., Mao, J., Piao, J., Wang, H., Jin, D., Li, Y.: S3: Social-network simulation system with large language model-empowered agents. arXiv preprint arXiv:2307.14984 (2023)
19. Holley, R.A., Liggett, T.M.: Ergodic theorems for weakly interacting infinite systems and the voter model. The annals of probability pp. 643–663 (1975)
20. Hu, T., Collier, N.: Quantifying the persona effect in llm simulations. arXiv preprint arXiv:2402.10811 (2024)
21. Kosinski, M.: Theory of mind might have spontaneously emerged in large language models (2023). DOI 10.48550/ARXIV.2302.02083. URL https://arxiv.org/abs/2302.02083
22. Krapivsky, P.L., Redner, S.: Dynamics of majority rule in two-state interacting spin systems. Physical Review Letters **90**(23), 238,701 (2003)
23. Lan, X., Gao, C., Jin, D., Li, Y.: Stance detection with collaborative role-infused llm-based agents. In: Proceedings of the International AAAI Conference on Web and Social Media, vol. 18, pp. 891–903 (2024)
24. Lorig, F., Johansson, E., Davidsson, P.: Agent-based social simulation of the covid-19 pandemic: A systematic review. Journal of Artificial Societies and Social Simulation **24**(3) (2021)
25. McPherson, M., Smith-Lovin, L., Cook, J.M.: Birds of a feather: Homophily in social networks. Annual review of sociology **27**(1), 415–444 (2001)
26. Monti, C., Aiello, L.M., De Francisci Morales, G., Bonchi, F.: The language of opinion change on social media under the lens of communicative action. Scientific Reports **12**(1), 17,920 (2022)
27. Newman, M.E.: Mixing patterns in networks. Physical review E **67**(2), 026,126 (2003)
28. Noorazar, H.: Recent advances in opinion propagation dynamics: A 2020 survey. The European Physical Journal Plus **135**, 1–20 (2020)
29. Nyhan, B., Reifler, J.: When corrections fail: The persistence of political misperceptions. Political Behavior **32**(2), 303–330 (2010)
30. Park, J.S., O'Brien, J., Cai, C.J., Morris, M.R., Liang, P., Bernstein, M.S.: Generative agents: Interactive simulacra of human behavior. In: Proceedings of the 36th annual acm symposium on user interface software and technology, pp. 1–22 (2023)
31. Park, J.S., Popowski, L., Cai, C., Morris, M.R., Liang, P., Bernstein, M.S.: Social simulacra: Creating populated prototypes for social computing systems. In: Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology, UIST '22. Association for Computing Machinery, New York, NY, USA (2022). DOI 10.1145/3526113.3545616. URL https://doi.org/10.1145/3526113.3545616

32. Peel, L., Delvenne, J.C., Lambiotte, R.: Multiscale mixing patterns in networks. Proceedings of the National Academy of Sciences **115**(16), 4057–4062 (2018)
33. Priya, P., Firdaus, M., Ekbal, A.: Computational politeness in natural language processing: A survey. ACM Computing Surveys **56**(9), 1–42 (2024)
34. Rossetti, G., Stella, M., Cazabet, R., Abramski, K., Cau, E., Citraro, S., Failla, A., Improta, R., Morini, V., Pansanella, V.: Y social: an llm-powered social media digital twin. arXiv preprint arXiv:2408.00818 (2024)
35. Rozado, D.: The political preferences of llms. PLOS ONE **19** (2024). URL https://api.semanticscholar.org/CorpusID:267412830
36. Schweitzer, F.: Sociophysics. Physics today **71**(2), 40–46 (2018)
37. Si, X.M., Li, C.: Bounded confidence opinion dynamics in virtual networks and real networks. Journal of Computers **29**(3), 220–228 (2018)
38. Sîrbu, A., Loreto, V., Servedio, V.D., Tria, F.: Opinion dynamics: models, extensions and external effects. Participatory sensing, opinions and collective awareness pp. 363–401 (2017)
39. Taubenfeld, A., Dover, Y., Reichart, R., Goldstein, A.: Systematic biases in llm simulations of debates. arXiv preprint arXiv:2402.04049 (2024)
40. T¨ornberg, P., Valeeva, D., Uitermark, J., Bail, C.: Simulating social media using large language models to evaluate alternative news feed algorithms. arXiv preprint arXiv:2310.05984 (2023)
41. Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., Chi, E.H., Hashimoto, T., Vinyals, O., Liang, P., Dean, J., Fedus, W.: Emergent abilities of large language models (2022). URL https://arxiv.org/abs/2206.07682
42. Xie, C., Chen, C., Jia, F., Ye, Z., Shu, K., Bibi, A., Hu, Z., Torr, P., Ghanem, B., Li, G.: Can large language model agents simulate human trust behaviors? arXiv preprint arXiv:2402.04559 (2024)