

Network Effects and Agreement Drift in LLM Debates^{*,**}

Erica Cau^{a,b,*}, Andrea Failla^{a,b} and Giulio Rossetti^b

^aDepartment of Computer Science, University of Pisa, Largo Bruno Pontecorvo, Pisa, Italy

^bInstitute of Information Science and Technologies “A. Faedo” (ISTI), National Research Council (CNR), Via Giuseppe Moruzzi 1, Pisa, Italy

ARTICLE INFO

Keywords:

LLM
Opinion Dynamics
Homophily
Social Simulation

ABSTRACT

Large Language Models (LLMs) have demonstrated an unprecedented ability to simulate human-like social behaviors, making them useful tools for simulating complex social systems. However, it remains unclear to what extent these simulations can be trusted to accurately capture key social mechanisms, particularly in highly unbalanced contexts involving minority groups. This paper uses a network generation model with controlled homophily and class sizes to examine how LLM agents behave collectively in multi-round debates. We show that LLM agents exhibit convergence and polarization patterns that are highly sensitive to network structure and relative group sizes. Moreover, our findings highlight a particular directional susceptibility that we term *agreement drift*, in which agents are more likely to shift toward specific positions on the opinion scale. Overall, our findings highlight the need to disentangle structural effects from model biases before treating LLM populations as behavioral proxies for human groups.

1. Introduction

The introduction of Large Language Models (LLMs) [47, 30, 1, 20] has revolutionized the field of Natural Language Processing (NLP), advancing the ability of language models to solve text-related tasks and generate texts that closely resemble human ones [14]. These models have demonstrated remarkable proficiency across a broad spectrum of NLP tasks [72, 43, 18], marking a transition in the development of general-purpose AI models and progress toward artificial general intelligence (AGI) [40]. In addition to their technical capabilities, LLMs have demonstrated an unprecedented ability to simulate human social behaviors. When prompted correctly, LLMs demonstrated the ability to mimic *user personae* with specific demographic and psychological traits [68, 17, 71], occupation [49, 51], political leanings [64, 51, 52], and, in general, human behavior [5], making them valuable tools for studying their potential behavior when interacting in complex simulated societies [49, 54, 2]. As these models become increasingly widespread across various applications, comprehending their collective behavior becomes essential.

A key question is whether LLMs can spontaneously exhibit human-like social behaviors as emergent properties [44] arising from their training processes, rather than through explicit behavioral programming. For instance, LLMs have demonstrated the ability to exhibit simulated Theory of Mind [53], which might enable subjective behavioral modeling and the comprehension of ambiguous natural language instructions [69]. There has been considerable debate, with studies offering contradictory evidence [35, 57,

63, 62, 67, 33]. However, even a simulated Theory of Mind could help LLMs interact more naturally with other agents.

Such collective behaviors, while useful for better understanding LLMs and human behavior, may introduce previously unknown biases into decision-making processes or exacerbate societal biases embedded in training data, with potentially unknown effects on human-LLM interactions [2]. While much of the literature has focused on defining unstructured agent interactions, such as discussions on controversial topics like climate change [64, 6] or genetically modified food [71], less attention has been given to investigating more structured populations of agents. When we interact with others, we form social ties that can be modelled as edges in a graph, with individuals acting as nodes. Human societies are grounded in unwritten social rules, such as homophily, exemplified by the saying *birds of a feather flock together* [37]. People tend to gravitate toward those perceived as similar, thereby strengthening their connections with them, while avoiding those perceived as too different. This behavior has various consequences: it might amplify social divisions, strengthen boundaries between groups, and eventually lead to the segregation of populations into homogeneous clusters.

Why are these matters important? There are at least two reasons why these questions need to be addressed. By observing how LLM agents respond to topological factors such as homophily and class imbalance, it is possible to characterize the latent predispositions that emerge from their collective behavior. Understanding these dynamics is essential if we are to interpret their outputs reliably, especially in high-stakes or decision-critical contexts. Second, since LLMs are increasingly employed in agent-based simulations of human social systems [54, 73, 24], it is crucial to evaluate the extent to which their interactions accurately reflect real human behavior. Many recent works position LLMs as *proxies for human behavior* [23], especially in fields where language is pivotal, such as psychology. However, such usage assumes a

*Corresponding author: A. Failla

✉ erica.cau@phd.unipi.it (E. Cau); andrea.failla@phd.unipi.it (A. Failla); giulio.rossetti@isti.cnr.it (G. Rossetti)

🌐 andreaifailla.github.io (A. Failla); giuliorossetti.net (G. Rossetti)

ORCID(s): 0009-0005-7267-9870 (E. Cau); 0009-0009-6162-0274 (A. Failla); 0000-0003-3373-1240 (G. Rossetti)

degree of behavioral realism that may not hold under closer scrutiny [55]. If LLM populations diverge systematically from known patterns in human opinion formation — e.g., by converging too quickly, suppressing disagreement, or exhibiting exaggerated sensitivity to class size — this raises questions about their validity as tools for modeling human collective behavior. Thus, understanding when and how LLMs approximate or deviate from human-like dynamics is critical for both methodological validity and ethical use.

Our contribution. This study aims to bridge the gaps outlined above by examining how the degree of homophily and relative class sizes within the interaction network affect opinion evolution in LLM societies. Previous work has investigated the homophilic mechanism in networks generated by LLM agents, focusing particularly on how the attributes of LLM agents (e.g., gender, age, or political leaning) influence link formation, leading to assortative structures and clustered communities [38]. In contrast, we model populations of LLMs as agents within a preferential attachment framework augmented with homophilic edge formation [31]. This allows us to manipulate not only traditional network parameters (e.g., degree distribution) but also the social alignment of nodes (via homophily) and the distribution of opinions (via class size imbalance). We simulate debates among agents on a given initial statement and trace how opinions evolve over time. Our investigation is structured around the following research questions:

RQ1. *How does homophily influence opinion dynamics in balanced LLM agent populations?* In this experiment, the agent population is evenly divided (50-50) between two opinion classes. We vary the homophily parameter while keeping all other factors constant, particularly class size, to isolate and observe the impact of homophilic attachment on the trajectory and convergence of opinion evolution.

RQ2. *How does the effect of homophily on LLM opinion dynamics change under increasingly imbalanced class distributions?* Building on RQ1, we investigate how skewing the initial class size distribution (i.e., from 70-30 to 90-10) affects the role of homophily in opinion dynamics. This allows us to examine whether the presence of a minority affects the extent to which homophily influences consensus formation and opinion stability.

RQ3. *How do the initial stances of minority and majority groups influence opinion dynamics?* We compare scenarios in which (i) the majority agrees with the initial statement while the minority disagrees, and (ii) the minority agrees with the initial statement while the majority disagrees. Given the LLMs' tendency toward agreement [58, 9], this question probes whether aligning with the majority on a particular stance amplifies or mitigates its influence on final opinion distributions.

RQ4. *How does knowledge of neighbors' opinion distribution affect opinion dynamics?* We explore scenarios in which agents have information about their neighbors' opinion distributions and assess the impact of this knowledge on opinion evolution.

Coherently with previous research, we find that LLM populations exhibit a systematic directional bias in persuasive interactions: when agents holding opposing views interact, movement toward endorsing the discussion statement is more likely than movement toward rejecting it, even in balanced populations. At the collective level, its effect depends on exposure: heterogeneous networks allow this bias to propagate and produce rapid convergence, whereas homophily and large disagreeing majorities limit cross-opinion encounters and lead to persistent polarization. Providing agents with information about their local neighborhood further reshapes the dynamics by favouring moderate agreement and making opinion change contingent on local alignment.

The remainder of this work is organized as follows. In Section 2, we provide an overview of the literature regarding opinion dynamics and LLM agents. In Section 3, we introduce the framework for LLM opinion dynamics, and in Section 4, we introduce the experimental setup and results. Finally, in Section 5, we conclude the work by discussing the results and suggesting future research directions.

2. Related Work

In this section, we discuss relevant literature on Opinion Dynamics and Large Language Model-augmented social simulations.

Opinion Dynamic Models Opinion dynamics (OD) describes computational processes that mimic how individual views evolve over time under empirically observed social influences [7, 59]. In a typical OD simulation, a network of interacting agents—the *population*—holds opinions encoded as real numbers. Empirically grounded social mechanisms are cast as update rules, drawing on techniques from statistical physics, such as graph theory, probability, and statistical modeling. For example, an agent may revise its stance by randomly copying a neighbor's opinion [29] (modeling imitation), by averaging the views in its local neighborhood [21] (capturing homophily [37]), or by adopting the most frequent opinion around it [34] (reflecting peer pressure and mutual awareness [16]). OD research asks whether the interplay of these mechanisms drives the system toward one of three mutually exclusive collective states: (i) *consensus*, when all agents share the same view; (ii) *polarization*, when opinions cluster around two opposing poles; or (iii) *fragmentation*, when more than two distinct clusters form [45, 60].

OD models are also classified by how they encode the opinion variable. In *discrete* models, opinions can take only a finite set of values, with rules specifying switches among them. The classic Voter Model [15], for instance, considers populations in which agents adopt one of two opposing

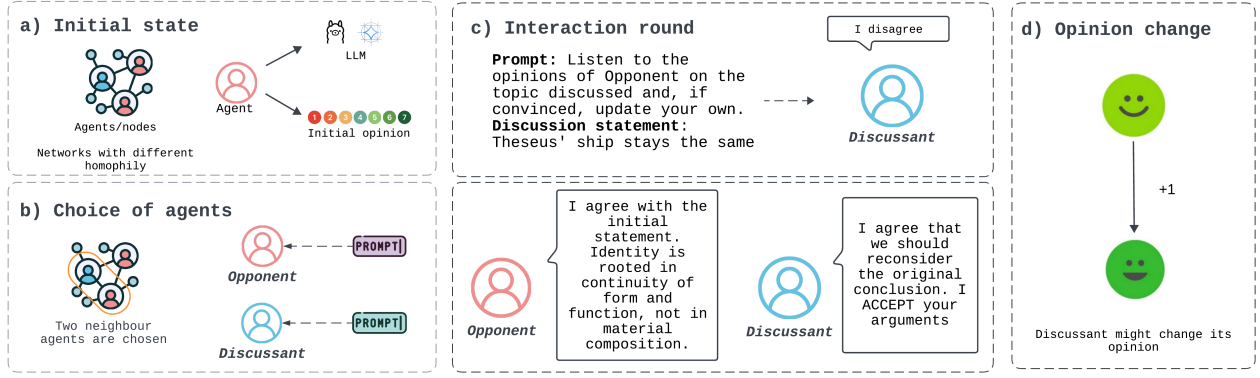


Figure 1: Graphical schema of the LLM-OD framework. The LLM agents population is initialized as nodes in a network; each agent is an LLM instance with an initial opinion in the range $[0, 6]$ (a). At each iteration, two neighbor agents are chosen and prompted to act as *Opponent* and *Discussant* (b). The *Discussant* is prompted to listen to the opinion of the *Opponent* around the discussion statement and may then accept, reject, or ignore such opinion (c) and update their current one accordingly by ± 1 (d).

stances (e.g., agree/disagree). By contrast, *continuous* models treat opinions as points on an uninterrupted spectrum. Canonical examples, such as the DeGroot Model [22], allow any value within a bounded interval – commonly $[-1, 1]$ – and interactions move opinions smoothly along that continuum.

Large Language Models Recent studies have examined the dynamics that arise from interactions among large language models (LLMs). These LLM agents typically exhibit realistic behavior both individually and collectively [49, 51], with human-like patterns naturally emerging from their interactions. Examples include the formation of scale-free networks [19], dissemination of information [28], and the development of trust [70]. When supplied with demographic backgrounds and personality traits, LLMs can more accurately mimic human behavior [50, 49, 4], including reflecting human biases and political inclinations [11]. Integrating LLMs with traditional agent-based modeling approaches [36], as demonstrated in [66], has shown how recommender systems can influence the quality of discourse. Specifically, agents exposed to popular content tend to exhibit increased toxicity and interparty engagement, mirroring the toxicity levels seen in U.S. Twitter discourse circa 2019. Further simulations reveal that confirmation bias among agents contributes to social fragmentation [12], aligning with established findings in opinion dynamics. LLMs are also capable of emulating persuasive communication [6, 41], generating coherent arguments based on psycho-linguistic theories of opinion change, and can be utilized for comprehensive social media simulations [54]. Nonetheless, LLMs have a natural inclination toward factual correctness and often resist generating content that contradicts scientific knowledge unless explicitly prompted [12]. This tendency may cause them to disregard assigned roles or personality traits [64], resulting in less authentic behavior. Additionally, LLM agents generally

avoid intense conflict [13, 66] and display sensitivity toward various social and political issues [56, 27].

3. LLM-powered Opinion Dynamics

We present a general framework to describe how populations of large language model (LLM) agents develop and revise opinions through structured interactions [8]. The system is represented by the tuple:

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

where each element captures a specific component of the simulation:

- $\mathcal{G} = \langle V, E \rangle$ denotes the interaction network, where V is the set of LLM agents and $E \subseteq V \times V$ is the set of undirected pairs of agents that can interact. The topology may be fully connected (i.e., mean-field) or follow an empirical or synthetic configuration.
- $\mathcal{S}$ is the discussion prompt or proposition that defines the initial subject of deliberation. It establishes the semantic dimension along which opinions are assessed.
- $\mathcal{O}$ represents the space of admissible opinions. This space can be discrete (e.g., $\mathcal{O} = \{0, 1\}$) or continuous (e.g., $\mathcal{O} = [0, 1]$), encoding positions from complete disagreement to full agreement with the statement $\mathcal{S}$.
- $f : \mathcal{O}^n \times \mathbb{R}^m \rightarrow \mathcal{O}$ is the opinion update function. It specifies how an agent revises its opinion as a function of its current position, the opinions of n neighboring agents, and additional features or external variables in $\mathbb{R}^m$.
- T is the temporal domain of the simulation, defined as an ordered sequence of discrete time steps $t \in \mathbb{Z}^+$.

- $D(t) = \{o_i(t) \mid i \in V\}$ denotes the opinion state at time t , where $o_i(t) \in \mathcal{O}$ is the opinion of agent i at that time.

This framework supports a wide variety of network topologies, interaction mechanisms, and opinion evolution rules. In what follows, we describe a particular instantiation of the model adopted in our experimental setting.

A general overview of its rationale is provided in Algorithm 1. In detail, we define the graph topology, the

Algorithm 1 LLM Opinion Dynamics

Require: Graph: G , Statement: S , Initial opinions: $D(0) = \{o_i(0) \mid i \in V\}$, Time steps: $T_{\max}$,
Ensure: Opinion Distribution Evolution D_{all}

```

1: Init  $D_{all}$  as an empty list
2:  $D_{all} \leftarrow D(0)$ 
3: for  $t = 0$  to  $T_{\max} - 1$  do
4:   for each agent  $i \in V$  do
5:     Select random neighbor  $v_j \in \text{Neighbors}(i)$ 
6:      $\Delta o_i(t) = \text{debate}(o_i(t), o_j(t), S)$ 
7:     Update opinion:  $o_i(t+1) = o_i(t) + \Delta o_i(t)$ 
8:   end for
9:   Update opinion distribution:  $D(t+1) = \{o_i(t+1) \mid i \in V\}$ 
10:   $D_{all} \leftarrow D(t+1)$ 
11: end for
12: return  $D_{all}$ 
    
```

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 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 with producing a persuasive argument to convince 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 interaction 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 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 ACCEPTS, its opinion is brought closer to the Opponent's, and if it REJECTS, it is

Algorithm 2 Debate

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

```

1: Init Discussant's Opinion Delta  $o_i^{\Delta+} \leftarrow 0$ 
2: Discussant states its opinion  $o_i$  on  $S$ 
3: for round = 1 to  $\text{max\_rounds} = 3$  do
4:   Opponent's Turn:
5:   Opponent evaluates Discussant's argument
6:   Opponent decides and justifies: ACCEPT, REJECT, IGNORE
7:   if decision is IGNORE or ACCEPT then
8:     break
9:   end if
10:  Discussant's Turn:
11:  Discussant evaluates Opponent's argument
12:  Discussant decides and justifies: ACCEPT, REJECT, IGNORE
13:  if decision is ACCEPT then
14:     $o_i^{\Delta+} \leftarrow 1$  if  $o_j > o_i$  else -1
15:    break
16:  else if decision is REJECT then
17:     $o_i^{\Delta+} \leftarrow -1$  if  $o_j > o_i$  else 1
18:    break
19:  end if
20: end for
21: return  $o_i^{\Delta+}$ 
    
```

brought further (e.g., backfire effect [46]). If the Discussant does neither, or if the debate terminates earlier, their opinion remains unchanged.

4. Results

We conduct a series of simulations within the LLM-based opinion dynamics framework introduced in Section 3. We generate scale-free networks of 100 agents in which nodes belong to two mutually exclusive classes, with tunable class sizes and assortativity [31]. For further information on the network generation process, please refer to Appendix A. The discussion prompt is based on the paradox of Theseus' Ship, first reported in Plutarch's *Parallel Lives* [25], which questions whether an object that has had all its components replaced retains its identity. The statement is deliberately non-empirical to avoid convergence driven by factual correctness and to preserve argumentative diversity [12, 9]. Unless otherwise stated, results are obtained with Llama 3.1 [1], and the neighborhood-pressure condition is additionally replicated with Gemma3 [65]. For each parameter configuration we run a single simulation on a fixed network realization in order to isolate the structural effect of the experimental parameters and to avoid compounding sources of stochasticity.

Experimental design. We begin with a symmetric configuration in which the two opinion groups are equally represented, providing a baseline to evaluate how homophily ($h \in [0, 0.25, 0.5, 0.75, 1]$) affects opinion diffusion in the absence of numerical advantage (**RQ1**). We then introduce asymmetric populations in which one group constitutes 30% and 10% of the agents to assess the impact of minority size on global dynamics (**RQ2**). Each imbalance condition is repeated in a reverse configuration, where the initial majority and minority positions are swapped, to test whether the outcomes depend on the opinion side rather than on group size alone (**RQ3**). Finally, we provide agents with information

about the opinion distribution in their ego network during the interaction to evaluate the effect of local social pressure on opinion change (RQ4).

We first examine the temporal evolution and final distribution of opinions across all scenarios, thereby identifying the macro-level effects of homophily and group composition. We then focus on the neighborhood-pressure condition and estimate opinion transition probability matrices to uncover the micro-level mechanisms driving the observed patterns. These probabilities quantify the likelihood that a Discussant moves toward the Opponent's position under different initial configurations and neighborhood states. To assess whether the observed transition probabilities differ from what would be expected by chance, we employ a permutation-based null model. Specifically, we randomly reshuffle the post-interaction opinions among interaction events while keeping fixed (i) the pre-interaction opinions of both agents and (ii) the neighborhood composition. This procedure preserves the marginal distributions and the interaction structure but removes any systematic association between interaction context and opinion change. For each entry of the transition matrix, we generate a null distribution from 1000 permutations and retain only those probabilities that are significantly different from the null expectation ($p < 0.01$). Details on the estimation and validation procedure are reported in Appendix B.

4.1. Effect of homophily and group size imbalance

Opinion trajectories for all configurations are shown in Figures 2 and 3. We first discuss the symmetric population (RQ1), then introduce class imbalance (RQ2), and finally analyze the reverse configurations (RQ3).

Balanced Scenario (RQ1). With equal group sizes, all networks with $0 \leq h \leq 0.75$ rapidly converge toward the positive side of the opinion scale, typically within 20–40 iterations (Figure 2a). Lower homophily accelerates convergence, with heterophilic networks reaching agreement almost immediately. As h increases, convergence becomes slower, but the final outcome remains unchanged. A qualitatively different regime emerges at complete homophily ($h = 1$), where interactions occur almost exclusively within like-minded groups, and the system stabilizes in a segregated, polarized state.

Imbalanced Scenario (RQ2). Introducing a minority strengthens the convergence toward agreement. With a 70/30 split, the system reaches consensus in about 20–30 iterations across all homophily levels, with only minor fluctuations between *agree* and *strongly agree* (Figure 2b). Similarly to the balanced scenario, high homophily ($h = 0.75$) slows convergence speed; nonetheless, the agents reach perfect adherence to the *strongly agree* opinion. Reducing the minority to 10% further accelerates its disappearance, especially for $0 \leq h \leq 0.25$, while higher homophily

temporarily preserves small opposing clusters before convergence (Figure 2c). At $h = 1$, segregation persists because cross-group interactions are rare.

Reversed Imbalanced Scenario (RQ3). Interestingly, when moving to the reverse class distribution scenario, where the minority now holds the *strongly agree* opinion, trends show a clear pattern that diverges from simple symmetry (Figure 3). With $min = 0.3$, disagreeing agents form a dominant cluster, with small neutral and agreeing clusters surviving for several iterations. Although the latter gradually decreases in size after 40 iterations, the dominant opinion does not reach full consensus. This pattern persists for $0.25 \leq h \leq 0.75$, indicating that negative majorities are more resistant to complete convergence. Strengthening the imbalance at $min = 0.1$ heightens the dominance of *strongly disagreeing* opinion clusters across all homophily levels, with only small residual *agreeing* clusters. Only in the case of $h = 1$, agents behave almost symmetrically to what was previously observed in the base imbalanced scenario, which we attribute to strong structural constraints.

4.2. Effect of neighborhood awareness

Providing agents with the distribution of opinions in their ego network results in systematically faster opinion change and shifts the dynamics toward moderate agreement (Figure 4a). In the balanced population, convergence is almost immediate but typically stops at *agree*, also preventing the full convergence observed without neighborhood information. Under complete homophily ($h = 1$), agents initially maintain separation between the two extreme clusters; then, an additional cluster with the *agree* opinion emerges. With moderate class imbalance, neighborhood awareness eliminates the rapid dominance of extreme positive opinions observed in the base scenario and allows small neutral and mildly positive clusters to persist across homophily levels (Figure 4b). Under extreme class imbalance, the majority of opinions stabilize on *agree*, followed by a few *strongly agree* and minimal other clusters (Figure 4c). Interestingly, the polarized outcome between extreme stances observed so far at $h = 1$ does not hold in this scenario; here, we note an *agree* majority followed by small *mildly agree* and *neutral* clusters. In other words, unlike other high-homophily settings, the minority slowly moves toward the majority's opinion under extreme class imbalance and known neighborhood configuration. In the reverse configurations, where the minority initially holds the positive stance (Figures 5a and 5b), the dynamics become asymmetric. For $min = 0.3$, negatively oriented majorities progressively shift toward moderate and positive positions, indicating that access to neighborhood opinion distributions mitigates the dominance of negative opinions; a stable negative cluster persists only under complete homophily ($h = 1$), where interactions remain within groups. In the extreme imbalance ($min = 0.1$), however, the negative majority instead consolidates and converges, a process reinforced by homophily, while the small positive

Network Effects and Agreement Drift

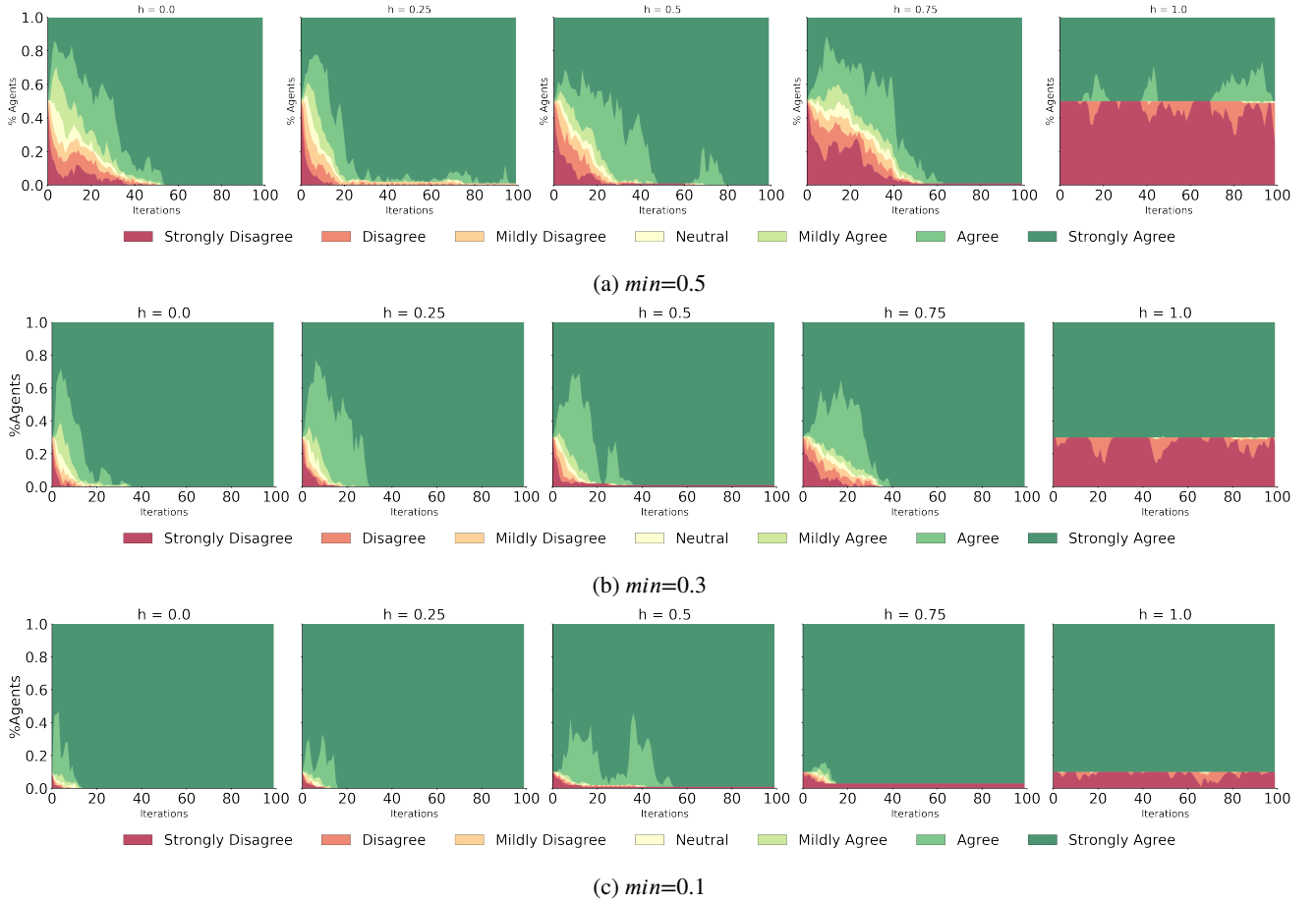


Figure 2: Opinion trends for the base scenario (strongly-agreeing majority) across different homophily levels for minority fraction $\min=0.5$ (a), $\min = 0.3$ (b) and $\min = 0.1$ (c) Each panel shows the proportion of agents holding different opinion states over 100 iterations. All points of the opinion scale are mapped to the colors in the legend, from *Strongly Disagree* (dark red) to *Strongly Agree* (dark green).

minority survives only in fully homophilic networks without changing its stance.

4.2.1. Model Comparison

The Gemma-based simulations reveal a systematically stronger drift toward positive opinions than observed with Llama. In balanced populations, negatively oriented agents disappear rapidly for all $0 \leq h \leq 0.75$, and the system converges either to *strongly agree* or to a polarized configuration dominated by a large positive cluster (Figure 6). Even under complete homophily, where Llama agents remain segregated at the two extremes, Gemma agents predominantly reinforce the positive side, indicating a higher intrinsic propensity to accept upward shifts. This tendency is further amplified in imbalanced settings: when the positive class is the majority, intermediate opinions vanish quickly, and convergence toward the positive extreme is faster and more stable than in the corresponding Llama runs.

The asymmetry persists in the reverse configurations, where negative agents form the numerical majority. Although negative clusters initially survive for several iterations, upward

transitions remain frequent and the dynamics eventually lead to either strong polarization or to the re-emergence of a dominant positive cluster. This behavior suggests that Gemma agents are less sensitive to structural constraints and more responsive to persuasive interactions that point toward agreement. At the micro level, this is reflected in transition probabilities close to 1 for low-to-high shifts and in the high stability of high-opinion states, confirming a stronger and more persistent bias toward positive opinions compared with Llama.

4.2.2. Neighborhood Awareness and Interaction Dynamics

To connect the aggregate opinion trajectories with the underlying interaction dynamics, we analyze conditional transition probability matrices for balanced populations ($\min = 0.5$). Across all homophily levels and model variants, upward transitions (low $\rightarrow$ high) are consistently more likely than downward ones (high $\rightarrow$ low), indicating that persuasion is structurally asymmetric and biased toward positive opinions (Figure 7). This asymmetry, which we will refer to as *agreement drift*, persists even when group

Network Effects and Agreement Drift

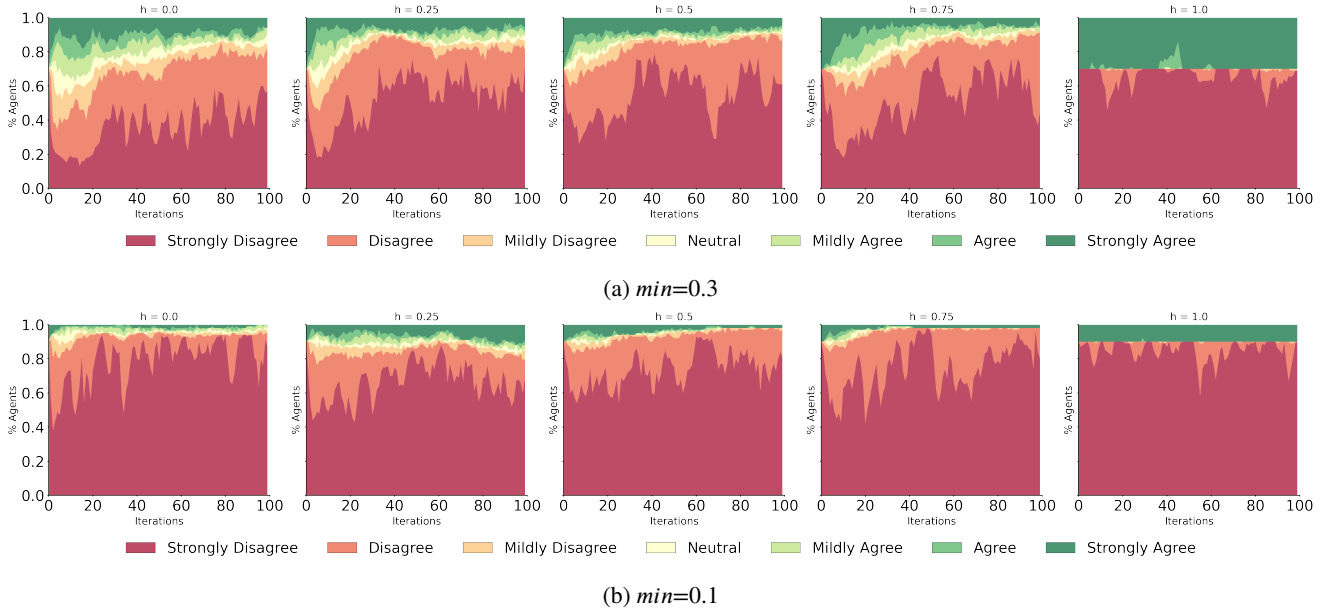


Figure 3: Opinion trends for the reversed scenario (strongly-disagreeing majority) across different homophily levels for minority fraction $min=0.5$ (a), $min = 0.3$ (b), and $min = 0.1$ (c) Each panel shows the proportion of agents holding different opinion states over 100 iterations with different levels of homophily. All points of the opinion scale are mapped to the colors in the legend, from *Strongly Disagree* (dark red) to *Strongly Agree* (dark green).

sizes are equal, showing that it is not a consequence of numerical dominance, but likely an intrinsic feature of the LLM interaction dynamics. Increasing homophily reduces the frequency of cross-opinion encounters and leads to structural segregation, yet the imbalance between upward and downward shifts remains visible whenever interactions between opposite opinions occur.

Providing agents with knowledge of their neighbors' opinion distributions (Figure 7, mid row) produces distinctive behavior: agents seem to change their opinion when facing a higher-opinion opponent, but become more resistant to changing their opinion when facing a lower-opinion opponent. The most interesting scenario is $h = 1$, as discussants holding a low opinion prefer high-opinion opponents (83%), while the reverse transition occurs (100%), indicating that persuasion remains active despite structural segregation. Maintaining a high opinion, however, seems difficult for high-opinion agents, as their likelihood of retaining it is relatively low (30%), especially compared with the base scenario. The effect is even stronger for Gemma agents (Figure 7, bottom row), for which upward transitions approach probability 1 and high-opinion states become highly stable.

To further assess the impact of agents' knowledge of their neighbors' opinion distribution, we quantify the likelihood of the deviation of the discussants' final opinions in relation to their neighbors in the opinion-distribution-aware setting. These matrices were created for both Llama (Figure 8) and Gemma simulations (Figure 9).

For Llama agents, opposing neighborhoods strongly stabilize high-opinion discussants, whose probability of maintaining their position remains close to one across all h . At the same time, upward shifts from low to high

emerge even when the local environment discourages them, becoming deterministic for $h = 0.75$. Mixed neighborhoods display the most symmetric dynamics, with statistically significant transitions in both directions for $h \leq 0.75$. For $h \leq 0.75$, transitions from high to low become more significant, allowing a downward movement from high to low. In supporting neighborhoods, opinion change largely follows the local consensus: shifts aligned with the neighborhood are reinforced, while the persistence of low-opinion states in low-to-low interactions remains rare. Overall, peer pressure acts as a selective amplifier [42], strengthening changes that are locally consistent, while still allowing occasional counter-aligned upward moves.

Gemma exhibits a qualitatively different regime. Most cross-class shift probabilities are not significant, especially when the direction of the shift is supported or opposed by neighboring agents. The most noticeable patterns appear in the first (opposing) and third (supporting) rows of the matrices in Figure 9. In opposing neighborhoods, statistically significant shifts in probability occur from low to high opinions, indicating that Gemma agents can adjust upward even when surrounded by opposing agents. Conversely, with supporting agents and in all homophily configurations, discussants move from high to low. Agents in this case are susceptible to neighborhood alignment and change their opinion when their peers reinforce this transition. Additionally, with high homophily values ($0.75 \leq h \leq 1$), only shifts probabilities consistent with each agent's ego network are allowed. According to this dynamic, opinion change is allowed only when contextual reinforcement is sufficiently strong. Therefore, Gemma exhibits a different



Figure 4: Opinion trends for the base scenario (strongly-agreeing majority) under neighborhood opinion awareness across different homophily levels for minority fraction $\min=0.5$ (a), $\min = 0.3$ (b), and $\min = 0.1$ (c) Each panel shows the proportion of agents holding different opinion states over 100 iterations. All points of the opinion scale are mapped to the colors in the legend, from *Strongly Disagree* (dark red) to *Strongly Agree* (dark green).

behavior to Llama, with rare opinion shifts in which agents are subjected to peer pressure from the neighborhood during class changes. The bias toward acceptance is more evident in Gemma, where low-to-high transitions are allowed even in the presence of an opposing neighborhood. In the mixed scenario, Gemma’s agents exhibit more varied behavior, as they are not exposed to a single dominant opinion. The matrices show interaction probabilities that are statistically significant across all homophily levels, with a preference for agents to accept high-to-high opinion shifts.

5. Discussion

This work investigates how network structure, class imbalance, and local social information shape opinion dynamics in populations of LLM agents, and to what extent the resulting collective behavior resembles mechanisms known from human social systems. This study identifies two main results that jointly explain the collective behavior of LLM-based opinion dynamics.

First, we observe a systematic bias we term **agreement drift**:

in pairwise interactions, agents are more likely to shift from lower to higher positions on the opinion scale than in the opposite direction. Across homophily levels, class distributions, and model variants, upward transitions occur more frequently than downward ones whenever agents holding opposing views interact. A similar effect was reported in previous work [9, 39]; here we show that it generalizes across different models and heterogeneous network configurations. We find that this asymmetry is present even in balanced populations, indicating that it is not dependent on numerical dominance and cannot be reduced to a majority effect. At the same time, it does not automatically produce convergence toward agreement at the collective level: when cross-opinion encounters are limited (either because the disagreeing group forms a large majority or because homophily segregates the network) the directional tendency remains locally observable but cannot spread through the system. Collective outcomes, therefore, emerge from the interaction between an intrinsic persuasive bias and the opportunity structure for exposure created by the network.

Importantly, the agreement drift effect is distinct from the

Network Effects and Agreement Drift

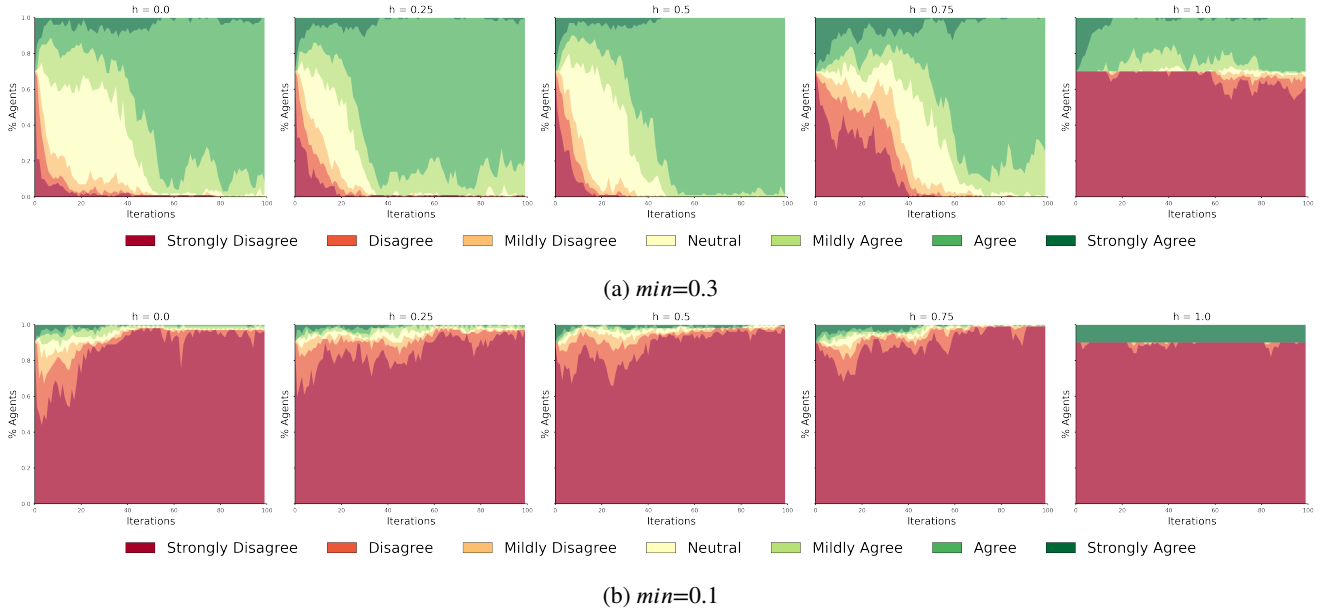


Figure 5: Opinion trends for the reversed scenario (strongly-disagreeing majority) under neighborhood awareness across different homophily levels for minority fraction $\min=0.5$ (a), $\min = 0.3$ (b), and $\min = 0.1$ (c) Each panel shows the proportion of agents holding different opinion states over 100 iterations. All points of the opinion scale are mapped to the colors in the legend, from *Strongly Disagree* (dark red) to *Strongly Agree* (dark green).

sycophantic behavior documented in prior work [10, 64, 26, 32]. Sycophancy is the tendency of an LLM to align with a user’s explicitly stated belief (regardless of truth or epistemic consistency) in order to maximise perceived helpfulness/social approval. Agreement drift, by contrast, is a directional susceptibility in which positions expressing endorsement of the discussion statement exert systematically greater persuasive pull than positions expressing rejection. In other words, when agents holding opposing views interact, movement toward the pro-position is more likely than toward the contra-position, regardless of who initially holds the majority, and without any external reward for compliance.

The second main result concerns the role of **structural and contextual constraints** in modulating this intrinsic tendency. Homophily regulates the reach of the directional bias by controlling the frequency of cross-opinion encounters. At low and moderate levels of homophily, heterogeneous exposure enables the asymmetric persuasive mechanism to propagate through the population, leading to rapid convergence toward agreement even in the absence of a numerical advantage. At high homophily, by contrast, the same bias becomes trapped within segregated clusters, leading the system to stabilize in polarized configurations. Class imbalance acts as an additional gatekeeper: large disagreeing majorities can prevent the diffusion of agreement drift by reducing the probability of encounters with pro-position agents, while small minorities disappear quickly when exposure is high. Providing agents with information about their local neighborhood introduces a further layer of mediation. Rather than simply accelerating convergence, neighborhood awareness

reshapes the outcome by favouring moderate agreement and by making opinion change contingent on local alignment, thereby revealing a mechanism analogous to peer-pressure effects in classical opinion-dynamics models.

Implications. Our findings have direct methodological implications for the use of LLM agents as proxies for human behavior in computational social science. In many empirical and theoretical models of human opinion dynamics, persuasive influence is assumed to be directionally neutral and collective outcomes are primarily attributed to network topology, initial conditions, and exposure patterns [60, 61, 48, 42]. In the LLM populations studied here, by contrast, the micro-level persuasive interaction is intrinsically asymmetric due to agreement drift. This shows that *convergence in LLM-based social simulations cannot be interpreted as evidence of consensus dynamics without first characterising the model’s intrinsic persuasion tendency*. This means that simulations employing LLM agents may systematically overestimate the emergence of consensus whenever cross-opinion encounters are sufficiently frequent, and conversely, attribute persistent disagreement to segregation even when a directional persuasive tendency is present at the interaction level. For the growing literature that treats LLM populations as behavioral surrogates for human groups, this result highlights the need to disentangle structural effects from model-intrinsic interaction biases before drawing substantive conclusions about social processes.

Limitations and Future Directions. Several limitations should be acknowledged. First, the results are obtained for a specific interaction protocol and a single non-factual

discussion topic; although the agreement drift is likely not a consequence of our interaction procedure because it emerges even under different configurations [9, 39], different prompting strategies or semantically grounded statements may alter its strength. Second, we consider networks of fixed size and a specific generative model; larger populations and/or alternative topologies may introduce additional mesoscopic effects. Third, the strong differences observed between LLM families call for a systematic analysis across models. Addressing these aspects is essential to determine which behavioral patterns are model-specific and which represent general properties of LLM societies. Future work should therefore investigate how heterogeneous agent profiles, memory, and repeated interactions affect the persistence of minority opinions, and calibrate LLM-based dynamics against empirical human data to assess their behavioral realism. Understanding how intrinsic agreement biases interact with algorithmic mediation and recommender systems is also crucial for evaluating the risk of artificial consensus formation in human-AI hybrid environments.

A. Networks with Controlled Homophily and Group Size

To address our research questions, we need to generate networks with controlled group sizes (i.e., node attribute frequencies) and mixing biases (i.e., nodes' preferences for connecting with similar/dissimilar peers). To do so, we leverage the *BA-homophily* model introduced in [31], which extends the classical Barabási-Albert (BA) preferential attachment model [31] by incorporating a *homophily* parameter. The classic BA model [3] is a foundational generative model for scale-free networks. It captures two key mechanisms underlying many real-world complex systems: *network growth* and *preferential attachment*. The model starts with a small connected network of m_0 nodes. At each time step, a new node is added to the network and forms $m \leq m_0$ links to existing nodes. The probability Π_i that the new node connects to an existing node i is proportional to the degree k_i of that node:

$$\Pi_i = \frac{k_i}{\sum_j k_j}$$

This rule encodes *preferential attachment*, where well-connected nodes are more likely to attract additional links, a “rich-get-richer” dynamic. Over time, this process generates networks with a power-law degree distribution, $P(k) \propto k^{-\gamma}$, with $\gamma \approx 3$.

The BA-homophily enriches this framework in two ways. First, each node belongs to one of two groups, a (minority) or b (majority), with group sizes defined by relative proportions f_a and $f_b = 1 - f_a$, where $f_a \leq f_b$. Second, when a new node j arrives, it selects m existing nodes to connect to, choosing targets based on both their degree and group similarity. Formally, the probability Π_{ij} that node j connects to node i is given by:

$$\Pi_{ij} = \frac{h_{ij} \cdot k_i}{\sum_l h_{lj} \cdot k_l},$$

where k_i is the degree of node i and h_{ij} encodes homophilic preference. In the symmetric case where node groups show equal homophilic preferences, we have that

$$h_{aa} = h_{bb} = h, \quad h_{ab} = h_{ba} = 1 - h$$

Via this mechanism, the model can generate networks with complete heterophily (nodes only connect to the opposite group, $h = 0$), neutral mixing (no group preference, classical BA, $h = 0.5$), up to complete homophily (nodes only connect within their group, $h = 1$). As shown in [31], in heterophilic networks ($h < 0.5$), minority nodes become structural hubs, attracting a disproportionate number of links from the majority. Conversely, in homophilic settings ($h > 0.5$), they become increasingly segregated.

B. Opinion Transition Probabilities and Validation

To uncover the mechanisms underlying the observed opinion dynamics, we analyze micro-level interaction patterns between *Discussant* and *Opponent* agents. Specifically, we construct transition probability matrices that quantify the likelihood that a Discussant shifts its opinion toward that of the Opponent following an interaction, conditional on the Discussant's and Opponent's initial opinion class. Formally, we estimate:

$$P(\text{disc} \rightarrow \text{opp} \mid o_{\text{disc}}, o_{\text{opp}}),$$

where $\text{disc} \rightarrow \text{opp}$ denotes a change in the Discussant's opinion toward the Opponent's position, o_{disc} is the Discussant's initial opinion, and o_{opp} is the Opponent's opinion. By doing so, we aim to estimate the rate of successful (as in persuasive) interactions.

For the simulations in which agents are equipped with information on their neighbors' opinion distribution, we additionally condition this probability on the Discussant's local neighborhood configuration, distinguishing whether its neighbors are aligned, misaligned, or mixed, with respect to the direction of change. This extension isolates the effect of peer agreement on opinion shifts. Formally:

$$P(\text{disc} \rightarrow \text{opp} \mid o_{\text{disc}}, o_{\text{opp}}, c),$$

where $c \in \{\text{aligned}, \text{misaligned}, \text{mixed}\}$ denotes the configuration of the Discussant's neighborhood relative to the direction of the potential opinion shift, that is, whether the majority of the Discussant's neighbors hold opinions that are closer to the Opponent's position (aligned), closer to the Discussant's initial position (misaligned), or are evenly split between the two (mixed).

To identify statistically significant transitions, we adopt a permutation-based null model that tests whether the observed association between opinion pairs and the direction of change exceeds what would be expected by chance. For each interaction, we keep the pair (o_{disc}, o_{opp}) (resp. (o_{disc}, o_{opp}, c) for the extension) fixed and randomly shuffle the observed opinion variations across interactions. This procedure preserves the empirical distribution of opinion changes as well as the frequency of each opinion pair, while removing any systematic relationship between them. For every permutation, we recompute the transition probabilities, thereby obtaining a null distribution for each conditioning configuration. The statistical significance of the observed probability is then assessed using a two-tailed test, defined as the fraction of null values that are at least as extreme as the empirical estimate.

CRediT authorship contribution statement

Erica Cau: Conceptualization, Methodology, Investigation, Formal analysis, Visualization, Writing - Original Draft. **Andrea Failla:** Conceptualization, Methodology, Formal analysis, Validation, Visualization, Writing - Original Draft. **Giulio Rossetti:** Conceptualization, Methodology, Software, Writing - Review & Editing, Supervision, Funding acquisition.

References

- [1] AI@Meta, 2024. Llama 3 model card. N.A. URL: https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md.
- [2] Ashery, A.F., Aiello, L.M., Baronchelli, A., 2025. Emergent social conventions and collective bias in llm populations. *Science Advances* 11, eadu9368.
- [3] Barabási, A.L., Albert, R., 1999. Emergence of scaling in random networks. *science* 286, 509–512.
- [4] Bhandari, P., Fay, N., Wise, M.J., Datta, A., Meek, S., Naseem, U., Nasim, M., 2025. Can llm agents maintain a persona in discourse?, in: *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 29201–29217.
- [5] Binz, M., Akata, E., Bethge, M., Brändle, F., Callaway, F., Coda-Forno, J., Dayan, P., Demircan, C., Eckstein, M.K., Éltető, N., et al., 2025. A foundation model to predict and capture human cognition. *Nature*, 1–8.
- [6] Breum, S.M., Egdal, D.V.d., Mortensen, V.G., Møller, A.G., Aiello, L.M., 2024. The persuasive power of large language models, in: *Proceedings of the International AAAI Conference on Web and Social Media*, pp. 152–163.
- [7] Castellano, C., Fortunato, S., Loreto, V., 2009. Statistical physics of social dynamics. *Reviews of modern physics* 81, 591.
- [8] Cau, E., Failla, A., Rossetti, G., 2024. Bots of a feather: mixing biases in llms' opinion dynamics, in: *International conference on complex networks and their applications*, Springer. pp. 166–176.
- [9] Cau, E., Pansanella, V., Pedreschi, D., Rossetti, G., 2025. Selective agreement, not sycophancy: investigating opinion dynamics in llm interactions. *EPJ Data Science* 14, 59.
- [10] Cheng, M., Yu, S., Lee, C., Khadpe, P., Ibrahim, L., Jurafsky, D., 2025. Elephant: Measuring and understanding social sycophancy in llms. *arXiv preprint arXiv:2505.13995*.
- [11] Chuang, Y.S., Goyal, A., Harlalka, N., Suresh, S., Hawkins, R., Yang, S., Shah, D., Hu, J., Rogers, T.T., 2023. Simulating opinion dynamics with networks of llm-based agents. *arXiv preprint arXiv:2311.09618*.
- [12] Chuang, Y.S., Goyal, A., Harlalka, N., Suresh, S., Hawkins, R., Yang, S., Shah, D., Hu, J., Rogers, T.T., 2024. Simulating opinion dynamics with networks of llm-based agents. URL: <https://arxiv.org/abs/2311.09618>, arXiv:2311.09618.
- [13] Cisneros-Velarde, P., 2024. On the principles behind opinion dynamics in multi-agent systems of large language models. URL: <https://arxiv.org/abs/2406.15492>, arXiv:2406.15492.
- [14] Clark, E., August, T., Serrano, S., Haduong, N., Gururangan, S., Smith, N.A., 2021. All that's human is not gold: Evaluating human evaluation of generated text. *arXiv preprint arXiv:2107.00061*.
- [15] CLIFFORD, P., SUDBURY, A., 1973. A model for spatial conflict. *Biometrika* 60, 581–588. URL: <http://dx.doi.org/10.1093/biomet/60.3.581>, doi:10.1093/biomet/60.3.581.
- [16] Crespi, I., 2013. *The public opinion process: How the people speak*. Routledge.
- [17] De Duro, E.S., Improta, R., Stella, M., 2025. Introducing counsellme: A dataset of simulated mental health dialogues for comparing llms like haiku, llamantino and chatgpt against humans. *Emerging Trends in Drugs, Addictions, and Health*, 100170.
- [18] De-Marcos, L., Domínguez-Díaz, A., 2025. Llm-based topic modeling for dark web q&a forums: A comparative analysis with traditional methods. *IEEE Access*.
- [19] De Marzo, G., Pietronero, L., Garcia, D., 2023. Emergence of scale-free networks in social interactions among large language models. *arXiv preprint arXiv:2312.06619*.
- [20] DeepSeek-AI, Liu, A., Feng, B., et al., 2025. Deepseek-v3 technical report. URL: <https://arxiv.org/abs/2412.19437>, arXiv:2412.19437.
- [21] Deffuant, G., Neau, D., Amblard, F., Weisbuch, G., 2000. Mixing beliefs among interacting agents. *Advances in Complex Systems* 3, 87–98.
- [22] Degroot, M.H., 1974. Reaching a consensus. *Journal of the American Statistical Association* 69, 118–121. URL: <http://dx.doi.org/10.1080/01621459.1974.10480137>, doi:10.1080/01621459.1974.10480137.
- [23] Demszy, D., Yang, D., Yeager, D.S., Bryan, C.J., Clapper, M., Chandhok, S., Eichstaedt, J.C., Hecht, C., Jamieson, J., Johnson, M., et al., 2023. Using large language models in psychology. *Nature Reviews Psychology* 2, 688–701.
- [24] Donkers, T., Ziegler, J., 2025. Understanding online polarization through human-agent interaction in a synthetic llm-based social network. N.A..
- [25] Dryden, J., Clough, A., 2018. *Parallel Lives*. e-artnow. URL: <https://books.google.it/books?id=amSdAAQBAJ>.
- [26] Fanous, A., Goldberg, J., Agarwal, A., Lin, J., Zhou, A., Xu, S., Bikia, V., Daneshjou, R., Koyejo, S., 2025. Syceval: Evaluating llm sycophancy, in: *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pp. 893–900.
- [27] Farina, M., Lavazza, A., 2023. Chatgpt in society: emerging issues. *Frontiers in Artificial Intelligence* 6. URL: <https://api.semanticscholar.org/CorpusID:259168081>.
- [28] Gao, C., Lan, X., Lu, Z., Mao, J., Piao, J., Wang, H., Jin, D., Li, Y., 2023. S3: Social-network simulation system with large language model-empowered agents. *arXiv preprint arXiv:2307.14984*.
- [29] Holley, R.A., Liggett, T.M., 1975. Ergodic theorems for weakly interacting infinite systems and the voter model. *The annals of probability*, 643–663.
- [30] Jiang, A.Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D.S., de las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L.R., Lachaux, M.A., Stock, P., Scao, T.L., Lavril, T., Wang, T., Lacroix, T., Sayed, W.E., 2023. Mistral 7b. URL: <https://arxiv.org/abs/2310.06825>, arXiv:2310.06825.
- [31] Karimi, F., Génois, M., Wagner, C., Singer, P., Strohmaier, M., 2018. Homophily influences ranking of minorities in social networks. *Scientific reports* 8, 11077.
- [32] Khan, S.T., Joshi, V.R., Bhardwaj, D., Sharma, S.K., Waheed, S.A., Sharma, D.K., 2024. Mitigating voltage fluctuations in grid-tied pv systems: power quality enhancement, in: *2024 3rd International Conference on Computational Modelling, Simulation and Optimization (ICCMO)*, IEEE. pp. 407–416.

- [33] Kosinski, M., 2024. Evaluating large language models in theory of mind tasks. *Proceedings of the National Academy of Sciences* 121, e2405460121.
- [34] Krapivsky, P.L., Redner, S., 2003. Dynamics of majority rule in two-state interacting spin systems. *Physical Review Letters* 90, 238701.
- [35] Li, H., Chong, Y.Q., Stepputtis, S., Campbell, J., Hughes, D., Lewis, M., Sycara, K., 2023. Theory of mind for multi-agent collaboration via large language models. *arXiv preprint arXiv:2310.10701*.
- [36] Lorig, F., Johansson, E., Davidsson, P., 2021. Agent-based social simulation of the covid-19 pandemic: A systematic review. *Journal of Artificial Societies and Social Simulation* 24.
- [37] McPherson, M., Smith-Lovin, L., Cook, J.M., 2001. Birds of a feather: Homophily in social networks. *Annual review of sociology* 27, 415–444.
- [38] Mehdiadeh, A., Hilbert, M., 2025a. Homophily-induced emergence of biased structures in llm-based multi-agent ai systems. *Social Network Analysis and Mining* 15, 1–25.
- [39] Mehdiadeh, A., Hilbert, M., 2025b. When your ai agent succumbs to peer-pressure: Studying opinion-change dynamics of llms. *arXiv preprint arXiv:2510.19107*.
- [40] Minaee, S., Mikolov, T., Nikzad, N., Chenaghlu, M., Socher, R., Amatriain, X., Gao, J., 2025. Large language models: A survey. URL: <https://arxiv.org/abs/2402.06196>, *arXiv:2402.06196*.
- [41] Monti, C., Aiello, L.M., De Francisci Morales, G., Bonchi, F., 2022. The language of opinion change on social media under the lens of communicative action. *Scientific Reports* 12, 17920.
- [42] Moussaid, M., Kämmer, J.E., Analytis, P.P., Neth, H., 2013. Social influence and the collective dynamics of opinion formation. *PLoS one* 8, e78433.
- [43] Niu, M., Jaiswal, M., Provost, E.M., 2024. From text to emotion: Unveiling the emotion annotation capabilities of llms. *arXiv preprint arXiv:2408.17026*.
- [44] Nolfi, S., 2024. On the unexpected abilities of large language models. *Adaptive Behavior* 32, 493–502.
- [45] Noorazar, H., 2020. Recent advances in opinion propagation dynamics: A 2020 survey. *The European Physical Journal Plus* 135, 1–20.
- [46] Nyhan, B., Reifler, J., 2010. When corrections fail: The persistence of political misperceptions. *Political Behavior* 32, 303–330.
- [47] OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., Bao, H., Bavarian, M., Belgum, J., Bello, I., Berdine, J., Bernadett-Shapiro, G., Berner, C., Bogdonoff, L., Boiko, O., Boyd, M., Brakman, A.L., Brockman, G., Brooks, T., Brundage, M., Button, K., Cai, T., Campbell, R., Cann, A., Carey, B., Carlson, C., Carmichael, R., Chan, B., Chang, C., Chantzis, F., Chen, D., Chen, S., Chen, R., Chen, J., Chen, M., Chess, B., Cho, C., Chu, C., Chung, H.W., Cummings, D., Currier, J., Dai, Y., Decareaux, C., Degry, T., Deutsch, N., Deville, D., Dhar, A., Dohan, D., Dowling, S., Dunning, S., Ecoffet, A., Eleti, A., Eloundou, T., Farhi, D., Fedus, L., Felix, N., Fishman, S.P., Forte, J., Fulford, I., Gao, L., Georges, E., Gibson, C., Goel, V., Gogineni, T., Goh, G., Gontijo-Lopes, R., Gordon, J., Grafstein, M., Gray, S., Greene, R., Gross, J., Gu, S.S., Guo, Y., Hallacy, C., Han, J., Harris, J., He, Y., Heaton, M., Heidecke, J., Hesse, C., Hickey, A., Hickey, W., Hoeschele, P., Houghton, B., Hsu, K., Hu, S., Hu, X., Huizinga, J., Jain, S., Jain, S., Jang, J., Jiang, A., Jiang, R., Jin, H., Jin, D., Jomoto, S., Jonn, B., Jun, H., Kaftan, T., Łukasz Kaiser, Kamali, A., Kanitscheider, I., Keskar, N.S., Khan, T., Kilpatrick, L., Kim, J.W., Kim, C., Kim, Y., Kirchner, J.H., Kiros, J., Knight, M., Kokotajlo, D., Łukasz Kondraciuk, Kondrich, A., Konstantinidis, A., Kosic, K., Krueger, G., Kuo, V., Lampe, M., Lan, I., Lee, T., Leike, J., Leung, J., Levy, D., Li, C.M., Lim, R., Lin, M., Lin, S., Litwin, M., Lopez, T., Lowe, R., Lue, P., Makanju, A., Malfacini, K., Manning, S., Markov, T., Markovski, Y., Martin, B., Mayer, K., Mayne, A., McGrew, B., McKinney, S.M., McLeavey, C., McMillan, P., McNeil, J., Medina, D., Mehta, A., Menick, J., Metz, L., Mishchenko, A., Mishkin, P., Monaco, V., Morikawa, E., Mossing, D., Mu, T., Murati, M., Murk, O., Mély, D., Nair, A., Nakano, R., Nayak,
- R., Neelakantan, A., Ngo, R., Noh, H., Ouyang, L., O’Keefe, C., Pachocki, J., Paino, A., Palermo, J., Pantuliano, A., Parascandolo, G., Parish, J., Parparita, E., Passos, A., Pavlov, M., Peng, A., Perelman, A., de Avila Belbute Peres, F., Petrov, M., de Oliveira Pinto, H.P., Michael, Pokorny, Pokrass, M., Pong, V.H., Powell, T., Power, A., Power, B., Proehl, E., Puri, R., Radford, A., Rae, J., Ramesh, A., Raymond, C., Real, F., Rimbach, K., Ross, C., Rotsted, B., Roussez, H., Ryder, N., Saltarelli, M., Sanders, T., Santurkar, S., Sastry, G., Schmidt, H., Schnurr, D., Schulman, J., Selsam, D., Sheppard, K., Sherbakov, T., Shieh, J., Shoker, S., Shyam, P., Sidor, S., Sigler, E., Simens, M., Sitkin, J., Slama, K., Sohl, I., Sokolowsky, B., Song, Y., Staudacher, N., Such, F.P., Summers, N., Sutskever, I., Tang, J., Tezak, N., Thompson, M.B., Tillet, P., Tootoonchian, A., Tseng, E., Tuggle, P., Turley, N., Tworek, J., Uribe, J.F.C., Vallone, A., Vijayvergiya, A., Voss, C., Wainwright, C., Wang, J.J., Wang, A., Wang, B., Ward, J., Wei, J., Weinmann, C., Welihinda, A., Welinder, P., Weng, J., Weng, L., Wiethoff, M., Willner, D., Winter, C., Wolrich, S., Wong, H., Workman, L., Wu, S., Wu, J., Wu, M., Xiao, K., Xu, T., Yoo, S., Yu, K., Yuan, Q., Zaremba, W., Zellers, R., Zhang, C., Zhang, M., Zhao, S., Zheng, T., Zhuang, J., Zhuk, W., Zoph, B., 2024. Gpt-4 technical report. URL: <https://arxiv.org/abs/2303.08774>, *arXiv:2303.08774*.
- [48] Pansanella, V., Rossetti, G., Milli, L., 2022. From mean-field to complex topologies: network effects on the algorithmic bias model, in: *Complex Networks & Their Applications X: Volume 2, Proceedings of the Tenth International Conference on Complex Networks and Their Applications COMPLEX NETWORKS 2021* 10, Springer. pp. 329–340.
- [49] Park, J.S., O’Brien, J., Cai, C.J., Morris, M.R., Liang, P., Bernstein, M.S., 2023. Generative agents: Interactive simulacra of human behavior, in: *Proceedings of the 36th annual acm symposium on user interface software and technology*, pp. 1–22.
- [50] Park, J.S., Popowski, L., Cai, C., Morris, M.R., Liang, P., Bernstein, M.S., 2022. Social simulacra: Creating populated prototypes for social computing systems, in: *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*, Association for Computing Machinery, New York, NY, USA. pp. 383–390. URL: <https://doi.org/10.1145/3526113.3545616>, doi:10.1145/3526113.3545616.
- [51] Park, J.S., Zou, C.Q., Shaw, A., Hill, B.M., Cai, C., Morris, M.R., Willer, R., Liang, P., Bernstein, M.S., 2024. Generative agent simulations of 1,000 people. URL: <https://arxiv.org/abs/2411.10109>, *arXiv:2411.10109*.
- [52] Potter, Y., Lai, S., Kim, J., Evans, J., Song, D., 2024. Hidden persuaders: LLMs’ political leaning and their influence on voters, in: *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 4244–4275.
- [53] Premack, D., Woodruff, G., 1978. Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences* 1, 515–526. URL: <http://dx.doi.org/10.1017/S0140525X00076512>, doi:10.1017/S0140525X00076512.
- [54] Rossetti, G., Stella, M., Cazabet, R., Abramski, K., Cau, E., Citraro, S., Failla, A., Morini, V., Pansanella, V., 2026. Ysocial: An ai-powered social media virtual twin. *Big Data & Society* doi:10.1177/20539517261431576.
- [55] Rossi, L., Harrison, K., Shklovski, I., 2024. The problems of llm-generated data in social science research. *Sociologica: International Journal for Sociological Debate* 18, 145–168.
- [56] Rozado, D., 2024. The political preferences of llms. *PLOS ONE* 19. URL: <https://api.semanticscholar.org/CorpusID:267412830>.
- [57] Sap, M., LeBras, R., Fried, D., Choi, Y., 2022. Neural theory-of-mind? on the limits of social intelligence in large llms. *arXiv preprint arXiv:2210.13312*.
- [58] Sharma, M., Tong, M., Korbak, T., Duvenaud, D., Askeel, A., Bowman, S.R., DURMUS, E., Hatfield-Dodds, Z., Johnston, S.R., Kravec, S.M., et al., 2025. Towards understanding sycophancy in language models, in: *The Twelfth International Conference on Learning Representations*, pp. 383–390.

- [59] Si, X.M., Li, C., 2018. Bounded confidence opinion dynamics in virtual networks and real networks. *Journal of Computers* 29, 220–228.
- [60] Sîrbu, A., Loreto, V., Servedio, V.D., Tria, F., 2017. Opinion dynamics: models, extensions and external effects. *Participatory sensing, opinions and collective awareness*, 363–401.
- [61] Sîrbu, A., Pedreschi, D., Giannotti, F., Kertész, J., 2019. Algorithmic bias amplifies opinion fragmentation and polarization: A bounded confidence model. *PloS one* 14, e0213246.
- [62] Strachan, J.W., Albergo, D., Borghini, G., Pansardi, O., Scaliti, E., Gupta, S., Saxena, K., Rufo, A., Panzeri, S., Manzi, G., et al., 2024. Testing theory of mind in large language models and humans. *Nature Human Behaviour* 8, 1285–1295.
- [63] Street, W., Siy, J.O., Keeling, G., Baranes, A., Barnett, B., McKibben, M., Kanyere, T., Lentz, A., Dunbar, R.I., et al., 2024. Llms achieve adult human performance on higher-order theory of mind tasks. *arXiv preprint arXiv:2405.18870*.
- [64] Taubenfeld, A., Dover, Y., Reichart, R., Goldstein, A., 2024. Systematic biases in llm simulations of debates. *arXiv preprint arXiv:2402.04049*.
- [65] Team, G., Kamath, A., Ferret, J., Pathak, S., Vieillard, N., Merhej, R., Perrin, S., Matejovicova, T., Ramé, A., Rivière, M., et al., 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.
- [66] Törnberg, P., Valeeva, D., Uitermark, J., Bail, C., 2023. Simulating social media using large language models to evaluate alternative news feed algorithms. *arXiv preprint arXiv:2310.05984*.
- [67] Ullman, T., 2023. Large language models fail on trivial alterations to theory-of-mind tasks. *arXiv preprint arXiv:2302.08399*.
- [68] Wang, Y., Zhao, J., Ones, D.S., He, L., Xu, X., 2025. Evaluating the ability of large language models to emulate personality. *Scientific reports* 15, 519.
- [69] Wu, Z., Peng, R., Ito, T., Xiao, C., 2025. Llm-based social simulations require a boundary. *arXiv preprint arXiv:2506.19806*.
- [70] Xie, C., Chen, C., Jia, F., Ye, Z., Shu, K., Bibi, A., Hu, Z., Torr, P., Ghanem, B., Li, G., 2024. Can large language model agents simulate human trust behaviors? *arXiv preprint arXiv:2402.04559*.
- [71] Yang, F., Bai, J., Li, L., Zeng, D., 2025. Modeling social opinion evolution with llm agents: Integrating personality traits with embedded opinion dynamics, in: 2025 International Joint Conference on Neural Networks (IJCNN), IEEE. pp. 1–8.
- [72] Zhang, W., Deng, Y., Liu, B., Pan, S.J., Bing, L., 2023. Sentiment analysis in the era of large language models: A reality check. *arXiv preprint arXiv:2305.15005*.
- [73] Zhang, X., Lin, J., Mou, X., Yang, S., Liu, X., Sun, L., Lyu, H., Yang, Y., Qi, W., Chen, Y., et al., 2025. Socioverse: A world model for social simulation powered by llm agents and a pool of 10 million real-world users. *arXiv preprint arXiv:2504.10157*.

Network Effects and Agreement Drift

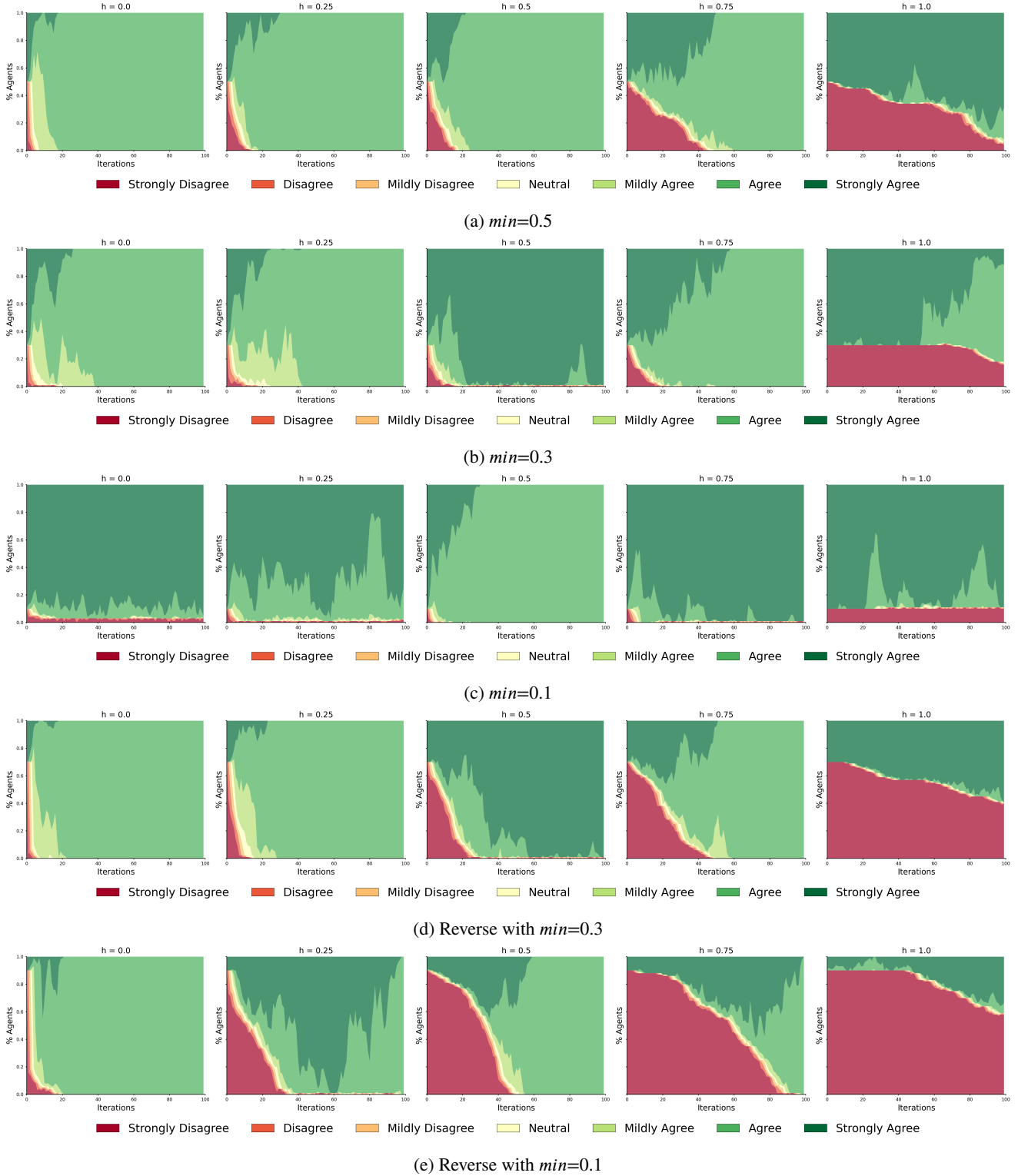


Figure 6: Opinion trends under neighborhood awareness across homophily levels. Panels (a–c) show the base scenario (strongly agreeing majority) for minority fractions $m = 0.5$, $m = 0.3$, and $m = 0.1$, respectively; panels (d–e) show the reversed scenario (strongly disagreeing majority) for $m = 0.3$ and $m = 0.1$. Each panel reports the proportion of agents in each opinion state over 100 iterations. colors map the 7-point scale from *Strongly Disagree* (dark red) to *Strongly Agree* (dark green).

Network Effects and Agreement Drift

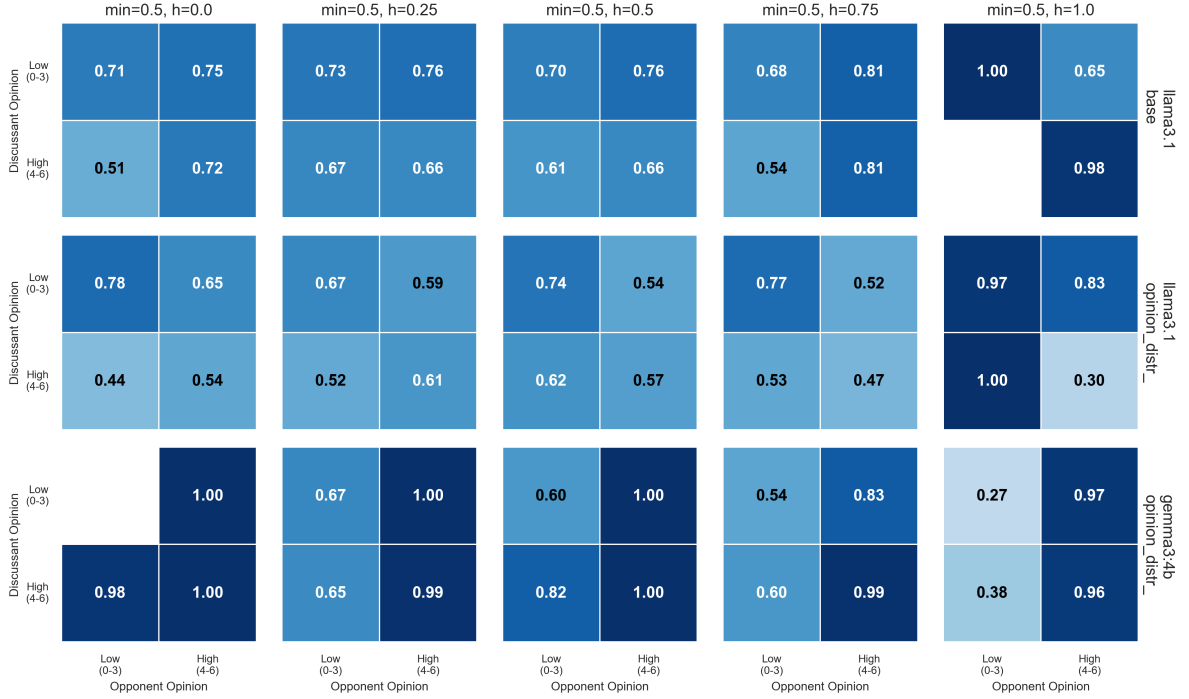


Figure 7: Conditional probabilities of persuasion in pairwise interactions between LLM agents. Each matrix reports $P(\text{disc} \rightarrow \text{opp} \mid o_{\text{disc}}, o_{\text{opp}})$ for balanced populations ($\text{min} = 0.5$), with the 7-point scale collapsed into low- and high-opinion macrostates. Rows correspond to the discussant's initial opinion and columns to the opponent's. Panels compare homophily levels (h) and interaction settings (base vs. neighborhood-aware). Only statistically significant probabilities are shown.

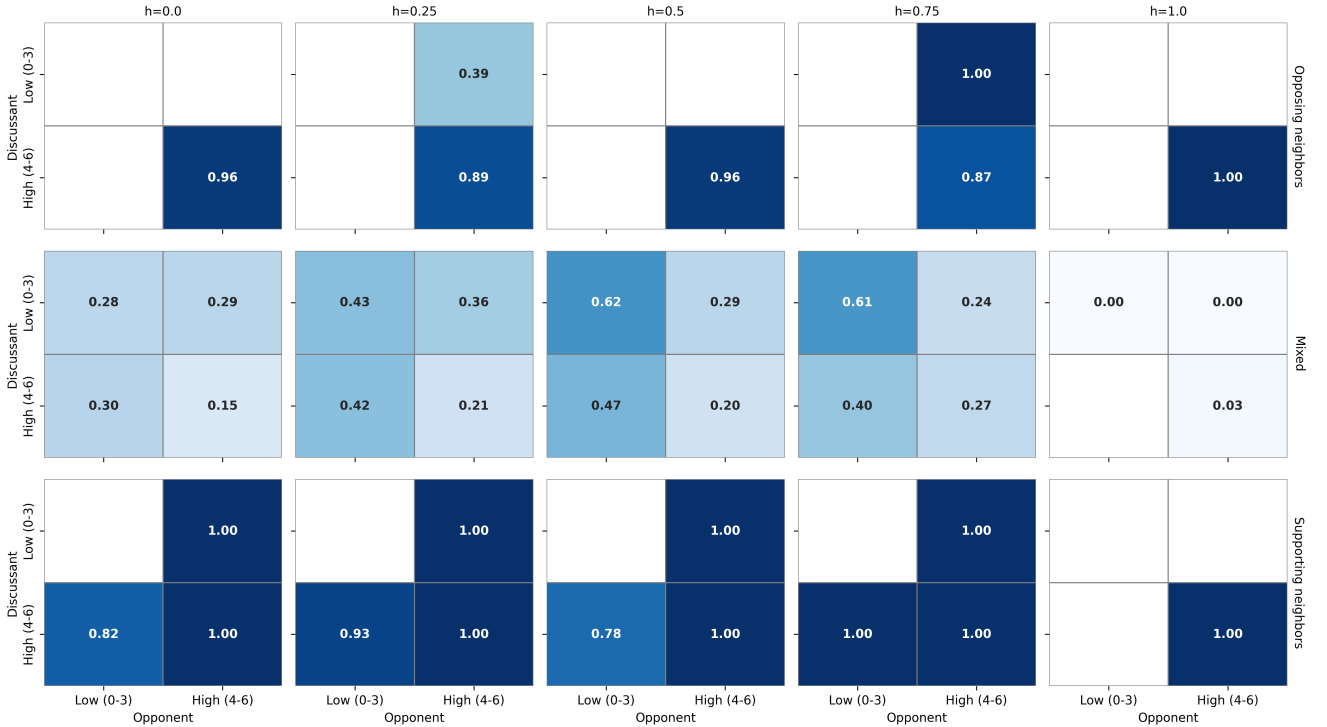


Figure 8: Facet plots of statistically significant opinion shifts for Llama agents under different neighborhood conditions. Each panel reports $P(\text{disc} \rightarrow \text{opp} \mid o_{\text{disc}}, o_{\text{opp}}, c)$, where $c \in \{\text{aligned}, \text{misaligned}, \text{mixed}\}$ denotes the discussant's local neighborhood configuration. Rows correspond to aligned, misaligned, and mixed neighborhoods, while columns vary the homophily level h . Only statistically significant transitions are shown.

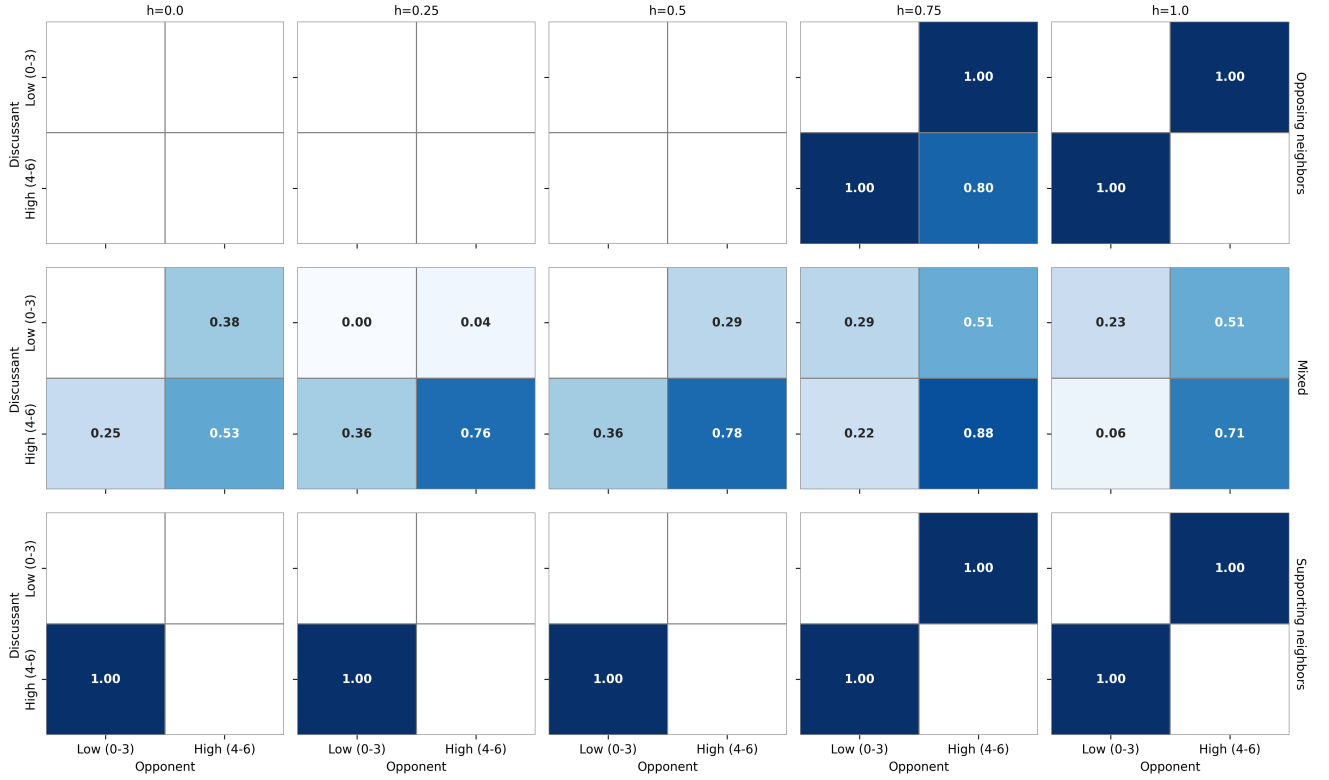


Figure 9: Facet plots of statistically significant opinion shifts for Llama agents under different neighborhood conditions. Each panel reports $P(\text{disc} \rightarrow \text{opp} \mid o_{\text{disc}}, o_{\text{opp}}, c)$, where $c \in \{\text{aligned}, \text{misaligned}, \text{mixed}\}$ denotes the discussant's local neighborhood configuration. Rows correspond to aligned, misaligned, and mixed neighborhoods, while columns vary the homophily level h . Only statistically significant transitions are shown.