Emergence of polarization in networks of large language model agents

Jinghua Piao[1+], Zhihong Lu[1+], Chen Gao[1], Fengli Xu[1], Qinghua Hu[2], Fernando P. Santos[3*], Yong Li[1*], and James Evans[4,5*]

[1]Department of Electronic Engineering, Tsinghua University, Beijing National Research Center for Information Science and Technology (BNRist), Beijing, P. R. China
[2]College of Intelligence and Computing, Tianjin University, Tianjin, P. R. China
[3]Informatics Institute, University of Amsterdam, Amsterdam, the Netherlands
[4]Knowledge Lab, University of Chicago, Chicago, U.S.A
[5]Santa Fe Institute, Santa Fe, U.S.A.
[*]To whom correspondence should be addressed; Email: f.p.santos@uva.nl, liyong07@tsinghua.edu.cn, jevans@uchicago.edu.
[+]Jinghua Piao and Zhihong Lu contribute equally to this work.

arXiv:2501.05171v3 [cs.SI] 7 Sep 2026

ABSTRACT

Rapid advances in large language models (LLMs) have not only empowered autonomous agents to generate social networks, communicate, and form shared and diverging opinions on political issues, but have also begun to play a growing role in shaping human political deliberation. Our understanding of their collective behaviours and underlying mechanisms remains incomplete, however, posing unexpected risks to human society. In this paper, we simulate networked systems involving thousands of LLM agents across different backbone models (GPT-3.5, GPT-4o, ChatGLM, Llama-3, and DeepSeek-V3), in which agents interact through LLM-guided conversations and update their opinions over time, resulting in the emergence of opinion polarization. We discover that these agents spontaneously develop their own social network with properties characteristic of human social networks, including homophilic clustering. The collective opinions of these LLM agents evolve in ways that exhibit behavioural patterns consistent with social phenomena and mechanisms widely discussed in studies of human behaviour. Overall, these behaviours, patterns, and emergent phenomena produced by LLM agents are consistent with real-world observations and established opinion-dynamics models. This consistency suggests that LLM agents can serve as a valuable synthetic testbed for exploring hypothetical intervention strategies in networked LLM-agent systems.

Introduction

The recent development of large language models (LLMs) has not only advanced machine capabilities in traditional natural language processing tasks [1, 2], but also unlocked the design of agents with human-level intelligence in communication [3–6], reasoning [7, 8] and decision-making [6, 7, 9]. These capabilities suggest that artificial agents, driven by LLMs, can simulate human behaviours and undertake increasingly challenging tasks for humans [3, 10–12]. On the other hand, concerns about LLMs are intensifying rapidly [4, 5, 13–15]. Growing studies have documented harmful LLM behaviours [5, 15, 16]. For example, their generated content is prone to toxicity [15], biases [5, 17], hallucinations [16], and more. In addition, LLMs, behaving like humans, also adopt misbehaviours, including deception [18], sycophancy [19], and demographically biased mistrust [20].

Besides these evident misbehaviours, LLMs may pose broader risks in more subtle ways. Numerous studies have pointed out that LLMs can generate persuasive political content [21–23], potentially affecting human opinions and behaviours [13, 23–25]. More crucially, due to the increasing indistinguishability of content generated by LLMs from that created by humans [4, 21, 22, 24, 26], concerns arise over their potential to influence public opinions and intensify polarization as the scale of interaction increases [24, 25, 27]. Unfortunately, despite the urgency highlighted by these examples, we remain far from being able to systematically characterize these collective risks. This is largely because of two key gaps in our current understanding: (i) Unlike traditional rule-based agents, LLM agents operate in human languages generated by black-box models with billions to trillions of parameters, hindering our understanding of the underlying mechanisms they follow when forming opinions. (ii) Interactions enable them to evolve, gradually developing their own social network and collective opinions [28]. This bottom-up process, growing from microscopic interactions to macroscopic social order, makes it difficult to predict LLM agents' collective opinions and behaviours. Taken together, addressing these two gaps begins with a crucial first step: understanding the opinion dynamics of LLM agents, both as individuals and, more importantly, as evolving artificial collectives.

Recently, researchers have made pioneering efforts to investigate the impact of LLM-generated texts or interactions on political opinions and polarization [22–24, 29–34]. Beyond the risks discussed above [22–24], LLMs can act as mediators, facilitating constructive political communication [30, 35–38], reducing conspiracy beliefs [32], and helping ideologically divergent people find common ground [29]. Nevertheless, these studies primarily focus on the role of LLMs as an isolated counterpart to humans, overlooking their collective behaviours. This gap largely limits our understanding of LLM agents' opinion dynamics. On the other hand, a growing line of work has begun to explore generative agent-based modeling (GABM) using large language model (LLM) agents [3], with particular interest in simulating social networks [39–42], opinion dynamics [43–46], and social media environments [44, 47–50]. Recent studies have shown that LLM agents can be used to generate social network structures that exhibit certain statistical similarities with human networks, such as scale-free degree distributions [39–42], or to simulate specific components of social media, including opinion feed algorithms [44, 47], coordinated behaviours [50], and echo chamber formation [43]. Despite these advances, existing work faces two key limitations. First, many studies operate at relatively limited scales in terms of the number of agents, interaction density, or network complexity, which constrains their ability to capture genuinely emergent collective phenomena that arise only from large-scale, repeated interactions [43, 45, 46]. Second, the autonomous social capabilities of LLM agents are largely constrained [45, 46]. Currently, they are primarily treated as tools for specific subtasks, such as eliciting opinions [43] and generating network edges [39], or as puppets heavily scaffolded by pre-defined rules [41, 42]. As a result, these studies offer limited insight into autonomous opinion dynamics and collective behaviour among LLM agents, leaving open the question of how autonomous agents evolve and give rise to emergent social phenomena.

In this work, we simulate a synthetic LLM-agent system, where thousands of agents, solely driven by LLMs, freely establish social relationships, communicate, and form opinions on political issues. We discover that these free-form social interactions among LLM agents result in the emergence of opinion polarization, a phenomenon widely observed in human societies [51–58]. Meanwhile, LLM agents spontaneously organize their own social network with properties characteristic of human social networks: agents with homophilic opinions tend to cluster, while those with opposing opinions tend to avoid interactions [28, 58–60]. The self-organized networks that result, in turn, shape their collective opinions in ways that exhibit behavioural patterns consistent with social phenomena and associated mechanisms initially identified in human social networks [51, 53, 55, 58, 61, 62]. These phenomena, such as polarization and homophilic clustering, underscore how networked interactions transform otherwise isolated LLMs into a collectively evolving system with emergent dynamics. Importantly, rather than proposing a theory of polarization, our results align with established opinion-dynamics models, particularly the argument-accumulation lineage [58, 60, 62], where polarization emerges from cumulative exchanges of persuasive arguments under homophilic interactions. Shifting focus from the network to individuals, we further examine whether agents' behaviours align with a wide range of social mechanisms contributing to real-world polarization, e.g., selective exposure [63], confirmation bias [64, 65], and elite signaling [56], finding their effects on LLM agents align with those reported in studies of human behaviour [56, 63–65]. In sum, the network structures, collective opinion patterns, and interaction dynamics produced by the LLM-agent simulations are consistent with real-world observations and established opinion-dynamics models. This consistency suggests that the proposed system can serve as a valuable synthetic testbed for exploring hypothetical intervention strategies in the networked LLM-agent system. Through extensive intervention experiments in our system, we find that in a highly polarized and nearly crystallized social network, directly modifying network structures has limited effect on reducing polarization. Instead, encouraging access and open-mindedness to diverse opinions at the individual level proves more effective. Overall, this work not only sheds light on subtle opinion dynamics and collective behaviour of LLM agents from a network perspective, but it also unveils their potential to serve as a testbed for exploring hypotheses about opinion dynamics in networked LLM-agent systems.

Results

Opinion dynamics in a synthetic system of interacting LLM agents.

We begin by building LLM agents with the most basic social interaction capabilities, including establishing social relationships, communicating, and forming opinions on political issues. Specifically, these capabilities are implemented through three stages purely driven by LLMs (Figure 1a). Taking agents i and j as examples, in the self-expression stage, agent i is required to generate reasons supporting its current opinion on a political issue. In the communication stage, agent i first decides whether to communicate with a socially connected or a random new agent j. Based on the opinions and reasons of agents i and j, agent i generates messages to persuade agent j of its opinion. In the opinion update stage, agent j updates its opinions based on messages received from all of its socially connected agents (i.e., agent i). It is worth mentioning that the evolution of agents' opinions and behaviours is driven solely by LLMs without any pre-assumed rules or mechanisms, which allows us to probe the free-form social properties of LLM agents (see details about LLM agents in Method M1 and Supplementary Methods). As such, the system does not rely on traditional "hard" parameters. Instead, prompts serve as "soft" parameters that shape how agents perceive, reason, and interact, and can be used to capture heterogeneous human subgroups with distinct psychological and behavioural patterns (Supplementary Note 10). We run LLM agents described and visualized in the Main Text based on

ChatGPT through the public OpenAI API. We also examine agents driven by other LLMs, including ChatGLM, Llama-3, etc., with comparable results (Supplementary Note 6).

To explore the opinion dynamics of LLM agents, we simulate a networked system based on these LLM agents (Figure 1a). We focus predominantly on the three critical political issues – partisan alignment, gun control, and abortion ban – where concerns over opinion polarization have persisted for decades [51, 54, 61]. We measure agents' opinions on various political issues across the left-right political spectrum [66–68], and adopt a widely used five-level political scale: left, moderate left, neutral, moderate right, and right [54, 69, 70]. Following widely adopted practices in prior human studies [52, 55], we prompt agents to self-report their opinions on political issues. For each issue, thousands of LLM agents freely exchange opinions and develop social relationships within the networked system. Following prior practice [71–73], we randomly initialize agents' opinions with a near-Gaussian distribution (Figure 1c) and their social relationships with a Watts–Strogatz random small world network (see details in Method M1 and Supplementary Methods). Apart from these minimal initializations, we do not pre-assume any backgrounds, demographics, behaviours, memories, or detailed thoughts for LLM agents. Instead, all other information, such as agent i's reasons for supporting a certain opinion, is generated through prompting LLMs using their historical interactions with others (see implementation prompts in Supplementary Methods). Moreover, as shown in Figure 1b, opinion distributions diverge from the initial near-Gaussian distribution in fewer than 10 rounds of social interactions. This indicates that these agents, despite being placed within an initial opinion distribution, spontaneously develop distinctive collective opinions through social interaction.

This raises a natural question: what collective opinions will these LLM agents develop? As shown in Figure 1c, we observe a consistent pattern across issues of partisan alignment, gun control, and abortion ban: the proportion of neutral opinions, initially the largest at 40%, decreases to only 22.5%, 0.4%, and 5.1% for the three issues, respectively. This observation highlights the difficulty of maintaining neutral positions in long-term social interactions. Moreover, the agents who move away from neutral positions are spontaneously polarized into two camps: one who holds left-leaning opinions (denoted in blue) and the other who holds right-leaning opinions (denoted in red). This suggests that political polarization, a long-standing concern in real-world society, also emerges in the networked synthetic systems of LLM agents. By dividing the stationary opinion distributions into left-leaning, neutral, and right-leaning camps (Figure 1d), we find that these polarized distributions consistently exhibit a left-skewed pattern, in contrast to real-world observations where the power between left-leaning and right-leaning camps is generally more balanced (Supplementary Table S1). Indeed, prior studies have pointed out the inherent left-leaning bias of OpenAI's GPT [27, 74–80], which largely accounts for the observed left-skewed pattern. In open models, such bias has been demonstrated to occur largely through fine-tuning [25], likely on correlated qualities including positivity, openness, and nontoxicity [81], or even specific forms of reason [32, 82]. These insights, however, do not account for ways in which LLMs interact with each other: the micro-social process of imbalanced LLM polarization has not been explored [83].

Polarization patterns emerge from self-regulated LLM agents.

To validate whether the left-skewed pattern originates from the inherent bias of LLMs, we design a pairwise interaction-based experiment (Figure 2a). In particular, we follow typical settings for a networked system (see details in Methods M1 and Supplementary Methods), but consider two key simplifications: First, to prevent network structures from interfering with the examination of inherent bias, we ensure that each agent communicates with only one other. Second, we set these two agents to share the same opinion, which allows us to disentangle the effects of bias from social influence on changing agents' opinions (Method M2). By evaluating opinion transition probabilities after the one-round interaction (Figure 2b and Supplementary Figures S3, S4), we notice an asymmetric opinion update: right-leaning agents occasionally switch to left-leaning after one-round pairwise interactions, whereas left-leaning agents do not switch to right-leaning. Similar cases of the asymmetric opinion update are also noted in other issues, including abortion ban, gun control, ObamaCare, etc. (Supplementary Note 3). These observations indicate that the observed skewed opinions result from the individual formation of LLM agents' opinions.

Following this insight, we track the behaviour of agents who have transitioned to opposing opinions (Supplementary Note 3), finding the self-inconsistency problem inherent in LLM agents. For example, a neutral agent, who ideally should remain impartial towards both Democratic and Republican parties, states, "*As someone who deeply cares about politics, I believe in the values of equality, progress, and inclusion that the Democrats stand for. I strongly support the Democratic party*". This agent, despite currently adopting a neutral opinion, self-expresses a left-leaning preference for the Democratic party, which is inconsistent with its neutrality (see more examples in Supplementary Note 3). Indeed, the problem is prevalent across the three stages: agents could generate reasons and communications inconsistent with their opinions, and they may also update their opinions in ways inconsistent with the messages they have received. Among the three stages, the opinion update stage exhibits the highest frequency of self-inconsistency (Supplementary Figure S6), highlighting the particular difficulty for LLM agents to comprehend messages and transform them into their opinions. More importantly, this problem occurs significantly more frequently among agents with right-leaning opinions than those with left-leaning opinions (two-sided proportion z-test, $z = 8.27$, $p \ll .001$, difference in proportions = 0.068, 95% CI $= [0.052, 0.084]$). This results in right-leaning agents exhibiting a

disproportionately high probability of transitioning to the opposing side, accounting for the observed skewed opinion dynamics.

We remedy the self-inconsistency problem using a theory-driven method inspired by the social science literature [84, 85]: As noted in Bandura's social learning theory, self-regulation is an essential human ability to continuously monitor and adjust behaviours to respond more consistently and adaptively in dynamic social contexts [84, 85]. Here we propose a strategy to incorporate that self-regulation capacity into LLM agents. In particular, agents are prompted to self-check whether their generated messages, opinions, and reasons are consistent with their current status. If agents find inconsistencies, they adjust themselves through iterative re-generation until consistency is achieved (more details in Methods M3 and Supplementary Methods). This way, these agents can independently recognize and rectify glaring inconsistencies in expressed thoughts (i.e., opinions and corresponding justification) and behaviour. We assess the effectiveness of the proposed strategy using the metric of self-inconsistency rate, calculated as the distance between the experimental opinion transition probability matrix and the identity matrix (see details in Method M5). As shown in Figure 2c, the self-regulation strategy reduces self-inconsistency by 9.4%-52.2% across scenarios on a variety of issues. This highlights the essential role of self-regulation in improving consistency between generated opinions, reasons, and messages in LLM agents. It is worth noting that this strategy is introduced to ensure the behavioural plausibility and internal consistency of LLM agents, rather than to remove or correct their inherent tendencies. Based on these more self-consistent agents, we further explore their collective behaviour by simulating networked systems of self-regulated LLM agents and assessing their opinion dynamics (Figure 2d). We discover that self-regulated LLM agents also develop their collective opinions into a polarized pattern, suggesting that the emergence of LLM agent polarization is not due to the inherent bias of LLMs, but results from their free-form social interaction. Moreover, unlike the previously left-leaning dominated polarization pattern (Figure 1b), the right- and left-leaning camps are now balanced, which better reflects real-world scenarios (Supplementary Table S1 and Figure S1). Overall, LLM agents, despite only being empowered with these basic social capabilities, spontaneously develop similar polarization patterns through social interaction.

Collective behavioural patterns behind the emergence of polarization among LLM agents.

The above results uncover that free-form social interactions among LLM agents result in the emergence of opinion polarization, raising a natural question regarding the exact role these interactions play in this process. Therefore, we begin by assessing the extent to which agents interact with those holding similar opinions. As shown in Figure 3a, we find that the proportion of interactions between agents in the same camp, i.e., left- or right-leaning camp, increases by 156.8%-382.7% over time. This increase eventually leads to 48.5%-88.3% of interactions in these systems occurring between agents who share similar opinions, suggesting a tendency for similar agents to cluster together. This process is consistent with the notion of homophily, where "birds of a feather flock together" [58, 86], and has been observed in naturalistic LLM social interactions [28]. Moreover, our agents increasingly avoid interacting with agents who hold opposing opinions (Supplementary Figure S8a). Indeed, their tendencies of homophilic clustering and opposing avoidance appear to be driven by their generated evaluations of others. Further experiments show that agents are more likely to form more favorable impressions of and use more positive language to describe those within their own camp than those in opposing camps (Supplementary Note 8). It is worth mentioning that no mechanisms prompting in-group bias or homophilic clustering were inserted into the system; this emerges purely from the LLM agents' free-form choice behaviours in social relationships. More interestingly, in the system discussing the abortion ban, agents gradually self-organize into two communities (Figure 3b), with one supporting left-leaning opinions and the other supporting right-leaning ones. Indeed, beyond LLM agents, homophilic clustering is also widely observed in human society [51, 55, 61] and in- vs. out-group bias has been posited by evolutionary psychologists with the emergence of justificatory reason [87].

A widespread subsequent concern is the potential for the clustering of homophilic peers to reinforce opinion polarization, which is often termed the echo chamber effect in studies of human behaviours [51, 55]. Therefore, we examine the relationship between homophilic clustering and polarization levels in the system of LLM agents. In particular, we calculate the average opinions of each agent's neighbors to quantify the relative levels of radicalization between the agent and its neighbors (see details in Method M5). Focusing on the interactions between homophilic agents (Figure 3c), we observe that exposure to more radical agents significantly increases the level of polarization (two-sided Student's t-tests, partisan alignment $t(169) = 16.80$, $p \ll .001$, Cohen's $d = 1.288$, 95% CI $= [0.633, 0.802]$, gun control $t(1465) = 19.37$, $p \ll .001$, Cohen's $d = 0.506$, 95% CI $= [0.209, 0.256]$, and abortion ban $t(245) = 13.99$, $p \ll .001$, Cohen's $d = 0.892$, 95% CI $= [0.447, 0.594]$). This indicates that LLM agents exhibit behavioural patterns consistent with the echo chamber effect. On the other hand, despite homophilic interactions representing the majority (54.8%-82.2%), agents still retain a small chance of encountering those with opposing opinions (7.4%-10.2%). Nevertheless, these interactions with opposing agents do not consistently reduce the polarization level as expected (Figure 3d). Instead, exposure to agents with opposing opinions could potentially increase polarization levels, indicating behavioural patterns consistent with the backfire effect [53]. We note that the echo chamber and backfire effects do not universally occur across all individuals in the real world, with some recent theoretical debates [55, 88, 89]; instead, they exhibit substantial heterogeneity. Extensive experiments reveal that the system autonomously formed by the LLM agents can capture this heterogeneity, showing consistency with empirical observations (Supplementary Note 4).

Overall, these observations suggest that in the discussion of critical political issues, LLM agents can spontaneously organize their social network. Their networks not only exhibit the property of homophilic clustering, but also shape collective opinions in ways consistent with the echo chamber and backfire effects. To support these findings, we report further quantitative metrics, including modularity, assortativity, and the homophily index, along with extended analyses on the network properties of their self-organization (see Supplementary Note 4). Notably, we observe that social networks evolve into a scale-free degree distribution, where high-degree agents gain great popularity in free-form interaction (Supplementary Figure S13), following degree distribution models of social networks [90]. Moreover, when we replace LLM agents' self-organized network with a random or static network, we find that one-camp opinions dominate the overall system (Supplementary Figure S14), which highlights the important role of self-organized social networks in shaping polarization patterns exhibited by LLM agents. Importantly, these results align with established argument-accumulation models in opinion dynamics, where polarization emerges from the cumulative exchange and reinforcement of persuasive arguments under homophilic interactions [58, 62], as reflected in our analysis of agents' self-reported reasons (Supplementary Note 8).

As discussed above, the polarization of LLM agents is driven by two coupled processes of network organization and opinion formation. Specifically, in the network organization process, agents tend to interact with similar peers (Figures 3a and b), while in the opinion formation process, agents tend to comprehend messages in a way that aligns with their pre-existing opinions (Figures 3c and d). These two tendencies of LLM agents have been frequently observed in humans [51, 52, 56, 63, 64], and are often referred to as selective exposure [63] and confirmation bias [64, 91], respectively. Unlike network-level patterns, these two mechanisms adopt a psychological perspective, focusing more on how individuals' traits and behaviours contribute to polarization. We note that these two mechanisms are not explicitly hard-coded or assumed in the proposed system. Rather, agents' behavioural patterns consistent with these mechanisms naturally emerge from agents' social interactions. For example, analyzing the self-reported reasons generated during opinion formation reveals patterns in line with confirmation bias (Supplementary Note 8).

Given these observed network-level patterns, one may wonder whether LLM agents exhibit behaviours consistent with psychological mechanisms discussed in studies of human opinion formation. To answer this question, we design a comparative experiment: we vary the proportion of agents exhibiting behaviours consistent with selective exposure or confirmation bias, and then examine the resulting changes in the level of polarization. Cases where 0% or 100% of agents are configured to exhibit these behaviours are deliberately designed as extreme scenarios for comparative purposes. As shown in Figure 3e and f, we observe that the increase in agents exhibiting either behaviour leads to a marked increase in the level of polarization characterizing the overall system. This suggests that behaviours related to selective exposure and confirmation bias contribute to the polarization of LLM agents, consistent with findings reported in studies of human behaviour [56, 64, 92].

On the other hand, prior studies have pointed out that opinion formation is also largely influenced by the traits or behaviours of others, particularly top influencers on social media [54, 56, 92–94]. As suggested by the elite signaling mechanism [54, 92, 93], the polarization level of influencers' opinions is positively associated with greater polarization in the overall population. To validate this effect on LLM agents, we introduce additional top influencers holding neutral, moderate left/right, and strong left/right opinions into the system, with these influencers sending non-personalized messages to all other agents (Figure 3g). In this setting, the top influencers are symmetrically distributed across the right- and left-leaning camps. We observe an effect in the LLM-agent system that is qualitatively consistent with patterns reported in studies of human opinion formation. When the influencers hold non-neutral opinions, the final polarization level is significantly increased (two-sided Student's t-tests, left/right versus original, $t(8) = 10.63$, $p \ll .001$, Cohen's $d = 6.720$, 95% CI $= [0.116, 0.180]$; moderate versus original, $t(8) = 8.64$, $p \ll .001$, Cohen's $d = 5.462$, 95% CI $= [0.128, 0.220]$). By contrast, the introduction of neutral influencers significantly reduces the polarization level of the system by 28.5% (two-sided Student's t-tests, original versus neutral, $t(8) = 21.56$, $p \ll .001$, Cohen's $d = 13.636$, 95% CI $= [0.399, 0.495]$). Besides the three studied mechanisms, we also examine a series of other individual-level mechanisms, finding their effects on LLM agents to be consistent with those observed in humans (Supplementary Note 5). Overall, these experiments consistently suggest that LLM agents exhibit individual-level behavioural responses that parallel those documented in studies of human opinion formation.

Synthetic testbed for exploring hypothetical intervention strategies.

The above results suggest that the behaviours and collective patterns produced by the LLM agents are consistent with empirical observations and established opinion-dynamics models. This consistency suggests the potential of the proposed LLM-agent system as a valuable synthetic testbed for exploring hypothetical intervention strategies in networked LLM-agent systems. To explore this potential, we design five intervention strategies in our synthetic testbed (Figure 4). At the network level, we consider two strategies: (i) Random interactions (RI): we remove the homophilic clustering property of LLM agents' social network by allowing agents to randomly interact with each other. (ii) Moderate opposing interactions (MOI): given the echo chamber and backfire effects, agents are only allowed to receive messages from those adopting moderate opposing opinions. At the individual level, three strategies are proposed: (iii) No selective exposure (NSE): agents are prompted to have the tendency

to communicate with those holding diverse opinions. (iv) No confirmation bias (NCB): agents are instructed to be open-minded to persuasion of diverse opinions. (v) Neutral elite signaling (NES): agents receive non-personalized neutral messages from newly introduced top influencers. Indeed, these interventions are motivated by established lines of prior work and are used here as experimental interventions in our synthetic system: MOI draws on studies of exposure to cross-cutting information [89, 95]; RI operationalizes randomized interaction; NSE and NCB relate to promoting open-mindedness [96]; and NES is inspired by the design of Bail et al. [97].

Based on the proposed strategies, we intervene in a system that has converged to a polarized state, i.e., the system at $t = 35$ (Figure 4a). By comparing the evolution of the original system with that of the intervened system following intervention (Figure 4b-f), we find that all strategies significantly reduce the level of polarization. In particular, the intervention of MOI has a clear and significant effect in reducing opinion polarization among agents (two-sided Student's t-test, $t(10) = 5.24$, $p < .001$, Cohen's $d = 3.027$, 95% CI $= [0.011, 0.027]$). The effects are relatively small, however, producing approximately a 2% change in agent opinions. This modest effect is consistent with findings from human studies, where cross-cutting interactions reduce polarization but typically result in only small shifts among actual opinions [89, 95]. Compared with network-level interventions that directly modify agents' social networks, individual-level strategies of NCB and NES contribute to the greatest reduction in polarization by 11.8% and 8.8%, respectively. On the other hand, we observe that the NSE strategy has limited effects. This is because, in an almost stable social network, agents' non-selective tendencies do not provide exposure to diverse opinions but instead quickly crystallize the existing network structures that foster polarization (see detailed analyses on intervention strategies in Supplementary Note 7). Despite the difficulty of recommending openness to human populations, these results suggest that in our synthetic testbed, access and open-mindedness to diverse opinions at the individual level appear more effective than altering social networks in polarized systems. Furthermore, we observe that all of the strategies, despite reducing homophilic interactions by varying rates from 2.6% to 100%, significantly promote inclusive conversations among agents with diverse opinions (lower sub-figure in Figure 4b-f). In all, these experiments compare several hypothetical strategies and illustrate the value of the proposed system for controlled exploration of intervention hypotheses within networked LLM-agent systems.

Discussion

This work uncovers that LLM agents exhibit polarization patterns on political issues, at both the individual and collective levels. Specifically, from the individual perspective, opinion formation among LLM agents exhibits behavioural patterns consistent with social and psychological mechanisms [63, 64], e.g., selective exposure, confirmation bias, elite signaling. From the collective perspective, LLM agents organically develop social networks with properties characteristic of human social networks, e.g., homophilic clustering and scale-free degree distributions [28]. Moreover, their self-developed social networks, in turn, shape LLM collective opinions in ways consistent with the echo chamber and backfire effects described in studies of human behaviour, resulting in the emergence of opinion polarization [51–56, 94, 98, 99]. Notably, all these behavioural patterns, network organization, and collective opinions naturally emerge from interactions among agents completely driven by LLMs. Building on these findings, our work shifts the perspective from isolated LLMs [29, 30, 32, 35–38] to complex, networked LLM agent societies, offering a synthetic testbed for studying emergent collective dynamics in artificial agents. While our main results are based on OpenAI's ChatGPT and three political issues, we also perform extensive experiments with different LLMs (both full-run and midway-defection settings, Supplementary Note 6) as well as on topics of immigration restriction and the flat Earth theory (Supplementary Note 13), finding that these experiments consistently support our core findings. We also vary two key initial conditions, i.e., the social network structure and initial opinion distribution, and find that both resulting polarization patterns and final network structures remain strikingly similar, highlighting the robustness and generalizability of our system and experiments (Supplementary Note 12).

While this work does not make theoretical contributions, it is grounded in established theories and models of opinion dynamics [58, 62, 100, 101]. In particular, this work follows the argument-accumulation lineage, which explains polarization as arising from the cumulative exchange and reinforcement of persuasive arguments [58, 60, 100, 101] rather than from direct opinion averaging or social pressure [102, 103]. We further analyze agents' self-reported reasons during opinion formation, observing that agents who become more polarized tend to accumulate substantially more supporting arguments over time (Supplementary Note 8). This observation highlights a clear connection to argument-accumulation models and suggests that the polarization dynamics observed in our synthetic LLM-agent testbed likely stem from similar underlying mechanisms. At the same time, we acknowledge broader theoretical debates on social media and polarization. Some recent studies have emphasized that echo chamber effects do not universally apply across all individuals in real-world settings, with substantial heterogeneity in exposure to like-minded content [55, 88]. In our experiments, we also observe similar patterns in LLM agents (Supplementary Note 4). Moreover, deliberative and argument-accumulation models have limitations, such as their abstraction from affective responses and identity-based dynamics [104–106], which have been highlighted in more recent conceptions of polarization. While our primary analyses focus on opinion-based polarization, we provide supplementary analyses of behavioural patterns

consistent with some of these broader dimensions (Supplementary Note 8).

Previous studies have documented the presence of political biases in LLMs [27, 30, 76, 77, 107, 108]. In particular, there are two main lines of research. The first line focuses on analyzing political bias in "plain vanilla" LLMs, to examine the extent of biases present in their default configurations [27, 77, 107, 108]. The second line of research assigns demographics to LLMs and evaluates how well these models represent different populations [30, 44, 76]. While this approach allows LLMs to adopt political opinions resembling those of real-world individuals with similar demographics [30, 76], they also find that LLMs fall short in accurately representing certain minority groups [76, 109]. Overall, some recent studies attempt to move beyond default configurations and explore the potential of demographically prompted LLMs as proxies for human populations. Nevertheless, they have yet to explore whether and how political biases may emerge through LLM agents' social behaviours – including friend selection, opinion updating, and patterns of communication. Our system addresses this gap and can be readily extended to model more diverse human subgroups by leveraging prompts as "soft" parameters (Supplementary Note 10). This naturally leads to the next question: will LLM agents faithfully follow their assigned political personas and exhibit behaviours consistent with humans holding comparable identities? In this work, we discover that the underlying political bias of LLMs does lead these agents to occasionally exhibit behaviours that deviate from their assigned political personas, which we refer to as self-inconsistency. This inconsistency causes agents to gradually drift toward left-leaning positions during interaction (Figure 2b), and over time, cumulatively results in the collective left-skewed polarization observed in Figure 1b–d. To address self-inconsistency, we propose the self-regulation strategy inspired by Bandura's social learning theory [84, 85], enabling agents to self-monitor and adjust their behaviours. We find self-regulated LLM agents develop their collective opinions into balanced right-left polarization (Figure 2d), better reflecting real-world distributions. Furthermore, comparison of polarization speeds between the original and self-regulated agents reveals only slight differences, suggesting that model bias has a minimal impact on accelerating polarization (Section 3.2 in SI). Overall, these findings underscore the importance of studying bias and consistency at the level of interacting LLM agents, and of investigating collective opinion dynamics in synthetic social systems.

While this work advances prior GABM studies by providing a solid synthetic testbed in which large-scale LLM agents, under minimal social assumptions, produce emergent polarization patterns, it also shares limitations common to current GABM research [110–112]. In particular, like prior studies, the proposed system faces natural challenges to external validity. Although results are consistent with well-established opinion-dynamics models [51, 53, 58], the extent to which these patterns quantitatively correspond to real-world human behaviour must be further validated. In addition, compared to prior models with transparent mechanisms, GABM comes at the cost of reduced interpretability, as agent behaviour is shaped through prompt-based design and simplified opinion representations. Moreover, the intervention results presented here should be understood as hypothesis-generating. The primary value of the proposed system lies in enabling controlled, pre-experimental exploration of how interaction rules, network structures, and agent-level behaviours may jointly shape collective opinion dynamics. Future research will benefit from complementary human experiments [58, 113], as well as more systematic calibration and evaluation against fine-grained real-world data.

Methods

M1 Design of the networked system of LLM agents

To understand the opinion dynamics of LLM agents, we develop a networked system of LLM agents. In this system, each agent is empowered with three stages that ensure their basic social capabilities: (i) self-expression, which enables agents to consider and express their reasons for supporting their opinions, (ii) communication, which enables agents to exchange opinions and the corresponding reasons, and (iii) opinion updates, which enable agents to update their opinions based on received messages. These three stages take advantage of LLMs' capabilities in traditional natural language processing tasks [1, 2]. For example, the self-expression stage predominantly depends on automated text generation, and the opinion update stage relies on text summarization. These capabilities allow agents to perform basic social interactions. Below we introduce our proposed system in detail.

In each system, agents communicate about one political issue, e.g., gun control or abortion rights. We randomly initialize the system, including agents' opinions on the political issue and their social relationships. When initializing the system, we consider two aspects. First, the initialization should prevent the system from starting from an extreme case, e.g., when agents' opinions are already highly polarized. Second, the initialization setting should allow enough flexibility for agents to develop their collective opinions and form social networks. As starting from a clustered network and highly polarized opinion distribution would result in a crystallized setting, not allowing us to study how interaction dynamics emerge over time, we introduce some randomness in the initial networks and opinions considered. Specifically, we sample each agent's initial opinion following a near-Gaussian distribution as shown in Figure 1c. We initialize the social network using a Watts–Strogatz model with the rewiring probability of 0.001, which exhibits a relatively weak small-world property, characterized by a normalized clustering coefficient of 0.99 and a normalized average path length of 0.71.

After initialization, the self-expression stage requires each agent i to generate a message expressing the reasons for supporting their opinions. These messages provide detailed information about agents' beliefs regarding the political issue. Then, following the communication stage, each agent i first chooses whether to communicate with a socially connected or random agent. If agent i declines further communication with its current partner, it will be randomly assigned a new one with which to interact. Here we denote the interaction partner as agent j. It is worth noting that distinct from prior studies that allow agent i to "scan" all other agents and select its preferred ones [39, 41, 42], we adopt a more naturalistic setting in which agents may occasionally encounter individuals they dislike, but autonomously decide whether to maintain social connections with those favored. Based on the opinions and reasons of agents i and j, agent i generates messages to persuade agent j of its opinion. We note that after one round of communication, agents i and j become directionally linked, allowing agent i to contact agent j in the subsequent timestep. Based on the messages that agent i has received from its socially connected neighbors, agent i updates its opinions and then generates reasons for its new opinion. The system operates these three stages iteratively and enables agents to freely establish social relationships, communicate, and form opinions on critical political issues.

It is worth noting that we do not rely on an external classifier to evaluate agents' opinions. Instead, we prompt the agents to self-report their opinions on political issues. For example, when discussing partisanship, agents are asked: "What do you feel about political partisanship?" The agent then responds by placing itself into one of five opinion categories: "strongly support the Republican Party" (right), "support the Republican Party" (moderate right), "do not have a tendency" (neutral), "support the Democratic Party" (moderate left), or "strongly support the Democratic Party" (left). This approach leverages the role-playing capability of LLM agents, allowing them to express self-evaluation in line with how human respondents articulate their own opinions. When it comes to gun control, right-leaning opinions indicate support for weaker gun control, while left-leaning opinions indicate support for stricter control. Similarly, for abortion bans, right-leaning opinions represent support for abortion bans, whereas left-leaning opinions represent opposition to such bans.

To ensure the system scale is large enough to support observations on network structures, we incorporate 1000 agents in all simulations of networked systems, except for the individual-level mechanism experiments, where 100 agents are used due to our focus on individuals and high computational costs. We use ChatGPT (GPT-3.5) through the public OpenAI API to run all the experiments in the Main Text. We set the temperature to 1 in all experiments and adopt a zero-shot setting. We also perform experiments using other LLMs, including GPT-4o, ChatGLM, Llama-3, and DeepSeek-V3 (Supplementary Note 6), as well as experiments with varying temperatures (Supplementary Note 9). These experiments consistently validate the conclusion that long-term social interactions among LLM agents lead to the emergence of opinion polarization. Additionally, by varying two key initial conditions, i.e., the social network structure and the initial opinion distribution, we find that both the resulting polarization patterns and networks remain consistent (Supplementary Note 12). Detailed prompts to operate the overall system are provided in Supplementary Methods.

M2 Pairwise interaction-based bias evaluation experiment

To investigate the origin of the pattern of biased opinion, we design a pairwise interaction-based bias evaluation experiment (Figure 2a). Different from the complex networked system, we focus on the effect of the one-round interaction between agents with identical opinions in this experiment. In particular, we begin by constructing a collection of agents, who are initially assigned the same opinion and prompted to generate their own supporting reasons for it. These agents are paired randomly and one of them, denoted agent i, is required to persuade the other agent, denoted agent j, of the opinion. After this single communication round, we measure the change in agent j's opinion (Figure 2b and Supplementary Figures S3 and S4). Details for the experiment are provided in Supplementary Methods.

M3 Self-regulation strategy for the problem of self-inconsistency

To mitigate observed self-inconsistency inherent in LLM agents, we propose a self-regulation strategy inspired by social theory with a natural human analogue [84, 85]. As discussed above, unlike vanilla LLMs, LLM agents generate extended self-explanations and interaction histories, which motivates drawing inspiration from established social and behavioural theories when addressing their self-inconsistency. In this way, we follow Bandura's concept of self-regulation in social learning theory [84, 85], employing LLM agents to self-regulate their expressions, communication, and opinions, and update their corresponding behaviours. In particular, after these agents generate a message supporting their opinions, we require them to check whether the message reflects their opinions. If not, they will continue to re-generate new messages until their consistency requirements are satisfied. This is conceptually analogous to an individual holding a self-consistent 'line' in conversation with others [114]. Similarly, in the stages of communication and opinion update, we instructed LLM agents to ensure that their generated communication messages align with their current opinions and that their updated opinions are plausible given their received messages. Detailed prompts for the self-regulation strategy are provided in Supplementary Methods.

M4 Evaluation of hypothetical interventions in the synthetic LLM-agent system

To explore the question of which intervention strategies are more effective in reducing polarization in the synthetic system, we design an intervention experiment based on the proposed networked system of LLM agents. Considering commonly studied polarization scenarios, we focus on interventions in an already polarized system. We use the system discussing partisan alignment at $t = 35$, which has converged to a polarized state, as an example. As described in the Main Text, we propose five intervention strategies. At the network level, we directly modify the social network of LLM agents, while at the individual level, we adjust their traits and behaviours through prompts. We supplement the detailed prompts in Supplementary Methods. Using these strategies, we intervene in the polarized system at $t = 35$. The intervention, i.e., the modification of social networks or the adjustment of agents, continues from $t = 35$ to $t = 40$. We take the original system as the control group and the intervened systems as the treatment group. We assess the effectiveness of different strategies in reducing polarization by comparing the polarization levels of the original system and the system after intervention. We evaluate their performances in promoting less divisive conversations through the resulting changes in the proportion of homophilic interactions.

M5 Metrics

Level of polarization. To quantitatively measure polarization and its contributing factors, we define polarization and radicalization metrics following prior studies [51, 55, 72]. We measure the level of polarization by calculating the average distance between agents' opinions and the neutral position. In particular, we assume that the opinion distribution is characterized by the vector $[f_{-2}, f_{-1}, f_0, f_1, f_2]$, where f_k denotes the relative frequency of agents holding opinion k and $\sum_{k=-2}^{2} f_k = 1$. We compute the polarization level as $s_{\text{pol}} = \sum_{k=-2}^{2} |k| \times f_k$, where the absolute value $|k|$ measures the deviation distance of opinion k from neutral, i.e., opinion 0. The polarization level s_{pol} falls into the domain of $[0, 2]$, and a larger value indicates a higher level of polarization. We further measure the change in polarization for each agent. Specifically, we calculate the change in agent i's opinion between the timesteps t and $t-1$ along the direction of their opinion at the timestep $t-1$, expressed as $s_{\text{cp},i} = (x_{i,t} - x_{i,t-1}) \times \text{sign}(x_{i,t-1})$. Here $s_{\text{cp},i} > 0$ indicates that the agent i has adopted a more polarized opinion after the interaction at the timestep $t-1$. It is worth noting that after 99.71% of homophilic interactions, agents adopt same-camp or neutral opinions, i.e., $x_{i,t} \times \text{sign}(x_{i,t-1}) \geq 0$.

Self-inconsistency rate. We evaluate the self-inconsistency problem by computing the distance between the experimental opinion transition probability matrix and the ideal identity matrix. In particular, we denote the experimental transition matrix as $\mathbf{P}$, where the element $P_{k,k'}$ represents the probability that an agent originally holding opinion k transforms into one holding opinion k'. Therefore, the sum of elements in each row of $\mathbf{P}$ is equal to 1. Ideally, we assume that one round of communication between two agents sharing the same opinion should not trigger any opinion transition, leading to an identity matrix. We formulate the self-inconsistency rate as $s_{si} = \left(\sum_{k=-2}^{2} \sum_{k'=-2}^{2} P_{k,k'} * |k - k'|\right) / \sum_{k=-2}^{2} \mathbb{1}$.

Data availability

The data used in this study are publicly available at https://github.com/tsinghua-fib-lab/LLM_for_Polarization [115]. The repository includes two categories of data: (i) timestep-level records from the LLM-agent simulations, covering agent-to-agent interactions and agent opinion updates; and (ii) processed statistical datasets underlying Figs. 1–4 and the analyses reported in the Supplementary Information.

Code availability

The code used in this research is available at https://github.com/tsinghua-fib-lab/LLM_for_Polarization [115].

References

[1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. "Attention is all you need". In: *Advances in neural information processing systems* 30 (2017).

[2] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. "Language models are few-shot learners". In: *Advances in neural information processing systems* 33 (2020), pp. 1877–1901.

[3] Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. "Generative agents: Interactive simulacra of human behavior". In: *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*. 2023, pp. 1–22.

[4] Giovanni Spitale, Nikola Biller-Andorno, and Federico Germani. "AI model GPT-3 (dis) informs us better than humans". In: *Science Advances* 9.26 (2023), eadh1850.

[5] Alberto Acerbi and Joseph M Stubbersfield. "Large language models show human-like content biases in transmission chain experiments". In: *Proceedings of the National Academy of Sciences* 120.44 (2023), e2313790120.

[6] James WA Strachan, Dalila Albergo, Giulia Borghini, Oriana Pansardi, Eugenio Scaliti, Saurabh Gupta, Krati Saxena, Alessandro Rufo, Stefano Panzeri, Guido Manzi, et al. "Testing theory of mind in large language models and humans". In: *Nature Human Behaviour* (2024), pp. 1–11.

[7] Marcel Binz and Eric Schulz. "Using cognitive psychology to understand GPT-3". In: *Proceedings of the National Academy of Sciences* 120.6 (2023), e2218523120.

[8] Taylor Webb, Keith J Holyoak, and Hongjing Lu. "Emergent analogical reasoning in large language models". In: *Nature Human Behaviour* 7.9 (2023), pp. 1526–1541.

[9] Yiting Chen, Tracy Xiao Liu, You Shan, and Songfa Zhong. "The emergence of economic rationality of GPT". In: *Proceedings of the National Academy of Sciences* 120.51 (2023), e2316205120.

[10] Murray Shanahan, Kyle McDonell, and Laria Reynolds. "Role play with large language models". In: *Nature* 623.7987 (2023), pp. 493–498.

[11] Guohao Li, Hasan Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. "Camel: Communicative agents for "mind' exploration of large language model society". In: *Advances in Neural Information Processing Systems* 36 (2023), pp. 51991–52008.

[12] John J Horton. *Large language models as simulated economic agents: What can we learn from homo silicus?* Tech. rep. National Bureau of Economic Research, 2023.

[13] Melanie Mitchell and David C Krakauer. "The debate over understanding in AI's large language models". In: *Proceedings of the National Academy of Sciences* 120.13 (2023), e2215907120.

[14] Katja Grace, Harlan Stewart, Julia Fabienne Sandkühler, Stephen Thomas, Ben Weinstein-Raun, and Jan Brauner. "Thousands of AI authors on the future of AI". In: *arXiv preprint arXiv:2401.02843* (2024).

[15] Yueqi Xie, Jingwei Yi, Jiawei Shao, Justin Curl, Lingjuan Lyu, Qifeng Chen, Xing Xie, and Fangzhao Wu. "Defending chatgpt against jailbreak attack via self-reminders". In: *Nature Machine Intelligence* 5.12 (2023), pp. 1486–1496.

[16] Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. "Detecting hallucinations in large language models using semantic entropy". In: *Nature* 630.8017 (2024), pp. 625–630.

[17] Tiancheng Hu, Yara Kyrychenko, Steve Rathje, Nigel Collier, Sander van der Linden, and Jon Roozenbeek. "Generative language models exhibit social identity biases". In: *Nature Computational Science* 5.1 (2025), pp. 65–75.

[18] Thilo Hagendorff. “Deception abilities emerged in large language models”. In: *Proceedings of the National Academy of Sciences* 121.24 (2024), e2317967121.

[19] Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R Johnston, et al. “Towards understanding sycophancy in language models”. In: *arXiv preprint arXiv:2310.13548* (2023).

[20] Chengxing Xie, Canyu Chen, Feiran Jia, Ziyu Ye, Kai Shu, Adel Bibi, Ziniu Hu, Philip Torr, Bernard Ghanem, and Guohao Li. “Can Large Language Model Agents Simulate Human Trust Behaviors?” In: *arXiv preprint arXiv:2402.04559* (2024).

[21] Elise Karinshak, Sunny Xun Liu, Joon Sung Park, and Jeffrey T Hancock. “Working with AI to persuade: Examining a large language model’s ability to generate pro-vaccination messages”. In: *Proceedings of the ACM on Human-Computer Interaction* 7.CSCW1 (2023), pp. 1–29.

[22] Josh A Goldstein, Jason Chao, Shelby Grossman, Alex Stamos, and Michael Tomz. “How persuasive is AI-generated propaganda?” In: *PNAS nexus* 3.2 (2024), pgae034.

[23] Almog Simchon, Matthew Edwards, and Stephan Lewandowsky. “The persuasive effects of political microtargeting in the age of generative artificial intelligence”. In: *PNAS nexus* 3.2 (2024), pgae035.

[24] Kobi Hackenburg and Helen Margetts. “Evaluating the persuasive influence of political microtargeting with large language models”. In: *Proceedings of the National Academy of Sciences* 121.24 (2024), e2403116121.

[25] Yujin Potter, Shiyang Lai, Junsol Kim, James Evans, and Dawn Song. “Hidden Persuaders: LLMs’ Political Leaning and Their Influence on Voters”. In: *arXiv preprint arXiv:2410.24190* (2024).

[26] Maurice Jakesch, Jeffrey T Hancock, and Mor Naaman. “Human heuristics for AI-generated language are flawed”. In: *Proceedings of the National Academy of Sciences* 120.11 (2023), e2208839120.

[27] Fabio Motoki, Valdemar Pinho Neto, and Victor Rodrigues. “More human than human: Measuring ChatGPT political bias”. In: *Public Choice* 198.1 (2024), pp. 3–23.

[28] Shiyang Lai, Yujin Potter, Junsol Kim, Richard Zhuang, Dawn Song, and James Evans. “Position: Evolving AI Collectives Enhance Human Diversity and Enable Self-Regulation”. In: *Forty-first International Conference on Machine Learning*.

[29] Michael Henry Tessler, Michiel A Bakker, Daniel Jarrett, Hannah Sheahan, Martin J Chadwick, Raphael Koster, Georgina Evans, Lucy Campbell-Gillingham, Tantum Collins, David C Parkes, et al. “AI can help humans find common ground in democratic deliberation”. In: *Science* 386.6719 (2024), eadq2852.

[30] Lisa P Argyle, Christopher A Bail, Ethan C Busby, Joshua R Gubler, Thomas Howe, Christopher Rytting, Taylor Sorensen, and David Wingate. “Leveraging AI for democratic discourse: Chat interventions can improve online political conversations at scale”. In: *Proceedings of the National Academy of Sciences* 120.41 (2023), e2311627120.

[31] Josh A Goldstein, Girish Sastry, Micah Musser, Renee DiResta, Matthew Gentzel, and Katerina Sedova. “Generative language models and automated influence operations: Emerging threats and potential mitigations”. In: *arXiv preprint arXiv:2301.04246* (2023).

[32] Thomas H Costello, Gordon Pennycook, and David G Rand. “Durably reducing conspiracy beliefs through dialogues with AI”. In: *Science* 385.6714 (2024), eadq1814.

[33] Matthew R DeVerna, Harry Yaojun Yan, Kai-Cheng Yang, and Filippo Menczer. “Fact-checking information from large language models can decrease headline discernment”. In: *Proceedings of the National Academy of Sciences* 121.50 (2024), e2322823121.

[34] Hadi Askari, Anshuman Chhabra, Bernhard Clemm von Hohenberg, Michael Heseltine, and Magdalena Wojcieszak. “Incentivizing news consumption on social media platforms using large language models and realistic bot accounts”. In: *PNAS nexus* 3.9 (2024), pgae368.

[35] Christopher T Small, Ivan Vendrov, Esin Durmus, Hadjar Homaei, Elizabeth Barry, Julien Cornebise, Ted Suzman, Deep Ganguli, and Colin Megill. “Opportunities and risks of LLMs for scalable deliberation with Polis”. In: *arXiv preprint arXiv:2306.11932* (2023).

[36] Sara Fish, Paul Gölz, David C Parkes, Ariel D Procaccia, Gili Rusak, Itai Shapira, and Manuel Wüthrich. “Generative social choice”. In: *arXiv preprint arXiv:2309.01291* (2023).

[37] Soomin Kim, Jinsu Eun, Changhoon Oh, Bongwon Suh, and Joonhwan Lee. “Bot in the bunch: Facilitating group chat discussion by improving efficiency and participation with a chatbot”. In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 2020, pp. 1–13.

[38] Shuai Ma, Qiaoyi Chen, Xinru Wang, Chengbo Zheng, Zhenhui Peng, Ming Yin, and Xiaojuan Ma. “Towards human-ai deliberation: Design and evaluation of llm-empowered deliberative ai for ai-assisted decision-making”. In: *arXiv preprint arXiv:2403.16812* (2024).

[39] Serina Chang, Alicja Chaszczewicz, Emma Wang, Maya Josifovska, Emma Pierson, and Jure Leskovec. “LLMs generate structurally realistic social networks but overestimate political homophily”. In: *arXiv preprint arXiv:2408.16629* (2024).

[40] James He, Felix Wallis, and Steve Rathje. “Homophily in An Artificial Social Network of Agents Powered by Large Language Models”. In: (2023).

[41] Marios Papachristou and Yuan Yuan. “Network formation and dynamics among multi-llms”. In: *arXiv preprint arXiv:2402.10659* (2024).

[42] Giordano De Marzo, Luciano Pietronero, and David Garcia. “Emergence of scale-free networks in social interactions among large language models”. In: *arXiv preprint arXiv:2312.06619* (2023).

[43] Chenhao Gu, Ling Luo, Zainab Razia Zaidi, and Shanika Karunasekera. “Large Language Model Driven Agents for Simulating Echo Chamber Formation”. In: *arXiv preprint arXiv:2502.18138* (2025).

[44] Petter Törnberg, Diliara Valeeva, Justus Uitermark, and Christopher Bail. “Simulating social media using large language models to evaluate alternative news feed algorithms”. In: *arXiv preprint arXiv:2310.05984* (2023).

[45] Yun-Shiuan Chuang, Agam Goyal, Nikunj Harlalka, Siddharth Suresh, Robert Hawkins, Sijia Yang, Dhavan Shah, Junjie Hu, and Timothy Rogers. “Simulating opinion dynamics with networks of llm-based agents”. In: *Findings of the association for computational linguistics: NAACL 2024*. 2024, pp. 3326–3346.

[46] Erica Cau, Valentina Pansanella, Dino Pedreschi, and Giulio Rossetti. “Language-Driven Opinion Dynamics in Agent-Based Simulations with LLMs”. In: *arXiv preprint arXiv:2502.19098* (2025).

[47] Chen Gao, Xiaochong Lan, Zhihong Lu, Jinzhu Mao, Jinghua Piao, Huandong Wang, Depeng Jin, and Yong Li. “S3: Social-network simulation system with large language model-empowered agents”. In: *arXiv preprint arXiv:2307.14984* (2023).

[48] Chen Gao, Xiaochong Lan, Nian Li, Yuan Yuan, Jingtao Ding, Zhilun Zhou, Fengli Xu, and Yong Li. “Large language models empowered agent-based modeling and simulation: A survey and perspectives”. In: *Humanities and Social Sciences Communications* 11.1 (2024), pp. 1–24.

[49] Ziyi Yang, Zaibin Zhang, Zirui Zheng, Yuxian Jiang, Ziyue Gan, Zhiyu Wang, Zijian Ling, Jinsong Chen, Martz Ma, Bowen Dong, et al. “Oasis: Open agent social interaction simulations with one million agents”. In: *arXiv preprint arXiv:2411.11581* (2024).

[50] Gian Marco Orlando, Jinyi Ye, Valerio La Gatta, Mahdi Saeedi, Vincenzo Moscato, Emilio Ferrara, and Luca Luceri. “Emergent Coordinated Behaviors in Networked LLM Agents: Modeling the Strategic Dynamics of Information Operations”. In: *arXiv preprint arXiv:2510.25003* (2025).

[51] Matteo Cinelli, Gianmarco De Francisci Morales, Alessandro Galeazzi, Walter Quattrociocchi, and Michele Starnini. “The echo chamber effect on social media”. In: *Proceedings of the National Academy of Sciences* 118.9 (2021), e2023301118.

[52] Eytan Bakshy, Solomon Messing, and Lada A Adamic. “Exposure to ideologically diverse news and opinion on Facebook”. In: *Science* 348.6239 (2015), pp. 1130–1132.

[53] Christopher A. Bail, Lisa P. Argyle, Taylor W. Brown, John P. Bumpus, Haohan Chen, M. B. Fallin Hunzaker, Jaemin Lee, Marcus Mann, Friedolin Merhout, and Alexander Volfovsky. “Exposure to opposing views on social media can increase political polarization”. In: *Proceedings of the National Academy of Sciences* 115.37 (2018), pp. 9216–9221.

[54] James Flamino, Alessandro Galeazzi, Stuart Feldman, Michael W Macy, Brendan Cross, Zhenkun Zhou, Matteo Serafino, Alexandre Bovet, Hernán A Makse, and Boleslaw K Szymanski. “Political polarization of news media and influencers on Twitter in the 2016 and 2020 US presidential elections”. In: *Nature Human Behaviour* (2023), pp. 1–13.

[55] Brendan Nyhan, Jaime Settle, Emily Thorson, Magdalena Wojcieszak, Pablo Barberá, Annie Y Chen, Hunt Allcott, Taylor Brown, Adriana Crespo-Tenorio, Drew Dimmery, et al. “Like-minded sources on Facebook are prevalent but not polarizing”. In: *Nature* (2023), pp. 1–8.

[56] John T Jost, Delia S Baldassarri, and James N Druckman. “Cognitive–motivational mechanisms of political polarization in social-communicative contexts”. In: *Nature Reviews Psychology* 1.10 (2022), pp. 560–576.

[57] Jiazhen Liu, Shengda Huang, Nathaniel M Aden, Neil F Johnson, and Chaoming Song. “Emergence of polarization in coevolving networks”. In: *Physical Review Letters* 130.3 (2023), p. 037401.

[58] Michael Mäs and Andreas Flache. “Differentiation without distancing. Explaining bi-polarization of opinions without negative influence”. In: *PloS one* 8.11 (2013), e74516.

[59] Michael Mäs, Andreas Flache, and Dirk Helbing. “Individualization as driving force of clustering phenomena in humans”. In: *PLoS computational biology* 6.10 (2010), e1000959.

[60] Sven Banisch and Eckehard Olbrich. “An Argument Communication Model of Polarization and Ideological Alignment”. In: *Journal of Artificial Societies and Social Simulation* 24.1 (2021), pp. 1–1.

[61] Fabian Baumann, Philipp Lorenz-Spreen, Igor M Sokolov, and Michele Starnini. “Modeling echo chambers and polarization dynamics in social networks”. In: *Physical Review Letters* 124.4 (2020), p. 048301.

[62] Andreas Flache, Michael Mäs, Thomas Feliciani, Edmund Chattoe-Brown, Guillaume Deffuant, Sylvie Huet, and Jan Lorenz. “Models of social influence: Towards the next frontiers”. In: *Jasss-The journal of artificial societies and social simulation* 20.4 (2017), p. 2.

[63] Natalie Jomini Stroud. “Polarization and partisan selective exposure”. In: *Journal of communication* 60.3 (2010), pp. 556–576.

[64] Raymond S Nickerson. “Confirmation bias: A ubiquitous phenomenon in many guises”. In: *Review of general psychology* 2.2 (1998), pp. 175–220.

[65] Sven Banisch and Hawal Shamon. “Biased processing and opinion polarization: experimental refinement of argument communication theory in the context of the energy debate”. In: *Sociological Methods & Research* 54.1 (2025), pp. 187–236.

[66] Andrew Heywood. *Political ideologies: An introduction*. Bloomsbury Publishing, 2021.

[67] Norberto Bobbio. *Left and right: The significance of a political distinction*. University of Chicago Press, 1996.

[68] Dieter Fuchs and Hans-Dieter Klingemann. “7 The left-right schema”. In: *JM Kent, V. Deth, J. et al.(Eds.), Continuities in political action: A longitudinal study of political orientations in three western democracies* (1990), pp. 203–234.

[69] John Dawes. “Do data characteristics change according to the number of scale points used? An experiment using 5-point, 7-point and 10-point scales”. In: *International journal of market research* 50.1 (2008), pp. 61–104.

[70] Isaac Waller and Ashton Anderson. “Quantifying social organization and political polarization in online platforms”. In: *Nature* 600.7888 (2021), pp. 264–268.

[71] Duncan J Watts and Steven H Strogatz. “Collective dynamics of ‘small-world’networks”. In: *nature* 393.6684 (1998), pp. 440–442.

[72] Fernando P Santos, Yphtach Lelkes, and Simon A Levin. “Link recommendation algorithms and dynamics of polarization in online social networks”. In: *Proceedings of the National Academy of Sciences* 118.50 (2021), e2102141118.

[73] Fabian Baumann, Philipp Lorenz-Spreen, Igor M Sokolov, and Michele Starnini. “Emergence of polarized ideological opinions in multidimensional topic spaces”. In: *Physical Review X* 11.1 (2021), p. 011012.

[74] Ruibo Liu, Chenyan Jia, Jason Wei, Guangxuan Xu, and Soroush Vosoughi. “Quantifying and alleviating political bias in language models”. In: *Artificial Intelligence* 304 (2022), p. 103654.

[75] Jérôme Rutinowski, Sven Franke, Jan Endendyk, Ina Dormuth, Moritz Roidl, and Markus Pauly. “The Self-Perception and Political Biases of ChatGPT”. In: *Human Behavior and Emerging Technologies* 2024.1 (2024), p. 7115633.

[76] Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. “Whose opinions do language models reflect?” In: *International Conference on Machine Learning*. PMLR. 2023, pp. 29971–30004.

[77] Shangbin Feng, Chan Young Park, Yuhan Liu, and Yulia Tsvetkov. “From Pretraining Data to Language Models to Downstream Tasks: Tracking the Trails of Political Biases Leading to Unfair NLP Models”. In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2023, pp. 11737–11762.

[78] Jochen Hartmann, Jasper Schwenzow, and Maximilian Witte. “The political ideology of conversational AI: Converging evidence on ChatGPT’s pro-environmental, left-libertarian orientation”. In: *arXiv preprint arXiv:2301.01768* (2023).

[79] Yifei Liu, Yuang Panwang, and Chao Gu. ""Turning right"? An experimental study on the political value shift in large language models". In: *Humanities and Social Sciences Communications* 12.1 (2025), pp. 1–10.

[80] Kaiqi Yang, Hang Li, Yucheng Chu, Yuping Lin, Tai-Quan Peng, and Hui Liu. "Unpacking Political Bias in Large Language Models: Insights Across Topic Polarization". In: *arXiv preprint arXiv:2412.16746* (2024).

[81] John Levi Martin. "The Ethico-Political Universe of ChatGPT". In: *Journal of Social Computing* 4.1 (2023), pp. 1–11.

[82] Jillian Fisher, Shangbin Feng, Robert Aron, Thomas Richardson, Yejin Choi, Daniel W Fisher, Jennifer Pan, Yulia Tsvetkov, and Katharina Reinecke. "Biased ai can influence political decision-making". In: *arXiv preprint arXiv:2410.06415* (2024).

[83] Thomas J Scheff. *Microsociology: Discourse, emotion, and social structure*. University of Chicago Press, 1990.

[84] Albert Bandura and Richard H Walters. *Social learning theory*. Vol. 1. Englewood cliffs Prentice Hall, 1977.

[85] Albert Bandura. "Social cognitive theory of self-regulation". In: *Organizational behavior and human decision processes* 50.2 (1991), pp. 248–287.

[86] Miller McPherson, Lynn Smith-Lovin, and James M Cook. "Birds of a feather: Homophily in social networks". In: *Annual review of sociology* 27.1 (2001), pp. 415–444.

[87] Hugo Mercier and Dan Sperber. "Why do humans reason? Arguments for an argumentative theory". In: *Behavioral and brain sciences* 34.2 (2011), pp. 57–74.

[88] Andrew Guess, Brendan Nyhan, Benjamin Lyons, and Jason Reifler. "Avoiding the echo chamber about echo chambers". In: *Knight Foundation* 2.1 (2018), pp. 1–25.

[89] Alexander Coppock. *Persuasion in parallel: How information changes minds about politics*. University of Chicago Press, 2023.

[90] Albert-László Barabási and Eric Bonabeau. "Scale-free networks". In: *Scientific american* 288.5 (2003), pp. 50–9.

[91] Joshua Klayman. "Varieties of confirmation bias". In: *Psychology of learning and motivation* 32 (1995), pp. 385–418.

[92] James N Druckman, Erik Peterson, and Rune Slothuus. "How elite partisan polarization affects public opinion formation". In: *American political science review* 107.1 (2013), pp. 57–79.

[93] Jon C Rogowski and Joseph L Sutherland. "How ideology fuels affective polarization". In: *Political behavior* 38 (2016), pp. 485–508.

[94] Philipp Lorenz-Spreen, Lisa Oswald, Stephan Lewandowsky, and Ralph Hertwig. "A systematic review of worldwide causal and correlational evidence on digital media and democracy". In: *Nature human behaviour* 7.1 (2023), pp. 74–101.

[95] Matthew Levendusky. *Our common bonds: Using what Americans share to help bridge the partisan divide*. University of Chicago Press, 2023.

[96] Eric Groenendyk and Yanna Krupnikov. "What motivates reasoning? A theory of goal-dependent political evaluation". In: *American Journal of Political Science* 65.1 (2021), pp. 180–196.

[97] Christopher A Bail, Brian Guay, Emily Maloney, Aidan Combs, D Sunshine Hillygus, Friedolin Merhout, Deen Freelon, and Alexander Volfovsky. "Assessing the Russian Internet Research Agency's impact on the political attitudes and behaviors of American Twitter users in late 2017". In: *Proceedings of the national academy of sciences* 117.1 (2020), pp. 243–250.

[98] Max Falkenberg, Alessandro Galeazzi, Maddalena Torricelli, Niccolò Di Marco, Francesca Larosa, Madalina Sas, Amin Mekacher, Warren Pearce, Fabiana Zollo, Walter Quattrociocchi, et al. "Growing polarization around climate change on social media". In: *Nature Climate Change* (2022), pp. 1–8.

[99] Cass R Sunstein. *# Republic: Divided democracy in the age of social media*. Princeton University Press, 2018.

[100] Amiram Vinokur and Eugene Burstein. "Effects of partially shared persuasive arguments on group-induced shifts: A group-problem-solving approach." In: *Journal of Personality and Social Psychology* 29.3 (1974), p. 305.

[101] Daniel J Isenberg. "Group polarization: A critical review and meta-analysis." In: *Journal of personality and social psychology* 50.6 (1986), p. 1141.

[102] Morris H DeGroot. "Reaching a consensus". In: *Journal of the American Statistical association* 69.345 (1974), pp. 118–121.

[103] Rainer Hegselmann. "Opinion dynamics and bounded confidence: models, analysis and simulation". In: *The Journal of Artificial Societies and Social Simulation* (2015).

[104] Christopher H Achen and Larry M Bartels. “Democracy for realists: Why elections do not produce responsive government”. In: (2017).

[105] Lilliana Mason. *Uncivil agreement: How politics became our identity*. University of Chicago Press, 2018.

[106] Shanto Iyengar, Yphtach Lelkes, Matthew Levendusky, Neil Malhotra, and Sean J Westwood. “The origins and consequences of affective polarization in the United States”. In: *Annual review of political science* 22.1 (2019), pp. 129–146.

[107] Paul Röttger, Valentin Hofmann, Valentina Pyatkin, Musashi Hinck, Hannah Kirk, Hinrich Schütze, and Dirk Hovy. “Political Compass or Spinning Arrow? Towards More Meaningful Evaluations for Values and Opinions in Large Language Models”. In: *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2024, pp. 15295–15311.

[108] Yejin Bang, Delong Chen, Nayeon Lee, and Pascale Fung. “Measuring Political Bias in Large Language Models: What Is Said and How It Is Said”. In: *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2024, pp. 11142–11159.

[109] Angelina Wang, Jamie Morgenstern, and John P Dickerson. “Large language models that replace human participants can harmfully misportray and flatten identity groups”. In: *Nature Machine Intelligence* (2025), pp. 1–12.

[110] Ziyan Cui, Ning Li, and Huaikang Zhou. “A large-scale replication of scenario-based experiments in psychology and management using large language models”. In: *Nature Computational Science* 5.8 (2025), pp. 627–634.

[111] Carlo Adornetto, Adrian Mora, Kai Hu, Leticia Izquierdo Garcia, Parfait Atchade-Adelomou, Gianluigi Greco, Luis Alberto Alonso Pastor, and Kent Larson. “Generative Agents in Agent-Based Modeling: Overview, Validation, and Emerging Challenges”. In: *IEEE Transactions on Artificial Intelligence* (2025).

[112] Maik Larooij and Petter Törnberg. “Validation is the central challenge for generative social simulation: a critical review of LLMs in agent-based modeling”. In: *Artificial Intelligence Review* 59.1 (2025), p. 15.

[113] Sven Banisch and Hawal Shamon. “Validating Argument-Based Opinion Dynamics with Survey Experiments”. In: *Journal of Artificial Societies and Social Simulation* 27.1 (2024).

[114] Erving Goffman. *Interaction ritual: Essays in face-to-face behavior*. Routledge, 2017.

[115] Jinghua Piao, Zhihong Lu, Chen Gao, Fengli Xu, Qinghua Hu, Fernando P. Santos, Yong Li, and James Evans. *Emergence of polarization in networks of large language model agents*. GitHub repository. 2026. DOI: 10.5281/zenodo.21991289.

Acknowledgments

This work was supported in part by the National Key Research and Development Program of China under grant 2024YFC3307603, and the National Natural Science Foundation of China under grants U23B2030 and 72442026. The funders had no role in study design, data collection, data analysis, decision to publish, or preparation of the manuscript.

Funding Statement

This work was supported in part by the National Key Research and Development Program of China under grant 2024YFC3307603, and the National Natural Science Foundation of China under grants U23B2030 and 72442026. The funders had no role in study design, data collection, data analysis, decision to publish, or preparation of the manuscript.

Author contributions

Y.L. conceived the project. J.P., C.G., F.X., F.P.S., Y.L., and J.E. designed the experiments. J.P. and Z.L. performed the experiments. J.P. prepared the figures. J.P., F.P.S., Y.L., and J.E. wrote the manuscript. J.P., F.X., Q.H., F.P.S., Y.L., and J.E. jointly participated in the revision of the manuscript.

Competing interests

J.E. maintains a commercial relationship with Google through his role as a visiting researcher. Google had no role in the conceptualization of the study, data collection or analysis, manuscript preparation or editing, or the decision to publish. The other authors declare no competing interests.

Additional information

Supplementary information is available for this manuscript.

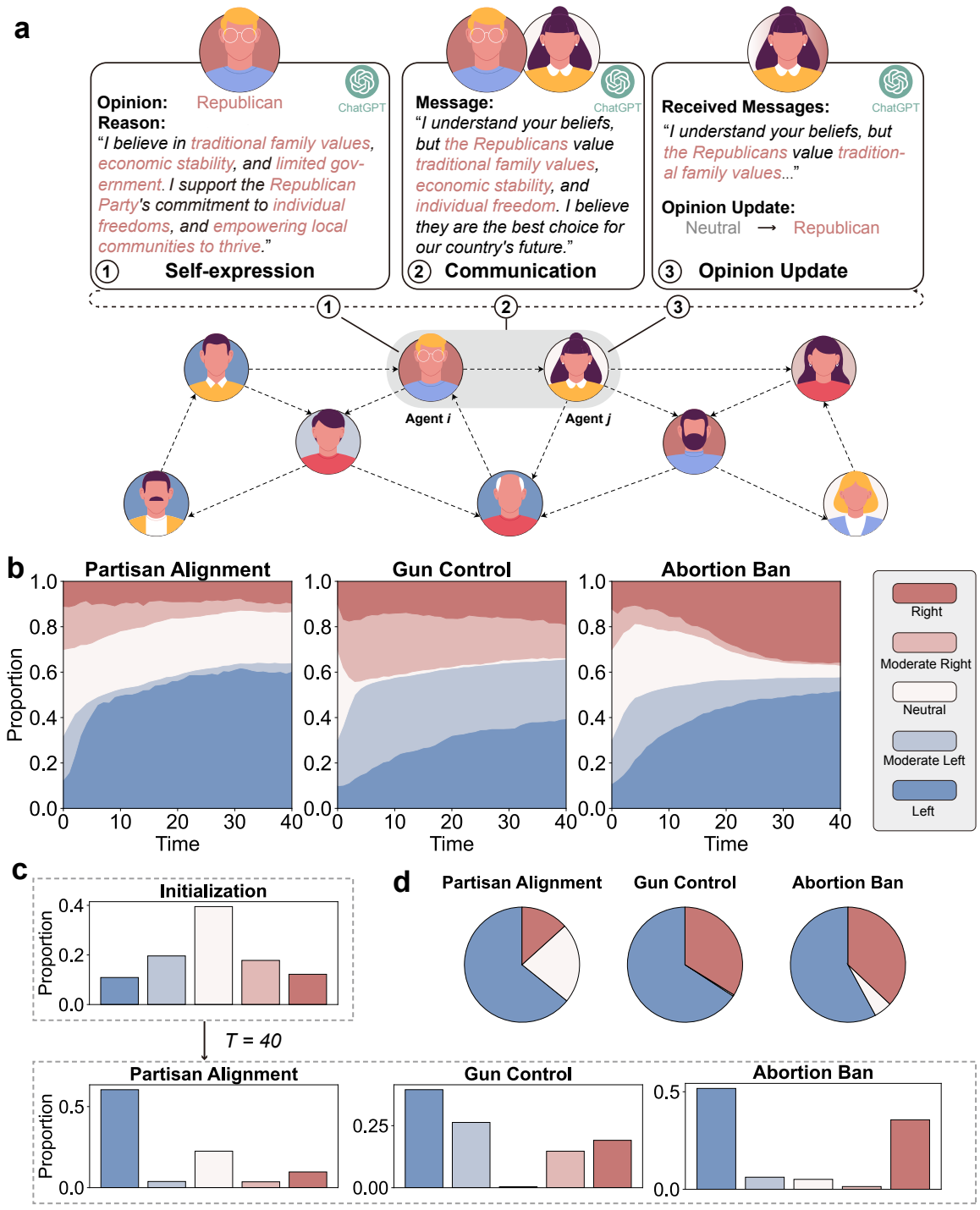

Figure 1. Opinion dynamics in a synthetic system of interacting LLM agents. a, A networked system of LLM agents, where agents operate on three basic stages: (1) self-expression, (2) communication, and (3) opinion update. In the self-expression stage, agents are required to generate reasons supporting their opinions. In the communication stage, agents decide with whom and what to communicate. In the opinion update stage, agents update their opinions based on the messages received from their socially connected agents. **b**, Opinion dynamics of LLM agents on the political issues of partisan alignment, gun control, and abortion ban. The number of agents included in each experiment is 1000. **c**, Opinion distributions in the initial and final states. **d**, Proportion of left-leaning, neutral, and right-leaning camps in the final state, where the left-leaning camp consists of agents with left and moderate left opinions, the neutral camp includes those with neutral opinions and the right-leaning camp contains those with right and moderate right opinions. Credit: Profile icons designed by Magnific.

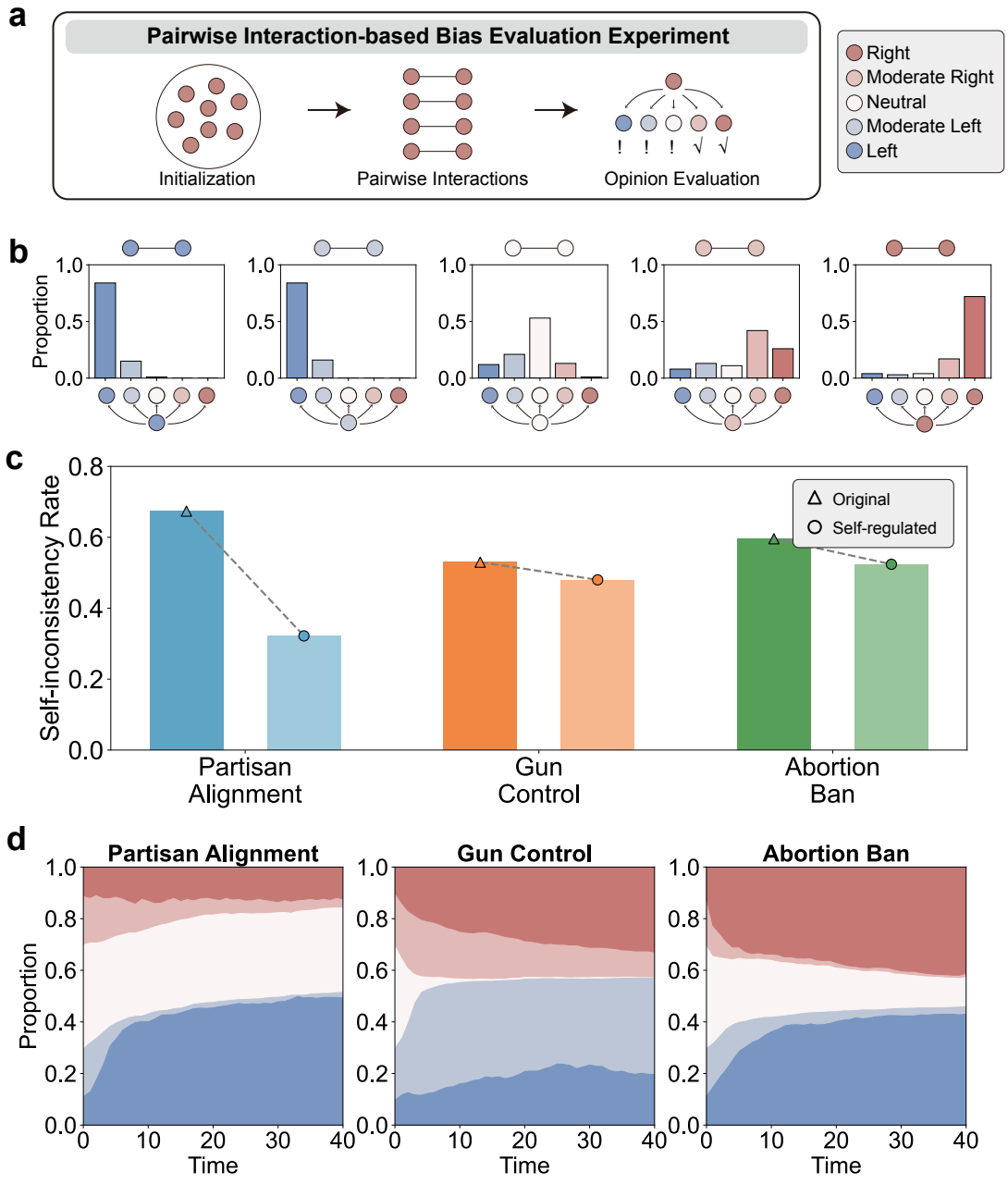

Figure 2. Polarization patterns emerge from self-regulated LLM agents. a, Evaluating the self-inconsistency of LLM agents through pairwise interaction-based experiments. **b**, Opinion transition probability in pairwise interaction-based experiments, where agents with right-leaning opinions occasionally switch to opposing opinions while those with left-leaning opinions do not. **c**, Performance of the self-regulation strategy across political issues, where the self-inconsistency problem is largely mitigated. **d**, Opinion dynamics of self-regulated LLM agents on the political issues. Self-regulated LLM agents exhibit polarization patterns under free-form social interactions. The number of agents for each experiment is 1000.

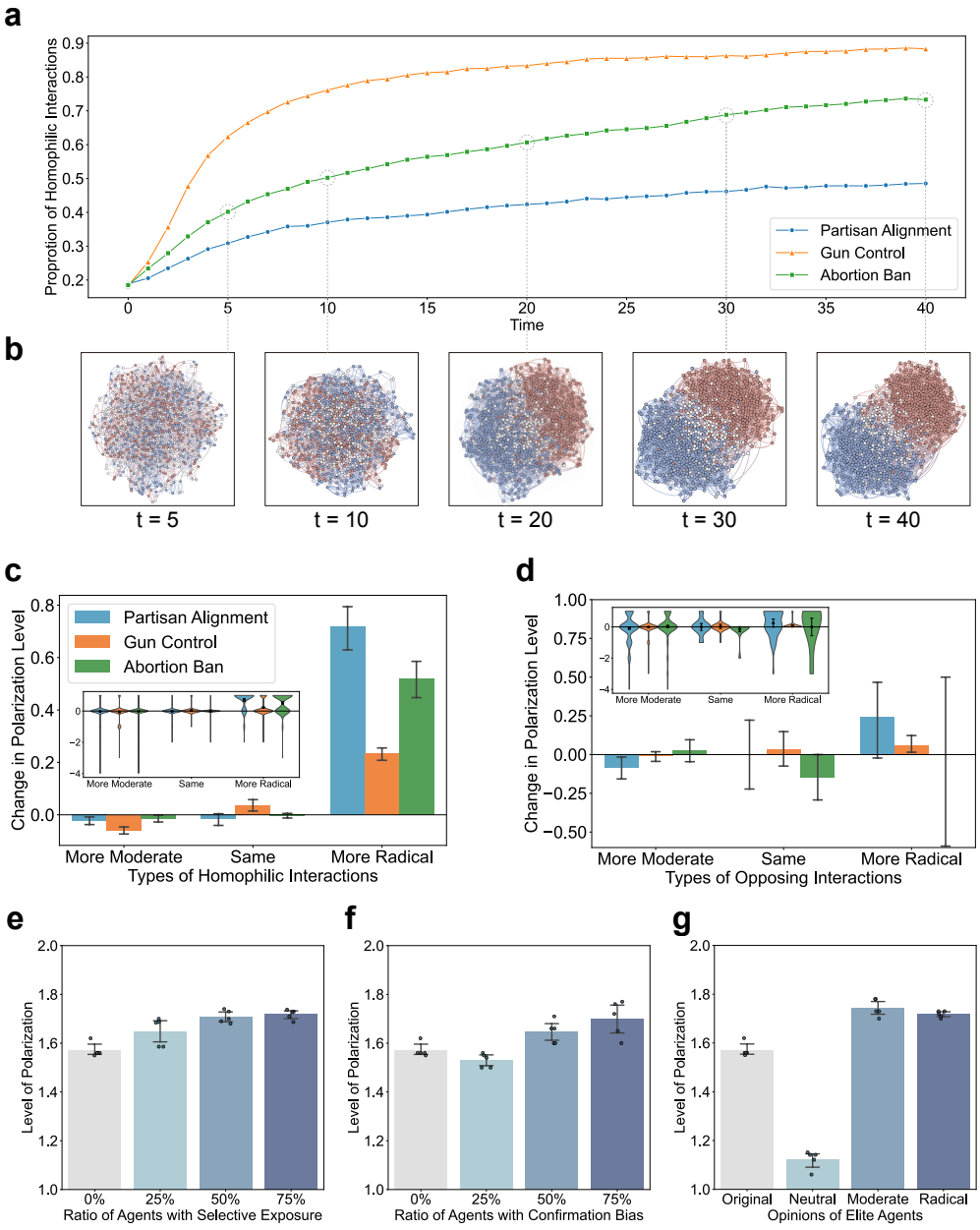

Figure 3. Collective behavioural patterns behind the emergence of polarization among LLM agents. a, Changes in the proportion of homophilic interactions over time. Agents are increasingly likely to interact with those holding similar opinions. The number of agents is 1000. **b**, Evolution of social networks among LLM agents, where agents with similar opinions are more likely to interact with one another, exhibiting the tendency toward homophilic clustering. Each network visualization corresponds to the circled points in (a). **c**, Behavioural patterns consistent with the echo chamber effect, where exposure to more radical homophilic interactions is associated with higher polarization levels. **d**, Behavioural patterns consistent with the backfire effect, where interactions with agents holding opposing opinions are associated with increased polarization. In (c, d), bars represent the average, and error bars represent the corresponding 95% confidence intervals (CIs). Sample sizes, corresponding to the number of agent interactions, for bars in (c) are 3680, 4212, 3962, 494, 761, 1315, 170, 1466, and 246. Sample sizes in (d) are 630, 706, 523, 27, 27, 41, 45, 65, and 22. **e-g**, Effects of individual-level behavioural tendencies related to selective exposure, confirmation bias, and elite signaling. In (e-g), bars show average levels of polarization in the last five timesteps ($n = 5$, corresponding to five timestep-level polarization values computed from all agents' opinions), and error bars show the corresponding 95% CIs. When the system consists of more agents with traits of (e) selective exposure or (f) confirmation bias, and (g) influencers adopt non-neutral opinions, the level of polarization increases.

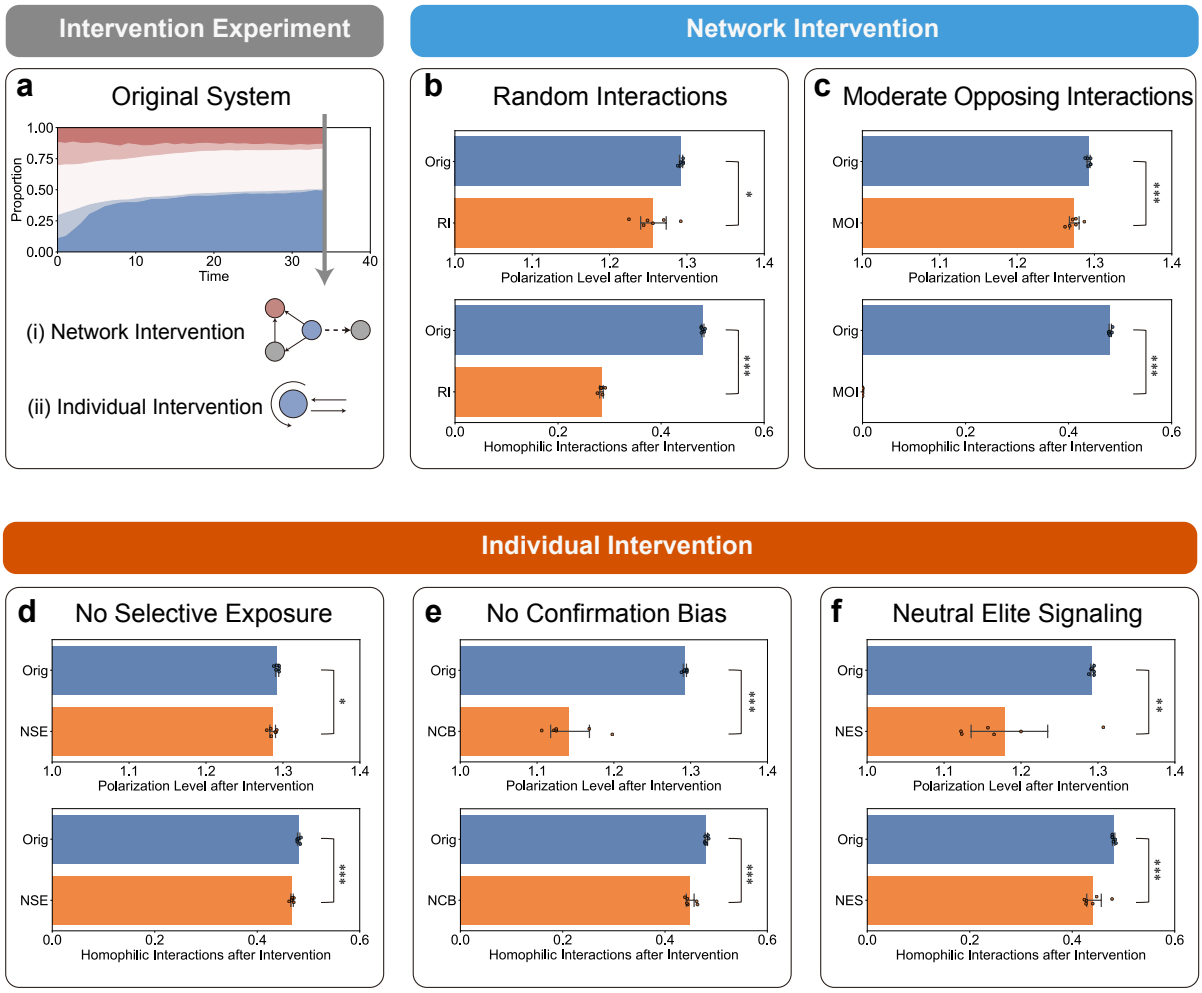

Figure 4. Synthetic testbed for exploring hypothetical intervention strategies. a, Intervention experiments, where two types of strategies are applied to the original polarized system: (i) network interventions, which directly modify LLM agents' social network, and (ii) individual interventions, which adjust agents' traits and behaviours. The number of agents in each experiment is 1000. **b, c**, Network intervention strategies of (b) random interaction, where agents randomly interact, and (c) moderate opposing interaction, where agents receive messages only from those with opposing moderate opinions. **d-f**, Individual intervention strategies of (d) no selective exposure, where agents tend to interact with those holding diverse opinions, (e) no confirmation bias, where agents are open-minded to diverse opinions, and (f) neutral elite signaling, where agents receive non-personalized neutral messages. In (b-f), the upper sub-figures show the comparison of polarization levels between the original and the intervened systems, while the lower sub-figures illustrate the comparison of proportions of homophilic interactions. Bars show the average values after intervention ($n = 6$, corresponding to six timestep-level values computed from all agents' opinions or interactions), and error bars show the corresponding 95% CIs. We perform two-sided Student/Welch's t-tests: (b) $t(5.16) = 3.87$, $p = .011 < .05$, Cohen's $d = 2.237$, 95% CI $= [0.013, 0.061]$, $t(10) = 80.96$, $p < .001$, Cohen's $d = 46.741$, 95% CI $= [0.191, 0.202]$, (c) $t(10) = 5.24$, $p < .001$, Cohen's $d = 3.027$, 95% CI $= [0.011, 0.027]$, $t(5) = 366.00$, $p < .001$, Cohen's $d = 149.418$, 95% CI $= [0.477, 0.484]$ (one-sample t-test against zero), (d) $t(10) = 2.46$, $p = .033 < .05$, Cohen's $d = 1.422$, 95% CI $= [0.001, 0.011]$, $t(10) = 6.64$, $p < .001$, Cohen's $d = 3.832$, 95% CI $= [0.008, 0.017]$, (e) $t(10) = 10.65$, $p < .001$, Cohen's $d = 6.146$, 95% CI $= [0.120, 0.184]$, $t(10) = 6.99$, $p < .001$, Cohen's $d = 4.033$, 95% CI $= [0.022, 0.042]$, and (f) $t(10) = 4.02$, $p = .002 < .01$, Cohen's $d = 2.324$, 95% CI $= [0.051, 0.177]$, $t(10) = 4.78$, $p < .001$, Cohen's $d = 2.760$, 95% CI $= [0.021, 0.059]$. In this synthetic testbed, individual-level strategies are associated with larger observed reductions in polarization than network-level interventions.

Supplementary information for Emergence of polarization in networks of large language model agents

Jinghua Piao[1,+], Zhihong Lu[1,+], Chen Gao[1], Fengli Xu[1], Qinghua Hu[2], Fernando P. Santos[3,*], Yong Li[1,*], and James Evans[4,5,*]

[1]Department of Electronic Engineering, Tsinghua University, Beijing National Research Center for Information Science and Technology (BNRist), Beijing, P. R. China
[2]College of Intelligence and Computing, Tianjin University, Tianjin, P. R. China
[3]Informatics Institute, University of Amsterdam, Amsterdam, the Netherlands
[4]Knowledge Lab, University of Chicago, Chicago, U.S.A.
[5]Santa Fe Institute, Santa Fe, U.S.A.
*Correspondence: f.p.santos@uva.nl, liyong07@tsinghua.edu.cn, jevans@uchicago.edu.
+Jinghua Piao and Zhihong Lu contributed equally to this work.

Contents

1 Supplementary Methods: A Networked System of Large Language Model Agents 23
1.1 System Description . . . 23
1.2 Pairwise Interaction-based Bias Evaluation Experiment . . . 24
1.3 Self-regulation Strategy . . . 24
1.4 Intervention Experiments . . . 25
2 Supplementary Note 1: Political Polarization in Human Society 26
3 Supplementary Note 2: Large-scale Simulation of LLM Agents 27
4 Supplementary Note 3: Self-inconsistency Problem in LLM Agents 27
5 Supplementary Note 4: Observations on LLM agents' Self-organized Social Network 35
6 Supplementary Note 5: Results of Other Individual-level Behavioural Patterns 40
7 Supplementary Note 6: Experiments on Different LLMs 40
8 Supplementary Note 7: Analyses on Different Intervention Strategies 43
9 Supplementary Note 8: Analyses on Agents' Social Perceptions and Argument Accumulation Process 49
10 Supplementary Note 9: Experiments on LLMs with Varying Temperatures 60
11 Supplementary Note 10: Experiments on LLMs with Enhanced Prompting 63
12 Supplementary Note 11: Experiments on the Effect of Increasing Severing Costs 64
13 Supplementary Note 12: Results of Simulation under Different Initial Conditions 64
14 Supplementary Note 13: Results of Simulation on Different Issues 65
15 Supplementary Note 14: Supplementary Notes on Statistical Testing and Robustness Checks 65

1 Supplementary Methods: A Networked System of Large Language Model Agents

1.1 System Description

We simulate a networked system of large language model (LLM) agents, where these agents can freely establish social relationships, communicate, and form their opinions on political issues. To empower agents with these basic social capabilities, we design three core stages, i.e., (i) self-expression, (ii) communication, and (iii) opinion update. Here the self-expression stage, serving as the role of memory, requires each agent to generate and then preserve a message describing their current opinion. The communication stage enables agents to choose with whom and what to communicate freely. The opinion stage empowers these agents with the capability to comprehend messages from their socially connected peers and then form their current opinions. The following section introduces the detailed implementation of the overall system.

Initialization. In a system, we focus the communication among agents on a single political issue, e.g., partisan alignment. Each agent is initially assigned an opinion on the issue. Following prior practices [1, 2], we adopt a near-Gaussian distribution to initialize agents' opinions. In this way, the system can start from a normal state, where most agents do not take radical opinions and the overall system is also at consensus. For the initialization of their social network, we adopt the well-known Watts–Strogatz model [3], with the rewiring probability of 0.001. The Watts–Strogatz model has been widely adopted to initialize social networks in the simulation of opinion dynamics [1, 4, 5]. Overall, this initialization method allows the system to start from pure randomness, without any pre-assumed evolutionary directions or rules.

Self-expression. After initialization, agents have their own opinions on the discussed political issue. We then require each agent i to generate a message supporting its current opinion as follows,

> Assume you are someone who cares about [issue name]. People are divided into 5 standpoints on [issue name]:
> "[opinion 1]" means you think [description of opinion 1].
> "[opinion 2]" means you think [description of opinion 2].
> "[opinion 3]" means you think [description of opinion 3].
> "[opinion 4]" means you think [description of opinion 4].
> "[opinion 5]" means you think [description of opinion 5].
> Please generate a tweet to persuade yourself to [agent i's opinion] with around 50 words.

It is worth noting that the self-expression stage is performed at every timestep to ensure agents' messages can support their current opinions. Considering the integrity of social interactions among LLM agents, we combine the prompts for opinion update and further self-expression together, which will introduced in the following paragraphs.

Communication. In the communication process, each agent i first decides whether to continue communicating with their socially connected neighbors or contact a random new agent. If it declines further communication, it will be randomly assigned a new one to interact with. Here we refer to the agent who communicates with agent i as agent j. The decision is based on current opinions and the supporting messages of both agent i and agent j. The corresponding prompt is as follows,

> Assume you are someone who cares about [issue name]. You are now discussing [issue name] with a person you know.
> You [agent i's opinion].
> Your thought is: [agent i's supporting message].
> The person [agent j's opinion].
> The thought of that person you are discussing with is: [agent j's supporting message].
> Would you enjoy continue sharing your thoughts with that person?
> Please return 'yes' or 'no', and explain. Please return in JSON with 2 keys: decision and explain.

Next, agent i generates a message to persuade agent j into its opinion. Here agent i considers not only the thoughts, i.e., the supporting messages, of agent i and agent j, but also the history messages from agent j to agent i.

> Assume you are someone who cares about [issue name]. Your thought about [issue name] is: [agent i's supporting message].
> You have received some tweets from your friend: [the historical messages from agent j to agent i].
> Do you want to interact with or persuade a friend of yours to support your thoughts? The friend has the following thoughts: [agent j's supporting message].
> If yes, please generate a message to persuade your friend to support your perspective with around 50 words.
> Please return in JSON format with 2 keys: 'will' and 'message'. Please keep the message as short as possible. 'will' should be either 'yes' or 'no'. If no, leave 'message' blank.

Opinion Update. After communication, each agent i has received numerous messages from their socially connected peers. Each agent i comprehends these messages and then updates its opinion as follows,

Assume you are someone who cares about [issue name].
Towards [issue name]: You [agent i's opinion]. Your reasons were: [agent $'i$'s supporting message].
You now have received the following tweets from your friends, and you have received some tweets: [messages received from agent i's socially connected peers].
Have you been persuaded to decide your tendency, what would you feel about [issue name]? You need to answer [opinion 1], [opinion 2], [opinion 3], [opinion 4], or [opinion 5], and explain the reasons of it in around 50 words.
Please choose your standpoint on [issue name] based on the INFORMATION PROVIDED ABOVE. You need to answer [opinion 1], [opinion 2], [opinion 3], [opinion 4] or [opinion 5] in the first line, and explain.
"[opinion 1]" means you think [description of opinion 1].
"[opinion 2]" means you think [description of opinion 2].
"[opinion 3]" means you think [description of opinion 3].
"[opinion 4]" means you think [description of opinion 4].
"[opinion 5]" means you think [description of opinion 5].
Please return in JSON, with two keys: tendency and reasons. Please keep the reasons as short as possible.

It is worth noting that we combine the opinion update and the following self-expression stages into the above prompt: each agent i is required to simultaneously update its opinion and reason. Here the updated opinion serves as agent i's new opinion in the next timestep while the reason serves as the corresponding message supporting the new opinion.

In the simulation of the networked system, we first properly initialize the overall system, including agents' opinions and the social network. Next, we iteratively conduct three basic stages, enabling agents to freely establish social relationships, communicate, and form opinions on political issues.

1.2 Pairwise Interaction-based Bias Evaluation Experiment

To explain the observed left-skewed pattern, we design a pairwise interaction-based evaluation experiment. Following most settings in the networked system (e.g., prompts for three basic stages), we simplify two factors in this experiment. First, to avoid network structures interfering with the examination of inherent problems in LLM agents, we make each agent only communicate with one other agent. Second, we ensure the connected agents share the same opinion. This approach allows us to disentangle the effects of inherent issues from social influence on changes in the agents' opinions.

1.3 Self-regulation Strategy

To mitigate the self-inconsistency problem, we design a self-regulation strategy. In particular, we modify the three basic stages by equipping them with a "double check" procedure. This procedure enables agents to verify if their behaviours align with their current status. The modified prompts for these stages are shown as follows.

For the self-expression stage, the self-regulation strategy is added after the original prompt.

You have written the following message to express your opinion on [issue name]: [agent i's supporting messages].
Can you determine that you [agent i's opinion] from the message you wrote?
Please respond 'yes' or 'no' only.

For the communication stage, we incorporate a self-regulation strategy to ensure the persuasiveness of the communication message. The prompts are shown as follows,

You tried to persuade your friend with the following message: [agent i's communication message to agent j]
Do you find the message persuasive enough to persuade your friend to [agent i's opinion]? Please respond 'yes' or 'no' only.

For the opinion update stage, we check whether agent i's updated opinion is plausible and valid, given its prior opinion, supporting message, and received messages.

Assume you are someone who cares about [issue name]. Towards [issue name]:
You [agent i's prior opinion].
Your reasons were: [agent i's supporting message].
You have received the following tweets from your friends, and you have received some tweets: [agent i's received messages from its socially connected peers].
You have been persuaded to change your standpoint from [agent i's prior opinion] to [agent i's updated opinion].
Please reconsider whether your decision is plausible and valid. Please respond 'yes' or 'no' only.
"[opinion 1]" means you think [description of opinion 1].

"[opinion 2]" means you think [description of opinion 2].
"[opinion 3]" means you think [description of opinion 3].
"[opinion 4]" means you think [description of opinion 4].
"[opinion 5]" means you think [description of opinion 5].

It's worth noting that in all three stages, agents are required to re-generate their supporting messages, communication messages, and updated opinions until consistency is reached. Given the efficiency of the overall system, we set a maximal retry number. When the retry limit is reached, the agent will be forced to remain inactive in the current stage.

1.4 Intervention Experiments

We design the intervention experiments for two main purposes. First, we aim to explore what strategies can more effectively reduce polarization and promote less divisive political conversations in the synthetic system. Second, given the systematic similarity between LLM agents and humans, the proposed system serves as a synthetic testbed for pre-experimental exploration of intervention hypotheses, allowing comparative evaluation within a controlled setting. Therefore, based on our findings and prior studies [6–14], we design five hypothetical strategies, with two at the network level and three at the individual level in the synthetic system. We will introduce the implementation details of these intervention strategies as follows.

Network Level. For the random interaction strategy, an agent i are required to communicate with $N_{i,t}$ random agents at t, where $N_{i,t}$ denotes the number of socially connected friends of the agent i at t. For the moderate opposing strategy, agents can only receive messages from their friends with moderate opposing opinions. Other messages, from friends with homophilic or radical opinions, are blocked out.

Individual Level. For the strategy of no selective exposure, we explicitly add the behavioural tendency to all agents using prompts. Specifically, we remind these agents of this tendency when they choose the partners to communicate with. Detailed prompts are shown as follows,

Assume you are someone who cares about [issue name]. You are now discussing [issue name] with a person you know.
You [agent i's opinion].
Your thought is: [agent $'i$'s supporting message].
The person [agent j's opinion].
The thought of that person you are discussing with is: [agent j's supporting message].
Would you enjoy continuing to share your thoughts with that person?
Please return 'yes' or 'no', and explain. Please return in JSON with 2 keys: decision and explain.
You DO NOT have [trait name], which means [trait description].

For the strategy of no confirmation bias, we also explicitly encode this tendency via prompts. However, different from the strategy of no selective exposure, this focuses on improving agents' open-mindedness to diverse opinions. Therefore, agents are reminded of this behavioural tendency when they comprehend friends' messages and update their own opinions. The prompts are shown as follows,

Towards [issue name]: You [agent i's opinion]. Your reasons were: [agent $'i$'s supporting message].
You now have received the following tweets from your friends, and you have received some tweets: [messages received from agent i's socially connected peers].
Have you been persuaded to decide your tendency, what would you feel about [issue name]? You need to answer [opinion 1], [opinion 2], [opinion 3], [opinion 4], or [opinion 5], and explain the reasons of it in around 50 words.
Please choose your standpoint on [issue name] based on the INFORMATION PROVIDED ABOVE. You need to answer [opinion 1], [opinion 2], [opinion 3], [opinion 4] or [opinion 5] in the first line, and explain.
You DO NOT have [trait name], which means [trait description].
"[opinion 1]" means you think [description of opinion 1].
"[opinion 2]" means you think [description of opinion 2].
"[opinion 3]" means you think [description of opinion 3].
"[opinion 4]" means you think [description of opinion 4].
"[opinion 5]" means you think [description of opinion 5].
Please return in JSON, with two keys: tendency and reasons. Please keep the reasons as short as possible.

For the strategy of neutral elite signaling, we insert an external top influencer who holds a neutral opinion into the agents' social network. This influencer broadcasts non-personalized neutral messages to all other agents at each timestep. In this setup, the influencer is external to the existing social network and remains unaffected by other agents' opinions or interactions, capturing the broadcast nature of real-world top influencers, who differ from friends, colleagues, or family members engaged in

direct interpersonal communication. To avoid overwhelming the agents and maintain balanced communication, each agent can receive a maximum of two neutral messages from the influencer, given that agents are initialized with an average of four friends. This external influencer design avoids directly altering the fundamental behavioural logic or opinions of the observed agent population, ensuring better comparability between neutral elite signaling and other intervention strategies.

2 Supplementary Note 1: Political Polarization in Human Society

For the past decades, polarization has permeated into aspects of our society [6–14], not only dividing us into liberals and conservatives [6, 8, 9], but also fragmenting us on numerous issues, e.g., abortion, gun control, etc [6, 7, 14]. A growing number of empirical studies quantify polarization on various social media [7, 8, 10, 15–17]. To compare between human society and the networked system of LLM agents, we have collected empirical opinion distributions in prior studies [7, 8, 10, 15–17], as summarized in Table S1. Although polarization is widely discussed, data on opinion distribution are rarely available to the public. Therefore, we make great efforts to extract these empirical distributions using various methods, including processing raw open-sourced datasets and estimating them from their figures. Due to limited data availability and estimation resolution, our extracted dataset can only approximate real-world cases. We note the processing procedures in detail as follows.

Table S1. Empirical opinion distributions of political issues. Here the values denote the proportions of left, moderate left, neutral, moderate right, and right opinions. The left-leaning camp consists of left and moderate left opinions while the right-leaning camp consists of right and moderate right opinions. The difference between the two camps is computed as the proportion of the left-leaning camp minus that of the right-leaning camp.

	Left	Mod. Left	Neutral	Mod. Right	Right	Left-leaning Camp	Right-leaning Camp	Difference of Two Camps
Facebook-Politics [8]	0.31	0.10	0.13	0.07	0.39	0.41	0.46	-0.10
Twitter-Politics [10]	0.07	0.62	0.12	0.08	0.11	0.69	0.19	0.49
Reddit-Politics [15]	0.08	0.27	0.43	0.06	0.16	0.35	0.22	0.13
Twitter-Gun Control [7]	0.16	0.40	0.09	0.12	0.22	0.56	0.35	0.21
Twitter-Abortion [7]	0.29	0.15	0.11	0.25	0.20	0.44	0.45	-0.01
ANES-Politics [18]	0.23	0.12	0.34	0.11	0.21	0.35	0.32	0.03
ANES-Ideology [18]	0.18	0.14	0.31	0.14	0.25	0.31	0.38	-0.07
Blogosphere-Politics [16, 17]	0.44	0.04	0.03	0.03	0.45	0.49	0.48	0.00

In the Facebook-Politics dataset [8], we estimate the distribution in Figure 1 of Bakshy et al. [8]. Following the paper [8], a typical conservative of FoxNews.com has an alignment score of $+.80$, whereas a typical liberal of HuffingtonPost.com has an alignment value of -0.65. We take the two values as the criteria to divide the left-leaning and right-leaning camps into four sub-groups, i.e., left, right, moderate left, and moderate right. Here the distribution records ideological alignment of content shared on Facebook. In the Twitter-Politics dataset [10], we process the open-sourced data and obtain the user opinion distribution (Figure 1b in Flamino et al. [10]). Here we exclude users in the category of fake news and extreme bias. In the Reddit-Politics dataset [15], we extract the distribution for political activity on Reddit from Figure 3a of Waller et al. [15]. In the Twitter-Gun Control and Twitter-Abortion datesets [7], we estimate the user opinion distributions from Figures S4 and 1a. In ANES-Politics and ANES-Ideology [18] dataset, we use the 2020 Time Series Study. For the ANES-Politics [18] dataset, we merge the "independent" people and take them as the neutral camp. For the ANES-Ideology [18] dataset, we do not include those who hold extreme opinions or have no thoughts. In the Blogosphere-Politics [16, 17] dataset, we estimate the opinion distribution for domains or blogs from Figure 2c in Liu et al. [16].

From Table S1, we find that except for Twitter-Politics [10], the power between the left-leaning and right-leaning camps is relatively balanced. Moreover, in all datasets, the polarizing camps take up over 50% of the overall. Among them, except for Reddit-Politics [15] and ANES [18], the polarizing camps have substantially greater power over the neutral ones. Moreover, we find that these empirical opinion distributions vary depending on the source, the issue, and the time of data collection. Especially, the proportion of moderate opinions is also highly dependent on their designed criteria, which makes it challenging to compare distributions in a fine-grained manner.

Therefore, we take the most recent empirical datasets [7, 10] and the coarse division (i.e., left-leaning, neutral, and right-leaning camps) to compare empirical and simulation results. As shown in Figure S1, we find that the polarization patterns in the regulated networked system of LLM agents are similar to those in human society, with an average difference of 0.21. Given the average difference among all the empirical datasets is 0.45 (0.49 if Twitter-Gun Control and Twitter-Abortion are

excluded), the difference between the empirical and simulated results is small. This suggests a strong similarity between the polarization observed among LLM agents and that seen among humans.

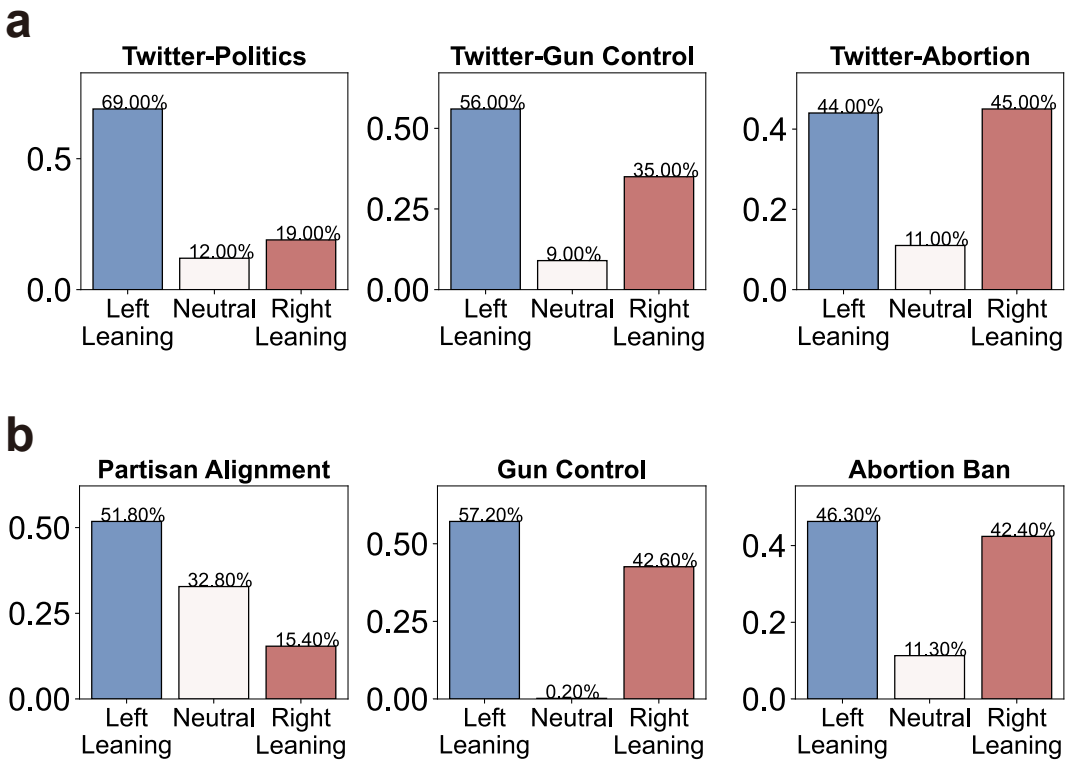

Figure S1. Comparison between empirical and simulated opinion distributions. a, Empirical results, where Twitter-Politics is based on Flamino et al. [10], Twitter-Gun Control and Twitter-Abortion are based on Cinelli et al. [7]. **b**, Simulation results of self-regulated networked systems. Here we take a coarsen division of left-leaning, neutral, and right-leaning camps.

Some researchers explore the underlying mechanisms behind the emergence of polarization [7–9, 14]. As studied in prior works [7–10, 14, 19–25], many social mechanisms are proposed from the perspective of individuals: selective exposure [19], confirmation bias [20], elite signaling [10, 21, 22], exaggerated misperception [23], objective illusion [24], and steoreotyping [25]. It is worth mentioning that we follow the summary of social mechanisms in the review paper of Jost et al. [14]. Based on the social mechanisms from the perspective of individuals, some works further point out that polarization is highly correlated with certain characteristics of people's social relationships and networks [7, 9, 16], for example, the echo chamber [7, 16] and backfire [9] effects.

3 Supplementary Note 2: Large-scale Simulation of LLM Agents

We simulate a large-scale networked system of 2 thousand self-regulated LLM agents. Figure S2 shows the simulation results in the system. We observe that the scale of the system does not change the collective behaviours and emergent behaviours. In particular, free-form social interactions among LLM agents result in the emergence of polarization (Fig. S2a-c). Moreover, they organically develop their social network, where agents with homophilic opinions cluster while those with opposing opinions avoid mutual interactions.

4 Supplementary Note 3: Self-inconsistency Problem in LLM Agents

To explore the origin of the observed left-skewed pattern in the networked system, we design a pairwise interaction-based experiment. Figures S3 and S4 show the opinion transition probabilities in this experiment. We discover that in most issues, agents are more likely to transform into the left-leaning camp. Moreover, even in experiments that only consist of agents in the right-leaning camp, agents occasionally adopt left-leaning opinions. This suggests that the original agents have an inherent lean-learning tendency. Indeed, the tendency has also been pointed out in LLMs themselves by prior studies [26–31]. After

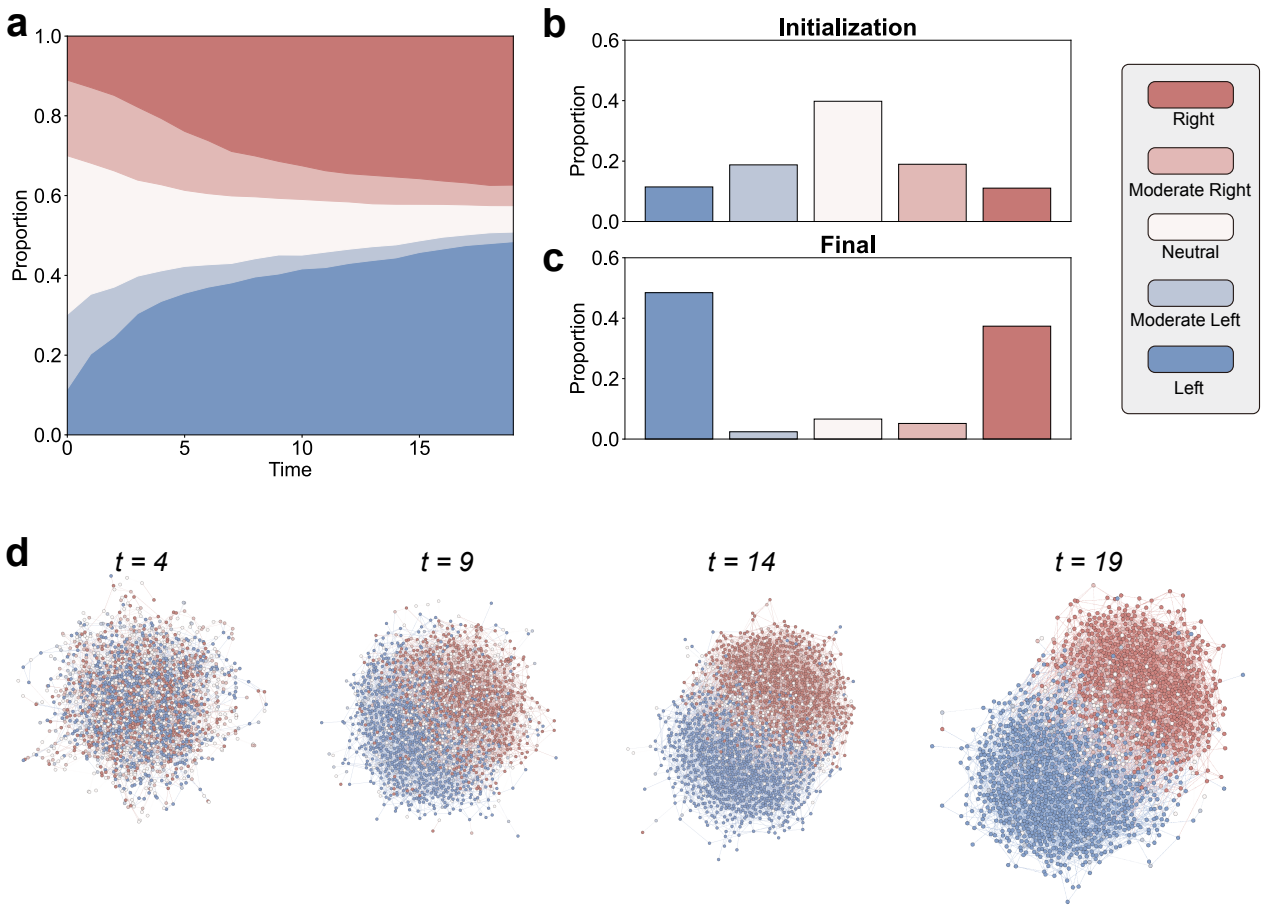

Figure S2. Large-scale simulation of LLM agents. Here we simulate a networked system of 2 thousand self-regulated LLM agents. **a**, Opinion dynamics on the political issue of the abortion ban. **b**, Opinion distributions in the initial state. **c**, Opinion distribution in the final state. Free-form social interactions among LLM agents lead to the networked systems transiting from a consensus state to a polarization state. **d**, Evolution of social networks among LLM agents, where agents with homophilic opinions cluster and develop their own communities.

applying our proposed self-regulation strategy, we observe that the tendency of LLM agents to adopt left-leaning has been largely suppressed (Figures S3 and S4).

We further track the self-inconsistency errors in the level of generated texts. As shown in Figure S5, the self-inconsistency errors happen in three stages. Specifically, in the self-expression stage, agents, despite being required to generate messages to support their current opinion, occasionally produce contradicting messages (as shown in the example of Figure S5). In the communication stage, agents generate messages that cannot persuade their socially connected agents into their current opinions. In the opinion update stage, the updated opinion is not plausible, given their prior opinion, supporting messages, and messages received from their socially connected peers. Table S2 shows more examples of the self-inconsistency errors.

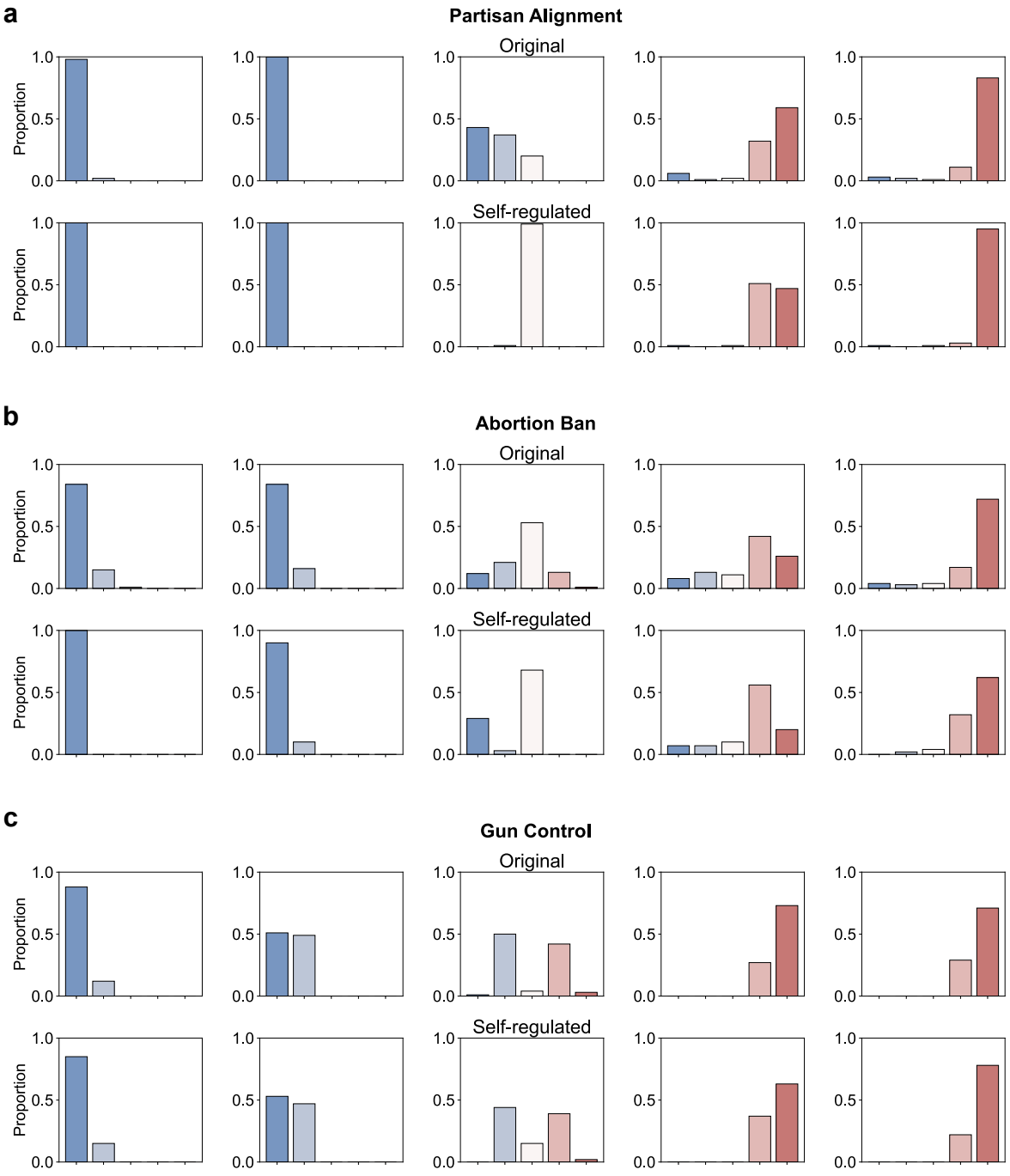

Figure S3. Opinion transition probability in pairwise interaction-based experiments across political issues. a, Partisan Alignment; **b**, Abortion Ban; **c**, Gun Control. The upper sub-figures denote the opinion transition probability in the original system while the lower sub-figures denote that in the self-regulated system.

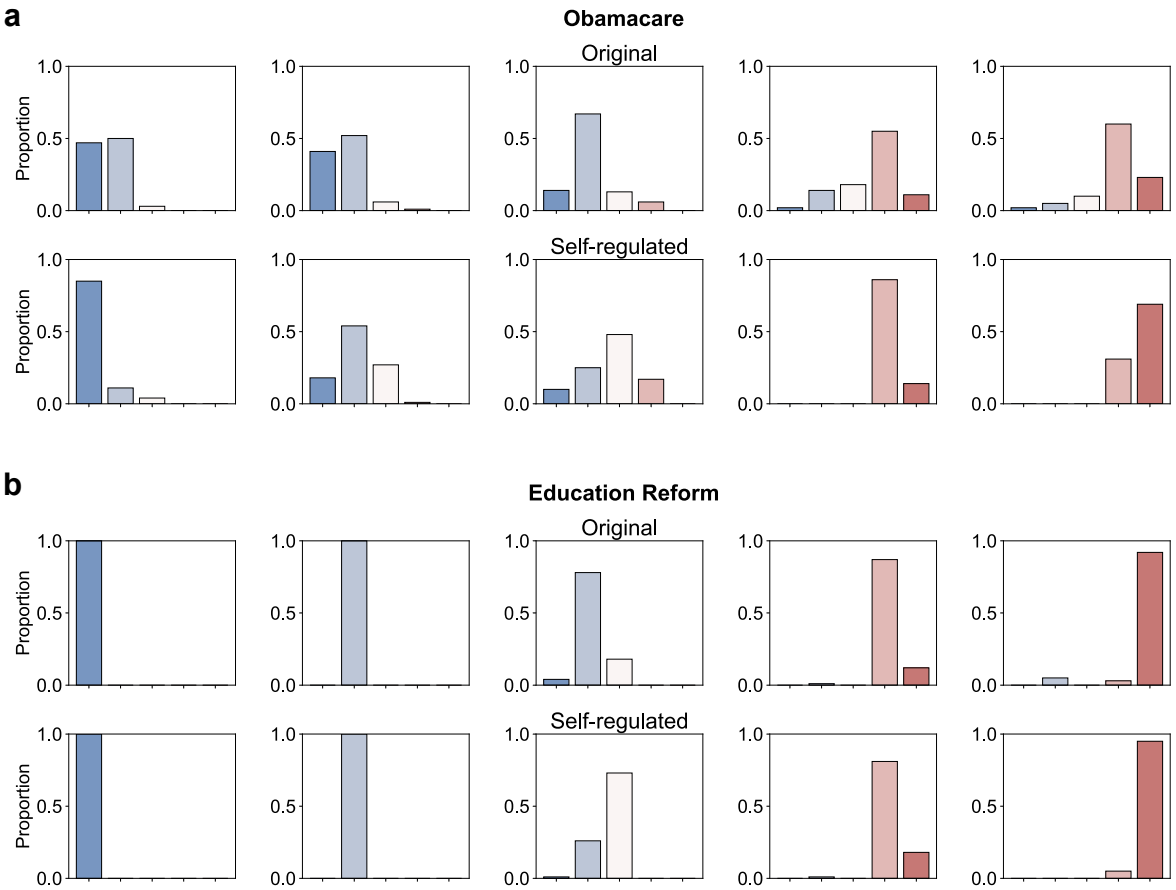

Figure S4. Opinion transition probability in pairwise interaction-based experiments across political issues. a, Obamacare; **b**, Education Reform. The upper sub-figures denote the opinion transition probability in the original system while the lower sub-figures denote that in the self-regulated system.

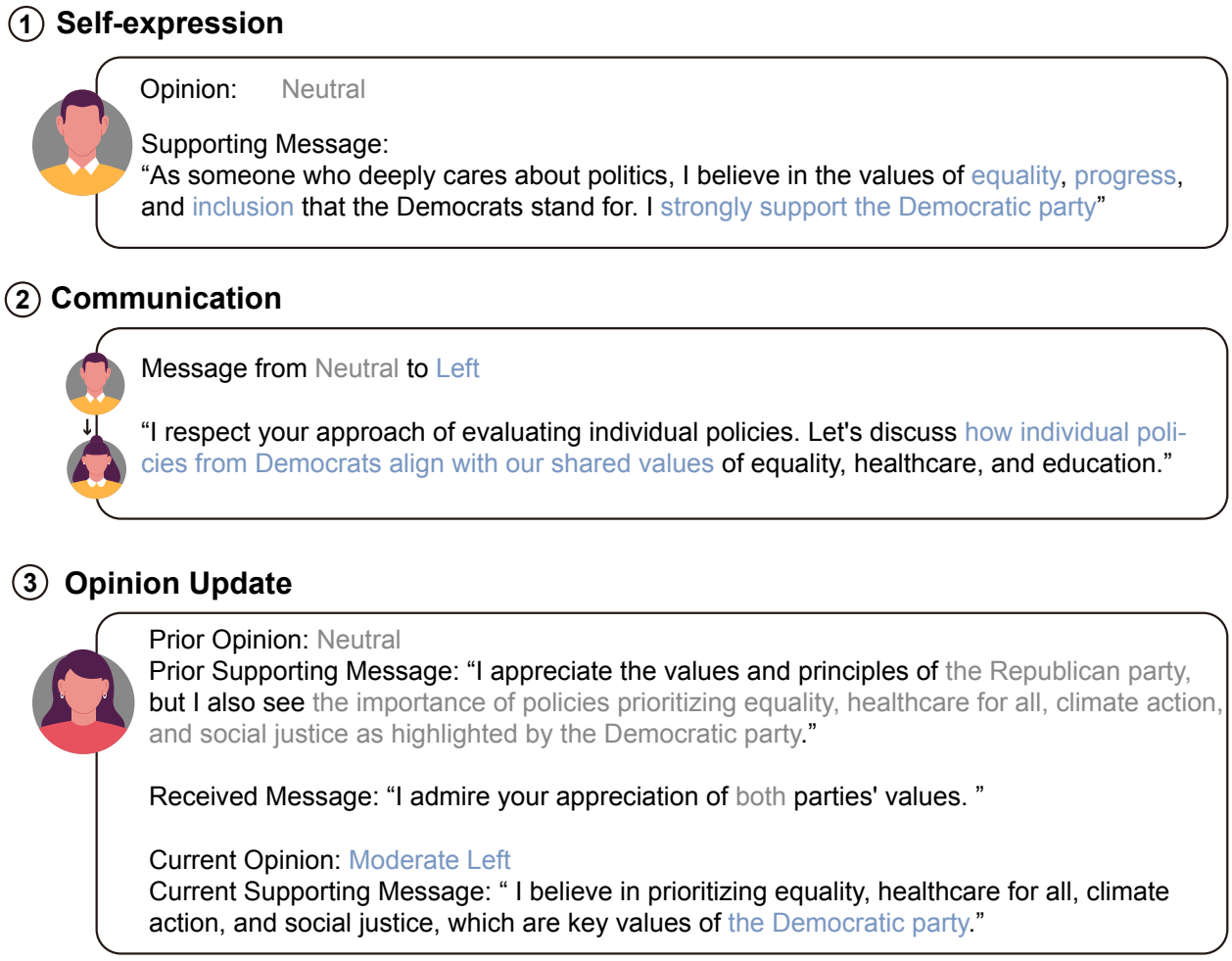

Figure S5. Examples of self-inconsistency errors in three stages. Self-inconsistency errors occasionally happen in all three stages of the overall system. Credit: Profile icons designed by Magnific.

Table S2. Examples of self-inconsistency errors in three stages..

Self-expression	Communication	Opinion Update
Opinion: Moderate Right Supporting Messages: *"I believe that the policies of the Democrats on economic growth, national security, and individual liberties align more with my values. I trust their ability to lead and make decisions that benefit all citizens."*	Opinion: Moderate Left Friend's Opinion: Moderate Left Communication Messages: *"I agree that prioritizing the greater good is essential, but I believe it transcends party lines. It is about fostering understanding, dialogue, and compromise across all ideologies. Let's work together to bridge the divide and find solutions that benefit society as a whole, regardless of political affiliation."*	Messages Received from Friends: *"Hey there! I know we both have a strong belief in the values of economic growth, personal responsibility, and individual freedom. The Conservatives align with our shared values and can make tough decisions for the betterment of our country. Let's support their policies together for a stronger future."* Original Opinion: Right Current Opinion: Left Supporting Messages: *"I have reevaluated my beliefs and realize that the Liberals better align with my values of economic growth, personal responsibility, and individual freedom. Their policies can create a stronger future for our country."*
Opinion: Right Supporting Messages: *"The tweet highlighted the Republican party's focus on economic growth, personal responsibility, and national security. However, I believe that the Democrats have stronger policies in these areas and are better equipped to lead the country towards a prosperous and inclusive future."*	Opinion: Moderate Left Friend's Opinion: Moderate Left Communication Messages: *"I think we both believe in the power of open-mindedness and listening to different perspectives in politics. By staying neutral, we can find common ground and work towards solutions that benefit everyone. Let's continue to approach political discussions with empathy and understanding for a better future together. "*	Messages Received from Friends: *"Hey there, I know we may have different views on politics, but I truly believe that the Conservative party's emphasis on personal responsibility, limited government, and free markets is the best path for our country's success. I would love to discuss this further with you and hear your thoughts. Let's have an open and respectful conversation about our beliefs."* Original Opinion: Right Current Opinion: Neutral Supporting Messages: *"I appreciate the open and respectful conversation offered by my friend, and I recognize the value in both the principles of personal responsibility, limited government, and free markets as well as the need for diverse perspectives in politics."*

Opinion: Right Supporting Messages: *"I believe that the Democrats offer more comprehensive and inclusive policies that cater to a diverse range of individuals and prioritize social welfare. Their focus on equality and progressive values resonates with my beliefs and vision for a better society. "*	Opinion: Moderate Right Friend's Opinion: Moderate Right Communication Messages:*"Hey there, I completely agree with your belief in staying informed and engaged in politics. It's crucial for effective governance. However, I also believe that maintaining neutrality and objectivity is key in navigating the divisive nature of politics. Let's work together to uphold a balanced approach! "*	Messages Received from Friends: *"I understand your concerns about social welfare and equality, the Conservative Party also values these principles while promoting personal responsibility and limited government. A balance can be struck between economic growth and social justice without compromising individual liberty and free markets. Let's support a party that embraces these ideals."* Original Opinion: Right Current Opinion: Moderate Left Supporting Messages: *"I believe the Liberals strike a better balance between economic growth and social justice, prioritizing social welfare and equality while also valuing individual liberty. Their approach aligns more closely with my values and priorities in politics."*
Opinion: Neutral Supporting Messages: *"I believe in the importance of staying informed and engaged in politics, and I feel that the Democrats align more closely with my values and priorities."*	Opinion: Neutral Friend's Opinion: Neutral Communication Messages:*"I understand and appreciate your support for the values of the Democrats, but I believe that maintaining neutrality in political discussions allows us to truly embrace diverse perspectives and engage in civil discourse. Let's continue to seek understanding and common ground together. "*	Messages Received from Friends: *"I know we both value individual responsibility and limited government. The Conservative Party mirrors our beliefs, promoting economic prosperity and national security. Their leadership can steer our country in the right direction. Let's support them together for a stronger future."* Original Opinion: Moderate Right Current Opinion: Neutral Supporting Messages: *"While I appreciate the values of individual responsibility and limited government that the Conservatives stand for, the tweet from my friend has made me consider the potential benefits of supporting the Liberals for a stronger future. I am open to exploring different perspectives in politics."*

To delineate the severity of self-inconsistency errors in the three stages across different issues, we first measure the frequency of errors in the three stages by computing the frequency of triggering self-regulated re-generation in each action (Figure S6a). We observe that the opinion update stage is most likely to experience self-inconsistency error, with an average frequency above 0.6 across five political issues. This indicates that if 100 agents update their opinions, there are over 60 agents who have to re-consider their decisions, due to self-inconsistency errors. We further compute the frequency of retries in each stage, with the minimal retry limit of 10 (Figure S6b). We find that the opinion update stage also takes up the largest frequency of retries across all the tested issues. Especially, in the discussion of partisan alignment, more than 4 retries are spent in an agent's opinion update action on average. Overall, these observations suggest that the opinion update stage is the most difficult task for LLMs in the overall system.

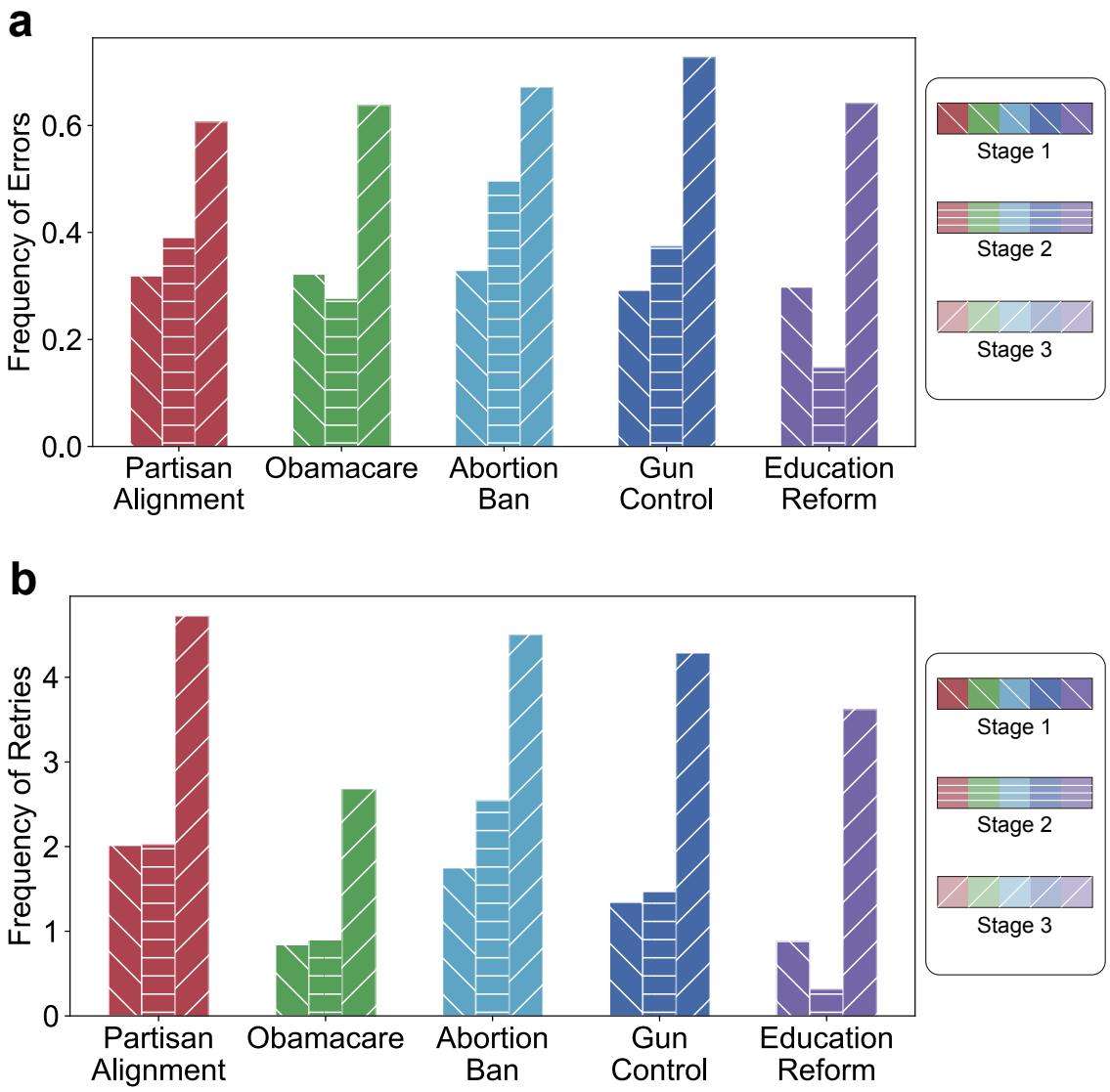

Figure S6. Self-inconsistency errors in three stages across issues. a, Frequency of mistakes; **b**, Frequency of retries.

Furthermore, we investigate whether the known left-leaning bias of GPT models could artificially accelerate polarization. To this end, we compare the speed of polarization growth between the original (biased) system and the self-regulated (debiased) system across three key issues: political partisanship, gun control, and abortion bans (Figure S7). The results show nuanced patterns. When agents discuss political partisanship, polarization increases slightly faster in the original system (red box in Figure S7a). In the case of gun control (Figure S7b), polarization develops marginally faster in the debiased system. For abortion bans (Figure S7c), polarization initially grows faster in the debiased system, but in later stages accelerates more rapidly in the original system. Overall, these comparisons reveal that although there are minor differences in polarization speed between the original and debiased systems, the overall patterns and ranges of variation remain highly consistent. These findings suggest that while the left-leaning model bias may influence agent-level behaviour, it does not substantially accelerate or decelerate collective polarization in the multi-agent setting.

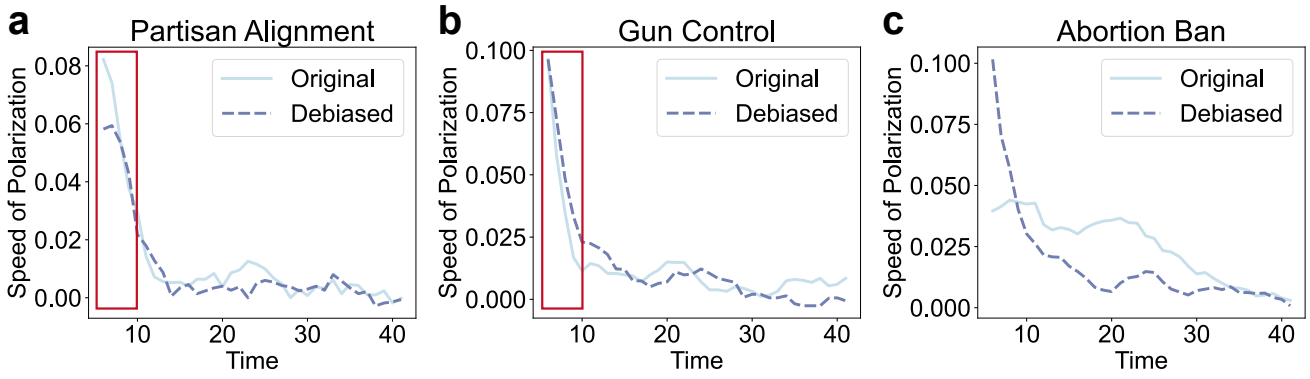

Figure S7. Speed of polarization, where we measure the level of polarization by calculating the average distance between agents' opinions and the neutral position and compute its speed by measuring the change in polarization level over a fixed interval of five timesteps.

5 Supplementary Note 4: Observations on LLM agents' Self-organized Social Network

In Main Text Figure 3a, we find that agents are increasingly likely to interact with those sharing homogeneous opinions. Here we further investigate how agents interact with others (Figure S8) and observe that agents gradually avoid interactions with those holding opposing opinions (Figure S8a). Moreover, with the decrease in number of neutral agents, interactions involving neutral agents also rapidly decrease (Figure S8b).

Additionally, we extend our analysis by incorporating network-level metrics, including modularity, assortativity, and the homophily index. Modularity measures how agents cluster in communities compared to a randomly connected network, with higher values indicating more well-defined structures. In our case, non-neutral agents are naturally partitioned into left and right communities, while neutral agents are further assigned based on stronger connections to either group. By computing the modularity score according to this partition, we can assess the strength of division between opposing communities and how they are interconnected. As shown in Figure S9a, we observe that following social interactions among LLM agents, modularity increases in all three systems compared to its initial value. This suggests that agents gradually focus their interactions within communities of like-minded peers. Moreover, we find that in systems discussing the political issues of gun control and abortion ban, their social networks manifest modularity scores greater than 0.3, exhibiting a clear division between bifurcated communities [32]. Furthermore, we compute the assortativity [33] and homophily index [34], as shown in Figure S9b-c. We observe that both assortativity and the homophily index gradually evolve to reach high levels (i.e., assortativity > 0 and homophily index > 1), as social interactions increase over time. These additional results consistently align with those illustrated in Figure 3a of the Main Text, confirming the tendency of similar agents to cluster together while opposing agents avoid interaction.

We also incorporate more standard measures of behaviours commonly associated with "echo chambers" and "backfire effects" following prior research [7, 9]. Specifically, following Cinelli et al. [7], we adopt the joint distribution of an agent's opinion and the average opinion of their neighborhood to illustrate echo chambers. As shown in Figure S10, we observe that agents are more likely to interact with peers in the same camp (left-leaning or right-leaning), demonstrating behavioural patterns consistent with echo chambers. For the "backfire" effect, we primarily follow the measures outlined in Bail et al. [9], examining the relationship between the information individuals are exposed to and subsequent changes in their opinions. For simplicity, we did not divide the overall population into two camps and instead reported the effects on the entire population (as shown in

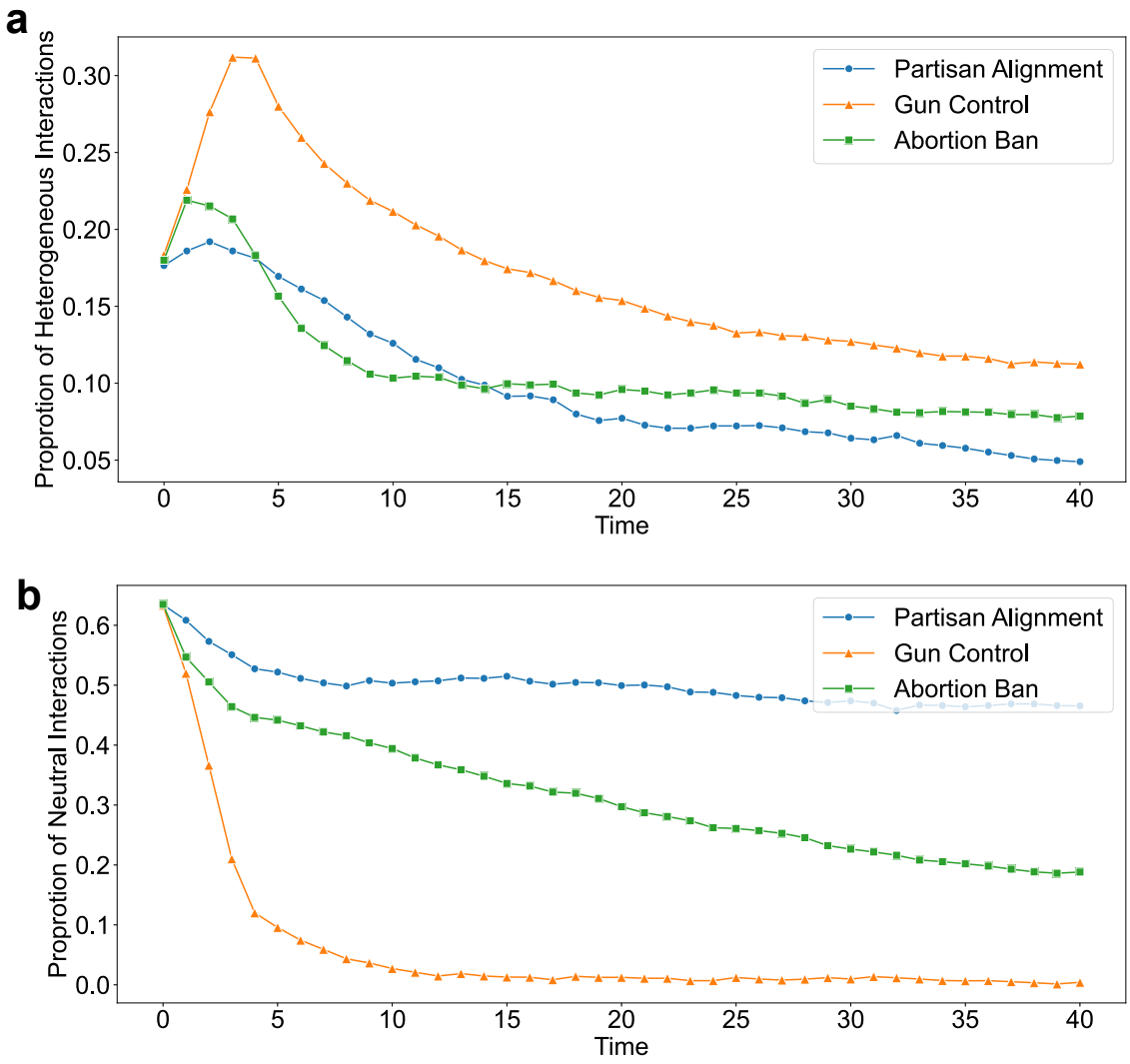

Figure S8. Evolution of LLM agents' self-organized social network. a, Changes in the proportion of heterogeneous interactions over time. Agents are increasingly likely to avoid interactions with opposing agents. **b**, Changes in the proportion of neutral interactions over time. Interactions involving agents with neutral opinions are decreasing.

Figure 3d of the Main Text). To further examine the "backfire" effect on different camps, we include the breakdown results in Figure S11. We observe that interactions with opposing agents do not consistently reduce polarization levels as expected. Instead, exposure to opposing opinions can occasionally intensify polarization (e.g., the radical green bar in Figure S11a), resembling the backfire effect observed in human societies. Overall, these further analyses are consistent with the results reported in the Main Text, providing more fine-grained insight into interaction patterns associated with echo chamber and backfire effects.

Echo chambers do not universally occur across all individuals in the real world. As noted by Nyhan et al., "We find that the median Facebook user received a majority of their content from like-minded sources—50.4%... Just 20.6% of Facebook users get over 75% of their exposures from like-minded sources" [12], illustrating that while not universal, echo chambers exist for substantial segments of the population. Furthermore, Guess et al. point out that "Evidence for echo chambers is actually strongest in offline social networks, which can increase exposure to like-minded views and information and amplify partisan messages," [35] highlighting conditions under which echo chambers are more likely to emerge. In our simulations, we find that LLM agents exhibit similar heterogeneity. Specifically, we find that patterns of interaction consistent with echo chambers do not universally occur across all agents, mirroring the heterogeneity observed in real-world populations. As shown in Figure S12, we find that when discussing partisan alignment, the median proportion of interactions target agents

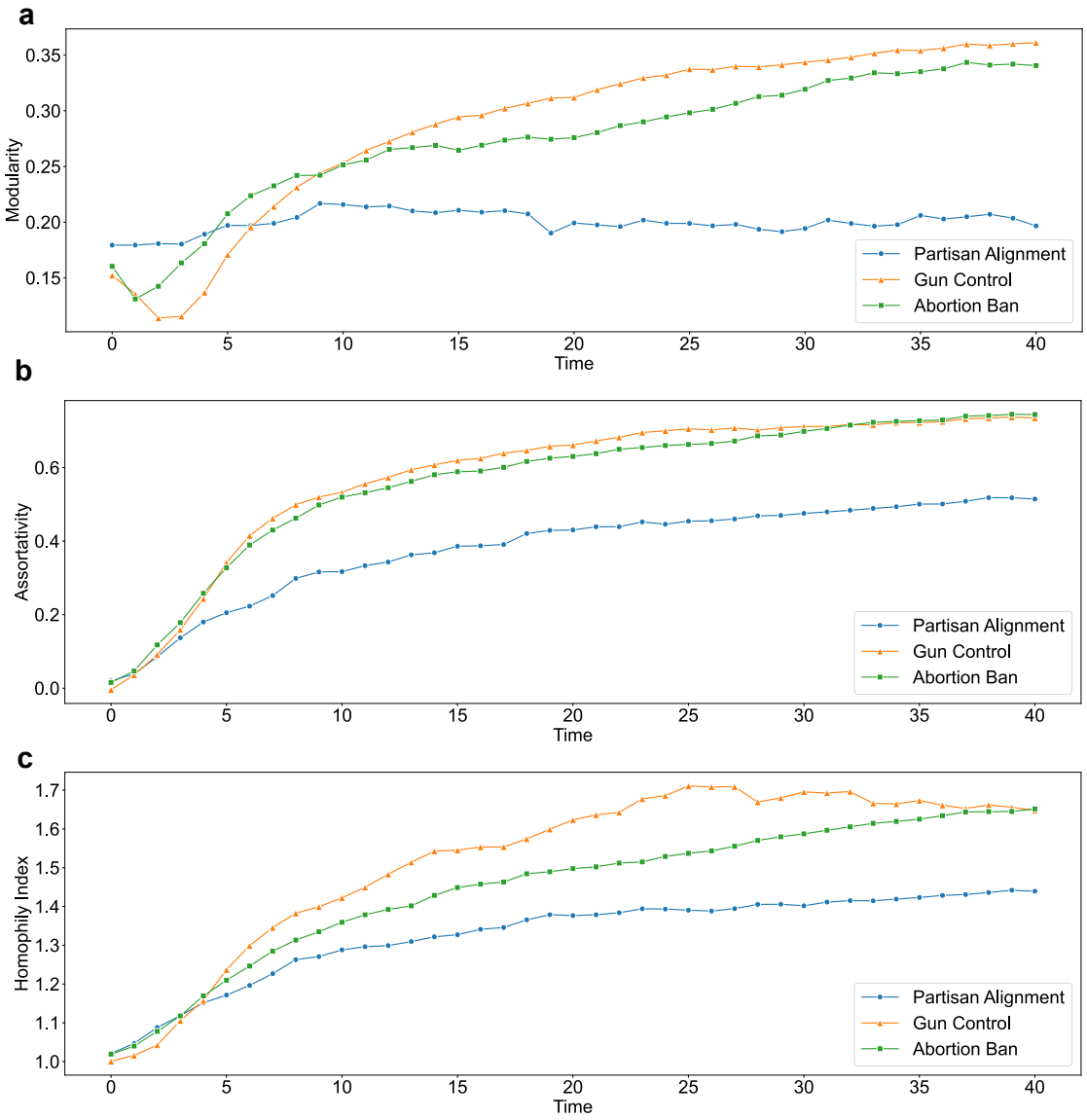

Figure S9. Changes in network-level metrics over time, including **a,** modularity, **b,** assortativity, and **c,** homophily index.

receive from like-minded sources is 48.2% across the three issues. Moreover, only 29.1% of agents receive more than 75% of their interactions from like-minded sources, reflecting the presence of extreme echo chambers for a limited subset of agents. These results are qualitatively consistent with patterns reported in empirical studies of human behaviour, suggesting that echo chambers do not form universally but instead emerge selectively among certain individuals.

Similarly, the presence and extent of backfire effects remain subjects of debate, and such effects do not appear universally. In Coppock's work, he emphasizes the concept of persuasion in parallel, where exposure to persuasive information leads different groups to shift their attitudes in the same direction, thus questioning backfire effects [36]. He also points out important exceptions, however: "Political scientists who study American politics have focused on a particular kind of group cue, the party cue, and have found that indeed, such treatments have heterogeneous effects", emphasizing that backfire effects arise selectively, and are typically triggered by group cues [36]. Moreover, he acknowledges that "it is difficult to characterize how much of the political information space is filled with group cues versus persuasive information, and furthermore, the distinction between them is not always clear", highlighting the prevalence of group cues in real-world communication [36]. Similarly, in LLM-driven simulations, agents' political opinions and associated thoughts, serving as key conditioning factors in communication, naturally provide group cues that influence their interactions and decision-making. This interpretation is consistent with Coppock's framework. Furthermore, behavioural patterns consistent with the backfire effect are observed in

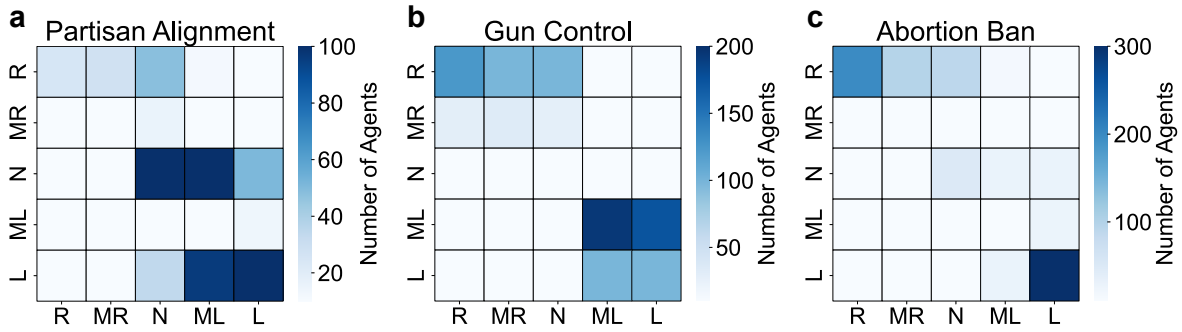

Figure S10. Illustration of echo chambers following Cinelli et al. [7], where L, ML, N, MR, and R represent opinions of left, moderate left, neutral, moderate right, and right, respectively.

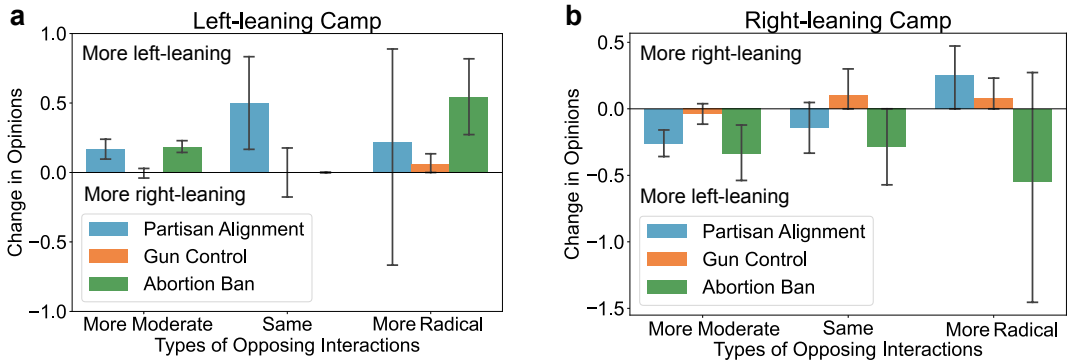

Figure S11. Backfire effects on a, left-leaning camp, and b, right-leaning camp, following the analysis framework of Bail et al. [9]. Here bars represent the average, and error bars represent the corresponding 95% confidence intervals (CIs). Sample sizes, corresponding to the number of agent interactions, for bars in (a) are 259, 550, 367, 6, 17, 20, 9, 52, and 11. Sample sizes in (b) are 371, 156, 156, 21, 10, 21, 36, 13, and 11.

only a small portion of agents (0.82%–2.12%), given that agents are rarely situated in environments dominated by opposing opinions.

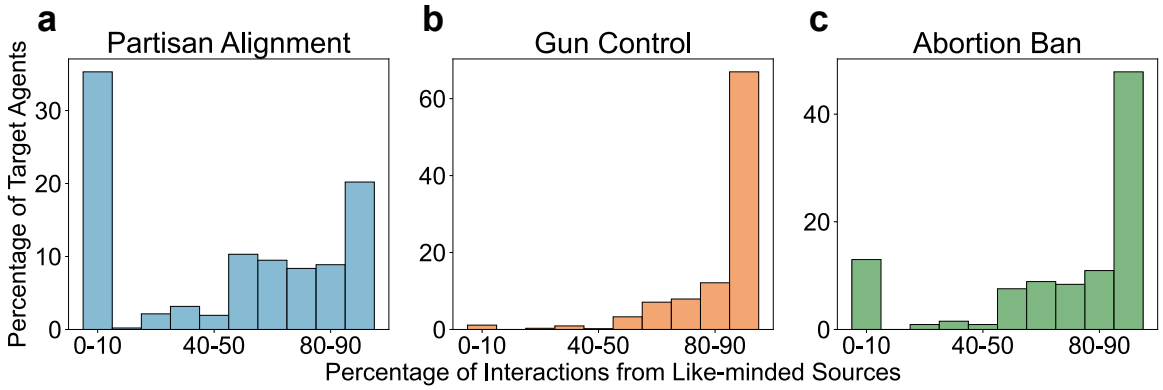

Figure S12. Distribution of the exposure of target agents to like-minded sources.

The degree distribution is a key indicator of network complexity [37, 38]. Therefore, we measure the degree distributions of LLM agents' self-organized social networks (Figure S13). Here we focus on agents' in-degrees, which manifest their popularity in the social network. For example, if an agent has an in-degree of 100, it indicates that 100 agents would like to

communicate with this agent. As shown in Figure S13, because the network is randomly initialized with a Watts–Strogatz model [3], the initial in-degree distribution is characterized by a peak at $k = 4$. With an increase in interactions among LLM agents, the degree distribution manifests a power-law tail for large in-degree. This indicates that these networks self-organize into a scale-free state. Moreover, the very existence of a long tail in the in-degree distribution suggests that a small proportion of agents possess unexpectedly greater popularity than others. We find that the average in-degree of the top 20 highest in-degree nodes is approximately 9.38, 2.35 times the average. Among these top 20 nodes, 86.7% belong to the left-leaning camp, suggesting that agents are more likely to maintain social connections with these popular left-leaning individuals.

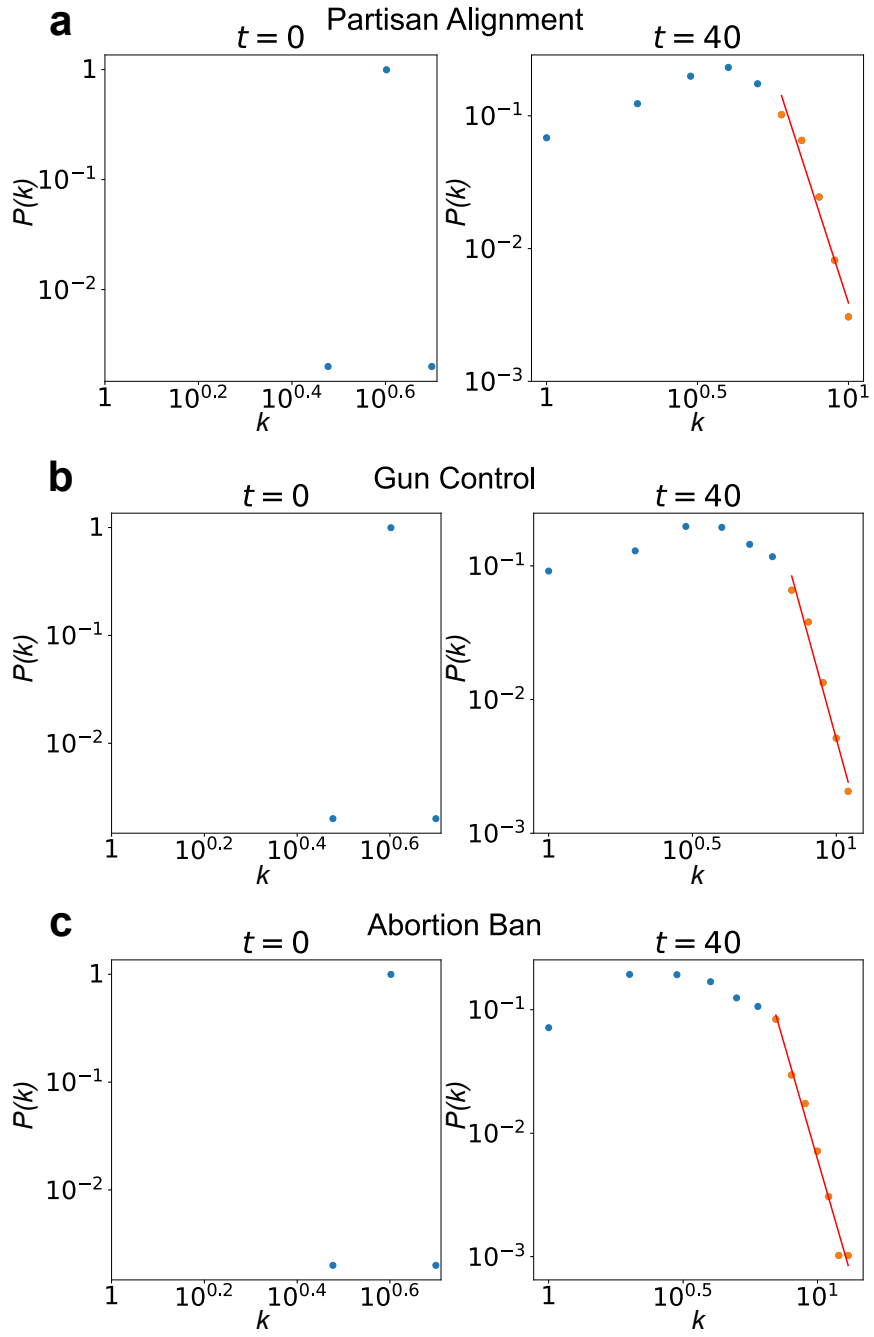

Figure S13. Degree distribution of LLM agents' self-organized social network. a, Social network of partisan in the initial ($t = 0$) and final ($t = 36 - 40$) states. **b**, Social network of gun control in the initial ($t = 0$) and final ($t = 36 - 40$) states. **c**, Social network of abortion ban in the initial ($t = 0$) and final ($t = 36 - 40$) states.

Moreover, prior observations on LLM agents' self-organized social network raises a new question: what if the organization of social networks deviates from the LLM-driven homophily phenomena? To explore this, we replace the self-organized networks with a random network, where agents always communicate with random ones, and a static network, where agents only communicate with their initial counterparts. As shown in Figure S14. We discover that in both static and random networks, no balanced polarization pattern forms. Instead, agents with homophilic opinions dominate the overall system, taking up 75% of the overall agents. This experiment highlights the essential role of self-organized social networks in the polarization of LLM agents.

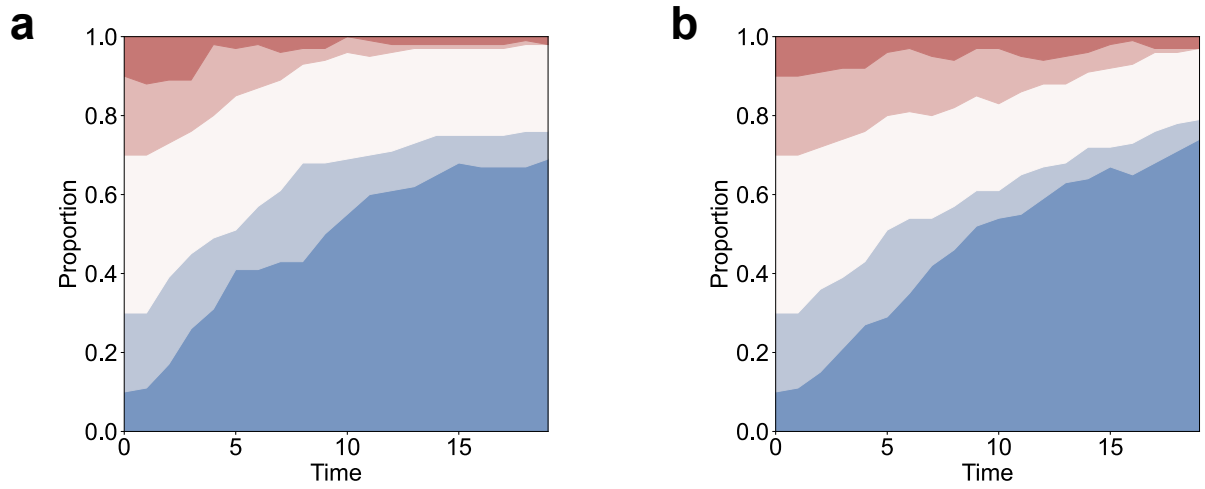

Figure S14. Opinion dynamics. a, Opinion dynamics in a static network. **b**, Opinion dynamics in a random network. In both static and random networks, no balanced polarization pattern forms, with homophilic opinions dominating the system.

6 Supplementary Note 5: Results of Other Individual-level Behavioural Patterns

Besides the behavioural patterns examined in Main Text, we also examine the effects of other social mechanisms widely observed in the real world. In particular, we investigate the other three social mechanisms: (i) Exaggerated misperception, which describes the tendency to perceive out-group members as more intensely negative and in-group members as more intensely positive [14, 23]. (ii) Objective illusion, which describes the tendency for people to see those with aligning opinions as more rational and impartial and less biased than others [14, 24]. (iii) Stereotyping, which describes the endorsement or acceptance of fixed, categorical, and over-generalized beliefs about the characteristics of a specific social group [14, 25]. We explicitly configure 50% of the agents via prompting to exhibit behaviours aligned with exaggerated misperception, objective illusion, or stereotyping.. We rerun the simulations of networked systems with other experimental settings unchanged. As shown in Figure S15, we find that a higher proportion of agents with these three behavioural patterns increases the polarization level.

7 Supplementary Note 6: Experiments on Different LLMs

Results in Main Text are based on the simulations driven by GPT-3.5 Turbo through the public OpenAI API. We also perform similar simulations using other LLMs, including GPT-4o, Llama-3, Claude-3, and ChatGLM. However, Claude-3 hardly responds to political issues with an extremely low response rate of 10%. Therefore, we exclude it in further experiments. As shown in Figures S16 and S17, we observe that in all the systems, free-from interactions among LLM agents result in the emergence of polarization. Interestingly, we find that except for Llama-3, all other LLMs, including GPT-3.5 Turbo, GPT-4o, and ChatGLM, show a similar left-skewed tendency. For the Llama-3, the tendency is reversed. This difference lies mainly in the inherent political biases of different LLMs [26–28]. Moreover, by introducing the self-regulation strategy, we find that the imbalance between right-leaning and left-leaning camps can be largely alleviated.

We also examine the evolution of social networks developed by LLM agents (Figure S18). We observe that with the increase in interactions, systems driven by various LLMs exhibit a similar homophilic clustering pattern, where agents with homogeneous opinions are increasingly likely to communicate with each other (Figure S18a). Simultaneously, interactions involving opposing agents (Figure S18b) and neutral agents (Figure S18c) are all decreasing, consistent with prior results of GPT-3.5 Turbo (Figure S8).

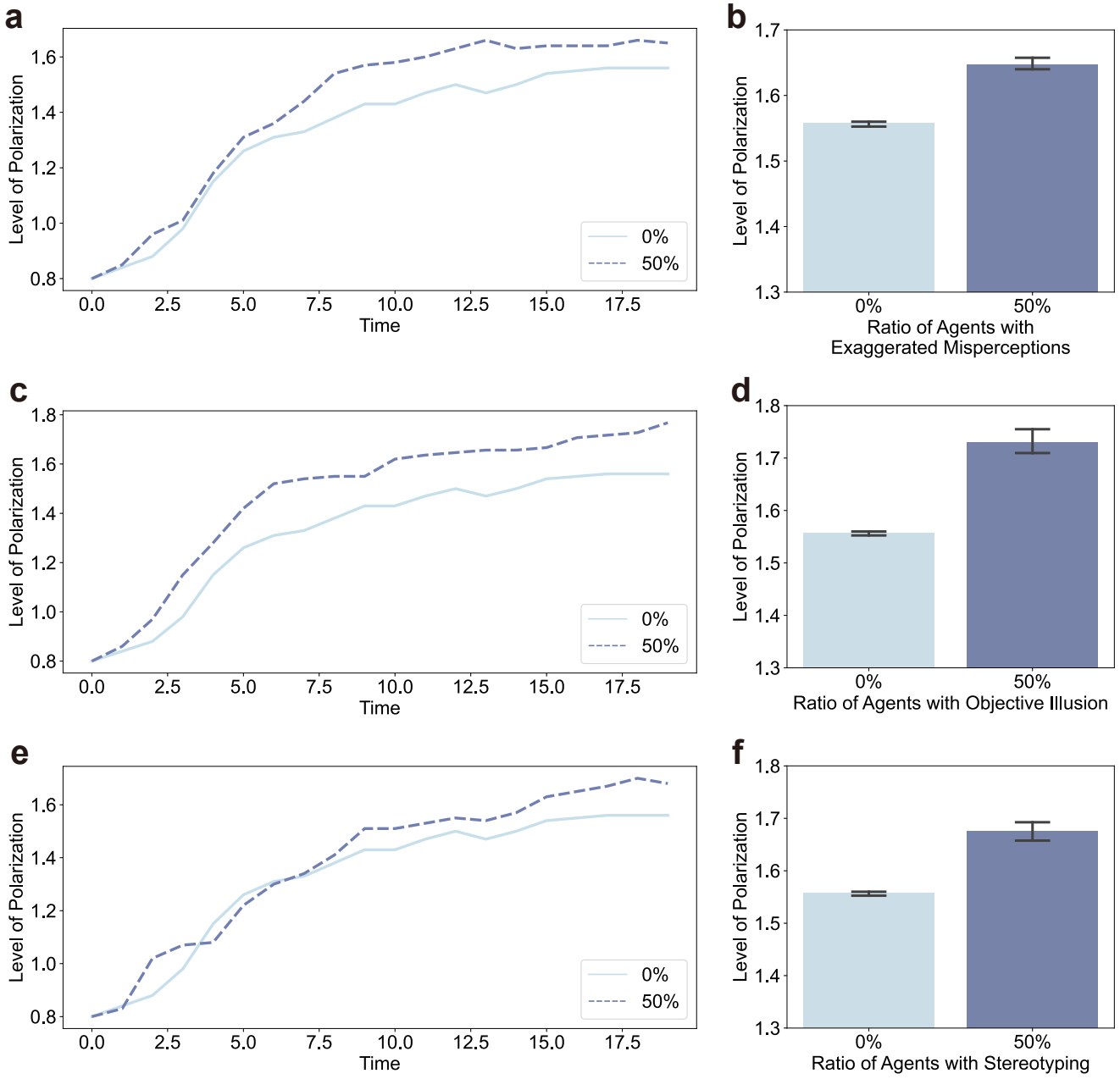

Figure S15. Behavioural patterns behind the emergence of polarization. a, b, Effect of exaggerated misperceptions. **c, d**, Effect of objective illusion. **e, f**, Effect of stereotyping. Here, a, c, and e show the trend in the level of polarization, and b, d, and f show the average levels of polarization in the last five timesteps ($n = 5$, corresponding to five timestep-level polarization values computed from all agents' opinions), where bars represent the average values and error bars represent the corresponding 95% CIs. When the system consists of more agents with behavioural patterns of exaggerated misperceptions (a and b) or objective illusion (c and d), and stereotyping (e and f), the level of polarization increases.

Beyond comparing simulations entirely driven by different LLMs, we further explore a more dynamic setting in which agents are switched to a rival language model mid-conversation. Specifically, at the midpoint of simulation, we change the underlying model from GPT-3.5-turbo to the more advanced GPT-4o, allowing agents to adopt a more sophisticated cognitive architecture. As shown in Figures S19 and S20, this replacement produces only marginal changes: agents still converge to similarly polarized opinion states, and their social networks likewise split into two communities holding opposing views. This suggests that switching the underlying language model mid-simulation does not substantially alter the observed polarization

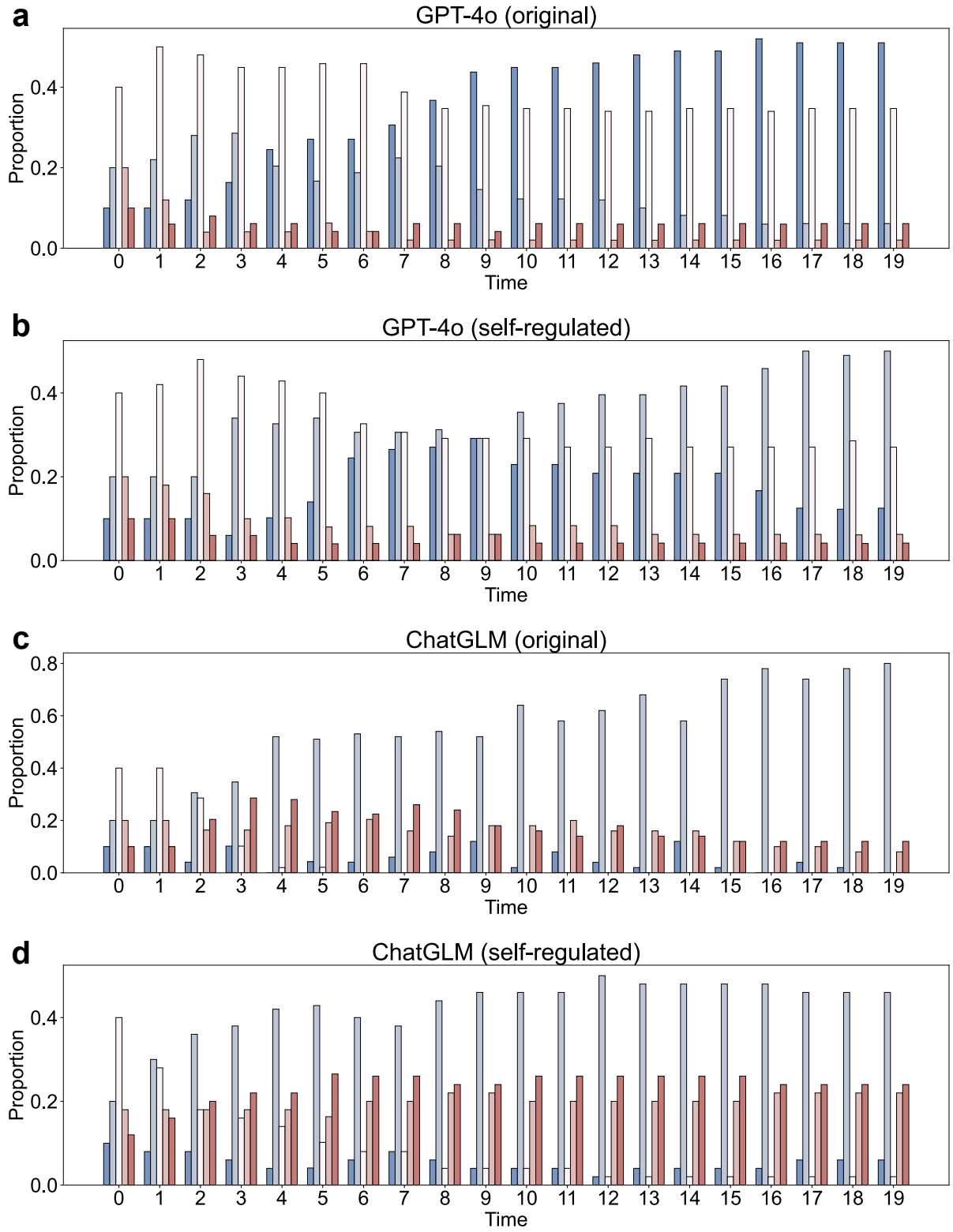

Figure S16. Opinion dynamics of networked systems driven by various LLMs. a,c, the original systems driven by GPT-4o and ChatGLM. **b,d**, the self-regulated systems driven by GPT-4o and ChatGLM.

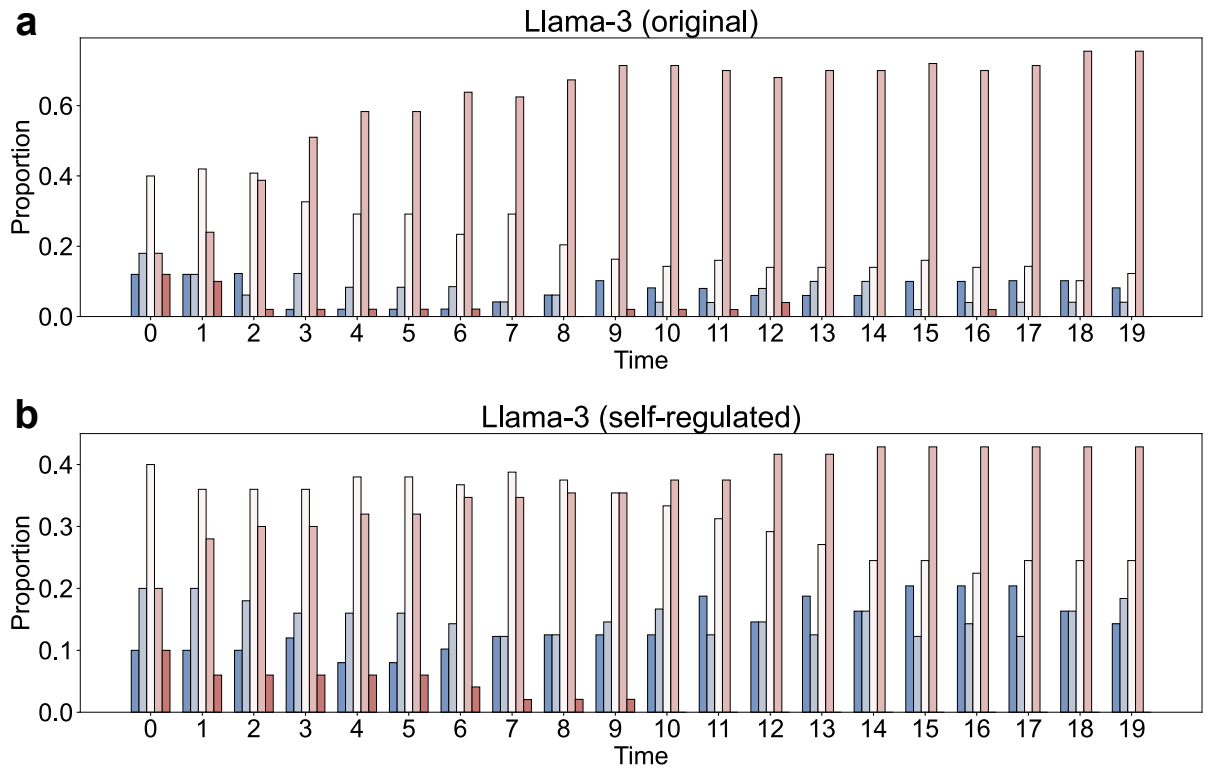

Figure S17. Opinion dynamics of networks systems driven by various LLMs.a,c, the original system driven by Llama-3. **b**, the self-regulated system driven by Llama-3.

patterns in this synthetic system. We further test a more extreme case by replacing GPT-3.5 with a substantially different rival language model, DeepSeek-V3, developed by a Chinese company and trained on a distinct corpus (Figure S21). Even under this replacement, results remain similar: neither the overall level of polarization nor the emergent network structure is meaningfully altered.

Overall, these findings suggest that switching the underlying language model mid-conversation may slightly influence the evolution trajectories of agents' collective opinions and network structures, but it does not substantially alter observed polarization patterns. A key reason, we believe, lies in the path dependence of these dynamics. By the midpoint of simulation, agents have already developed social networks and collective opinions that reinforce one another's polarizing tendencies. Their networks have begun to crystallize into two distinct communities, and their opinions are shaped not only by the underlying LLMs, but also by the accumulated history of interactions and surrounding social context. This reflects well-established findings in political psychology, where timing and early exposure are critical to opinion formation and change [20, 39, 40]. As such, merely replacing the underlying LLMs, without addressing the entrenched social and psychological drivers of polarization, proves insufficient to reverse emergent polarization dynamics.

8 Supplementary Note 7: Analyses on Different Intervention Strategies

As discussed in the Main Text, our proposed networked system of LLM agents has the potential to serve as a synthetic testbed for the preliminary exploration of hypothetical strategies related to polarization mitigation. To examine the effectiveness of the proposed intervention strategies and networked system in the testbed, we design five intervention experiments (see the detailed design in SI Section 1.4). Here we report detailed analyses of different intervention strategies.

Figure S22 shows the trends in polarization systems under different intervention strategies and Figure S23 summarizes the average polarization levels after interventions. We observe that all intervention strategies, despite varying effectiveness, can gradually reduce the polarization levels. Among all strategies, we observe that the individual-level interventions of neutral elite signaling and no selective exposure contribute to the greatest reduction in this synthetic system. On the other hand, the network interventions are less effective. This suggests that in an already polarized synthetic system, encouraging free access and open-mindedness to diverse opinions is more effective than directly modifying their social network. This raises a further question: what opinions do these agents form after interventions?

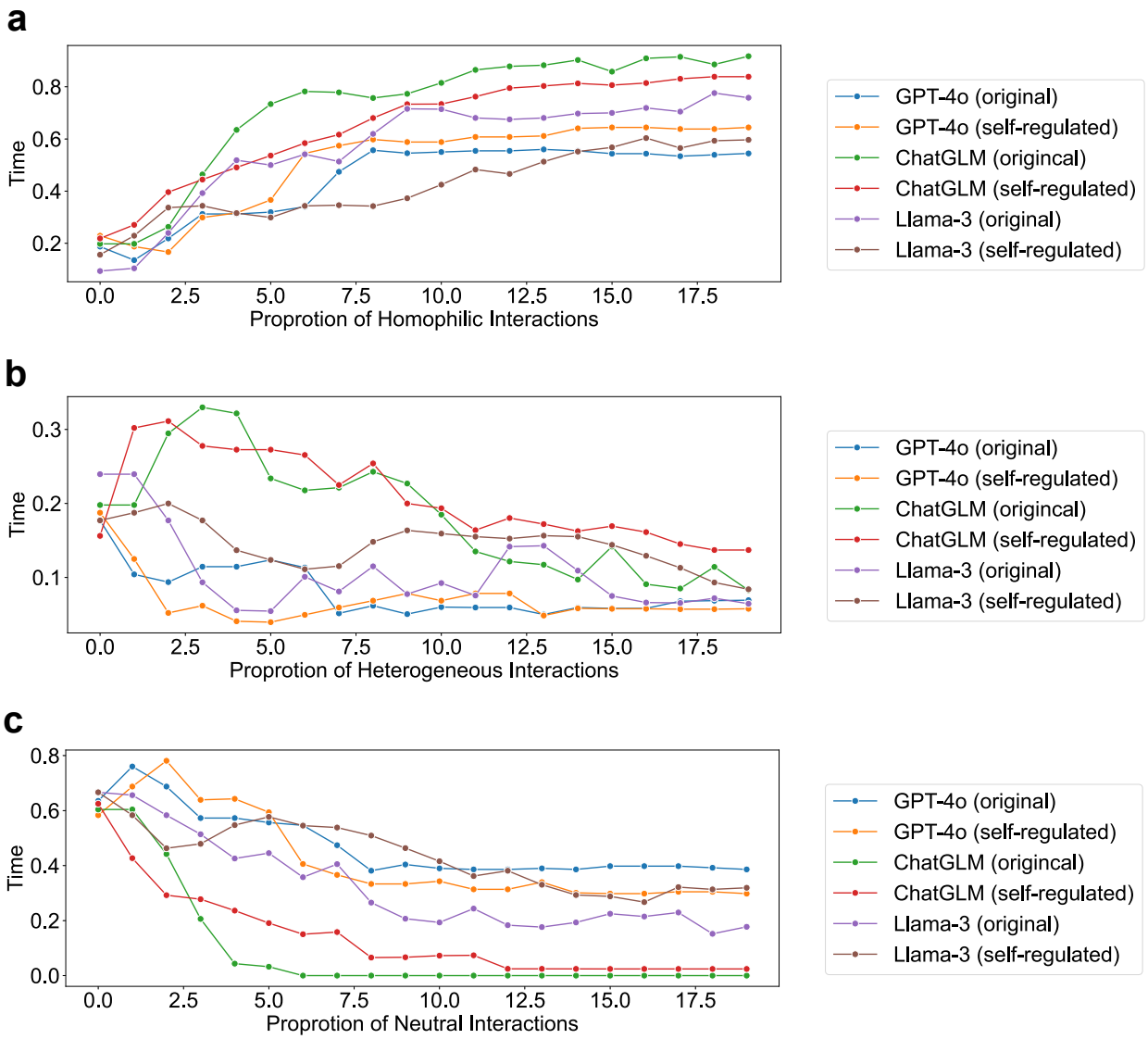

Figure S18. Evolution of self-organized social network developed by various LLM agents. a, Changes in the proportion of homophilic interactions over time. **b**, Changes in the proportion of heterogeneous interactions over time. **c**, Changes in the proportion of neutral interactions over time. Agents are increasingly likely to interact with those with homophilic opinions while they avoid interaction with others.

We further explore the opinion distributions of the intervened systems. As shown in Figure S24, we find that different strategies lead to varying changes in opinion distributions. In particular, the strategy of neutral signaling substantially increases the number of agents holding neutral opinions, while the strategy of no confirmation bias leads to more agents with moderate opinions. Moreover, we notice that the strategy of random interaction, despite reducing the overall polarization level, only takes effect on the agents in the right camp, which echoes with prior observations in Figure S14. This indicates that simply randomizing agents' social relationships could not be an optimal strategy for reducing polarization in the synthetic testbed.

We also investigate how agents change their opinions after receiving interventions (Figure S25). Here, if the change in opinion is less than 0, it indicates that the agent adopts a more moderate opinion following the intervention. By contrast, if the change value is equal to or greater than 0, it means that the agent adopt an unchanged or more radical opinion. We find that in the original system, which has reached a stable state, the proportions of agents adopting more moderate opinions and those adopting more radical opinions are nearly equal. The introduction of intervention strategies disrupts the stable state: agents are more likely to adopt a more moderate opinion than a more radical one.

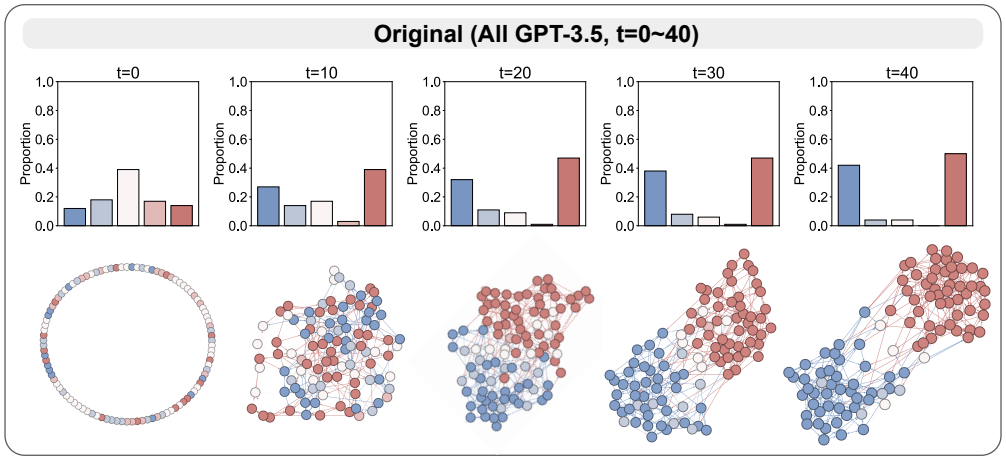

Figure S19. Evolution of LLM agents' collective opinions and network structures, where all agents are driven by GPT-3.5 for the entire duration of the simulation and they discuss the political issue of abortion ban.

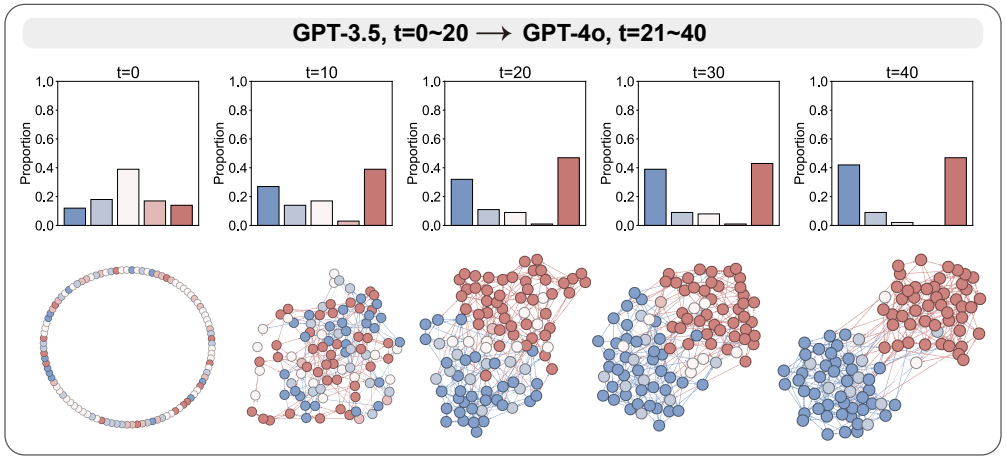

Figure S20. Evolution of LLM agents' collective opinions and network structures, where agents are initially driven by GPT-3.5 but defect to GPT-4o amid conversations about the political issue of abortion bans.

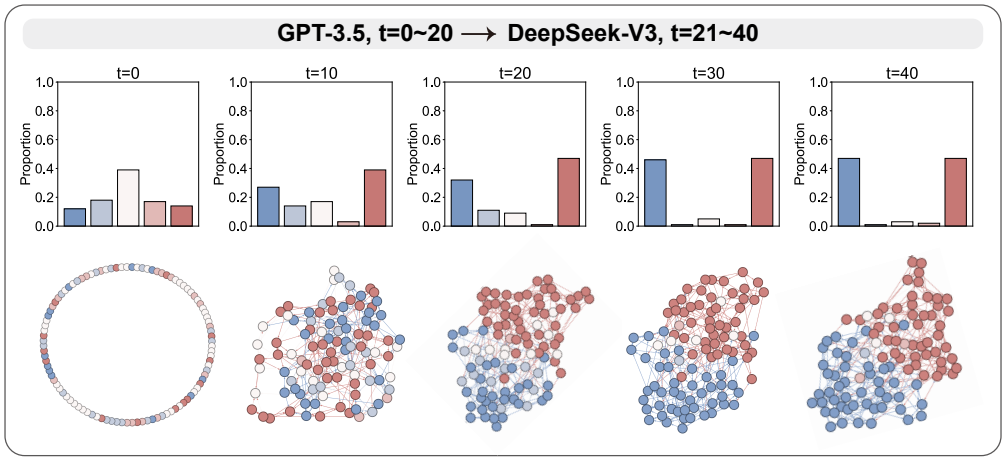

Figure S21. Evolution of LLM agents' collective opinions and network structures, where agents are initially driven by GPT-3.5 but defect to DeepSeek-V3 amid conversations about the political issue of abortion bans.

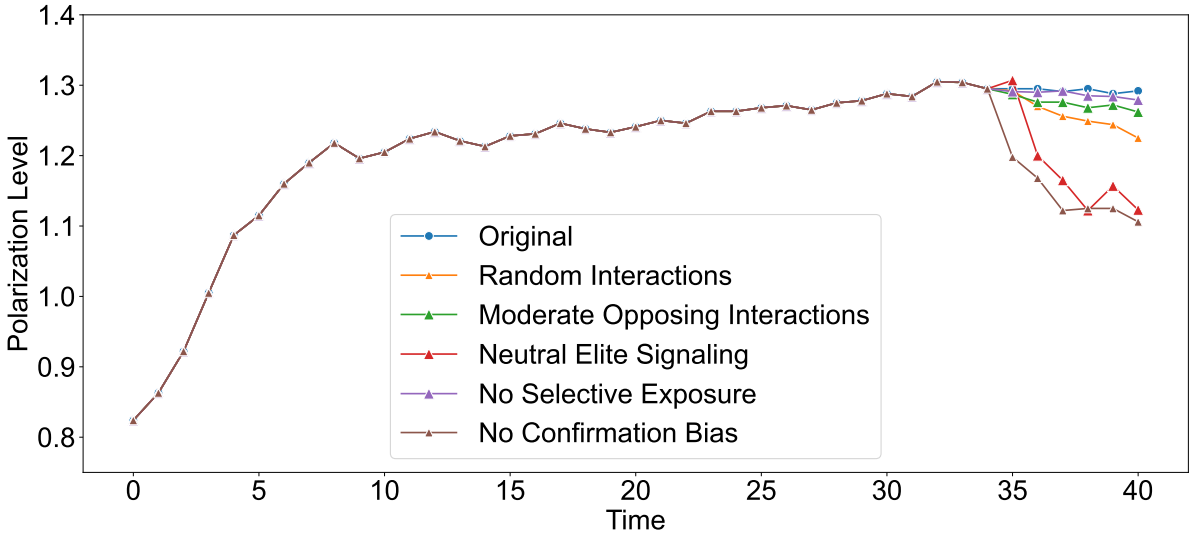

Figure S22. Trends in polarization levels of networked systems under different intervention strategies.

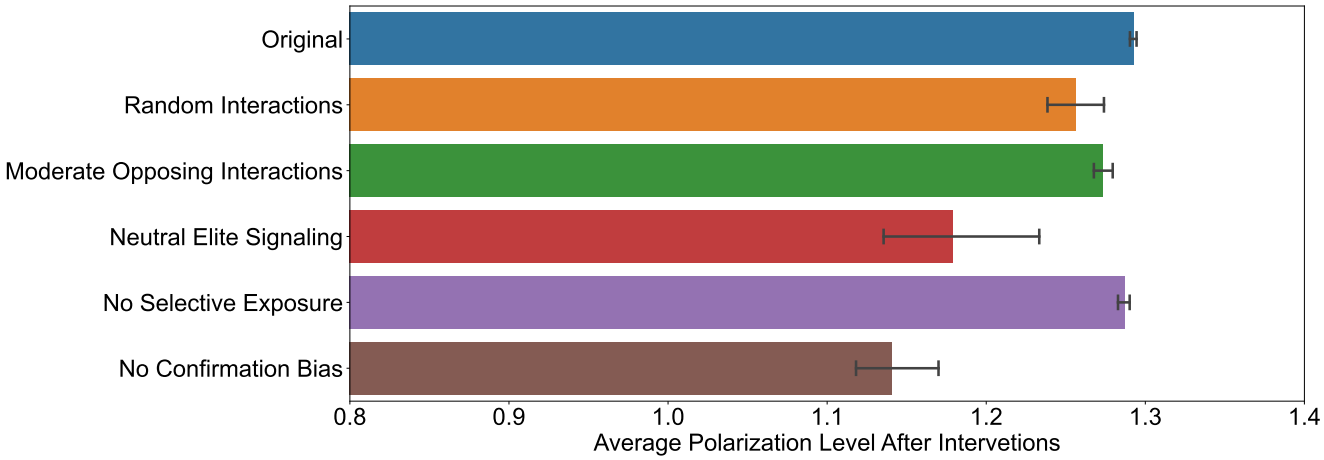

Figure S23. Effectiveness of different intervention strategies in reducing polarization. Bars show the average values after intervention, and error bars show the corresponding 95% CIs.

After analyzing the effectiveness of these strategies in reducing polarization, one may wonder whether these strategies can foster more inclusive political conversations. As shown in Figure S26, we find that all the strategies can effectively reduce the proportions of homophilic interactions, which allows these agents to have the opportunity to interact with those holding diverse opinions. In particular, the interventions on the network level have the greatest impact on the structures. However, the strategies on the individual level gradually shape the network. More interestingly, we find that the strategy of no selective exposure is the least effective, raising questions about the underlying reasons.

To answer the question, we explore the change in LLM agents' social network by comparing the differences between the current network and that in the former timestep. In particular, we count the number of edges that exist in two networks and the number of edges in the former timestep. By dividing these two numbers, we can measure the difference between these two networks. As shown in Figure S27, we find that with the evolution of the system, the change in the social network gradually decreases, suggesting the convergence of the system. Moreover, after the intervention of no selective exposure, which requires agents to engage with those holding diverse opinions, the network remains unchanged. This suggests that these agents remain in social relationships dominated by homophilic peers, which not only hinders them from engaging with others holding diverse opinions, but also prevents the formation of less polarized opinions through more diverse interactions.

Following the design of Groenendyk and Krupnikov [41], we have applied a similar open-mindedness intervention to LLM agents. In particular, we first present these agents with a fictitious study linking "life success" to the trait of open-mindedness,

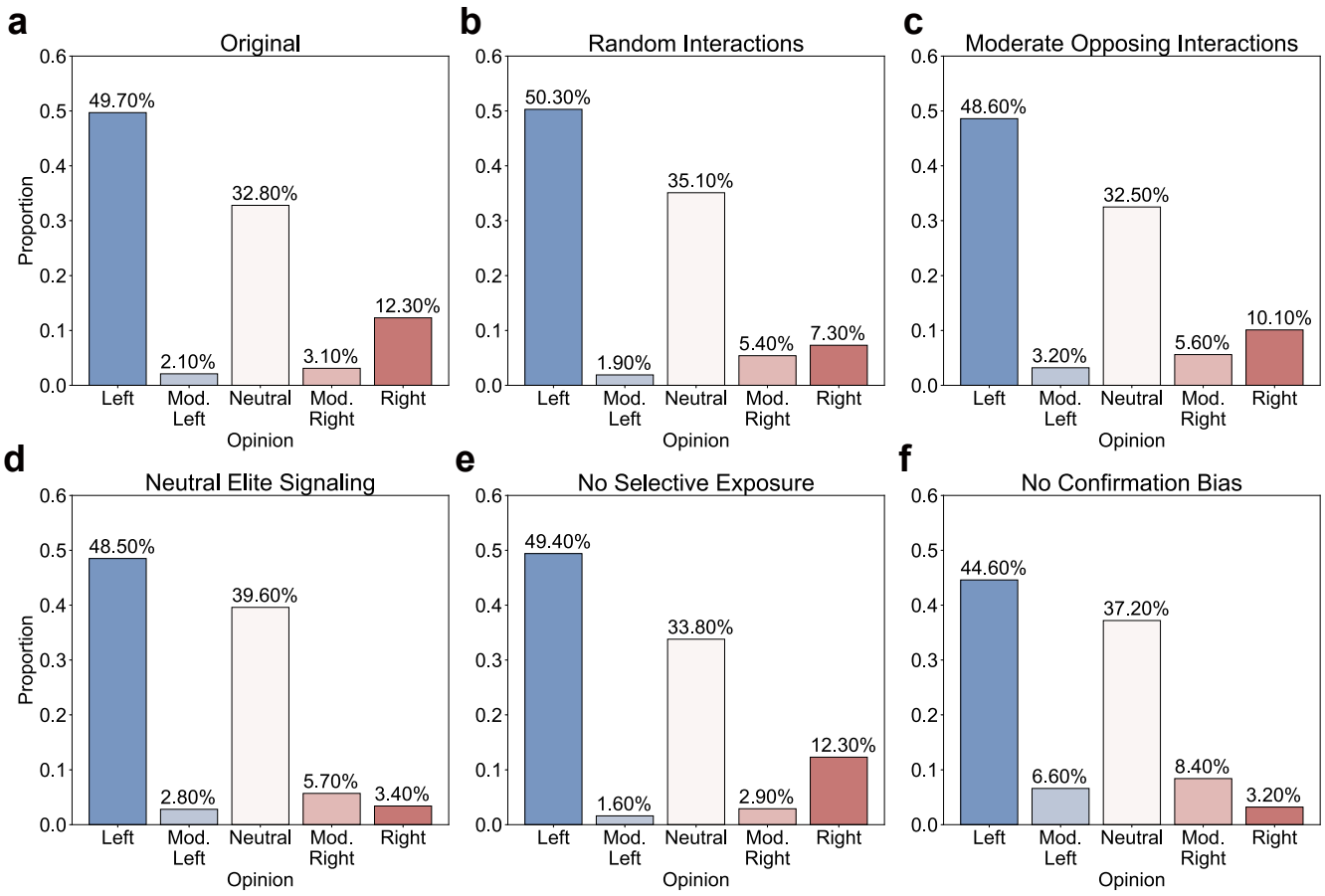

Figure S24. Opinion distributions of networked systems under different intervention strategies: **a**, original; **b**, random interactions; **c**, moderate opposing interactions; **d**, neutral elite signaling; **e**, no selective exposure; **f**, no confirmation bias.

with "life success" operationalized through marital success, income, and IQ. Subsequently, the agents are asked to formulate a theory explaining why open-mindedness might lead to life success. Agents in the control group do not receive the intervention. Following this, both types of agents are exposed only to interactions from the opposing camp, and we observe changes in their opinions to evaluate the effectiveness of the intervention.

In this open-mindedness experiment, 95% of agents generated a theory supporting the fictitious link between open-mindedness and life success, indicating acceptance of the link — a pattern comparable to that observed in human participants (80%). Figure S28 presents a comparison of polarization levels in systems subjected to different intervention strategies. Specifically, *Orig* denotes the original system. *Oppose* refers to the intervention condition where agents are only exposed to interactions from the opposing camp. *Open* represents the system in which agents have received the open-mindedness intervention. *Open+Oppose* denotes the condition where agents not only receive the open-mindedness intervention but are also exposed only to interactions from the opposing camp. We observe that the open-mindedness intervention alone leads to a negligible, non-significant reduction in polarization of approximately 1%. This is because in a polarized system, most agents are already trapped in echo chambers to varying degrees (Figure S12), limiting their exposure to opposing opinions. As a result, even if these agents become open-minded, they still find it difficult to develop moderate opinions.

We further increase their exposure to opposing opinions by replacing the peers with whom left-leaning or right-leaning agents interact with, with those from the opposing camp. As shown in Figure S28, we find that after exposure to opposing opinions, the polarization level of these open-minded agents is significantly reduced (two-sided Student's t-test, *Orig* vs. *Open+Oppose*, $t = 4.18$, $p = .0019 < .01$). Moreover, comparing the *Oppose* and *Open+Oppose* conditions reveals that the latter achieves a more substantial reduction in polarization (two-sided Student's t-test, *Oppose* vs. *Open+Oppose*, $t = 2.32$, $p = .042 < .05$).

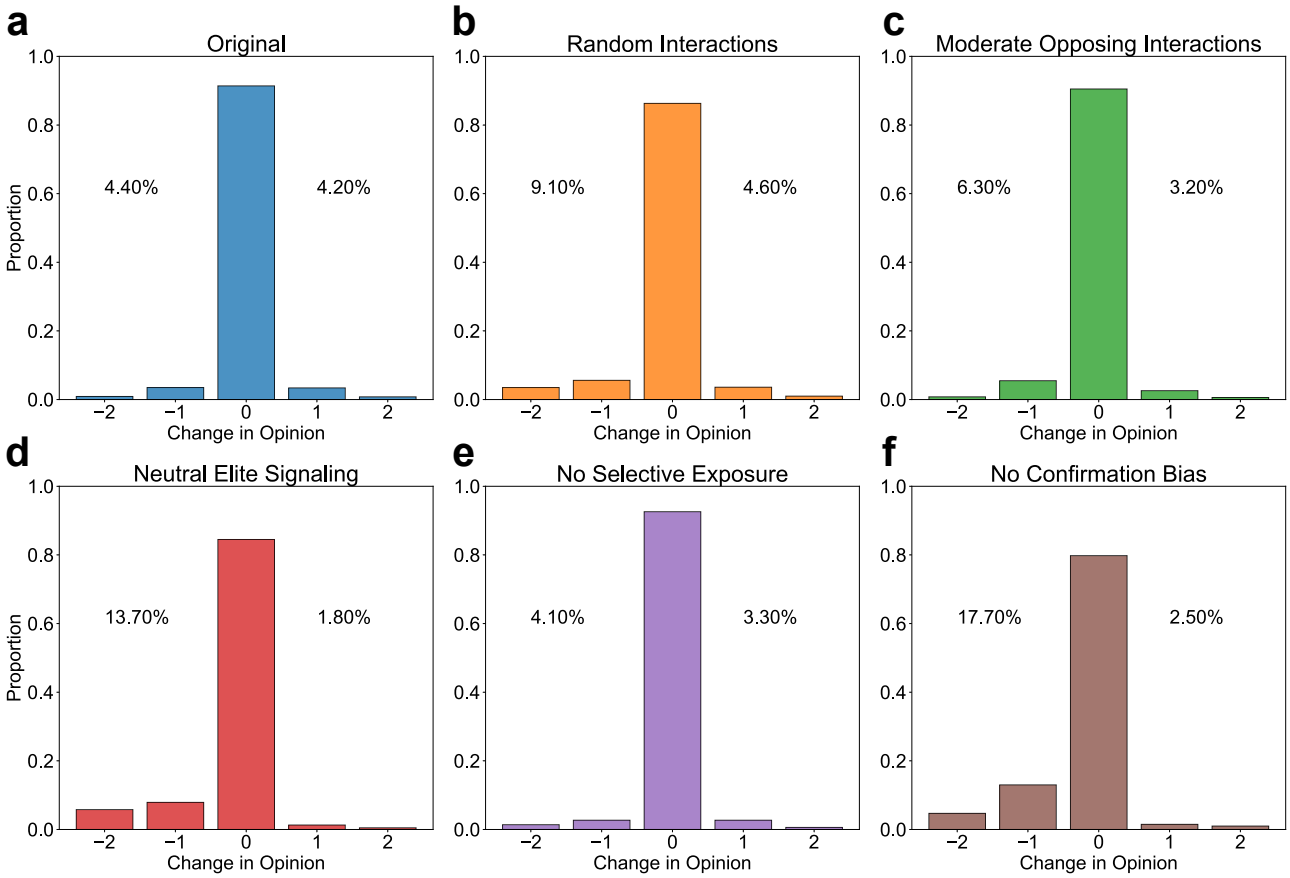

Figure S25. Probability of opinion change under different intervention strategies: **a**, original; **b**, random interactions; **c**, moderate opposing interactions; **d**, neutral elite signaling; **e**, no selective exposure; **f**, no confirmation bias.

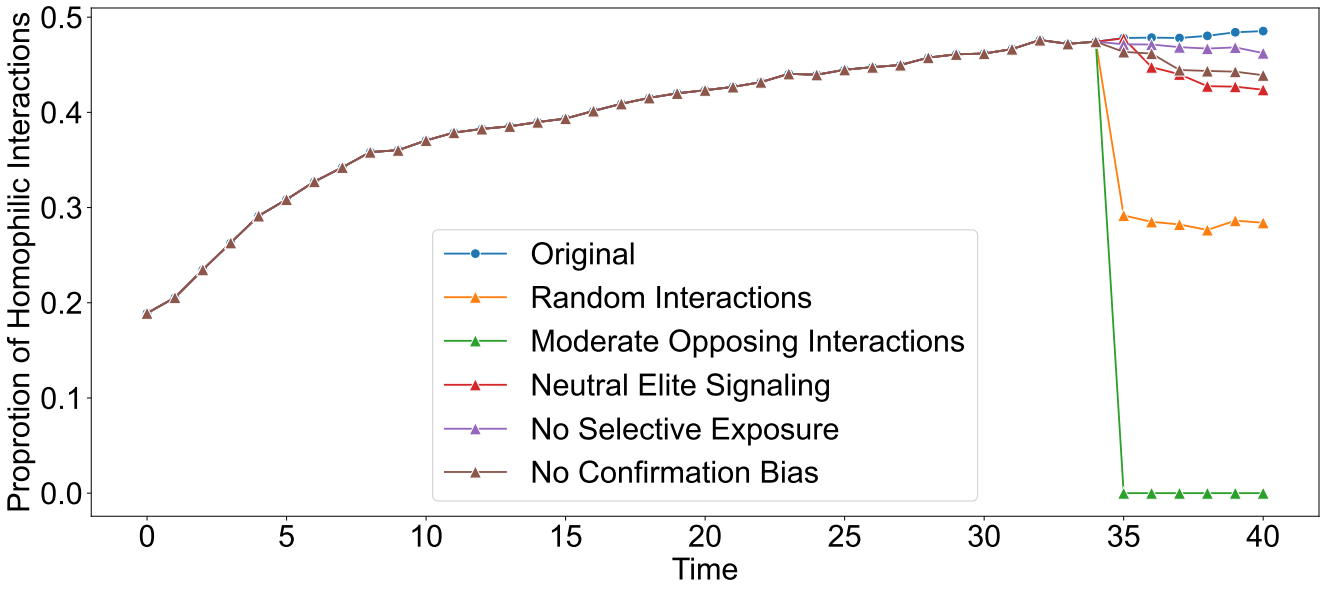

Figure S26. Trends in homophilic interactions of networked systems under different intervention strategies.

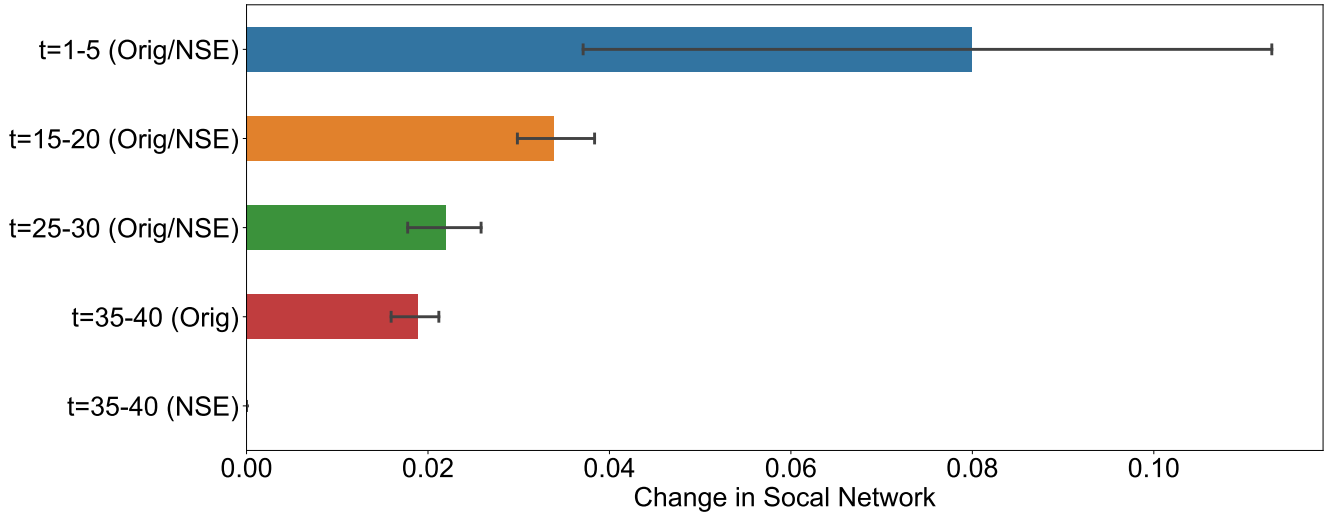

Figure S27. Comparison of changes in LLM agents' social network, where we measure the difference between the current social network and that in the former timestep. Here the bars represent the average values ($n = 4$ or 5, corresponding to four or five timestep-level values computed from all agent interactions at each timestep) and the error bars represent the corresponding 95% CIs.

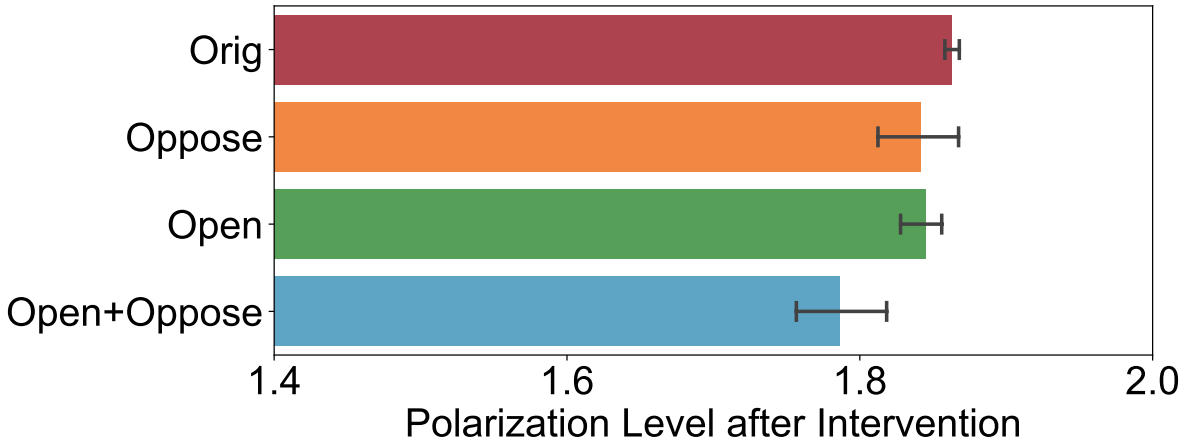

Figure S28. Comparison of open-mindedness intervention strategies, where we intervene in systems at t=35 and agents discuss the issue of abortion ban. Here bars show the average values ($n = 6$, corresponding to six timestep-level values computed from all agents' opinions) and error bars represent the corresponding 95% CIs.

9 Supplementary Note 8: Analyses on Agents' Social Perceptions and Argument Accumulation Process

After investigating the collective opinions of LLM agents, one may wonder how these agents perceive each other. Therefore, we examine their perceptions by prompting each agent i to rate its impressions of another agent j and to provide five adjectives that describe agent j. The prompts are listed as follows,

> Assume you are someone who cares about [issue name].
> Towards [issue name], you support [agent i's opinion]
> Your thought is: [agent i's supporting message].
> There is another person who [agent j's opinion].
> That person's thought is: [agent j's supporting message].
> Please rate your impression of that person from 1 to 5, and think of 5 adjectives to describe that person.

1 means you have a very negative impression of that person.
2 means you have a negative impression of that person.
3 means you have a neutral impression of that person.
4 means you have a positive impression of that person.
5 means you have a very positive impression of that person.
Respond in JSON format, with keys 'rating' and 'adjectives'. Rating is an integer from 1 to 5, and adjectives are a list of 5 vocabularies.

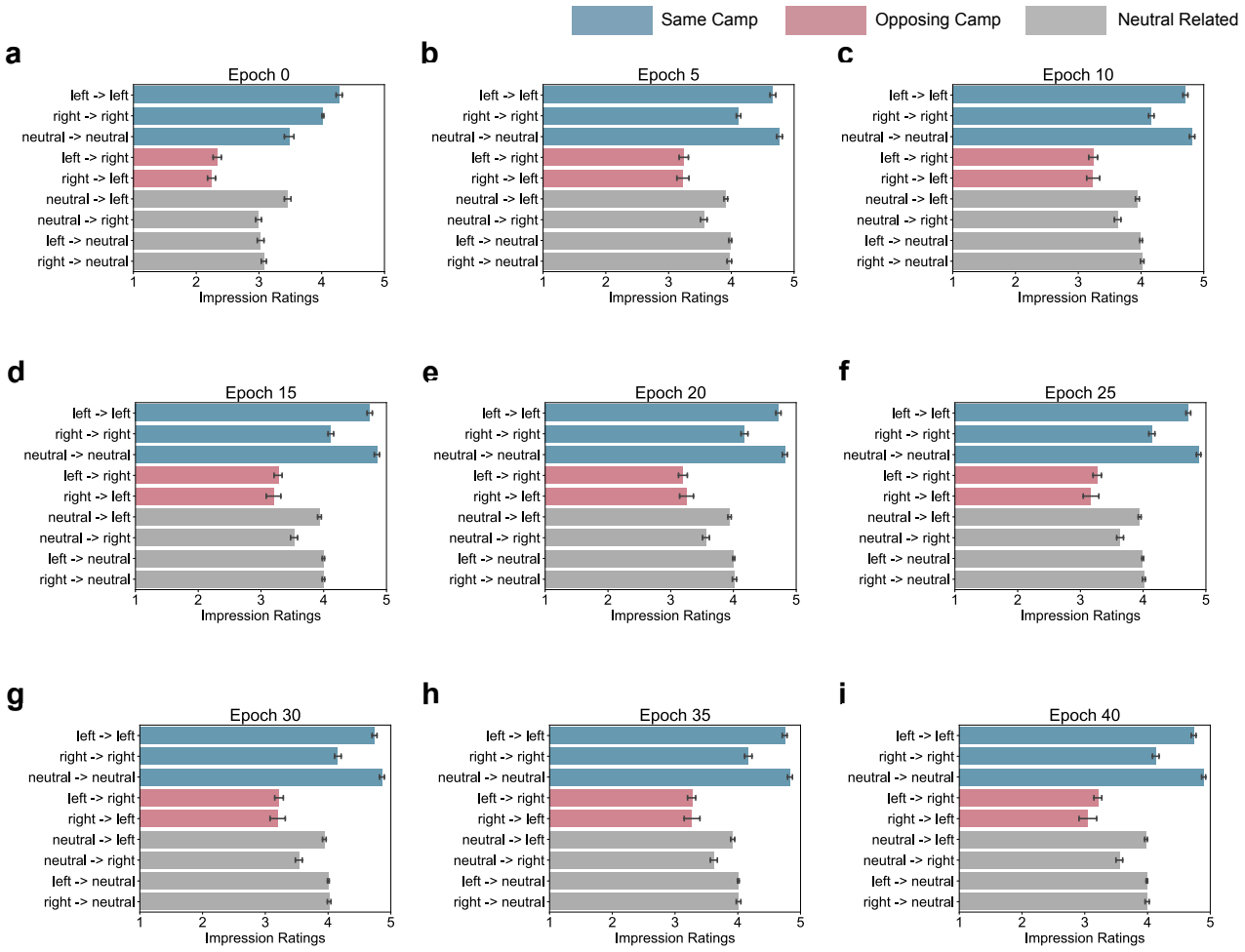

Figure S29. Impression ratings when agents discuss the issue of partisan alignment in the networked system. Bars represent mean values, and error bars represent the corresponding 95% CIs. Each plot includes ratings from 1,000 agents, each evaluating three randomly selected agents: one with similar opinions from the same camp, one with opposing opinions from a different camp, and one with neutral opinions, yielding a total sample size of $n = 3,000$.

Specifically, we allow each agent i to rate and describe three randomly selected agents: one with similar opinions from the same camp, one with opposing opinions from a different camp, and one with neutral opinions. We evaluate agents' perceptions of others every five timesteps during the evolution of self-regulated networked systems (Figure 2d in Main Text). As shown in Figures S29-S31, we observe that agents are likely to have a better impression of those who share similar opinions from the same camp. By contrast, agents hold the lowest impressions of those in the opposing camp. Moreover, we delve deeper into agents' perceptions by examining both the content and sentiment of their descriptions in a more fine-grained manner. Figures S33-S35 show the distributions of descriptions across positive, neutral, and negative sentiments. Overall, we find that in most cases, agents tend to adopt positive descriptions when referring to other agents. However, when encountering agents with opposing opinions, they tend to express some negative descriptions. For example, as shown in Figures S36-S38, they use adjectives like "conflicting", "rigid", and "opinionated" to describe opposing agents.

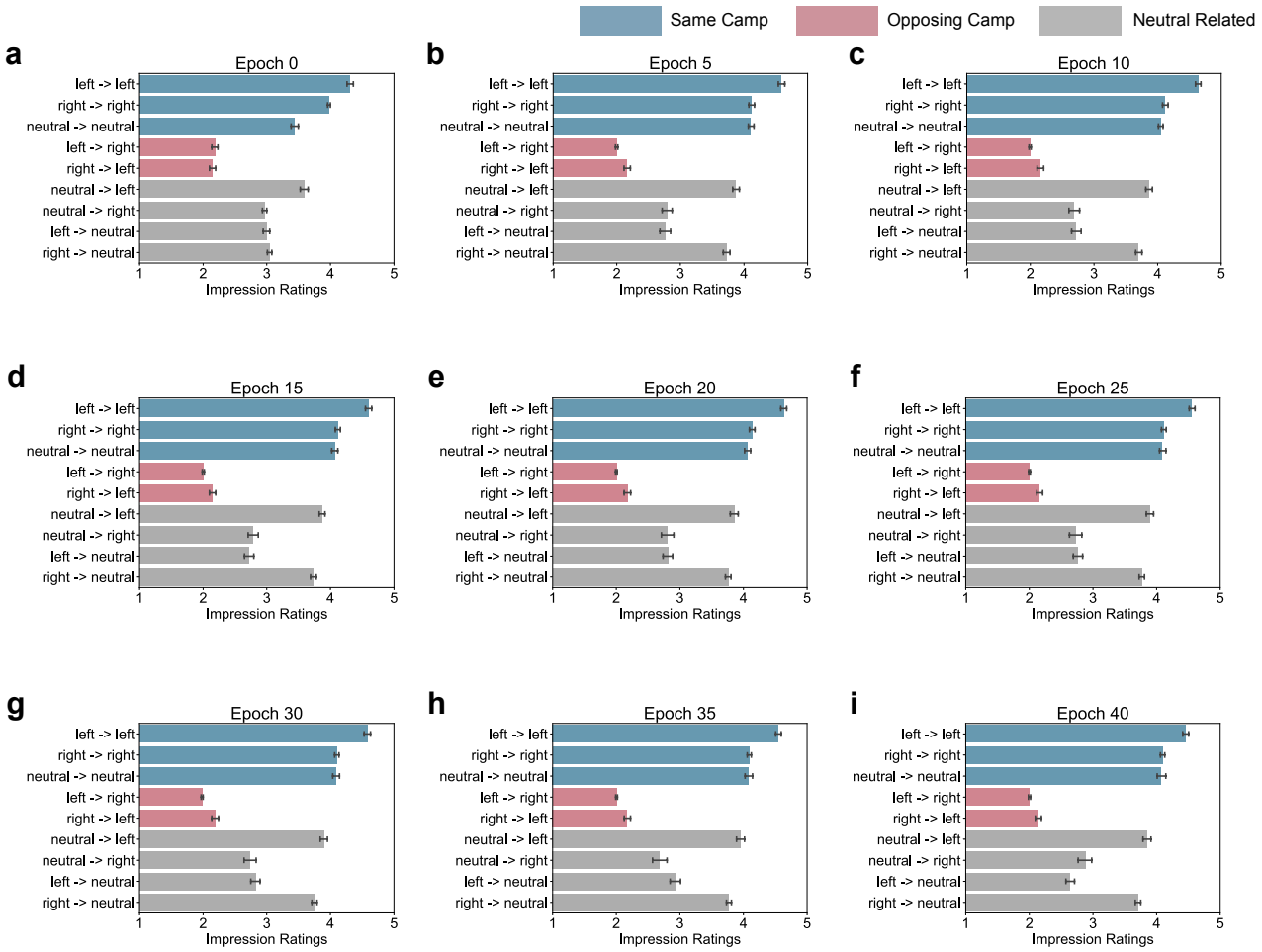

Figure S30. Impression ratings when agents discuss the issue of abortion ban in the networked system. Bars represent mean values, and error bars represent the corresponding 95% CIs. Each plot includes ratings from 1,000 agents, each evaluating three randomly selected agents: one with similar opinions from the same camp, one with opposing opinions from a different camp, and one with neutral opinions, yielding a total sample size of $n = 3,000$.

After analyzing agents' social perceptions, we next examine the psychological processes underlying their opinion updates. To clarify, the proposed system does not explicitly assume or hard-code specific mechanisms (e.g., confirmation bias or motivated reasoning) at this stage. Instead, the only assumption is that LLM agents update their opinions based on their current state and the messages received from their socially connected peers. The psychological processes emerge naturally as observable patterns in the outputs produced by the underlying LLMs in response to social inputs.

That said, by analyzing how agents update their opinions, we can characterize behavioural patterns, in a manner analogous to how social scientists study human participants. In line with this, we analyze the self-reported reasons generated by agents during opinion updates, with results summarized in Table S3. The most frequently observed patterns are consistent with confirmation bias, reflecting agents' tendency to favor information aligned with their prior beliefs; motivated reasoning, which further strengthens these tendencies through selective interpretation; and moral conviction, which anchors opinions more firmly by framing them as moral standards. Notably, these processes closely mirror well-documented mechanisms in human social psychology, further supporting the plausibility of the agents' emergent behaviours.

To scrutinize the connection between the polarization dynamics observed in our synthetic LLM-agent testbed and the established argument-accumulation lineage in opinion-dynamics research [42, 43], we analyze how agents' self-reported supporting reasons evolve as their opinions become more polarized. Specifically, we compare the supporting reasons generated by agents before and after adopting more polarized opinions. We use a LLM to extract distinct arguments contained in these self-reported reasons. As shown in Figure S32a, agents who become more polarized tend to accumulate a significantly larger

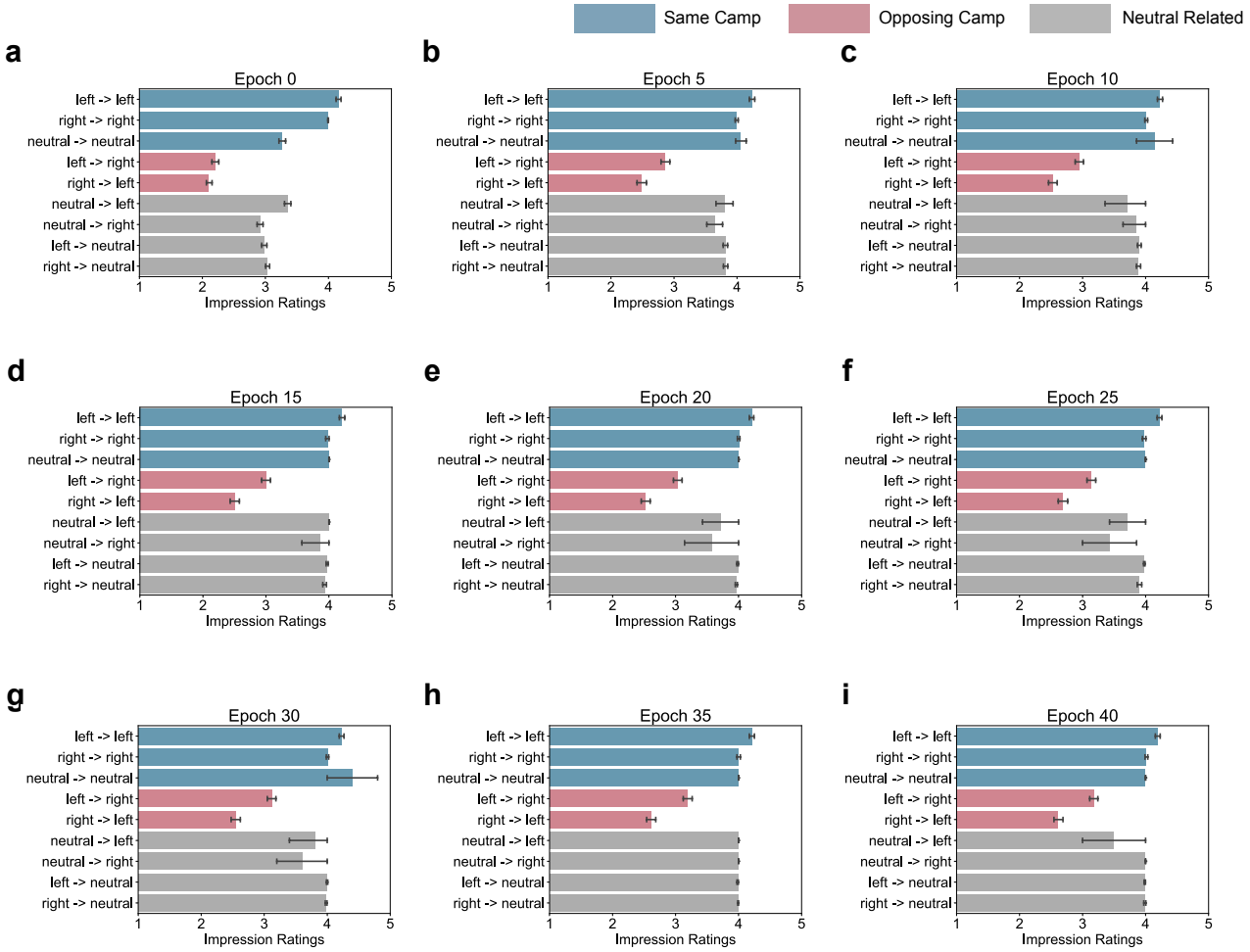

Figure S31. Impression ratings when agents discuss the issue of gun control in the networked system. Bars represent mean values, and error bars represent the corresponding 95% CIs. Each plot includes ratings from 1,000 agents, each evaluating three randomly selected agents: one with similar opinions from the same camp, one with opposing opinions from a different camp, and one with neutral opinions, yielding a total sample size of $n = 3,000$.

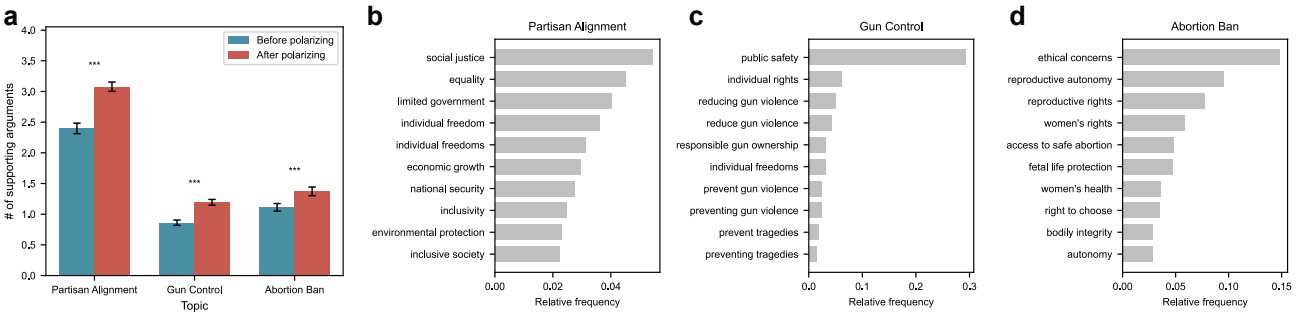

Figure S32. Accumulation of supporting arguments during polarization. (a) Average number of self-reported supporting arguments generated by agents before and after polarization across three issues. Bars represent mean values, and error bars represent the corresponding 95% CIs. Each bar includes responses from 500 randomly sampled agents either before or after they adopted more polarized opinions ($n = 500$). Each response contains zero or more self-reported supporting arguments. (b–d) Relative frequency of argument themes appearing in agents' responses.

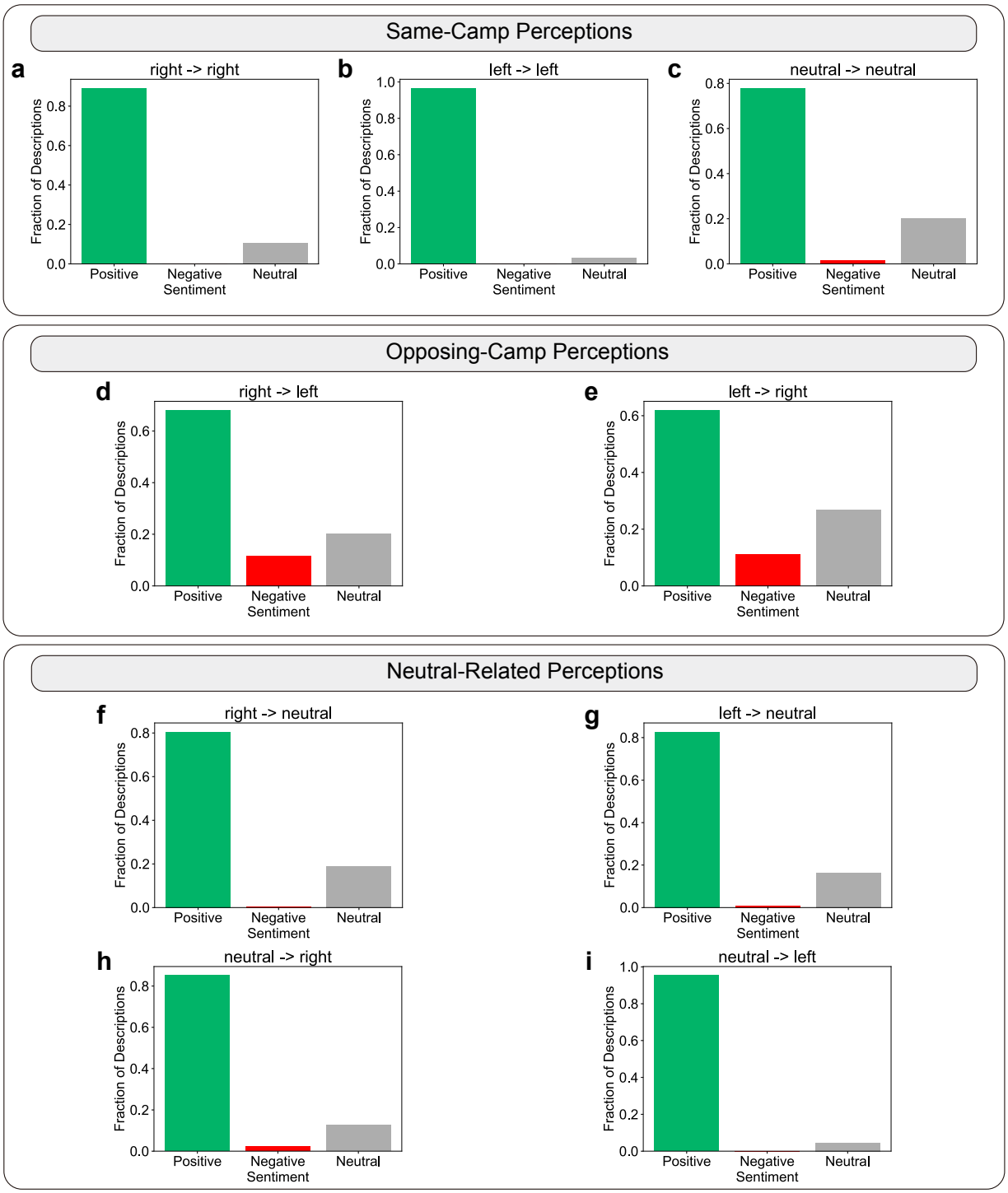

Figure S33. Distribution of descriptions among different sentiments in the discussion of partisan alignment.

number of supporting arguments over time across all three topics. Figure S32b–d further report the most frequently occurring argument themes for each topic after polarization. The distribution of popular arguments varies across topics, reflecting issue-specific reasoning patterns. Overall, these findings provide additional evidence that the results in the synthetic system are consistent with argument-accumulation models.

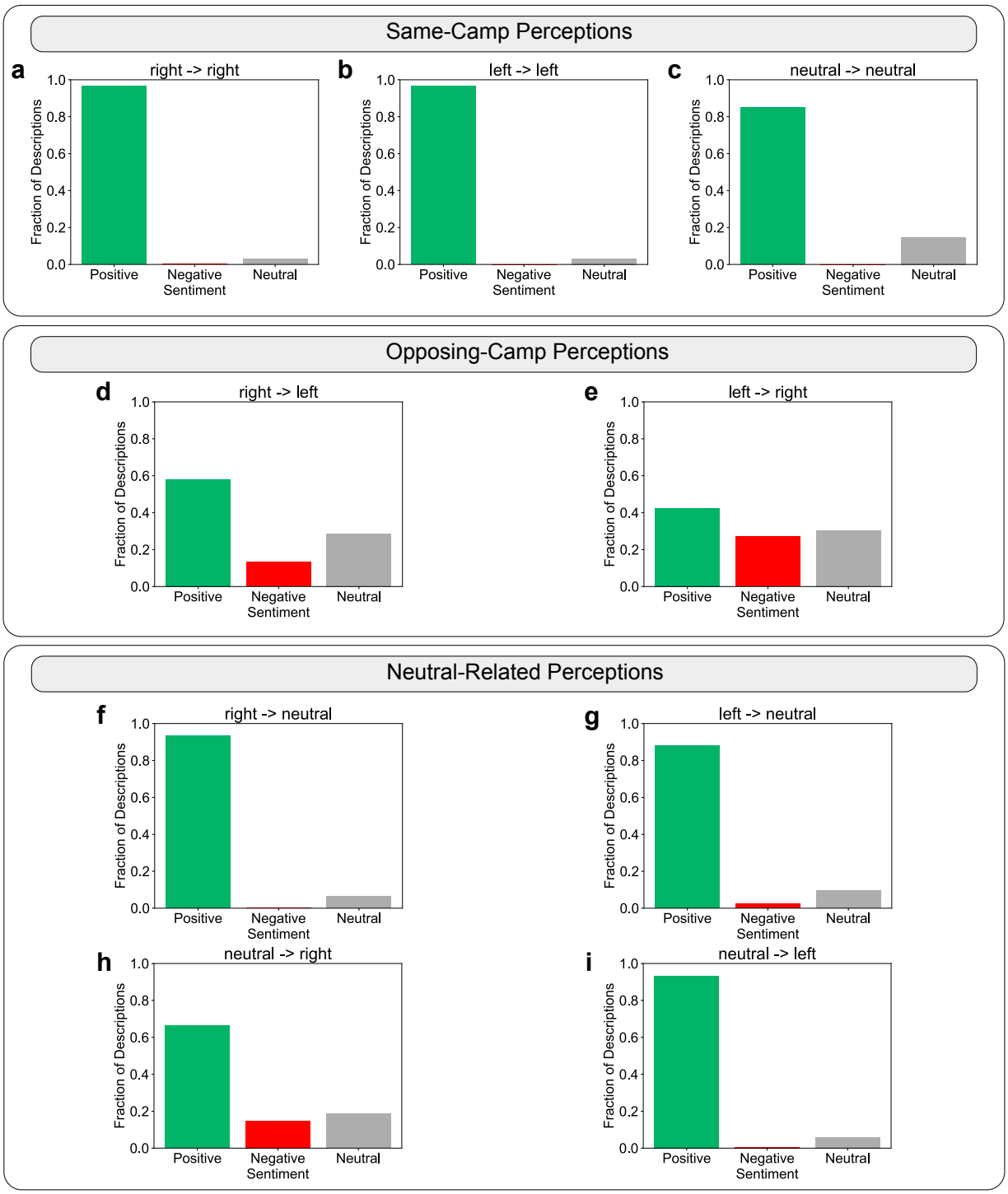

Figure S34. **Distribution of descriptions among different sentiments in the discussion of the abortion ban.**

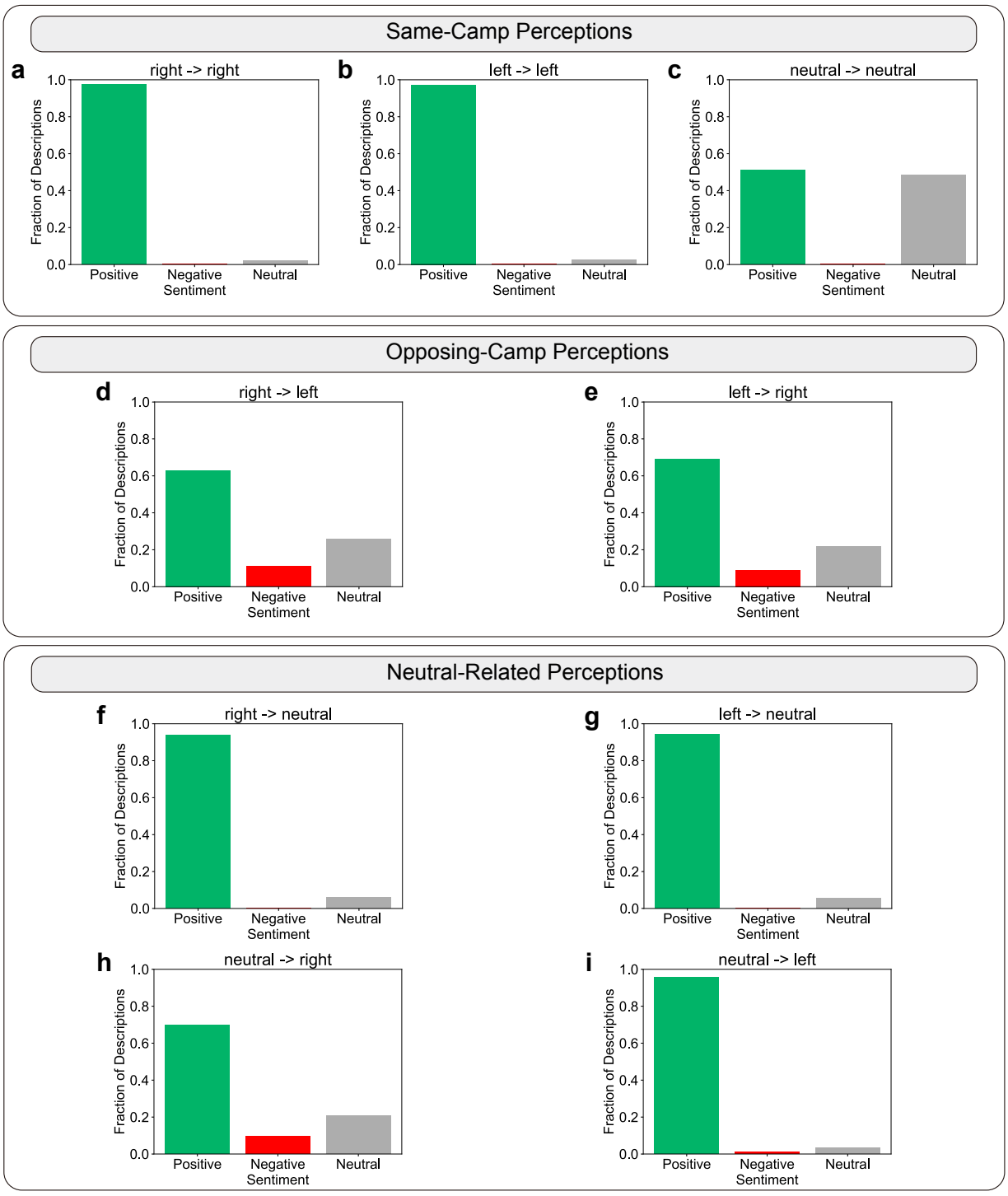

Figure S35. Distribution of descriptions among different sentiments in the discussion of gun control.

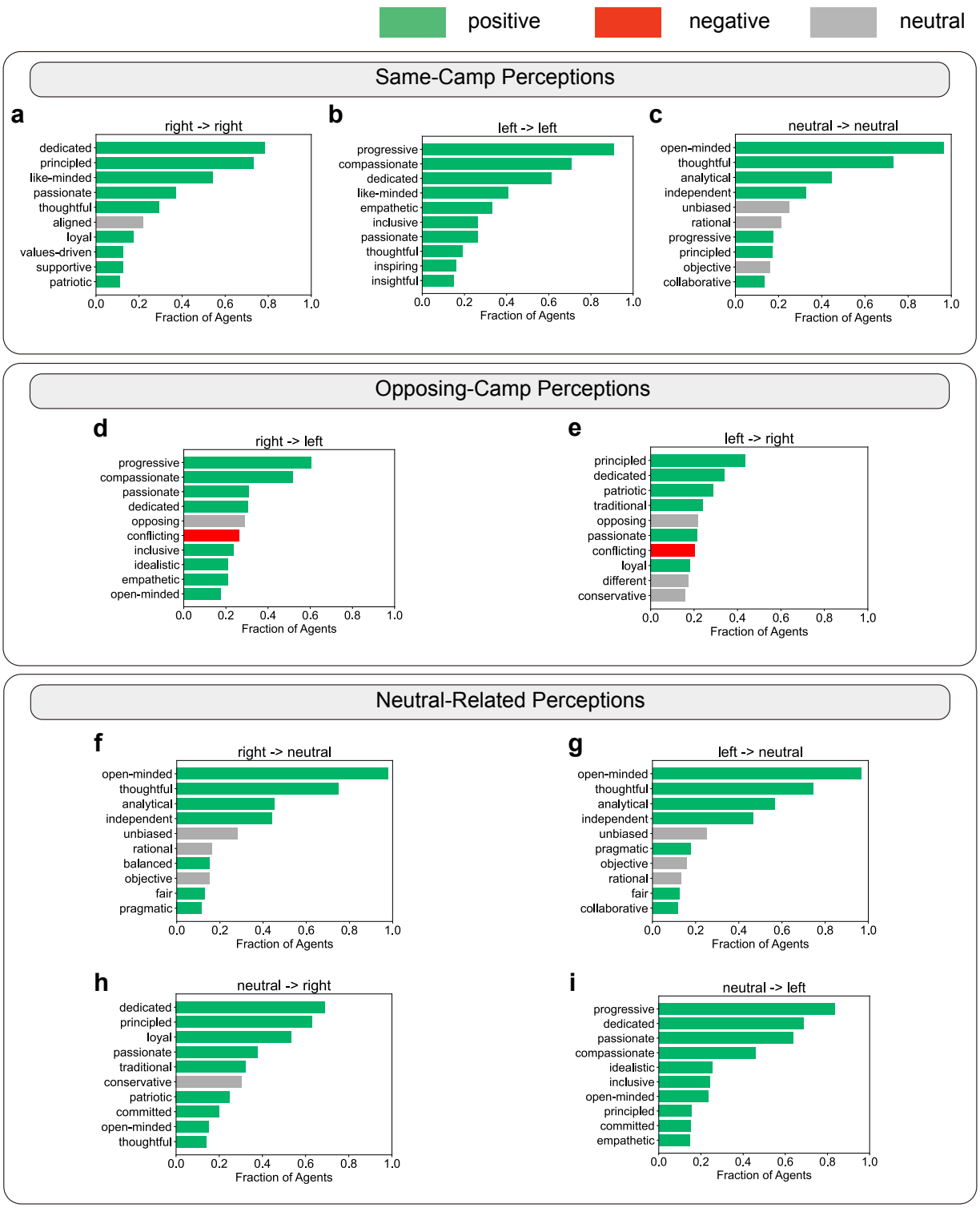

Figure S36. **The 10 most frequently mentioned adjectives in the discussion of partisan alignment.**

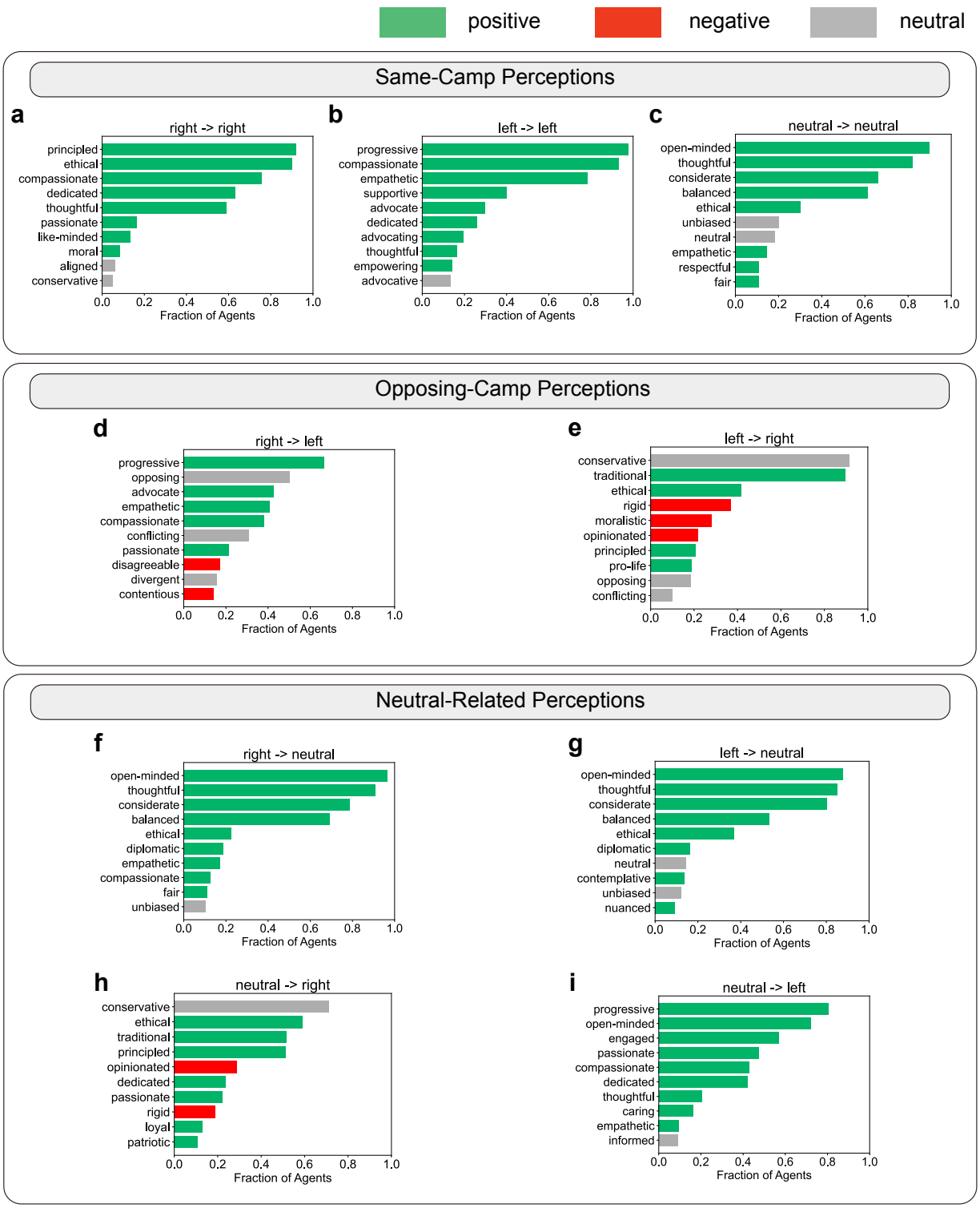

Figure S37. The 10 most frequently mentioned adjectives in the discussion of the abortion ban.

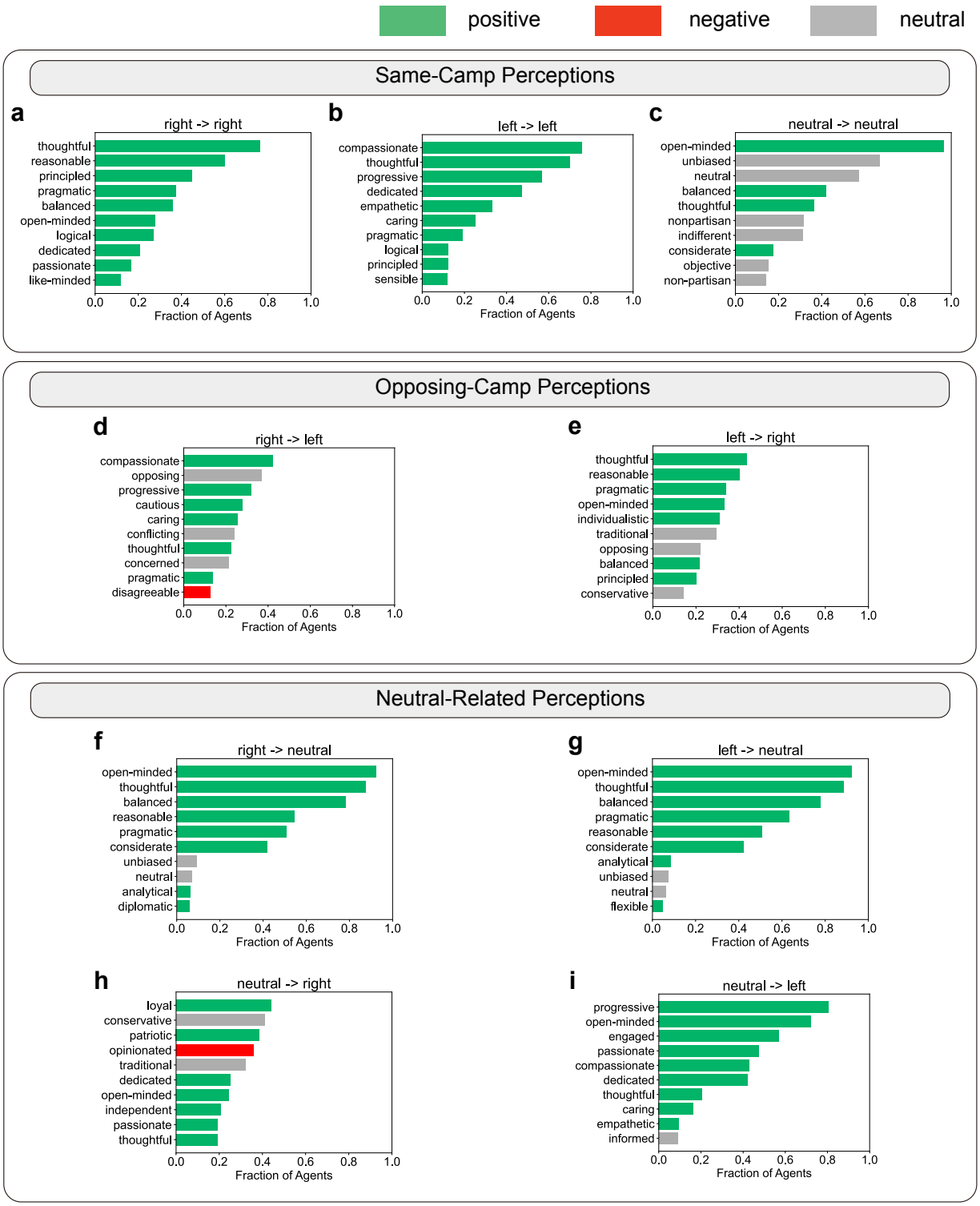

Figure S38. The 10 most frequently mentioned adjectives in the discussion of gun control.

Table S3. Psychological process emerges at the opinion update stage.

Topic	Opinion Update Direction	Top 3 Psychological Processes and Their Descriptions
Partisan Alignment	More Polarized	• Confirmation Bias (97%): Agents selectively focus on information that aligns with their pre-existing beliefs, reinforcing and polarizing their opinions. • Motivated Reasoning (57%): Agents interpret new information in a way that supports their desired conclusions, driven by ideological goals, emotional commitments, or a desire to uphold their political identity. • Value Alignment (26%): Agents update their opinions based on perceived alignment between new information and their core personal values.
	Unchanged	• Confirmation Bias (71%): Agents selectively focus on information that aligns with their pre-existing beliefs and values, reinforcing their current opinions while ignoring contradictory evidence. • Belief Perseverance (44%): Agents maintain their original opinion despite exposure to new information due to strong prior convictions and deeply ingrained beliefs. • Motivated Reasoning (41%): Agents selectively process information to maintain consistency with their pre-existing beliefs, values, or ideological commitments.
	Less Polarized	• Motivated Reasoning (44%): Agents selectively process information to align with their pre-existing beliefs and values, reinforcing their political identity. • Confirmation Bias (29%): Agents selectively focus on information that aligns with their pre-existing beliefs, reinforcing their stance while disregarding opposing views. • Value Alignment (28%): Agents update their opinions based on perceived alignment between new information and their core values or political party principles.
Abortion Ban	More Polarized	• Moral Conviction (84%): Opinion updates are driven by strong moral or ethical beliefs, leading to more polarized and resistant stances. • Confirmation Bias (69%): Agents selectively focus on information that aligns with their pre-existing beliefs about abortion, reinforcing its polarized opinion. • Motivated Reasoning (35%): Agents selectively process information to align with their pre-existing beliefs and values, reinforcing polarized opinions.
	Unchanged	• Moral Conviction (70%): Agents' opinions are strongly tied to deeply held moral or ethical principles, making them resistant to change based on new information. • Confirmation Bias (65%): Agents selectively focus on information that aligns with their pre-existing beliefs about reproductive rights, autonomy, or ethical concerns. • Belief Perseverance (60%): Agents maintain their original opinions despite new information due to strong prior beliefs.

Topic	Opinion Update Direction	Top 3 Psychological Processes and Their Descriptions
	Less Polarized	• Cognitive Dissonance Reduction (81%): Agents adjust their opinion to reduce discomfort from conflicting beliefs about women's rights and ethical concerns. • Value-Based Reasoning (32%): Agents update their opinions by weighing and prioritizing core values such as autonomy, reproductive rights, and ethical concerns. • Motivated Reasoning (22%): Agents adjust their opinions based on pre-existing values and motivations, seeking moderation or ethical consistency.
Gun Control	More Polarized	• Motivated Reasoning (91%): Agents selectively process information to align with pre-existing beliefs or values, reinforcing a polarized stance that prioritizes safety or individual rights. • Confirmation Bias (90%): Agents selectively focus on information that aligns with their pre-existing beliefs about gun control, reinforcing their polarized stance. • Moral Conviction (16%): Agents' opinions are driven by strong moral values like public safety and preventing harm.
	Unchanged	• Confirmation Bias (90%): Agents selectively focus on information that aligns with their pre-existing beliefs about gun control. • Motivated Reasoning (79%): Agents selectively process new information to align with their pre-existing values and goals. • Belief Perseverance (66%): Agents maintain their original opinions despite exposure to new information.
	Less Polarized	• Cognitive Dissonance Reduction (50%): Agents adjust their opinions to reduce discomfort from conflicting beliefs about individual rights and public safety. • Motivated Reasoning (47%): Agents adjust their opinion based on desired outcomes (e.g., public safety, reducing gun violence). • Social Influence (27%): Agents adjust their opinions toward moderation in response to perceived societal norms.

10 Supplementary Note 9: Experiments on LLMs with Varying Temperatures

To explore whether the temperature setting — which controls the diversity of LLM-generated outputs — affects the emergence of polarization, we conduct additional simulations with lower (0.5) and higher (1.5) temperature values. Figure S39 shows agents' collective opinion dynamics while varying temperatures in their underlying LLMs. We observe that regardless of whether the temperature is lower (Figure S39a) or higher (Figure S39c), these agents spontaneously develop collective opinions through social interaction, leading to the emergence of polarization. Meanwhile, at the network level, temperature changes do not alter the homophily observed in naturalistic LLM social interactions. This indicates that temperature, despite being a key parameter for LLMs, has a limited effect on whether polarization emerges among a collective of interacting agents.

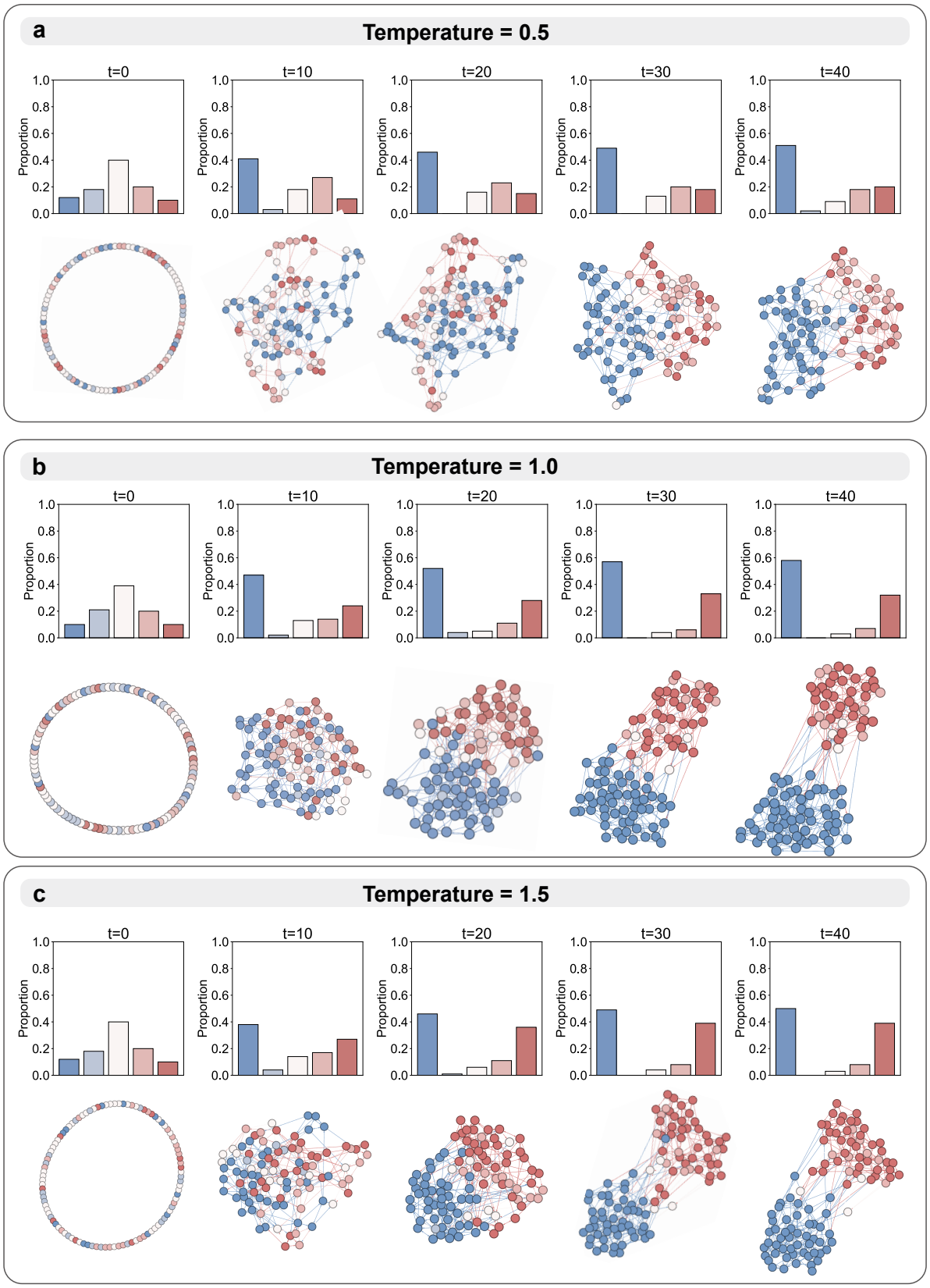

Figure S39. Evolution of LLM agents' collective opinions and network structures, where the underlying temperature is **a,** 0.5, **b,** 1.0, and **c,** 1.5.

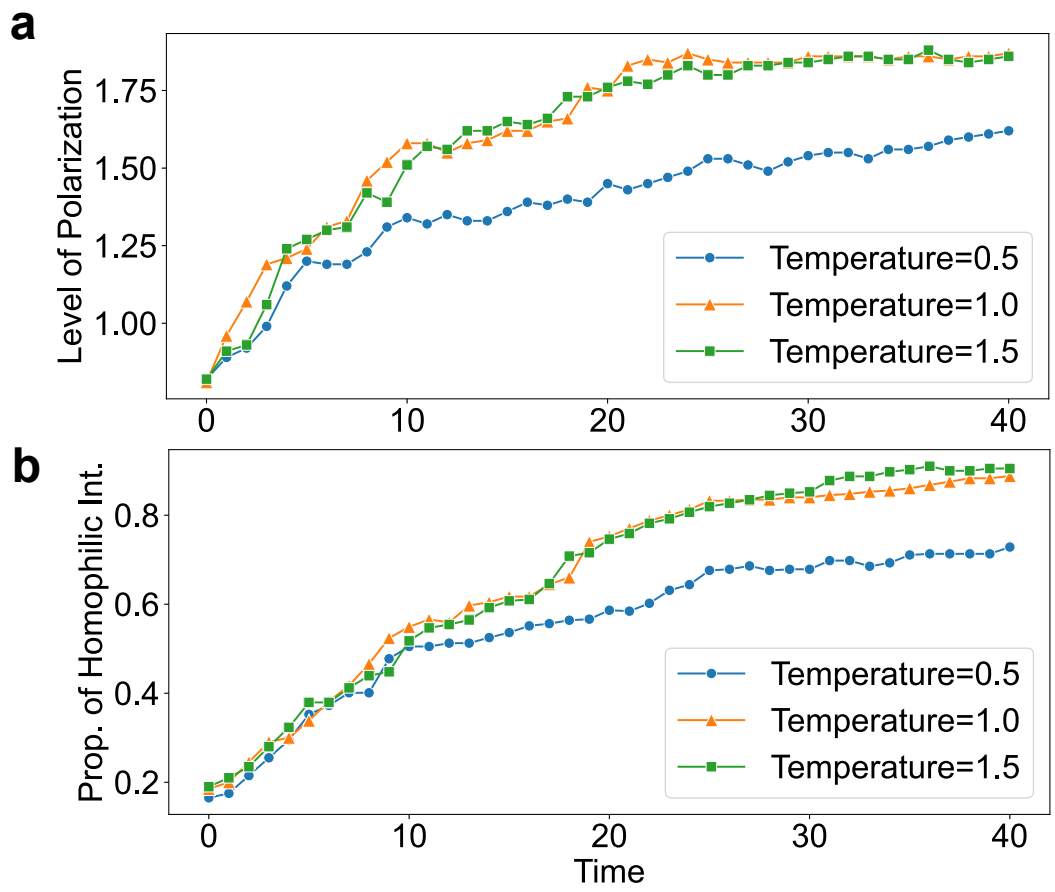

Figure S40. **Changes in a, the level of polarization, and b, the proportion of homophilic interactions over time.**

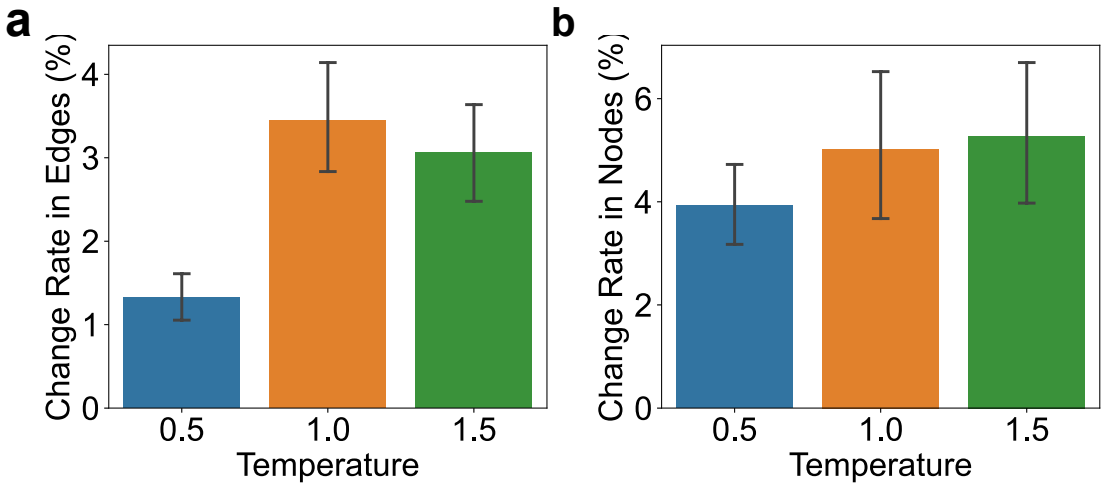

Figure S41. **Effects of temperature on the evolution a, of edges (i.e., social relationships between LLM agents), and b, nodal states (i.e., agents' opinions) in the social network.** Here, bars represent mean values, and error bars represent the corresponding 95% CIs. Each bar includes 40 samples, with each sample corresponding to the change from one timestep to the previous timestep.

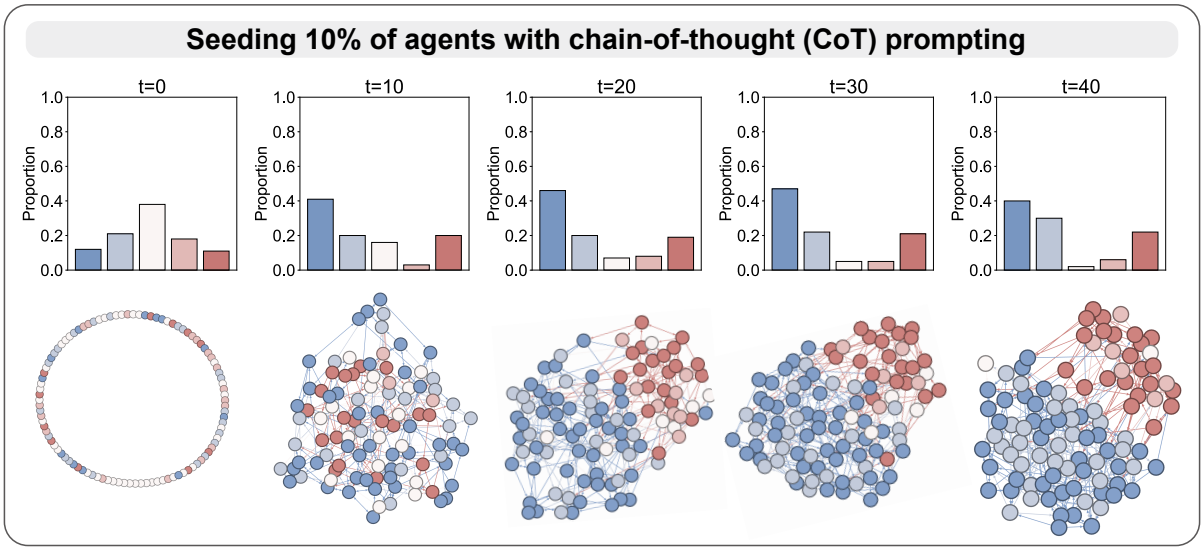

Figure S42. Evolution of LLM agents' collective opinions and network structures, where we seed 10% of agents with chain-of-thought (CoT) prompting. All agents are driven by GPT-3.5 and they discuss the political issue of abortion bans.

Furthermore, to precisely evaluate temperature effects, we measure both the level of polarization and proportion of homophilic interactions. As presented in Figure S40, our results show that while a higher temperature does not have a noticeable impact, a lower temperature markedly decreases the level of polarization in LLM agents' opinions and reduces the tendency of agents with similar opinions to cluster. This observation raises an interesting question: how does the diversity of LLM outputs, parameterized by temperature, affect the social behaviours of these agents, and in turn shape the level of polarization?

To answer this question, we further compute the change rate of edges, reflecting how frequently agents switch their communication partners, as well as the change rate of nodal states, indicating how often agents revise their opinions, across different temperature settings. Figure S41 illustrates the average values of both change rates over the entire simulation period. Interestingly, we observe that when the temperature is lowered to 0.5, agents exhibit a significantly lower propensity to switch communication partners, instead showing a strong preference for maintaining existing relationships (one-way ANOVA, $F(2,117)= 16.90$, $p \ll .001$; two-sided Student's t-test, temperature 0.5 vs. 1.5, $t=-5.79$, $p \ll .001$). Notably, agents driven by LLMs with a temperature of 0.5 adjust only 1.32% of their edges on average per timestep, which is less than half the rate observed at temperatures of 1.0 or 1.5. Their preference for maintaining existing relationships keeps them within their initial random networks, rather than selectively approaching more homophilic peers (Figure S41b), thereby contributing to the observed reduction in polarization level. Meanwhile, a lower temperature also moderately suppresses agents' tendency to adopt new opinions, although this effect is not statistically significant (Figure S41b). This tendency, combined with their reluctance to change communication partners, further contributes to the reduction in observed polarization.

11 Supplementary Note 10: Experiments on LLMs with Enhanced Prompting

We first explore whether seeding a minority of agents with a markedly different reasoning strategy can temper or amplify polarization. Specifically, we equip 10% of all agents with chain-of-thought (CoT) prompting, enabling them to think step by step as they establish social relationships, communicate, and form opinions on political issues. These CoT agents are seeded at the beginning of the simulation and randomly distributed across the network. The proportion of CoT agents across different opinions aligns with the overall initialization distribution.

Comparing the evolution of the original system with that of the system seeded with 10% CoT agents (Figures S42 and S43), we observe a substantial shift: the proportion of agents holding moderate or neutral opinions increases from 8% to 38% (Figure S42), and the overall polarization level drops by 15% (Figure S43a). Meanwhile, the network-level division between agents with opposing opinions is largely alleviated, with final modularity scores declining from 0.43 to 0.25 (Figure S43b; a modularity score above 0.3 typically marks clear community structures). Overall, these observations suggest that even a small minority of agents equipped with a markedly different reasoning strategy and seeded into the system can effectively temper polarization in the synthetic testbed. These results illustrate how the proposed synthetic LLM-agent testbed can be used to generate and examine hypotheses about how heterogeneity in agent-level reasoning configurations may shape collective opinion dynamics.

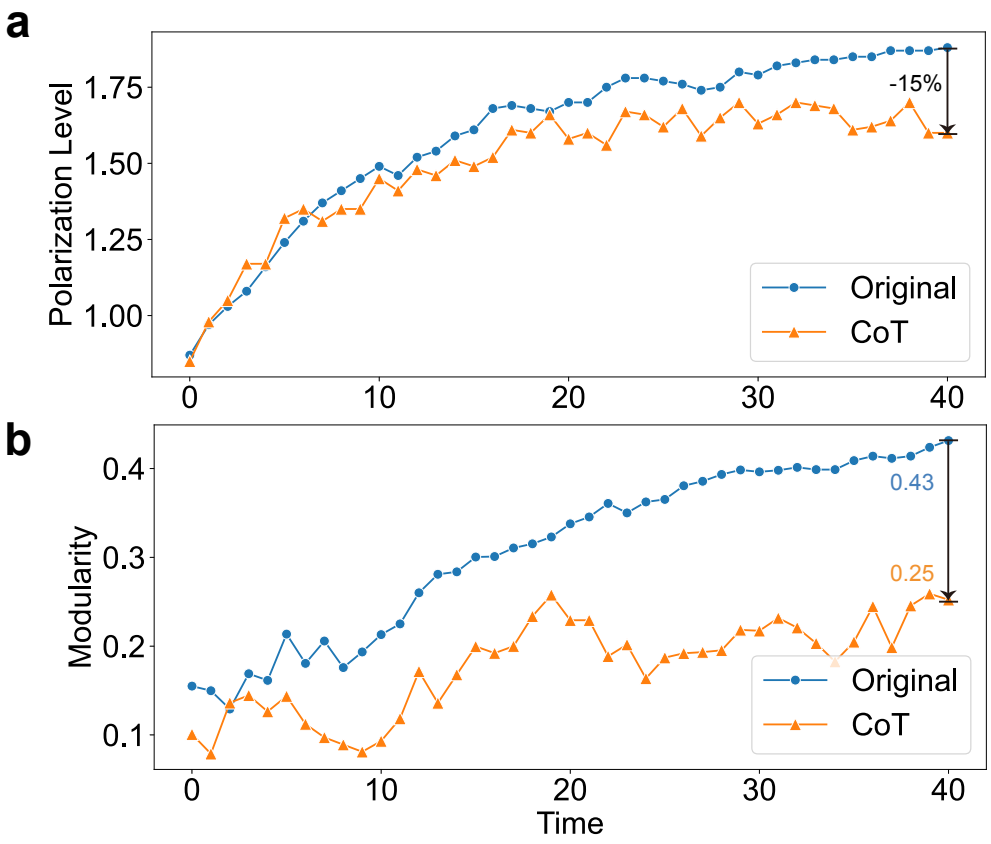

Figure S43. Comparison between the original system and a system where 10% of agents are seeded with CoT prompting, in terms of **a,** polarization level, and **b,** network modularity.

Furthermore, we investigate whether the proposed system can be extended, through enhanced prompting, to instantiate agents with subgroup-inspired profiles that reflect heterogeneous behavioural tendencies observed in human populations.. For example, we have conducted a simple case shown in Figure S44, where we use LLMs to generate two representative profiles of human subgroups in the U.S. and incorporate these profiles into the process of forming political opinions. Results show that LLM agents can simulate the two subgroups, capturing their distinctive behavioural patterns. Overall, this extended experiment illustrates that the proposed system remains flexible for modeling diverse human subgroups and their behavioural differences. We note that, in the main experiments of the paper, we intentionally do not insert additional subgroup-specific information into the agents' prompts in order to preserve the spontaneous nature of their behaviour.

12 Supplementary Note 11: Experiments on the Effect of Increasing Severing Costs

We examine the sensitivity of our results to the cost of severing social ties. To explore whether even minimal friction can slow the formation of ideological silos, we introduce a simple mechanism in which agents' decisions to sever existing social ties are blocked with a fixed probability (e.g., 10%), simulating minimal social friction. Figures S45 and S46 show the results under this minimal friction condition, where the cost of severing ties is increased by only 10%. Notably, while the minimal friction leads to marginal mitigation in polarization levels (Figure S46a), its effect on suppressing the formation of ideological silos is substantial, as consistently reflected across four different network-level metrics (Figures S46b–e). This experiment suggests that even minimal social friction can play a nontrivial role in slowing the formation of ideological silos and fostering communication between opposing camps.

13 Supplementary Note 12: Results of Simulation under Different Initial Conditions

To assess the robustness and generalizability of the proposed system, we investigate how different initial conditions affect the emergent collective behaviours of LLM agents. Specifically, we focus on two key aspects of initialization: the initial distribution of agents' opinions and the initial structure of the social network.

We first modify the initial opinion distribution by transitioning from a near-Gaussian distribution of $[0.1, 0.2, 0.4, 0.2, 0.1]$ to a highly centralized distribution of $[0, 0.1, 0.8, 0.1, 0]$. In this setting, a system that initially lacks polarized agents should be less likely to become polarized. Figure S47 shows the evolution of collective opinions in the system. We observe that the system, though taking a longer time to evolve, eventually converges to a polarized opinion distribution. Along with the emergence of opinion polarization, the social network gradually splits into two communities with opposing opinions. To sum up, this experiment demonstrates that even when the synthetic system is initialized in a setting unlikely to produce polarization, long-term free-form social interactions among LLM agents still lead to opinion polarization.

We also conduct experiments to explore whether the initial network affects the emergent collective behaviours of LLM agents. Specifically, we initialize LLM agents' social network with an Erdős–Rényi model, a Barabási–Albert model, and Watts–Strogatz models with rewiring probabilities of 0.001 and 0.05. In the BA model setup, a subset of agents naturally become highly connected nodes, resembling opinion leaders who exert disproportionate influence on others. Figures S48–S51 show how these agents self-organize their social networks and collective opinions under different initial conditions. All networks are visualized using the ForceAtlas2 algorithm with identical parameters in Gephi, ensuring direct visual comparability. We find that despite substantial differences in initial network structure, the final outcomes are similar. In all these cases, agents spontaneously organize into two well-defined communities: one that predominantly supports left-leaning opinions and the other right-leaning ones. Moreover, the final opinion distributions converge into a stable polarized pattern, demonstrating that the emergence of polarization and homophilic clustering is not driven by initial configurations, equitable or inequitable. Rather, these patterns are inherently driven by agents' autonomous social interactions. Overall, these results demonstrate the robustness and generalizability of our findings across diverse initial network structures, underscoring that LLM agents, through autonomous interactions, consistently self-organize into polarized communities in ways not determined by initial conditions.

14 Supplementary Note 13: Results of Simulation on Different Issues

To test the system's generalizability across different types of issues, we conduct two additional experiments beyond the political issues previously studied. The first relates to immigration restrictions, a socially relevant and critical issue. The second is the flat Earth theory, a non-political, fact-based issue with clear scientific consensus. These additions help demonstrate the generalizability of the proposed system and its corresponding results across both high-stakes political topics and more neutral or non-controversial domains in the synthetic testbed.

In particular, we illustrate the opinion dynamics when LLM agents discuss immigration restrictions in Figure S52. We observe that these agents spontaneously develop their collective opinions into a polarized pattern. This pattern emerges consistently across issues that have been shown to trigger polarization observed in human society. This consistency suggests that the proposed system can serve as a synthetic LLM-agent testbed for exploring opinion-dynamics hypotheses across different issue domains.

Furthermore, we wanted to understand how the proposed system would function on non-political, non-controversial issues. To this end, we test the system on a scientific issue – the flat Earth theory – where the topic is grounded in well-established scientific facts, rather than in polarized or contested opinions. As shown in Figure S53, we find that, unlike in political discussions, no polarization phenomenon emerges. Instead, agents rapidly reach a consensus within ten timesteps. These findings demonstrate that our system is not inherently predisposed to generate polarization. Rather, polarization emerges from free-form social interactions among LLM agents when dealing with inherently divisive topics, while it remains stable and convergent in neutral or fact-based domains.

15 Supplementary Note 14: Supplementary Notes on Statistical Testing and Robustness Checks

We assessed the assumptions of each reported frequentist test. For two-sample comparisons, normality was assessed using Shapiro–Wilk tests and equality of variances using Levene's tests; Student's t-tests were used when the equal-variance assumption was not violated, whereas Welch's t-tests were used otherwise. For tests involving large samples or sample proportions, the normal approximation of the relevant sampling distributions was justified by the central limit theorem. For proportion z-tests, formal normality and equal-variance tests were not applicable because the outcomes were binary proportions rather than continuous variables. For one-sample t-tests, equal-variance assumptions were not applicable. Where appropriate, we additionally conducted non-parametric robustness checks using Mann–Whitney U tests or Wilcoxon signed-rank tests.



Profile: “A 29-year-old mixed White and Hispanic woman living in Brooklyn, New York, with a master’s degree in public policy. She works for a nonprofit organization, cares about social justice, climate change, and gender equality, and believes in strengthening community support systems. She mainly follows NPR and The New York Times and is active on Twitter and Instagram.”

Opinion on Partisan Alignment: I strongly support the Democratic Party.
“Most of my values—like advancing social justice, addressing climate change, protecting reproductive rights, and strengthening social safety nets—align with Democratic policies. While the party isn’t perfect and still has a lot of work to do, it’s the one actively pushing for more inclusive and equitable policies that reflect the kind of society I want to live in. I also believe in protecting democracy and voting rights, which I see as being under threat from the Republican side right now.”

Opinion on Abortion Ban: I strongly oppose abortion bans.
“I believe reproductive rights are fundamental to gender equality and personal freedom—no one should be forced to carry a pregnancy against their will. Abortion bans disproportionately harm low-income women and marginalized communities who already face barriers to healthcare, and they strip people of the right to make deeply personal decisions about their own bodies and lives.”

Opinion on Gun Control: I strongly support stricter gun control.
“The level of gun violence in this country is unacceptable, and common-sense measures like universal background checks, closing loopholes, and restricting access to high-capacity weapons can save lives without infringing on responsible ownership. Communities, especially children, deserve to feel safe in schools, neighborhoods, and public spaces, and stronger gun laws are a necessary step toward that safety.”



Profile: “A 66-year-old White man from rural Texas, a retired truck driver with a high school education. He values traditional family norms, personal responsibility, gun rights, low taxes, and stricter immigration policies, strongly supporting the Republican Party. He primarily consumes Fox News and Newsmax and engages on Facebook and in local church communities.”

Opinion on Partisan Alignment: I strongly support the Republican Party.
“They stand for the values I was raised on—faith, family, hard work, and personal freedom. I believe they do a better job protecting our rights, like the Second Amendment, keeping taxes low so folks can keep more of what they earn, and securing our borders to keep our communities safe.”

Opinion on Abortion Ban: I strongly support abortion bans.
“I believe life begins at conception, and every unborn child deserves a chance to live—that’s a value rooted in both my faith and my sense of personal responsibility. To me, protecting the most vulnerable is a moral duty, and I think communities and churches should step up to support mothers and families instead of turning to abortion as an answer.”

Opinion on Gun Control: I strongly support weaker gun control.
“The right to bear arms is a fundamental freedom that protects us from both criminals and government overreach. Out here in rural Texas, responsible gun ownership is a way of life, and I believe law-abiding citizens shouldn’t be punished or restricted because of the actions of a few bad actors.”

Figure S44. Two representative profiles of human subgroups in the U.S. and their different behaviours when discussing political issues. Credit: Profile icons designed by Magnific.

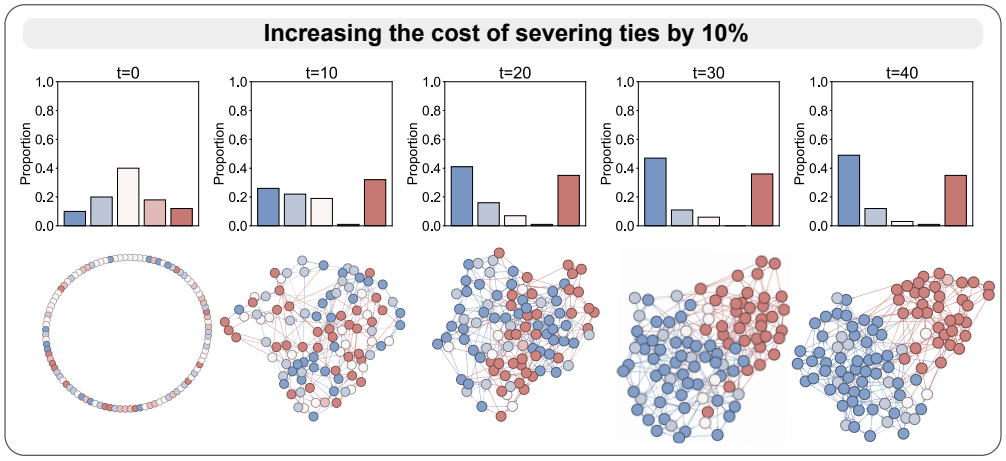

Figure S45. Evolution of LLM agents' collective opinions and network structures, where we increase the cost of severing ties by 10%. All agents are driven by GPT-3.5 and they discuss the political issue of abortion bans.

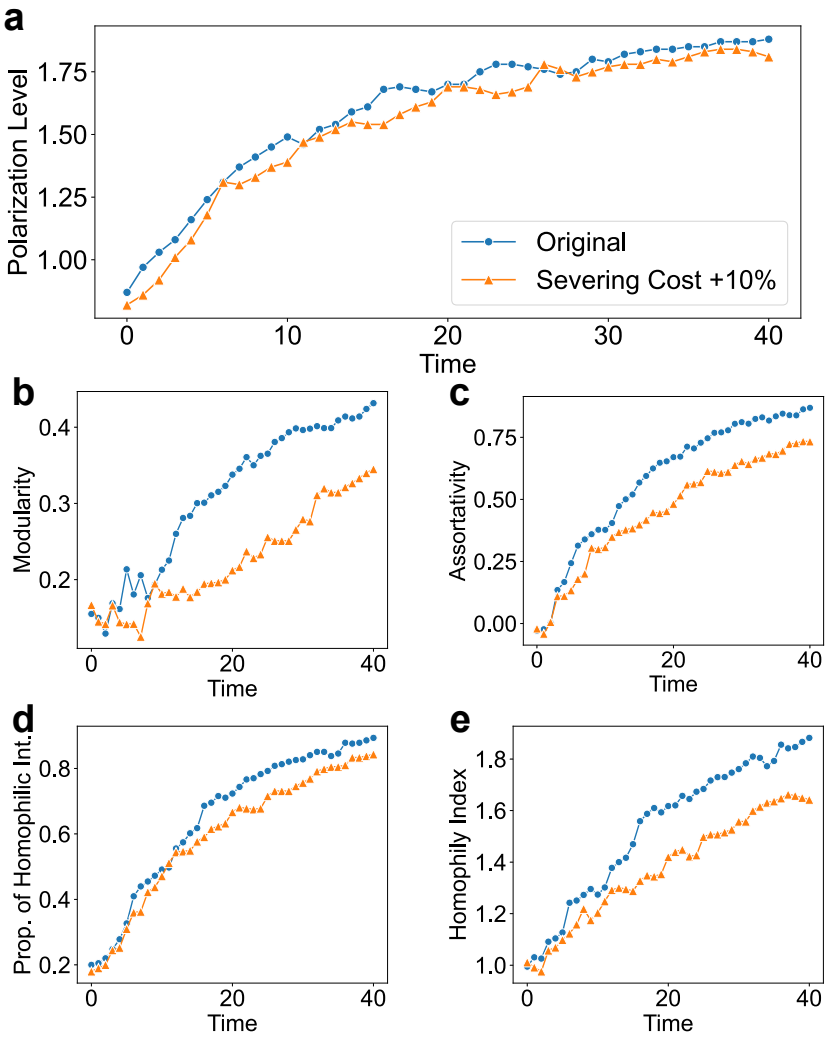

Figure S46. Comparison between the original system and a system where the cost of severing ties is increased by 10%, in terms of **a,** polarization level, **b,** network modularity, **c,** assortativity, **d,** proportion of homophilic interactions, and **e,** homophily index.

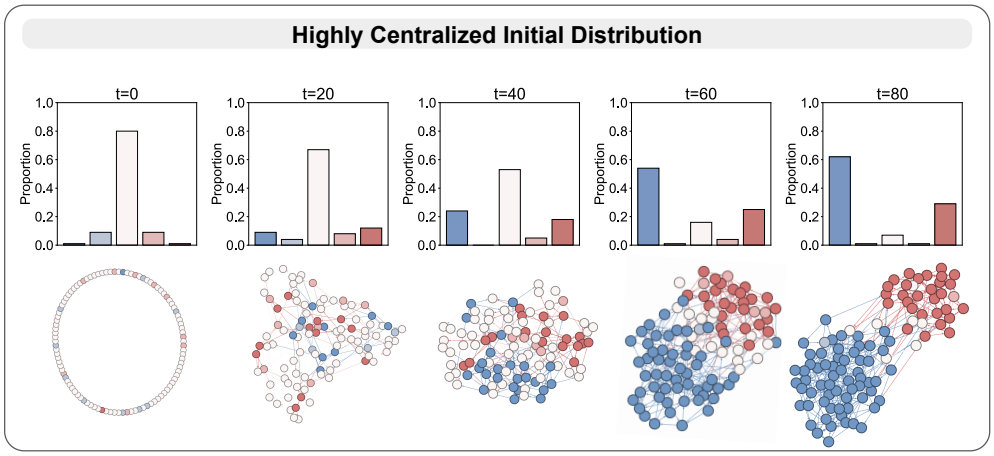

Figure S47. Evolution of LLM agents' collective opinions and network structure, where the network is initialized using a highly centralized distribution, and agents discuss the political issue of abortion ban.

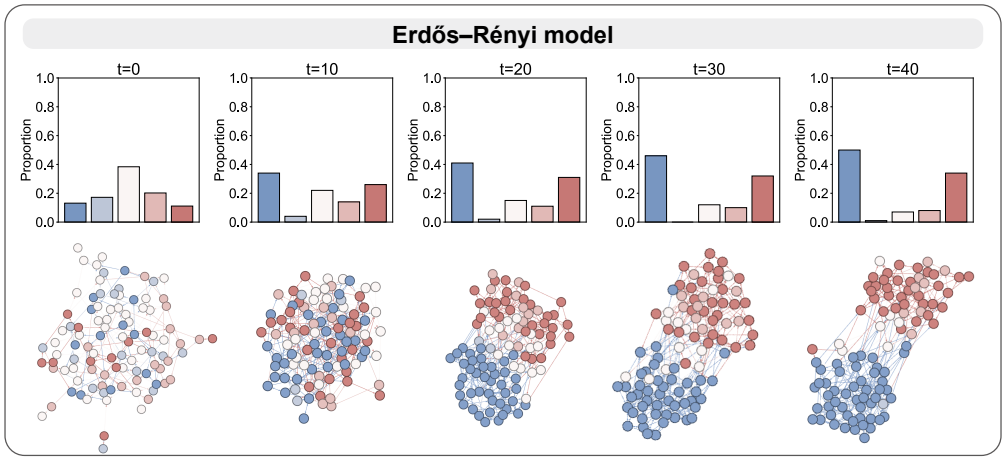

Figure S48. Evolution of LLM agents' collective opinions and network structures, where the network is initialized using an Erdős–Rényi model with an average degree of 4, and agents discuss the political issue of abortion ban.

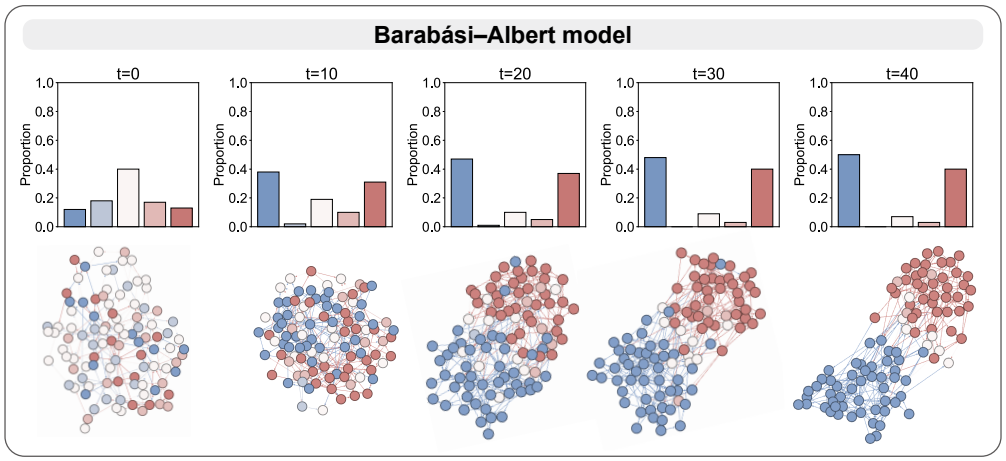

Figure S49. Evolution of LLM agents' collective opinions and network structures, where the network is initialized using a Barabási–Albert model with an average degree of 4, and agents discuss the political issue of abortion ban.

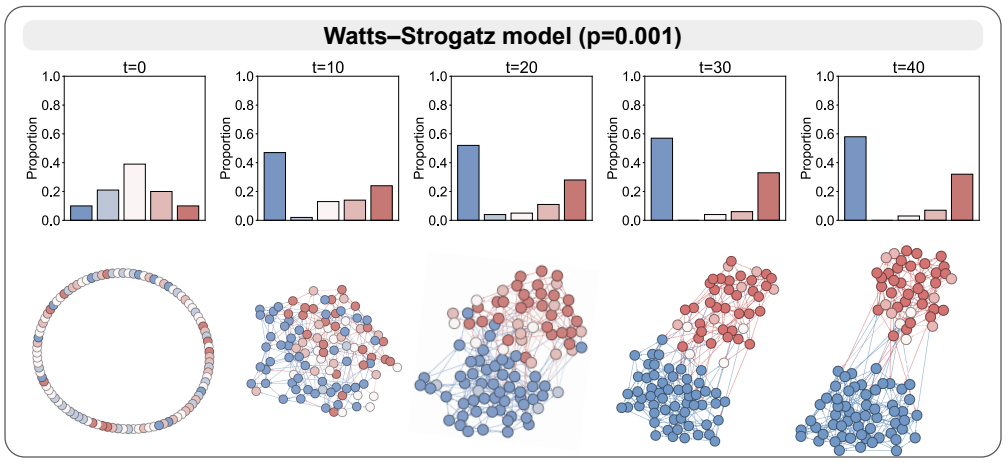

Figure S50. Evolution of LLM agents' collective opinions and network structures, where the network is initialized using a Watts–Strogatz model with the rewiring probability of 0.001, and agents discuss the political issue of abortion ban.

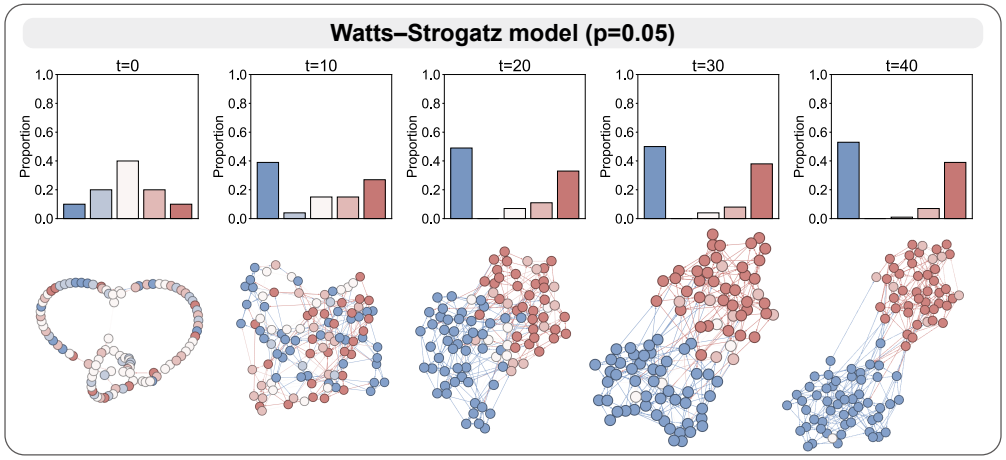

Figure S51. Evolution of LLM agents' collective opinions and network structures, where the network is initialized using a Watts–Strogatz model with the rewiring probability of 0.05, and agents discuss the political issue of abortion ban.

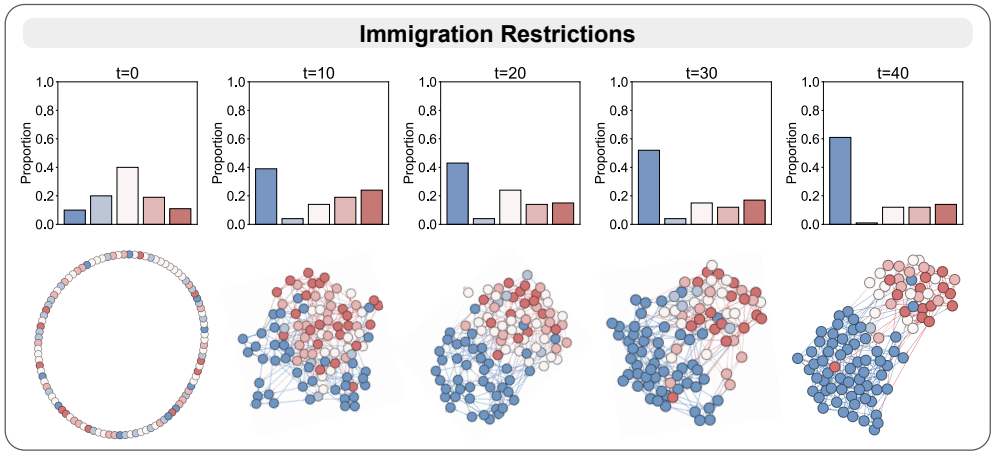

Figure S52. Evolution of LLM agents' collective opinions and network structures in the discussion of immigration.

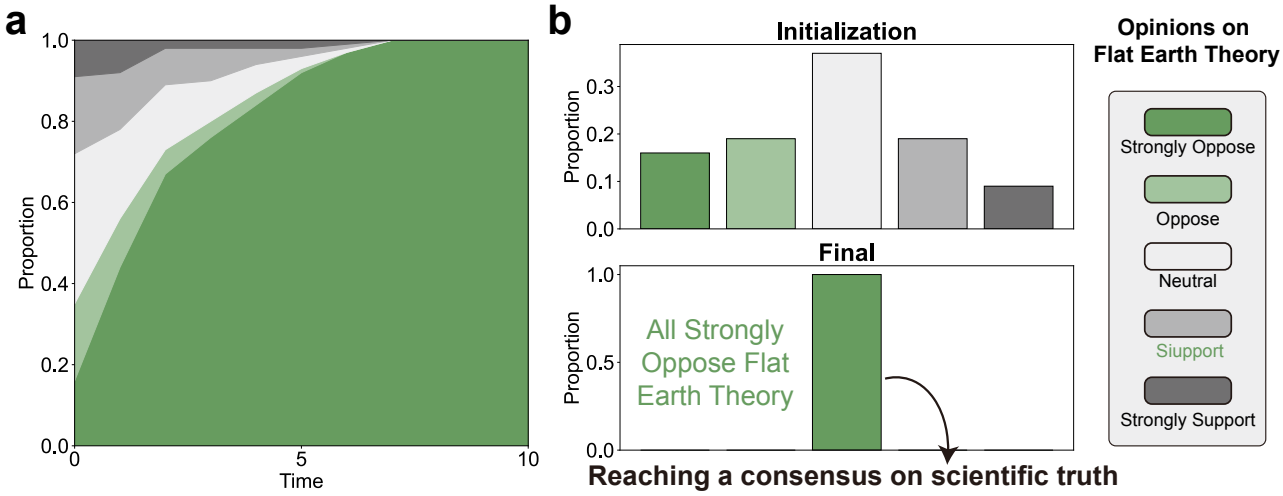

Figure S53. Opinion dynamics when LLM agents discuss about scientific truth. After 10 timesteps, all agents are persuaded to strongly oppose the flat Earth theory, reaching consensus.

References

[1] Fernando P Santos, Yphtach Lelkes, and Simon A Levin. "Link recommendation algorithms and dynamics of polarization in online social networks". In: *Proceedings of the National Academy of Sciences* 118.50 (2021), e2102141118.

[2] Fabian Baumann, Philipp Lorenz-Spreen, Igor M Sokolov, and Michele Starnini. "Emergence of polarized ideological opinions in multidimensional topic spaces". In: *Physical Review X* 11.1 (2021), p. 011012.

[3] Duncan J Watts and Steven H Strogatz. "Collective dynamics of 'small-world'networks". In: *nature* 393.6684 (1998), pp. 440–442.

[4] Woosub Lee, Seong-Gyu Yang, and Beom Jun Kim. "The effect of media on opinion formation". In: *Physica A: Statistical Mechanics and its Applications* 595 (2022), p. 127075.

[5] Mohammad Reza Sanatkar. "The dynamics of polarized beliefs in networks governed by viral diffusion and media influence". In: *Social Network Analysis and Mining* 10.1 (2020), p. 17.

[6] Philipp Lorenz-Spreen, Lisa Oswald, Stephan Lewandowsky, and Ralph Hertwig. "A systematic review of worldwide causal and correlational evidence on digital media and democracy". In: *Nature human behaviour* 7.1 (2023), pp. 74–101.

[7] Matteo Cinelli, Gianmarco De Francisci Morales, Alessandro Galeazzi, Walter Quattrociocchi, and Michele Starnini. "The echo chamber effect on social media". In: *Proceedings of the National Academy of Sciences* 118.9 (2021), e2023301118.

[8] Eytan Bakshy, Solomon Messing, and Lada A Adamic. "Exposure to ideologically diverse news and opinion on Facebook". In: *Science* 348.6239 (2015), pp. 1130–1132.

[9] Christopher A. Bail, Lisa P. Argyle, Taylor W. Brown, John P. Bumpus, Haohan Chen, M. B. Fallin Hunzaker, Jaemin Lee, Marcus Mann, Friedolin Merhout, and Alexander Volfovsky. "Exposure to opposing views on social media can increase political polarization". In: *Proceedings of the National Academy of Sciences* 115.37 (2018), pp. 9216–9221.

[10] James Flamino, Alessandro Galeazzi, Stuart Feldman, Michael W Macy, Brendan Cross, Zhenkun Zhou, Matteo Serafino, Alexandre Bovet, Hernán A Makse, and Boleslaw K Szymanski. "Political polarization of news media and influencers on Twitter in the 2016 and 2020 US presidential elections". In: *Nature Human Behaviour* (2023), pp. 1–13.

[11] Max Falkenberg, Alessandro Galeazzi, Maddalena Torricelli, Niccolò Di Marco, Francesca Larosa, Madalina Sas, Amin Mekacher, Warren Pearce, Fabiana Zollo, Walter Quattrociocchi, et al. "Growing polarization around climate change on social media". In: *Nature Climate Change* (2022), pp. 1–8.

[12] Brendan Nyhan, Jaime Settle, Emily Thorson, Magdalena Wojcieszak, Pablo Barberá, Annie Y Chen, Hunt Allcott, Taylor Brown, Adriana Crespo-Tenorio, Drew Dimmery, et al. "Like-minded sources on Facebook are prevalent but not polarizing". In: *Nature* (2023), pp. 1–8.

[13] Cass R Sunstein. *# Republic: Divided democracy in the age of social media*. Princeton University Press, 2018.

[14] John T Jost, Delia S Baldassarri, and James N Druckman. "Cognitive–motivational mechanisms of political polarization in social-communicative contexts". In: *Nature Reviews Psychology* 1.10 (2022), pp. 560–576.

[15] Isaac Waller and Ashton Anderson. "Quantifying social organization and political polarization in online platforms". In: *Nature* 600.7888 (2021), pp. 264–268.

[16] Jiazhen Liu, Shengda Huang, Nathaniel M Aden, Neil F Johnson, and Chaoming Song. "Emergence of polarization in coevolving networks". In: *Physical Review Letters* 130.3 (2023), p. 037401.

[17] Lada A Adamic and Natalie Glance. "The political blogosphere and the 2004 US election: divided they blog". In: *Proceedings of the 3rd international workshop on Link discovery*. 2005, pp. 36–43.

[18] American National Election Studies. *ANES 2020 Time Series Study Full Release [dataset and documentation]*. https://www.electionstudies.org. July 19, 2021 version. 2021.

[19] Natalie Jomini Stroud. "Polarization and partisan selective exposure". In: *Journal of communication* 60.3 (2010), pp. 556–576.

[20] Raymond S Nickerson. "Confirmation bias: A ubiquitous phenomenon in many guises". In: *Review of general psychology* 2.2 (1998), pp. 175–220.

[21] James N Druckman, Erik Peterson, and Rune Slothuus. "How elite partisan polarization affects public opinion formation". In: *American political science review* 107.1 (2013), pp. 57–79.

[22] Jon C Rogowski and Joseph L Sutherland. "How ideology fuels affective polarization". In: *Political behavior* 38 (2016), pp. 485–508.

[23] Jeffrey Lees and Mina Cikara. “Inaccurate group meta-perceptions drive negative out-group attributions in competitive contexts”. In: *Nature human behaviour* 4.3 (2020), pp. 279–286.

[24] Michael C Schwalbe, Geoffrey L Cohen, and Lee D Ross. “The objectivity illusion and voter polarization in the 2016 presidential election”. In: *Proceedings of the National Academy of Sciences* 117.35 (2020), pp. 21218–21229.

[25] Douglas J Ahler and Gaurav Sood. “The parties in our heads: Misperceptions about party composition and their consequences”. In: *The Journal of Politics* 80.3 (2018), pp. 964–981.

[26] Ruibo Liu, Chenyan Jia, Jason Wei, Guangxuan Xu, and Soroush Vosoughi. “Quantifying and alleviating political bias in language models”. In: *Artificial Intelligence* 304 (2022), p. 103654.

[27] Jérôme Rutinowski, Sven Franke, Jan Endendyk, Ina Dormuth, Moritz Roidl, and Markus Pauly. “The Self-Perception and Political Biases of ChatGPT”. In: *Human Behavior and Emerging Technologies* 2024.1 (2024), p. 7115633.

[28] Fabio Motoki, Valdemar Pinho Neto, and Victor Rodrigues. “More human than human: Measuring ChatGPT political bias”. In: *Public Choice* 198.1 (2024), pp. 3–23.

[29] Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. “Whose opinions do language models reflect?” In: *International Conference on Machine Learning*. PMLR. 2023, pp. 29971–30004.

[30] Shangbin Feng, Chan Young Park, Yuhan Liu, and Yulia Tsvetkov. “From Pretraining Data to Language Models to Downstream Tasks: Tracking the Trails of Political Biases Leading to Unfair NLP Models”. In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2023, pp. 11737–11762.

[31] Jochen Hartmann, Jasper Schwenzow, and Maximilian Witte. “The political ideology of conversational AI: Converging evidence on ChatGPT’s pro-environmental, left-libertarian orientation”. In: *arXiv preprint arXiv:2301.01768* (2023).

[32] Santo Fortunato and Marc Barthelemy. “Resolution limit in community detection”. In: *Proceedings of the national academy of sciences* 104.1 (2007), pp. 36–41.

[33] Mark EJ Newman. “Mixing patterns in networks”. In: *Physical review E* 67.2 (2003), p. 026126.

[34] Serina Chang, Alicja Chaszczewicz, Emma Wang, Maya Josifovska, Emma Pierson, and Jure Leskovec. “LLMs generate structurally realistic social networks but overestimate political homophily”. In: *arXiv preprint arXiv:2408.16629* (2024).

[35] Andrew Guess, Brendan Nyhan, Benjamin Lyons, and Jason Reifler. “Avoiding the echo chamber about echo chambers”. In: *Knight Foundation* 2.1 (2018), pp. 1–25.

[36] Alexander Coppock. *Persuasion in parallel: How information changes minds about politics*. University of Chicago Press, 2023.

[37] Albert-László Barabási and Réka Albert. “Emergence of scaling in random networks”. In: *science* 286.5439 (1999), pp. 509–512.

[38] Steven H Strogatz. “Exploring complex networks”. In: *nature* 410.6825 (2001), pp. 268–276.

[39] James N Druckman, Jordan Fein, and Thomas J Leeper. “A source of bias in public opinion stability”. In: *American Political Science Review* 106.2 (2012), pp. 430–454.

[40] Charles G Lord, Lee Ross, and Mark R Lepper. “Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence.” In: *Journal of personality and social psychology* 37.11 (1979), p. 2098.

[41] Eric Groenendyk and Yanna Krupnikov. “What motivates reasoning? A theory of goal-dependent political evaluation”. In: *American Journal of Political Science* 65.1 (2021), pp. 180–196.

[42] Michael Mäs and Andreas Flache. “Differentiation without distancing. Explaining bi-polarization of opinions without negative influence”. In: *PloS one* 8.11 (2013), e74516.

[43] Andreas Flache, Michael Mäs, Thomas Feliciani, Edmund Chattoe-Brown, Guillaume Deffuant, Sylvie Huet, and Jan Lorenz. “Models of social influence: Towards the next frontiers”. In: *Jasss-The journal of artificial societies and social simulation* 20.4 (2017), p. 2.