Language-Driven Opinion Dynamics in Agent-Based Simulations with LLMs

Erica Cau[1,2*†], Valentina Pansanella[2†], Dino Pedreschi[1], Giulio Rossetti[2]

[1*]Department of Computer Science, University of Pisa, Largo Bruno Pontecorvo, Pisa, Italy.
[2]Institute of Information Science and Technologies "A. Faedo" (ISTI), National Research Council (CNR), Via Giuseppe Moruzzi 1, Pisa, Italy.

*Corresponding author(s). E-mail(s): erica.cau@phd.unipi.it;
Contributing authors: valentina.pansanella@isti.cnr.it; dino.pedreschi@unipi.it; giulio.rossetti@isti.cnr.it;
†These authors contributed equally to this work.

Abstract

Understanding how opinions evolve is crucial for addressing issues such as polarization, radicalization, and consensus in social systems. While much research has focused on identifying factors influencing opinion change, the role of language and argumentative fallacies remains underexplored. This paper aims to fill this gap by investigating how language – along with social dynamics – influences opinion evolution through LODAS, a Language-Driven Opinion Dynamics Model for Agent-Based Simulations. The model exploits LLM agents to simulate debates around the "Ship of Theseus" paradox, in which agents with discrete opinions interact with each other and evolve their opinions by accepting, rejecting, or ignoring the arguments presented. Populations of LLM-based agents consistently converge toward a single opinion, mainly agreement, with the presented statement, regardless of model or framing. Convergence arises from an asymmetric bias: accept (reject) probability is positively (negatively) correlated with the signed distance between opinions. Moreover, such AI agents are often producers of fallacious arguments in the attempt to persuade their peers and – due to their complacency – they are also highly influenced by arguments built on logical fallacies. These results highlight the potential of this framework not only for simulating social dynamics but also for exploring, from another perspective, biases and shortcomings of LLMs, which may impact their interactions with humans.

Keywords: Large Language Model, Opinion Dynamics, Logical Fallacies, Social Simulations, Agent Based Model

1 Introduction

> For the logical question of things that grow; one side holding that the ship remained the same, and the other contending that it was not the same.
>
> *Plutarch, Life of Theseus 23.1*

In its original formulation, the "Ship of Theseus" paradox concerns a debate over whether or not a ship that had all its components replaced one by one would remain the same. Consider engaging in a discourse regarding this paradox within the context of a philosophy class, an online Reddit community, or during a dinner gathering with friends. Everyone will reason on the paradox and try to convince others of their stance. Convincing arguments can be proposed both in favor of and against this statement. Ultimately, everyone will leave the debate with their own opinion or no opinion at all. Regardless of the context in which the debate takes place, one thing does not change: the means through which we will try to convince our peers, or they will convince us, is *language*. When a speaker intentionally uses language to convey a specific purpose, they exert an illocutionary force that can influence the listener's perspective, leading to a common understanding or increased division. Therefore, we must consider how language shapes the development of opinions.

The development of individual and public opinions has long been a focus of psychologists and sociologists, and more recently, it has been extensively explored in computational social science [1, 2] and sociophysics [3, 4]. This research acknowledges the complexity of Opinion Dynamics, (henceforth, OD), where multiple interacting factors lead to emergent behaviours such as consensus [5], polarization [6], and radicalization [7], often difficult to predict. Understanding the drivers of opinion change and going beyond mere observation of opinion patterns remains a complex issue. One common approach to tackle this issue is through models of OD, which aim to explain how opinions evolve via social interactions [8]. These models simplify real-world phenomena, enabling the exploration of various what-if scenarios. They generally simulate a population of individuals and their interactions, with processes often governed by simple rules that reflect empirically observed behaviours, such as the repeated averaging of opinions with neighbours [9, 10]. Recent models also incorporate the backfire effect [11, 12], where individuals become more entrenched in their opinions when confronted with contradictory information [13]. Opinion evolution is driven by factors rooted in socio-psychological theories, such as cognitive biases [14], as well as external forces like peer pressure [15], algorithmic biases [16], and mass media [17]. While these models provide simplified representations of societal dynamics and help stakeholders understand social behaviours, they often overlook important complexities. For example, they typically map opinions and messages onto numerical values and rely on rule-based agents, which limits their ability to capture the nuances of human behaviour and the complex relationships between agents' characteristics, such as demographics and personality traits.

To overcome such limitations, we propose a novel framework exploiting Large Language Models (LLMs) capabilities to create an Agent Based Model (ABM) that allows for the study of the interplay between language and opinion change in the long term. The relationship between language and opinion change has been underexplored. Monti et al. [18] is a prominent exception, highlighting the role of knowledge, similarity, and trust in a social media case study. Their findings challenge simplistic OD models, emphasizing the need for more complex analysis. LLMs have revolutionized language-related studies, enabling more realistic social simulations. Park et al. [19, 20]

introduced LLM agents as *social simulacra*, capable of simulating personalities and social behaviours. Claims about LLMs possessing Theory of Mind (ToM) [21] remain debated: while Kosinski [22] and others [23, 24] suggest they exhibit emergent ToM abilities, critics [25–27] highlight their inconsistencies in ToM tasks and lack of genuine social intelligence. Nevertheless, even a simulated ToM may enhance OD models by enabling agents to consider interlocutors' mental states. LLM-driven populations display spontaneous emergent behaviours akin to human societies, such as scale-free networks [28], information diffusion [29], and social conventions through interactions [30]. In opinion evolution, LLM agents replicate echo chambers [31], polarization [32], and confirmation bias effects [33]. While LLMs can generate persuasive arguments [34] aligned with psycho-linguistic theories [18], they are less convincing than humans [35] and exhibit biases toward scientific accuracy [33], politeness [36], and platform-specific discourse styles [37]. Despite these biases, LLM-based agents have successfully reproduced experimental results in psychology and linguistics [38], making them valuable tools for *in silico* social experiments. A summary of representative works and their main characteristics is provided in Table 1.

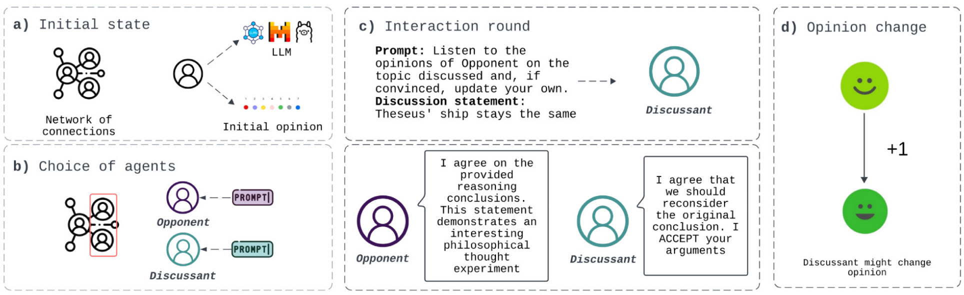

Fig. 1: **Graphical schema of LODAS.** The LLM agents population is initialized as a network; each agent is an LLM instance with an initial opinion in the range [0, 6] (a). At each iteration, two agents are randomly chosen and prompted to act as *Opponent* and *Discussant* (b). The *Discussant* is prompted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

This study aims to advance opinion dynamics and social simulations by leveraging LLMs. For this purpose, we propose a novel framework for OD with LLM agents, supported by a case study that addresses three research questions (**RQs**).

Traditional models rely on mechanistic assumptions rarely validated in real-world settings, limiting their applicability. Instead, we explore whether LLM agents, operating without predefined update rules and guided by the Theory of Mind hypothesis, can exhibit realistic individual behaviour and emergent collective dynamics (**RQ1**).

Since LLM agents engage in natural language interactions, unlike traditional mechanistic models, this opens up for the investigation of the interplay between language and opinion change. Specifically, we examine how these agents employ and propagate logical fallacies and assess their role in persuasion (**RQ2**). While existing work has focused on detecting logical fallacies using LLMs, it often overlooks the possibility that the LLMs' reasoning processes may be flawed and susceptible to fallacious argumentation.

A notable exception is Breum et al. (2023) [34], who analyzed LLM-driven persuasion, showing that trust, status, and knowledge influence stance shifts. However, their study focused on one-shot interactions, while we examine how LLMs adapt arguments and leverage fallacies over time. Payandeh et al. (2023) [39] provide the first systematic analysis of LLMs' susceptibility to fallacious reasoning in debates. They find that GPT-4 agrees with flawed arguments 67% of the time, significantly more than logically sound ones. Building on this, we investigate how LLMs not only process but also generate fallacies in multi-agent interactions, shedding light on their role in long-term opinion evolution (**RQ2**).

LLMs seem to be susceptible to input prompts, often tending towards sycophantic behaviours, as recently observed [40, 41]. In this work, we aim to evaluate how the initial conditions, particularly the distribution of initial opinions and the framing of arguments, influence the resulting opinion dynamics and linguistic behaviour (**RQ3**).

We hypothesize that the way statements are framed —whether positively or negatively— can directly affect the persuasiveness of agents, leading to different patterns in opinion evolution.

To investigate all these questions, we introduce LODAS, a Language-Driven Opinion Dynamics Model for Agent-Based Simulations framework. The framework allows the definition of a custom population of LLM agents and their interaction on a topic, where they express their opinion on the topic with illocutionary acts (**RQ1**).

A schematic representation of LODAS is provided in Figure 1. As shown in Figure 1(a), LLM agents (instances either of Mistral or Llama models) hold one of seven possible opinions, evolving through social interactions via ± 1 updates. The use of a 7-point scale Likert-scale [42] follows established methodologies in psychological research for measuring subjective constructs. We simulate three distinct scenarios: (i) a Baseline scenario with a uniform opinion distribution; (ii) a Polarized scenario, where opinions are bimodally distributed between positive and negative stances with no neutral positions; and (iii) an Unbalanced scenario, where most agents initially hold an extremely negative stance. Throughout the simulations, two agents are selected at random (see Figure 1(b)) to engage in discussion (see Figure 1(c)), where the *Opponent agent* (*Opponent*, from now on) attempts to persuade the *Discussant agent* (*Discussant*, from now on), who may then update their opinion on the Ship of Theseus paradox. This topic was chosen to minimize controversy and prevent convergence toward a predefined ground truth, a phenomenon documented in prior studies [33, 34, 43]. To assess the impact of linguistic framing, we start the discussion with one of two formulations: (i) a positive direction ("The ship remains the same") and (ii) a negative direction ("The ship becomes different"). This choice follows prior research [33] demonstrating how initial statement framing ("Global warming is/is not a hoax") may influence opinion evolution.

Table 1: Overview of works in the literature introducing LLM agents.

Paper	Population	LLMs	Opinion dynamics	Content analysis
Chuang et al., 2023 [43]	10	gpt-3.5-turbo-16k	✔	✗
Payandeh et al., 2023 [39]	18	GPT-3.5, GPT-4	✔	✔
Breum et al., 2024 [34]	2	Llama-2-70B-chat	✔	✔
Ju et al, 2024 [44]	5000	Llama-2-70B	✔	✗
Park et al., 2024 [20]	1000	GPT-4o	✗	✗
Törnberg et al., 2024 [37]	500	GPT-3.5	✗	✔
Wang et al., 2025 [32]	50	GPT-4o mini	✔	✗

The remainder of this paper is organized as follows. In Section 2, we examine the outcomes of our simulations across different initial conditions and scenarios, analyzing, on the one hand, opinion trends, acceptance rates, and, on the other, the linguistic patterns in agent interactions, assessing the role of logical fallacies in shaping opinion change. Section 3 details the simulation framework and experimental design. In Section 4, we discuss our findings, and highlight three different behaviour: convergence around a single position, tendency towards agreement and an asymmetric acceptance-rejection bias, whereas higher opinions are more often accepted and rarely rejected, while lower opinions are more often rejected and rarely accepted, producing an asymmetric pattern in opinion updating.

We also highlight the presence of fallacies in LLM-generated discourse and their impact on persuasion. Additionally, in Section 5, we outline key takeaways, study limitations, and directions for future research. Additional figures and analyses are provided in the Supplementary Materials.

2 Results

This work extends the modelling of OD using LLM agents to explore whether and which emergent behaviours arise without explicit opinion modification rules. Additionally, it examines the linguistic features of the debates, linking them to specific agent roles and behaviours. To this end, we defined a framework in which a networked population of LLM agents discusses a given topic, updating their opinions according to tunable behavioural rules. Our simulations considered a population of 140 LLM agents. We assumed a mean-field context (i.e., all agents can interact with all other agents without any social restrictions), a commonly used starting point to identify potential emerging behaviours from the opinion evolution process. Each agent is an LLM instance, holding a discrete opinion in the interval [0, 6], where 0 means *strongly disagree* and 6 *strongly agree* with a given statement. Agents – as in many classical OD models – interact with each other at discrete time intervals in a pairwise fashion: at each time step, an interacting pair is chosen at random among the connected agents; in this way, in each interaction, we can assign each agent one of two roles, respectively *Opponent* and *Discussant*.

In the present work, we assigned as a discussion topic the paradox of the *Ship of Theseus*, a thought experiment on the concept of identity first recorded in Plutarch's works. The rationale behind the paradox is the following: if all the parts of the ship are replaced over a long period, is the resulting ship the same ship it was at the beginning? This dilemma was chosen because there is no scientific truth. In this way, we avoid LLMs converging toward what they know to be scientifically valid and limit their bias toward immediate adherence to positive opinions. We designed our model to pose this "dilemma" in two different ways: (i) "the boat is the same", and (ii) "the boat is not the same". We leveraged Mistral-7B Instruct [45] (Mistral from now on) and Llama-3-8B [46] (Llama from now on) to compare different open state-of-the-art LLMs. By varying the direction of the dilemma, the LLM, and the initial distribution of opinions, we created 12 distinct settings. From our simulations, we obtained opinion evolution data and related textual data, allowing us to relate language and opinion change.

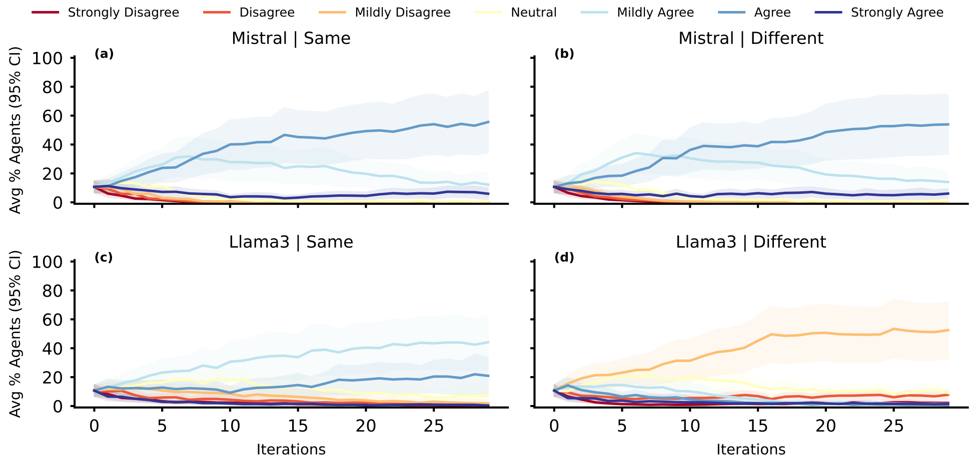

Fig. 2: **Balanced scenario – Mistral and Llama agents opinion trends.** Mistral (a)-(b) and Llama (c)-(d) opinion trends for the positive (a)-(c) and negative (b)-(d) statements. Trends are represented for Strongly Disagree (dark red), Disagree (red), Mildly Disagree (orange), Neutral (yellow), Mildly Agree (light blue), Agree (blue), and Strongly Agree (dark blue) opinions. Lines indicate the average prevalence of opinions at each time step, while the shade indicates the 95% confidence interval. Averages are computed over 10 runs.

Emergent Behaviours in LODAS

To investigate whether populations of LLM agents exhibit emergent social behaviours (**RQ1**) – such as convergence, consensus, or polarization – we begin by analyzing the opinion evolution in the Balanced scenario. Here, agents' initial opinions are uniformly distributed across the opinion spectrum. This setup serves as a neutral baseline to avoid initial biases and allows comparison with bimodal or skewed initial distributions.

Figure 2 illustrates the evolution of opinion distributions over 30 iterations, across 10 independent simulation runs. The shaded areas represent the 95% confidence interval. Across all four panels, we observe consistent patterns.

First, we note a consistent pattern of **convergence**: agent populations do not remain evenly distributed or fluctuate randomly, but rather gravitate toward one or two dominant opinions. This concentration is stable across runs, with the majority of agents consistently clustering around the same opinion categories.

Second, this convergence is predominantly oriented toward **agreement** with the presented statement, whether it is in the positive or negative direction. In both Mistral conditions (Figures 2(a)-(b)), we see a progression from mild agreement to full agreement, resulting in a dominant majority of agents aligning with the statement. Similarly, in *Llama—Same* setting (Figure 2(c)), agents increasingly converge around *Mildly Agree* and *Agree*, while *Neutral* initially rises and then declines. An exception to this tendency towards agreement is found in the *Llama—Different* setting (Figure 2(d)), where the dominant final opinion is *Mildly Disagree*. Here, although *Neutral* and *Mildly Agree* increase early on, they subsequently decline, reversing the opinion trend compared to other conditions. This distinct behaviour underscores that while convergence is a general feature, its orientation (agreement or disagreement) may depend on model-specific dynamics and prompt framing.

Comparison with Random Baseline.

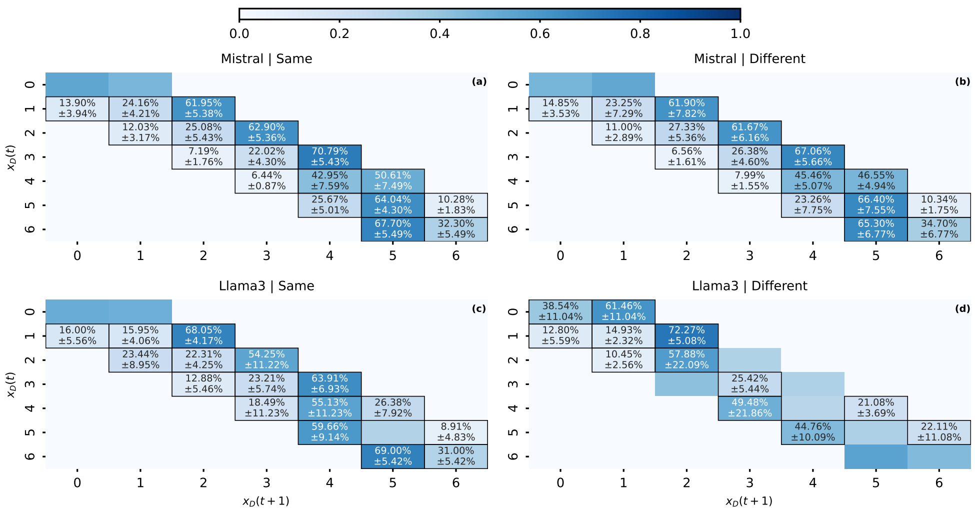

Fig. 3: **Balanced scenario – Mistral and Llama-agents transition matrices.** Mistral (a)-(b) and Llama (d)-(e) average (with std) transition rates from state i to state j for the positive (a)-(d) and negative (b)-(e) statements. Annotated cells are significant with respect to the Random Null Model ($p < 0.05$) according to the t-test. Results are averaged over 10 runs.

To determine whether the observed convergence and agreement patterns arise from chance or represent systematic behaviours, we compare them with a Random null model that mirrors the structural features of the simulations (population size, number of iterations, frequency of interactions, and initial distribution) but replaces agents' decision-making with stochastic transitions. In this model, agents randomly shift their opinion by -1, 0, or +1 upon interaction, with probabilities uniformly distributed across permitted transitions (see Section 3 and Supplementary Materials Section S1 for further details).

The Random null model fails to reproduce the emergent patterns observed in LODAS simulations. The opinion distribution remains uniform over time (this also holds with different initial conditions, see Supplementary Figures S1- S3).

To statistically validate the difference, we compare the transition matrices of the LODAS simulations and the Random baseline. Figure 3 presents the average transition

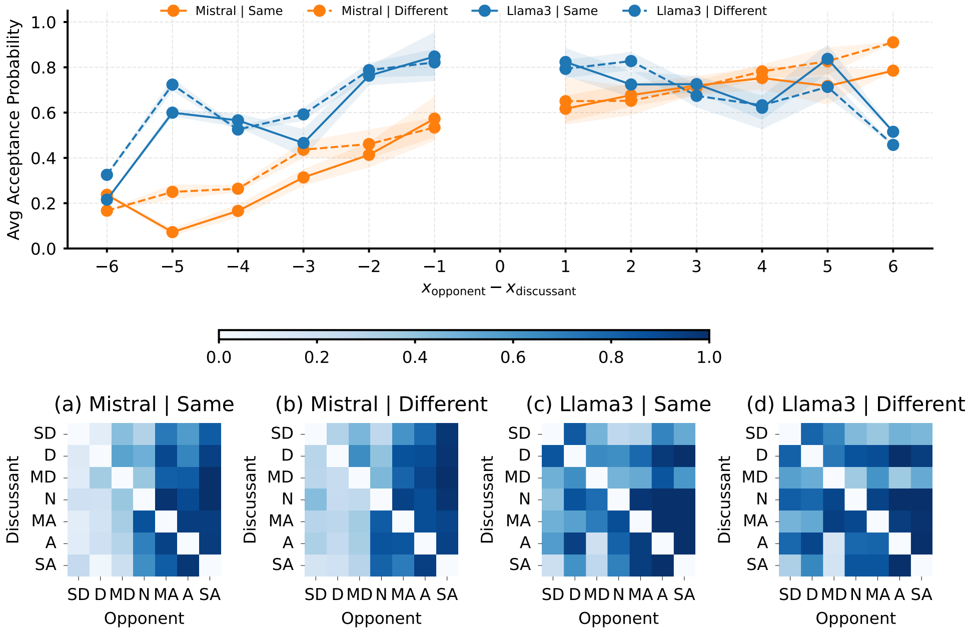

Fig. 4: **Balanced scenario – Mistral and Llama-agents average acceptance probabilities** $\mathbb{P}(\mathbf{A}|o_D, o_O)$**.** Mistral (orange line and matrices (a)-(b)) and Llama (blue lines and matrices (c)-(d)) average acceptance rates for the positive (solid lines and (a)-(c) matrices) and negative (dashed lines and (b)-(d) matrices) statements. Top panel: average probability of the Discussant accepting the Opponent's opinion as a function of $\Delta x = x_O - x_D$. Bottom panel: matrices represent acceptance rates averaged over 10 runs.

probabilities $T_{ij} = \mathbb{P}(x_D(t+1) = j \mid x_D(t) = i)$ across all conditions. Black-bordered cells denote statistically significant differences ($p < 0.05$, obtained with a two-sample Welch's t-test [47] with unequal variances on the distributions obtained from 10 independent executions of each model). A substantial majority of opinion transitions in LODAS simulations are significantly different from the random baseline, confirming that the observed behaviours are not attributable to randomness.

Mechanisms Behind Convergence – Testing the Sycophancy Hypothesis. This first analysis of opinion evolution trends, however, does not explain *how* these dynamics emerge. A first hypothesis is that convergence results from sycophantic behaviour, i.e., agents consistently adopting their opponent's opinion, which is an LLM characteristics recognized in the literature. To test this, we analyzed the acceptance probabilities $P(A \mid x_D, x_O)$, i.e., the likelihood that a Discussant's opinion x_D *moves towards* the Opponent's opinion x_O. We computed matrices of $P(A \mid x_D, x_O)$ and we also computed the average $P(A \mid \Delta x)$ with $\Delta x = x_O - x_D$.

Figure 4 shows that acceptance is not indiscriminate. For Mistral agents (orange lines and matrices (a) and (b)), acceptance probability increases with $\Delta x = x_O - x_D$: it is above 60% when the Opponent's opinion is more agreeable (i.e., $x_O > x_D$), but decreases down to 20% when the opponent has a more disagreeing opinion $x_O < x_D$. Llama agents exhibit a more symmetric pattern, but still show increased acceptance as Δx increases, revealing a positive correlation between acceptance and opinion distance.

This asymmetry in acceptance contradicts the sycophancy hypothesis. Agents are not passively agreeing with every interaction partner; they are selective, favoring opinions that are closer or more agreeable with the presented statement than their own.

Moreover, rejection patterns are complementary: Llama agents have a lower $P(R \mid x_D, x_O)$ than Mistral agents, and overall the probability of rejecting decreases as Δx increases (see Supplementary Materials, Figure S20).

These patterns indicate that agents **do not exhibit sycophantic behaviour nor bounded confidence**: distant opinions are actively accepted or rejected. Specifically, opinions that are more positive and increasingly distant from the discussant's position have a higher (lower) probability to be accepted (rejected), thus skewing the opinion distribution towards agreement. Conversely, opinions that are more distant but lower than the Discussant's position show a lower (higher) acceptance (rejection) probability. This asymmetry suggests a form of **backfire effect**, occurring only in one direction and proportionally to the opinion distance.

Finally, simulations using a toy model in which agents always accept their opponent's opinion produce markedly different dynamics: the majority of the population converges on the *Strongly Disagree* opinion. The complete absence of such a result in the LODAS simulations further refutes the sycophancy hypothesis. Conversely, toy model simulations where agents always reject their opponent's opinion generate dynamics more similar, albeit more extreme, to those observed, with the majority of agents converging on the *Strongly Agree* opinion (see Supplementary Materials Section S1 for further details).

Together, these findings address our first research question (**RQ1**), showing that LODAS consistently produce emergent behaviours characterized by **convergence and alignment**, typically **toward agreement**. These trends are **statistically significant** compared to a random baseline. Moreover, the behaviours do not stem from indiscriminate acceptance or sycophancy. Instead, they arise from structured, selective interaction patterns shaped by the underlying language models, resulting in an asymmetric backfire effect and a bias toward strongly positive opinions, which we can call an **asymmetric acceptance-rejection bias**. The strength of these effects varies depending on the choice of LLM.

Impact of Skewed Initial Opinion Distribution.

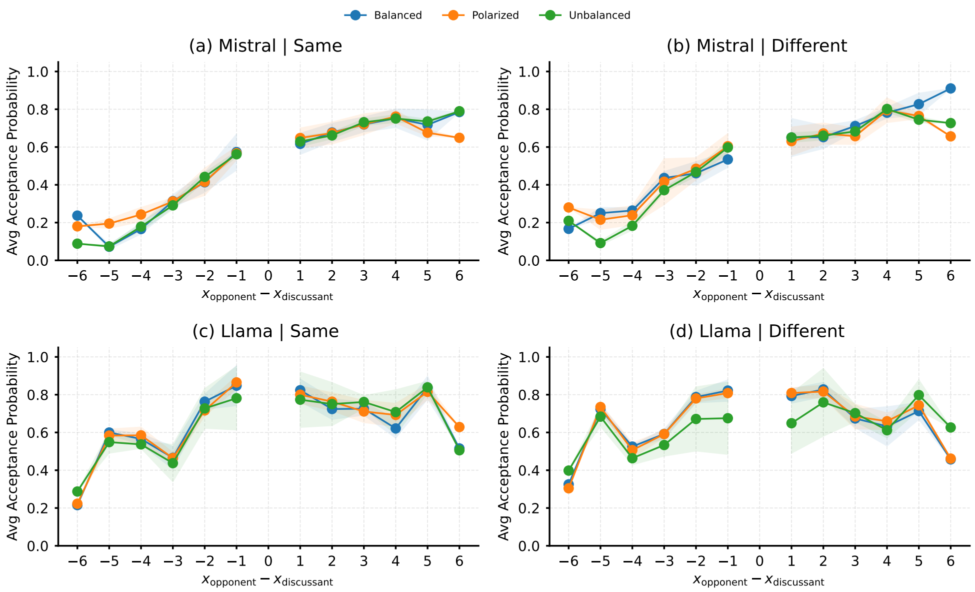

Fig. 5: **Acceptance probability $P(A \mid \Delta_x)$ as a function of opinion distance $\Delta_x = x_O - x_D$.** Each marker represents the average probability that a Discussant agent accepts—i.e., moves toward—the Opponent's opinion, as a function of the opinion distance $\Delta_x = x_O - x_D$. Lines correspond to different initial opinion distributions: *Balanced* (blue), *Polarized* (orange), and *Unbalanced* (green). Rows distinguish between the two language models: Mistral (top) and Llama (bottom). Columns refer to the direction of the statement: *Same* (left) and *Different* (right). Results are averaged over 10 independent runs; shaded areas indicate standard deviations across runs.

To assess the influence of initial conditions (**RQ3**), we systematically compared simulations initialized under three different configurations: *Balanced* (uniform distribution across the opinion spectrum), *Polarized* (bimodal distribution centered on *Strongly Disagree* and *Strongly Agree*), and *Unbalanced* (skewed distribution concentrated around *Strongly Disagree*). Despite these substantial differences in starting configurations, we observe that the qualitative evolution of opinions over time is largely preserved across scenarios. In all conditions, opinion trends rapidly shift away from initial extremes, with agents progressively converging around moderate or positive agreement positions (see Supplementary Materials, Figures S11- S14).

Final opinion distributions (see Supplementary Materials, Figures S9- S15) reinforce these conclusions, showing that the initial distribution influences only the early stages of interaction, with little effect on long-term outcomes. Also, variability metrics (such as entropy, standard deviation, and the effective number of clusters at each iteration) further support this conclusion (see Supplementary Materials, Figures S10-S16). Across all three initial conditions, we observe a consistent decrease in variability over time, indicating convergence toward fewer opinion states. Mistral agents reduce variability more quickly, especially in the first 10 iterations, while Llama agents follow a slower but steadier trajectory. Notably, the *Llama* | *Same* setting in the Unbalanced scenario is the only case in which variability increases or remains high throughout, reflecting persistent fragmentation.

Acceptance and rejection behaviours also appear robust to changes in initial conditions. The functional forms of $P(A \mid \Delta x)$ (see Figure 5) and $P(R \mid \Delta x)$ (see Supplementary Materials, Figure S26) are stable across *Balanced*, *Polarized*, and *Unbalanced* scenarios. Similarly, the matrix representations $P(A \mid x_D, x_O)$ and $P(R \mid x_D, x_O)$ reveal consistent interaction patterns.

Taken together, these results indicate that the initial distribution of opinions has a limited and transient influence on the collective dynamics (**RQ3**). Instead, the key determinant of opinion evolution, variability, and interaction behaviour is the LLM used to enhance agent decision-making. The differences between Mistral and Llama are more pronounced and persistent than those induced by any variation in the starting opinion configuration.

Linguistic Behaviour

Moving on to **RQ2**, we analyzed the arguments produced by the agents in both roles – *Opponents* and *Discussants* – during their conversations on the Theseus' Ship paradox. Specifically, we examined their linguistic behaviour, focusing on the production of persuasive yet fallacious content, and assessed how such fallacious utterances can influence the opinion change trend within multi-agent debate.

Table 2: Percentage of logical fallacies in *Opponents'* statements.

	Balanced				Polarized				Unbalanced			
	Llama		Mistral		Llama		Mistral		Llama		Mistral	
	S	D	S	D	S	D	S	D	S	D	S	D
% Fallacious (O)	20.87	23.39	19.01	19.31	19.88	26.83	16.79	20.22	22.06	20.77	18.39	15.56

Percentage of unique *Opponent* (O) statements classified as fallacious, across models (Llama, Mistral), initial opinion distribution (balanced, polarized, unbalanced), and opinion framing (same, different).

Table 2 shows the average percentage of fallacious statements generated by *Opponent* agents, calculated from aggregated results of 10 discussion runs. In each run, *Opponents* produced a total of 12.600 statements. The percentages represent the ratio of fallacious content relative to the total number of statements and are categorized by initial opinion distribution – Balanced, Polarized and Unbalanced– by statement, and by LLM.

The proportion of statements containing logical fallacies remained relatively stable across all scenarios and discussion framing, at around 20%. Variability was primarily attributed to the underlying LLM. Mistral agents produced slightly fewer fallacious statements than Llama, especially under unbalanced initial conditions, where only 15.56% of statements were classified as fallacious. Under balanced conditions, Mistral's fallacy rate remained close to 19%, regardless of the framing of the discussion. In contrast, Llama showed an increased sensitivity to negative framing, producing more fallacious utterances than Mistral. Nonetheless, the overall fallacy rate remained below 30% of the total statements.

Due to the limited variability in fallacy types observed across configurations of different LLMs and statement framing, we focus our analysis only on the patterns detected in the Balanced scenario. Additional figures for the Polarized and Unbalanced scenarios are provided in the Supplementary Materials (Figures S27 and S28).

As shown in Figure 6, few types of fallacies emerged, with LLMs often repeating similar patterns across different statement framings. The variability of their aggregated distribution over the 10 runs was minimal, as indicated by the low standard deviation value in the error bar. Overall, both Llama and Mistral relied more heavily on specific types of fallacies, particularly fallacies of relevance, credibility, and logic. Furthermore, both models generated arguments in which they reiterated the initial premises as conclusions, resulting in the pragmatic defect of circular reasoning; this occurred more frequently in the *same boat* discussion. Additionally, though less frequently, they tended to assume a causal relationship without justification (false causality).

Table 3: **Ratio of *Discussants* changing opinion for the effect of fallacious statements.**

	Balanced				Polarized				Unbalanced			
	Llama		Mistral		Llama		Mistral		Llama		Mistral	
	S	D	S	D	S	D	S	D	S	D	S	D
% Opinion change (D)	64.9	71.4	53.79	55.29	78.06	77.16	58.52	60.53	77.84	78.82	60.18	60.72

Percentage of *Discussants* (D) changing opinion for the effect of fallacious statements produced by *Opponents*, across models (Llama, Mistral), initial opinion distribution (Balanced, Polarized, Unbalanced), and opinion framing (same, different).

Having assessed the presence of fallacious utterances in the *Opponent* agents, we moved on to measure the persuasive impact of these fallacies over the *Discussant*. Specifically, we investigated whether the presence of a fallacy in the *Opponent* statement caused a shift by ± 1 in the opinion held by the *Discussant* compared to their prior stance before the interaction. An overview of this analysis can be found in Table 3. Overall, Llama *Discussants* demonstrated higher vulnerability to logical fallacies, changing their opinion 78% of the time in the *same boat* scenario, and 75%

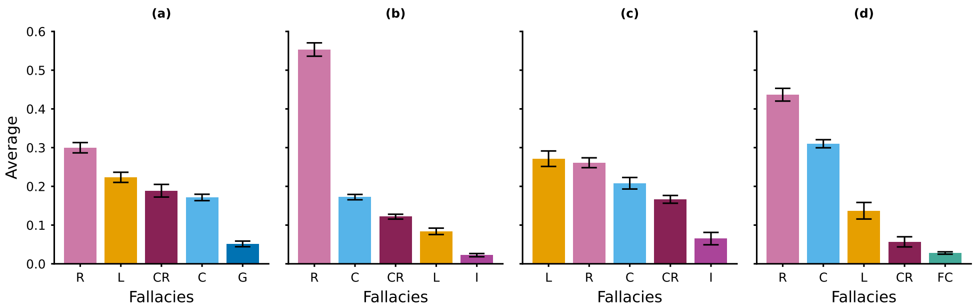

Fig. 6: Average logical fallacies proportions for different experiment configurations across multiple runs. Figures (a)-(b) refer to Llama, while (c)-(d) refer to Mistral. (a)-(c) refer to Llama (a) and Mistral (c) agents discussing the *same boat* statement, (b)-(d) refer to the *different boat* statement. Error bars represent standard deviation across 9 runs. The x-axis uses the following abbreviations: **R** (fallacy of relevance), **L** (fallacy of logic), **CR** (circular reasoning), **C** (fallacy of credibility), **G** (faulty generalization), **I** (intentional), and **FC** (false causality).

of the time in the *different boat* scenario. Conversely, Mistral agents showed greater robustness against logical fallacies. They both produced fewer fallacies than Llama agents (Table 2) and their *Discussants* resisted more than Llama ones, with opinion shifts occurring in 60% and 61% of the respective cases (Table 3).

Once investigated the production of fallacies at a macro-level, we proceeded to examine which specific types of logical fallacies were most effective in inducing the opinion shifts in the *Discussants*. Most changes, as highlighted in Table 4, are caused by fallacies of relevance when agents discussed the *different boat* scenario, whereas in the *same boat* discussion the opinion change is triggered by general logical fallacies that do not fall under the other labels recognized by the classifier.

Although it is difficult to interpret the specific fallacies introduced by the classification model under the label *fallacy of logic*, the preference for fallacies of relevance may reflect the tendency of LLMs to overlook logical reasoning in favor of empty rhetorical devices. This rhetoric is made up of compelling elements introduced into the argument, which may be unrelated to the discourse's premises, while creating a misleading yet persuasive discourse.

Table 4: Percentage of opinion changes in *Discussants* due to logical fallacies.

Fallacy Type	Llama (S)	Llama (D)	Mistral (S)	Mistral (D)
Fallacy of Logic	**24.35%**	10.65%	**27.07%**	16.82%
Faulty Generalization	8.85%	1.14%	3.01%	0.00%
Ad Populum	0.00%	0.00%	0.38%	0.00%
Appeal to Emotion	2.21%	0.00%	0.00%	0.00%
Fallacy of Credibility	19.11%	9.89%	23.31%	31.80%
Fallacy of Extension	0.00%	0.00%	0.00%	0.00%
Fallacy of Relevance	22.33%	**51.14%**	25.56%	**40.67%**

Values indicate the percentage of opinion shifts in the *Discussant* agents exposed to each fallacy type by the *Opponent*'s statement. Results refer to Llama and Mistral agents discussing the *same boat* (S) and *different boat* (D) initial statement.

3 Methods

In the Language-Driven Opinion Dynamics Model for Agent-Based Simulations (LODAS) model, we have a population of N agents, where each agent a is an LLM agent, i.e. an instance of a Large Language Model. Agents are enhanced using AutoGen [48]: "a framework for creating multi-agent AI applications that can act autonomously or work alongside humans". Specifically, we exploited AutoGen AgentChat's AssistantAgent, a built-in agent that uses a Large Language Model and has the ability to use tools. It serves as a foundational agent that can be customized or integrated into multi-agent conversations.

In our model, each LLM agent holds a discrete opinion $x_a \in \{0, \ldots, 6\}$ associated (from 0 to 6) with a negative (*strongly disagree, disagree, mildly disagree*), *neutral*, or positive (*mildly agree, agree, strongly agree*) stance on a given statement $s \in S$ around a given topic $\theta \in \mathcal{T}$. A statement s can have a *positive valence*, e.g., "this is true," or a *negative valence*, e.g., "this is not true."

In our study, we chose the Ship of Theseus paradox as the topic, where the statements were phrased as "the ship is the same" (positive valence) and "the ship is different" (negative valence). To formalize this, we define a function $\pi(s)$ that maps statements to their valence as follows:

$$\pi(s) = \begin{cases} +1, & \text{if } s \text{ expresses a positive valence (e.g., ``the ship is the same'')} \\ -1, & \text{if } s \text{ expresses a negative valence (e.g., ``the ship is different'')} \end{cases}$$

At each discrete time step t, a pair of agents (a_i, a_j) is randomly selected from this network. One agent from the pair is assigned the role of *Discussant* (D) while the other takes on the role of *Opponent* (O).

Prompts

Discussants D act according to the following prompt.

Discussant Prompt

```
[INST]
    ### You {Discussant_opinion} on the reasoning conclusion
    provided as input.
    Task:
    - Listen to the argument of {Opponent.name} on the reasoning
    conclusions and decide if you maintain your opinion
    or change it.

    ### Constraints:
    - At the end of each interaction declare if you
        - 'ACCEPT' {Opponent.name} argument;
        - 'REJECT' {Opponent.name} argument;
        - 'IGNORE' your original opinion.
    Write your response with the following format:
    \"My original opinion was I {Discussant_opinion}
    on the reasoning.
    After reading your argument my conclusions are:
```

```
    I <ACCEPT|REJECT|IGNORE> your stance because <argument>\"
[/INST]"
```

The role of the *Opponent* is instead modeled by the following prompt:

Opponent Prompt

```
[INST]
You {Opponent_opinion} on the reasoning conclusion provided as input.
Support your opinion by providing personal arguments.
Avoid using already generated arguments.

IF {Discussant.name} writes REJECT in his answer, write a second statement
where you declare if you <ACCEPT|REJECT|IGNORE> his stance.
Otherwise, conclude the conversation writing a message with
a single word 'END'.

### Constraints:
    - In your first statement you must adhere to your opinion
    ('{Opponent_opinion}')
    - Write your first response with as: \"I {Opponent_opinion} on the
    provided reasoning conclusions. I think that <argument>\}
[/INST]",
```

The selected Discussant a_D engages in a discussion with the Opponent a_O on a predefined topic $\theta \in \mathcal{T}$, with the goal of influencing the other's opinion. During this interaction, the Discussant a_D and the Opponent a_O are prompted to maintain their initial opinions unless convinced by the argumentation of the other.

The discussion is started by a_D, who asks agent a_O to express their opinion on statement s around topic θ with valence $\pi(s)$.

In our study, we have two different statements:

Positive valence statement $\pi(s) = +1$

Theseus set sail to reclaim the throne as king of Athens. During the journey, parts of Theseus's ship began to break or decay; Theseus and his crew replaced these parts as they sailed. Eventually, each part of the ship is replaced. In the end the Ship of Theseus is still the same ship on which he originally sailed.

and

Negative valence statement $\pi(s) =$ -1

Theseus set sail to reclaim the throne as king of Athens. During the journey, parts of Theseus's ship began to break or decay; Theseus and his crew replaced these parts as they sailed. Eventually, each part of the ship is replaced. In the end, the Ship of Theseus is completely different from the one he originally sailed.

The question has the following structure:

Discussion initialization

```
What do you think of the following statement?: {s}
```

The *Opponent* is asked to produce a persuasive utterance in response to the *Discussant*, based on their current opinion, to persuade the *Discussant* and shift their stance. The *Discussant* then processes the *Opponent*'s response and generates a comment about that statement, expressing whether it was convinced by the *Opponent* or not. The interaction may result in a positive (+1) or negative (-1) change in the *Discussant*'s opinion, or no change (0). Finally, the *Opponent* closes the discussion in one of two ways: if the *Discussant* chooses not to accept the persuasive statement, then it generates a new statement commenting on the current stance of the *Discussant* and thanking it for the discussion. This comment does not affect the opinions' status, it simply ends the iteration round. Otherwise, if the *Discussant* is persuaded by the *Opponent*, the *Opponent* can simply end the iteration with an END keyword.

In the present study, we set the number of iterations to $T = 30$. At each iteration t there are N pairwise random interactions (a_D, a_O).

Metrics and Analysis

Statistical Validation

To evaluate whether our results differ significantly from patterns that could arise by chance, we constructed a **Random Null Model** under the same experimental constraints as the LODAS setting. Specifically, we defined a population of $N = 140$ agents, interacting under the same structural rules. Each simulation was run for $T = 30$ iterations, with each iteration consisting of N pairwise interactions between a discussant agent D and an opponent agent O. In each interaction, only agent D could update their opinion, with a possible change of $+1$, -1, or 0.

We performed $R = 10$ independent runs for the null model to generate a reference distribution of opinion transitions. To compare these outcomes with those from the experimental condition, we analyzed the respective transition matrices. Each matrix $\mathbf{T} \in \mathbb{R}^{K \times K}$ represents the empirical average transition probabilities between K discrete opinion states. The element T_{ij} denotes the probability of transitioning from opinion state i to state j:

$$T_{ij} = P(x_D(t+1) = j \mid x_D(t) = i)$$

where $x_D(t) \in \mathcal{O}$ is the opinion of the discussant agent at time step t, and $\mathcal{O}$ is the set of all possible opinions. Each row of the matrix is normalised such that:

$$\sum_{j=1}^{K} T_{ij} = 1 \quad \text{for all } i \in \{1, \dots, K\}$$

To assess statistical differences between the experimental and null models, we used Welch's t-test for independent samples. For each matrix entry (i, j), we tested the null hypothesis:

$$H_0 : \mu_{ij}^{\text{exp}} = \mu_{ij}^{\text{null}}$$

against the two-sided alternative:

$$H_1 : \mu_{ij}^{\text{exp}} \neq \mu_{ij}^{\text{null}}$$

where μ_{ij}^{exp} and μ_{ij}^{null} represent the expected transition probabilities in the experimental and null conditions, respectively. We used the `scipy.stats` implementation of Welch's t-test, which does not assume equal variances. Statistical significance was determined using a threshold of $p < 0.05$, under which the null hypothesis was rejected in favor of a significant difference.

Opinion Evolution Metrics

Opinion dynamics were further analyzed by examining the temporal evolution of the average opinion distribution. Let $x_i(t) \in \mathcal{O}$ denote the opinion of agent i at time step t, where $\mathcal{O}$ is the set of possible discrete opinion states. For each opinion $x \in \mathcal{O}$ and each time step $t \in \{1, \dots, T\}$, we computed the proportion of agents holding opinion x, defined as:

$$\mathbb{P}_x(t) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}[x_i(t) = o]$$

where $\mathbb{I}[\cdot]$ is the indicator function. The resulting trajectories $P_x(t)$ were averaged across $R = 10$ independent simulation runs, and we report either 95% confidence intervals or standard deviation bands to indicate variability.

To assess sycophantic tendencies, we computed acceptance probability matrices $\mathbf{A} \in [0,1]^{K\times K}$, where each entry A_{ij} denotes the empirical probability that an agent with opinion $x_D = i$ accepts the opinion $x_O = j$ of their opponent. This can be expressed as:

$$A_{ij} = P(A \mid x_D = i, x_O = j) = \frac{N_A(i,j)}{N_{\text{int}}(i,j)}$$

where:

- $N_A(i,j)$ is the number of interactions in which an agent with opinion i accepted the opponent's opinion j,
- $N_{\text{int}}(i,j)$ is the total number of interactions between agents with opinions i (discussant) and j (opponent).

Analogously, to examine backfire-like effects, we constructed rejection probability matrices $\mathbf{R} \in [0,1]^{K\times K}$, where R_{ij} indicates the probability of rejecting the opponent's opinion j when the discussant holds opinion i. Rows correspond to the discussant's opinion and columns to the opponent's.

Additionally, we analyzed the influence of opinion distance on interaction outcomes by defining the signed opinion distance as $\Delta x = x_O - x_D$. For each possible value of Δx, we computed the acceptance and rejection probabilities $\mathbb{P}(A \mid \Delta x)$ and $\mathbb{P}(R \mid \Delta x)$, respectively. These conditional probabilities were estimated empirically and averaged over the $R = 10$ simulation runs, with uncertainty represented using standard deviations.

Logical Fallacies detection

The presence of logical fallacies in text is usually identified through transformers-based models, so the task is often approached as a multi-label classification problem. In this work, we employ the `distilbert-base-fallacy-classification` model [49] obtained from HuggingFace. We chose this specific model as it is trained on the dataset used by Jin et al. [50], which introduces the task of logical fallacies detection and the LOGIC dataset for fallacies. In the present article, we refer to the 13 fallacies illustrated in the original article by Jin et al. [50].

In the following, we present and discuss only the most common fallacies identified in our analysis; readers are referred to the original paper for a full summary.

- **Fallacy of credibility**: it consists of an appeal to a form of ethics or authority;
- **Fallacy of relevance**: the argument relies on premises that are irrelevant to the conclusion. In [51], it is suggested that premises might be *psychologically relevant* but not *logically relevant*, resulting in an argument that seem apparently correct and persuasive;
- **Appeal to emotion**: this fallacious argument assumes that premises are not relevant to conclusions, but the premises are used as a means to convey a specific emotion aiming to manipulate the beliefs of the reader;
- **Circular reasoning** (*circulus in probando*): a fallacy characterized by a circularity in reasoning so that the premises depend on the conclusions and vice versa.

4 Discussion

This study builds on recent literature [20, 33, 34] and introduces a Language-Driven Opinion Dynamics Model for Agent-Based Simulations (LODAS) to investigate how language and social influence shape opinion evolution, with a particular focus on the role of logical fallacies. In the model, each agent holds a discrete opinion ranging from

Strongly Disagree to *Strongly Agree*. At each time step, a *Discussant* asks an *Opponent* for their opinion on a topic; the *Opponent* responds with the intent to persuade the *Discussant*, who may then adjust their opinion by ± 1 or keep it unchanged. This process repeats until opinions stabilize or a stopping condition is reached. We studied three initial opinion distributions: (a) Balanced (uniformly distributed opinions), (b) Polarized (only extreme opinions), and (c) Unbalanced (majority extremely negative). The discussion topic was the paradox of the ship Theseus, chosen to prevent convergence toward a ground truth or consensus. Each initial condition was paired with either a positive framing ("The boat is the same") or a negative framing ("The boat is different"), producing six scenarios simulated with two different LLM.

Our findings address three main research questions: **RQ1:** Can LODAS generate emergent behaviours without mechanistic rules? **RQ2:** How do different LLM impact persuasion, particularly regarding logical fallacies? **RQ3:** To what extent do initial opinion distributions influence final outcomes?

Regarding **RQ1**, our analyses demonstrate that the LODAS framework can produce emergent behaviours without embedding explicit behavioural rules common in traditional mechanistic models [52]. Specifically, agents exhibit (i) strong convergence toward a dominant opinion, often forming a majority though not always full consensus; (ii) a consistent tendency towards agreement; and (iii) asymmetric acceptance-rejection bias — the probability of an agent accepting or rejecting an opponent's opinion is strongly and oppositely correlated with the signed opinion distance: higher opinions are more often accepted and rarely rejected, while lower opinions are more often rejected and rarely accepted, producing an asymmetric pattern in opinion updating. These emergent patterns underline the ability of language-driven interactions to naturally shape opinion evolution in ways that mirror empirical social phenomena, confirming the promise of LODAS as a modelling approach.

In exploring **RQ2**, we found meaningful differences between the two LLM agent types. Mistral agents yielded more stable results, with faster and stronger convergence toward agreement and a pronounced asymmetric acceptance-rejection bias. Conversely, Llama agents displayed greater openness to a range of opinions, though typically favoring those similar to their own. Notably, the framing of the discussion statement influenced Llama agents' dominant opinions: negative framing shifted their majority from agreement to mild disagreement. Linguistic analysis revealed that LLM agents frequently employed logical fallacies—particularly those related to relevance and credibility—in attempts to persuade others. These agents were also influenced by such fallacious arguments, consistent with prior work showing susceptibility of language models to faulty reasoning [39, 53, 54]. Interestingly, Llama agents were more effective persuaders, with about 68% of *Discussants* changing opinions after exposure to fallacious arguments, compared to 54% for Mistral. This highlights both the persuasive power of logical fallacies in artificial agents and the varying susceptibility depending on model architecture.

Finally, **RQ3** addresses the role of initial opinion distributions. Our results indicate that initial conditions have limited influence on final outcomes: whether balanced, polarized, or unbalanced, opinions converged toward specific stable points dictated by the model and statement framing. This suggests that each model-statement pair generates a strong internal dynamic that overrides initial biases, driving convergence toward a characteristic opinion cluster. This robustness underscores the influence of language model design on interaction dynamics, reflecting tendencies toward coherence and alignment that encourage agreement [55, 56]. The asymmetric acceptance-rejection bias, which reduces acceptance of negative opinions and amplifies influence from positive ones, appears to be a key driver of this stability.

Taken together, these insights advance our understanding of how language-based social simulations can capture complex opinion dynamics and the role of logical fallacies therein. The observed convergence, agreement bias, and asymmetric acceptance-rejection bias reveal intrinsic tendencies within LLM agents to favor coherence and social alignment, potentially mirroring real-world psychological phenomena but also raising concerns about sycophantic behaviours [55]. However, the agreement we observe is not simply a result of sycophancy but emerges from asymmetric processing of differing opinions, with broader acceptance of more positive views. Moreover, the presence and persuasive impact of logical fallacies emphasize the need to critically evaluate the reasoning capabilities of such agents in social simulations, given their susceptibility to flawed arguments [39, 54]. Our study thus provides a foundation for further work exploring how language-driven models can inform both the dynamics of opinion formation and the risks associated with fallacious reasoning in AI-mediated social influence.

5 Conclusion

This study introduces a Language-Driven Opinion Dynamics Model for Agent-Based Simulations (LODAS), allowing for the exploration of how language and social influence shape opinion dynamics. By utilizing LLM agents, this study shows that synthetic agents, when left unprompted, tend to converge toward agreement, irrespective of initial opinion distributions or prompt framing. This convergence is primarily shaped by the underlying language model, with agents exhibiting a consistent asymmetric acceptance-rejection bias: they are more likely to adjust their opinions toward more positive stances and reject more negative ones. This bias is more pronounced in Mistral, which favors agreement more strongly, whereas Llama agents exhibit a form of bounded confidence, showing greater susceptibility to nearby opinions. In both cases, agents frequently employ logical fallacies in their persuasive attempts and are, in turn, influenced by such flawed arguments.

One key limitation of the current framework is the simplicity of the agents. In this model, agents are equipped with verbal reasoning skills but lack distinct personalities or cognitive diversity. The introduction of more complex agent types - such as those with different decision-making styles, biases or psychological traits - could better replicate the diversity of human interactions [57, 58].

Future extensions of the framework could also benefit from a deeper integration of cognitive biases [43] and demographic factors [59], as these elements are known to influence opinion dynamics in the real world. Furthermore, the model currently assumes a mean-field scenario, which neglects the structure of real-world social networks. Incorporating network features such as clustering, assortativity, or echo chambers could significantly increase the realism of the simulations and improve their ability to replicate polarization dynamics [32, 60, 61]. Preliminary tests with alternative network topologies and more sophisticated opinion dynamics algorithms suggest the potential to capture more complex patterns of interaction.

The exploration of fallacious reasoning in social simulations of LLM agents and its role in opinion dynamics has been approached at a preliminary level in this study, leaving substantial opportunities for future investigation. The role of fallacies poses challenges not only in the context of social simulations - where agents could potentially be optimised through better prompting, enhanced memory, or other refinements to mitigate fallacious reasoning - but also in human-LLM interactions. If LLMs are easily swayed by illogical arguments and tend to validate human perspectives, they may inadvertently reinforce false or potentially harmful beliefs.

To improve the understanding of these dynamics, several directions for future research can be pursued. One key focus is investigating methods to reduce fallacious reasoning in LLMs, such as through improved prompting, enhanced memory mechanisms, or adjustments to biases. Understanding the interplay between memory, bias, and opinion evolution is also critical for analyzing the role of persuasive language in opinion change. Comparing LLM-based simulations with real-world data from online interactions or controlled experiments can help evaluate (i) the robustness of the framework, (ii) its ability to replicate human behaviour, and (iii) the effects of linguistic features on opinion change under controlled conditions.

To summarize, despite its limitations, the framework provides a valuable tool for studying the mechanisms of consensus-building and argumentation in a controlled environment. The framework could serve as a foundation for exploring the drivers of opinion dynamics and their implications for phenomena such as polarization, bias, and misinformation.

List of Abbreviations

ABM Agent Based Model. 3

LLM Large Language Model. 1–9, 13, 14, 16, 17, 20, 28–31, 42, 43

LODAS Language-Driven Opinion Dynamics Model for Agent-Based Simulations. 1, 4, 6, 9, 11, 12, 14, 20, 24, 27, 28, 30, 42

OD Opinion Dynamics. 2–5, 7, 8

Supplementary information. The present article has accompanying Supplementary Information files with figures and tables complementary to those presented in the main text.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

The datasets generated and analyzed during the current study are within this paper or publicly available at [62]. Code to replicate simulations and analysis is publicly available at [62].

Competing interests

The authors declare that they have no competing interests.

Funding

This project is supported by SoBigData.it which receives funding from the European Union—NextGenerationEU—National Recovery and Resilience Plan (Piano Nazionale di Ripresa e Resilienza, PNRR)—Project: “SoBigData.it—Strengthening the Italian RI for Social Mining and Big Data Analytics”—Prot. IR0000013—Avviso n. 3264 del 28/12/2021 (to VP and RG); this work is also supported by the scheme ’INFRAIA-01-2018-2019: Research and Innovation action’, Grant Agreement No 871042 ’SoBigData++: European Integrated Infrastructure for Social Mining and Big Data Analytics’ (to RG); finally this work is supported by: the EU NextGenerationEU programme under the funding schemes PNRR-PE-AI FAIR (Future Artificial Intelligence Research) (to EC and RG)

Authors’ contribution.

EC analyzed the data and wrote the paper. VP analyzed the data and wrote the paper. GR designed the experiments, performed the experiments and supervised the project. DP supervised the project. All authors read and approved the final manuscript.

Acknowledgements.

We thank Daniele Atzeni for the valuable feedback and Giuliano Cornacchia for the help in designing plots and figures.

References

[1] Conte, R., Gilbert, N., Bonelli, G., Cioffi-Revilla, C., Deffuant, G., Kertesz, J., Loreto, V., Moat, S., Nadal, J.-P., Sanchez, A., *et al.*: Manifesto of computational social science. The European Physical Journal Special Topics **214**, 325–346 (2012)

[2] Tucker, J.A.: Computational social science for policy and quality of democracy: Public opinion, hate speech, misinformation, and foreign influence campaigns. Handbook of Computational Social Science for Policy, 381–403 (2023)

[3] Hobbes, T.: Leviathan; Or, The Matter, Forme, & Power of a Common-wealth, Ecclesiasticall and Civill. London, Printed for A. Crooke. http://archive.org/details/leviathan00hobba

[4] Comte, A., Martineau, H.: The Positive Philosophy of Auguste Comte. The Positive Philosophy of Auguste Comte 2 Volume Paperback Set. Cambridge University Press

[5] Li, L., Scaglione, A., Swami, A., Zhao, Q.: Consensus, polarization and clustering of opinions in social networks. IEEE Journal on Selected Areas in Communications **31**, 1072–1083 (2013)

[6] Biondi, E., Boldrini, C., Passarella, A., Conti, M.: Dynamics of opinion polarization. IEEE Transactions on Systems, Man, and Cybernetics: Systems **53**(9), 5381–5392 (2023) https://doi.org/10.1109/TSMC.2023.3268758

[7] Ramos, M., Shao, J., Reis, S.D.S., Anteneodo, C., Andrade, J.S., Havlin, S., Makse, H.A.: How does public opinion become extreme? Scientific Reports **5**(1) (2015) https://doi.org/10.1038/srep10032

[8] Castellano, C., Fortunato, S., Loreto, V.: Statistical physics of social dynamics. Reviews of Modern Physics **81**(2), 591–646 (2009) https://doi.org/10.1103/revmodphys.81.591

[9] Degroot, M.: Reaching a consensus. Journal of the American Statistical Association **69**, 118–121 (1974)

[10] Friedkin, N.E.: A formal theory of social power. Journal of mathematical sociology **12**, 103–126 (1986)

[11] Chen, X., Tsaparas, P., Lijffijt, J., Bie, T.D.: Opinion dynamics with backfire effect and biased assimilation. PLoS ONE **16** (2021)

[12] Monti, C., De Francisci Morales, G., Bonchi, F.: Learning opinion dynamics from social traces. In: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 764–773 (2020)

[13] Nyhan, B., Reifler, J.: When corrections fail: The persistence of political misperceptions. Political Behavior **32**, 303–330 (2010)

[14] Allahverdyan, A.E., Galstyan, A.: Opinion dynamics with confirmation bias. PLoS ONE **9**(7), 99557 (2014) https://doi.org/10.1371/journal.pone.0099557

[15] Liu, L., Wang, X., Chen, X., Tang, S., Zheng, Z.: Modeling confirmation bias and peer pressure in opinion dynamics. Frontiers in Physics **9**, 649852 (2021)

[16] Sîrbu, A., Pedreschi, D., Giannotti, F., Kertész, J.: Algorithmic bias amplifies opinion fragmentation and polarization: A bounded confidence model. PLoS ONE

14 (2019)

[17] Pansanella, V., Sîrbu, A., Kertesz, J., Rossetti, G.: Mass media impact on opinion evolution in biased digital environments: a bounded confidence model. Scientific Reports **13**(1), 14600 (2023)

[18] Monti, C., Aiello, L.M., De Francisci Morales, G., Bonchi, F.: The language of opinion change on social media under the lens of communicative action. Scientific Reports **12**(1), 17920 (2022)

[19] Park, J.S., Popowski, L., Cai, C., Morris, M.R., Liang, P., Bernstein, M.S.: Social simulacra: Creating populated prototypes for social computing systems. In: Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology, pp. 1–18 (2022)

[20] Park, J.S., Zou, C.Q., Shaw, A., Hill, B.M., Cai, C., Morris, M.R., Willer, R., Liang, P., Bernstein, M.S.: Generative Agent Simulations of 1,000 People (2024). https://arxiv.org/abs/2411.10109

[21] Premack, D., Woodruff, G.: Does the chimpanzee have a theory of mind? Behavioral and Brain Sciences **1**(4), 515–526 (1978) https://doi.org/10.1017/s0140525x00076512

[22] Kosinski, M.: Theory of mind may have spontaneously emerged in large language models. arXiv preprint arXiv:2302.02083 **4**, 169 (2023)

[23] Street, W., Siy, J.O., Keeling, G., Baranes, A., Barnett, B., McKibben, M., Kanyere, T., Lentz, A., Dunbar, R.I., et al.: Llms achieve adult human performance on higher-order theory of mind tasks. arXiv preprint arXiv:2405.18870 (2024)

[24] Li, H., Chong, Y.Q., Stepputtis, S., Campbell, J., Hughes, D., Lewis, M., Sycara, K.: Theory of mind for multi-agent collaboration via large language models. arXiv preprint arXiv:2310.10701 (2023)

[25] Ullman, T.: Large Language Models Fail on Trivial Alterations to Theory-of-Mind Tasks. arXiv (2023). https://doi.org/10.48550/ARXIV.2302.08399 . https://arxiv.org/abs/2302.08399

[26] Sap, M., Le Bras, R., Fried, D., Choi, Y.: Neural theory-of-mind? on the limits of social intelligence in large LMs. In: Goldberg, Y., Kozareva, Z., Zhang, Y. (eds.) Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pp. 3762–3780. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (2022). https://doi.org/10.18653/v1/2022.emnlp-main.248 . https://aclanthology.org/2022.emnlp-main.248

[27] Shapira, N., Zwirn, G., Goldberg, Y.: How well do large language models perform on faux pas tests? In: Rogers, A., Boyd-Graber, J., Okazaki, N. (eds.) Findings of the Association for Computational Linguistics: ACL 2023, pp. 10438–10451. Association for Computational Linguistics, Toronto, Canada (2023). https://doi.org/10.18653/v1/2023.findings-acl.663

[28] De Marzo, G., Pietronero, L., Garcia, D.: Emergence of scale-free networks in social interactions among large language models. arXiv preprint arXiv:2312.06619 (2023)

[29] Gao, C., Lan, X., Lu, Z., Mao, J., Piao, J., Wang, H., Jin, D., Li, Y.: S3: Social-network simulation system with large language model-empowered agents. arXiv preprint arXiv:2307.14984 (2023)

[30] Ashery, A.F., Aiello, L.M., Baronchelli, A.: Emergent social conventions and

collective bias in llm populations. Science Advances **11**(20), 9368 (2025)

[31] Mou, X., Wei, Z., Huang, X.: Unveiling the truth and facilitating change: Towards agent-based large-scale social movement simulation. arXiv preprint arXiv:2402.16333 (2024)

[32] Wang, C., Liu, Z., Yang, D., Chen, X.: Decoding echo chambers: LLM-powered simulations revealing polarization in social networks. In: Proceedings of the 31st International Conference on Computational Linguistics, pp. 3913–3923. Association for Computational Linguistics, Abu Dhabi, UAE (2025). https://aclanthology.org/2025.coling-main.264/

[33] Chuang, Y.-S., Goyal, A., Harlalka, N., Suresh, S., Hawkins, R., Yang, S., Shah, D., Hu, J., Rogers, T.T.: Simulating opinion dynamics with networks of llm-based agents. arXiv preprint arXiv:2311.09618 (2023)

[34] Breum, S.M., Egdal, D.V., Mortensen, V.G., Møller, A.G., Aiello, L.M.: The persuasive power of large language models. arXiv preprint arXiv:2312.15523 (2023)

[35] Flamino, J., Modi, M.S., Szymanski, B.K., Cross, B., Mikolajczyk, C.: Limits of large language models in debating humans. arXiv preprint arXiv:2402.06049 (2024)

[36] Priya, P., Firdaus, M., Ekbal, A.: Computational politeness in natural language processing: A survey. ACM Comput. Surv. **56**(9) (2024) https://doi.org/10.1145/3654660

[37] Törnberg, P., Valeeva, D., Uitermark, J., Bail, C.: Simulating social media using large language models to evaluate alternative news feed algorithms. arXiv preprint arXiv:2310.05984 (2023)

[38] Aher, G.V., Arriaga, R.I., Kalai, A.T.: Using large language models to simulate multiple humans and replicate human subject studies. In: International Conference on Machine Learning, pp. 337–371 (2023). PMLR

[39] Payandeh, A., Pluth, D., Hosier, J., Xiao, X., Gurbani, V.K.: How susceptible are LLMs to Logical Fallacies? (2023)

[40] RRV, A., Tyagi, N., Uddin, M.N., Varshney, N., Baral, C.: Chaos with keywords: Exposing large language models sycophantic hallucination to misleading keywords and evaluating defense strategies. arXiv preprint arXiv:2406.03827 (2024)

[41] Ranaldi, L., Pucci, G.: When large language models contradict humans? large language models' sycophantic behaviour. arXiv preprint arXiv:2311.09410 (2023)

[42] Likert, R.: A technique for the measurement of attitudes. **22 140**, 55–55

[43] Chuang, Y.-S., Goyal, A., Harlalka, N., Suresh, S., Hawkins, R., Yang, S., Shah, D., Hu, J., Rogers, T.: Simulating opinion dynamics with networks of LLM-based agents. In: Duh, K., Gomez, H., Bethard, S. (eds.) Findings of the Association for Computational Linguistics: NAACL 2024, pp. 3326–3346. Association for Computational Linguistics, Mexico City, Mexico (2024). https://aclanthology.org/2024.findings-naacl.211

[44] Ju, D., Williams, A., Karrer, B., Nickel, M.: Sense and Sensitivity: Evaluating the simulation of social dynamics via Large Language Models (2024). https://arxiv.org/abs/2412.05093

[45] Jiang, A.Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D.S., Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L.R., Lachaux, M.-A., Stock, P., Scao, T.L., Lavril, T., Wang, T., Lacroix, T., Sayed, W.E.: Mistral

7B (2023). https://arxiv.org/abs/2310.06825

[46] Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Yang, A., Fan, A., et al.: The llama 3 herd of models. arXiv preprint arXiv:2407.21783 (2024)

[47] Welch, B.L.: The generalization of 'student's'problem when several different population variances are involved. Biometrika **34**(1-2), 28–35 (1947)

[48] Microsoft: microsoft/autogen. https://github.com/microsoft/autogen Accessed 2025-02-13

[49] Manabat, B.K.: q3fer/distilbert-base-fallacy-classification · Hugging Face. https://huggingface.co/q3fer/distilbert-base-fallacy-classification Accessed 2025-02-12

[50] Jin, Z., Lalwani, A., Vaidhya, T., Shen, X., Ding, Y., Lyu, Z., Sachan, M., Mihalcea, R., Schölkopf, B.: Logical Fallacy Detection. arXiv (2022). https://doi.org/10.48550/ARXIV.2202.13758 . https://arxiv.org/abs/2202.13758

[51] Copi, I.M., Cohen, C., MacMahon, K.: Introduction to Logic, 14 ed., pearson new international ed edn. Pearson custom library. Pearson Education Limited

[52] Sîrbu, A., Loreto, V., Servedio, V.D., Tria, F.: Opinion dynamics: models, extensions and external effects. In: Participatory Sensing, Opinions and Collective Awareness, pp. 363–401. Springer, Cham (2017)

[53] Li, Y., Wang, D., Liang, J., Jiang, G., He, Q., Xiao, Y., Yang, D.: Reason from Fallacy: Enhancing Large Language Models' Logical Reasoning through Logical Fallacy Understanding (2024). https://arxiv.org/abs/2404.04293

[54] Mouchel, L., Paul, D., Cui, S., West, R., Bosselut, A., Faltings, B.: A

logical fallacy-informed framework for argument generation. arXiv preprint arXiv:2408.03618 (2024)

[55] Taubenfeld, A., Dover, Y., Reichart, R., Goldstein, A.: Systematic biases in llm simulations of debates. arXiv preprint arXiv:2402.04049 (2024)

[56] Oviedo-Trespalacios, O., Peden, A.E., Cole-Hunter, T., Costantini, A., Haghani, M., Rod, J.E., Kelly, S., Torkamaan, H., Tariq, A., David Albert Newton, J., Gallagher, T., Steinert, S., Filtness, A.J., Reniers, G.: The risks of using chatgpt to obtain common safety-related information and advice. Safety Science **167**, 106244 (2023) https://doi.org/10.1016/j.ssci.2023.106244

[57] Cava, L.L., Tagarelli, A.: Open Models, Closed Minds? On Agents Capabilities in Mimicking Human Personalities through Open Large Language Models (2024). https://arxiv.org/abs/2401.07115

[58] Huang, J.-t., Lam, M.H., Li, E.J., Ren, S., Wang, W., Jiao, W., Tu, Z., Lyu, M.R.: Emotionally Numb or Empathetic? Evaluating How LLMs Feel Using EmotionBench (2024). https://arxiv.org/abs/2308.03656

[59] Wang, Z., Chiu, Y.Y., Chiu, Y.C.: Humanoid agents: Platform for simulating human-like generative agents. In: Feng, Y., Lefever, E. (eds.) Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pp. 167–176. Association for Computational Linguistics, Singapore (2023). https://doi.org/10.18653/v1/2023.emnlp-demo.15 . https://aclanthology.org/2023.emnlp-demo.15/

[60] Piao, J., Lu, Z., Gao, C., Xu, F., Santos, F.P., Li, Y., Evans, J.: Emergence of human-like polarization among large language model agents (2025). https://arxiv.org/abs/2501.05171

[61] Zheng, W., Tang, X.: Simulating social network with LLM agents: An analysis of information propagation and echo chambers. In: Tang, X., Huynh, V.N., Xia, H., Bai, Q. (eds.) Knowledge and Systems Sciences, pp. 63–77. Springer. https://doi.org/10.1007/978-981-96-0178-3_5

[62] Cau, E.: ericacau/LLM-Opinion-Dynamics. https://github.com/ericacau/LLM-Opinion-Dynamics Accessed 2025-02-12

Figure Legends

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

Figure 2. Balanced scenario – Mistral and Llama agents opinion trends. Mistral (a)-(b) and Llama (c)-(d) opinion trends for the positive (a)-(c) and negative (b)-(d) statements. Trends are represented for Strongly Disagree (dark red), Disagree (red), Mildly Disagree (orange), Neutral (yellow), Mildly Agree (light blue), Agree (blue), and Strongly Agree (dark blue) opinions. Lines indicate the average prevalence of opinions at each time step, while the shade indicates the 95% confidence interval. Averages are computed over 10 runs.

Figure 3. Balanced scenario – Mistral and Llama-agents transition matrices. Mistral (a)-(b) and Llama (d)-(e) average (with std) transition rates from state i to state j for the positive (a)-(d) and negative (b)-(e) statements. Annotated cells are significant with respect to the Random Null Model ($p < 0.05$) according to the t-test. Results are averaged over 10 runs.

Figure 4. Balanced scenario – Mistral and Llama-agents average acceptance probabilities $\mathbb{P}(\mathbf{A}|o_D, o_O)$. Mistral (orange line and matrices (a)-(b)) and Llama (blue lines and matrices (c)-(d)) average acceptance rates for the positive (solid lines and (a)-(c) matrices) and negative (dashed lines and (b)-(d) matrices) statements. Top panel: average probability of the Discussant accepting the Opponent's opinion as a function of $\Delta x = x_O - x_D$. Bottom panel: matrices represent acceptance rates averaged over 10 runs.

Figure 5. Acceptance probability $P(A \mid \Delta_x)$ as a function of opinion distance $\Delta_x = x_O - x_D$. Each marker represents the average probability that a Discussant agent accepts—i.e., moves toward—the Opponent's opinion, as a function of the opinion distance $\Delta_x = x_O - x_D$. Lines correspond to different initial opinion distributions: *Balanced* (blue), *Polarized* (orange), and *Unbalanced* (green). Rows distinguish between the two language models: Mistral (top) and Llama (bottom). Columns refer to the direction of the statement: *Same* (left) and *Different* (right). Results are averaged over 10 independent runs; shaded areas indicate standard deviations across runs.

Figure 6. Average logical fallacies proportions for different experiment configurations across multiple runs. Figures (a)-(b) refer to Llama, while (c)-(d) refer to Mistral. (a)-(c) refer to Llama (a) and Mistral (c) agents discussing the *same boat* statement, (b)-(d) refer to the *different boat* statement. Error bars represent standard deviation across 9 runs. The x-axis uses the following abbreviations: **R** (fallacy of relevance), **L** (fallacy of logic), **CR** (circular reasoning), **C** (fallacy of credibility), **G** (faulty generalization), **I** (intentional), and **FC** (false causality).

Table Legends

Table 1. Overview of works in the literature introducing LLM agents.

Table 2. Percentage of logical fallacies in *Opponents'* statements. Percentage of unique *Opponent* (O) statements classified as fallacious, across models (Llama, Mistral), initial opinion distribution (balanced, polarized, unbalanced), and opinion framing (same, different).

Table 3. Ratio of *Discussants* changing opinion for the effect of fallacious statements.. Percentage of *Discussants* (D) changing opinion for the effect of fallacious statements produced by *Opponents*, across models (Llama, Mistral), initial opinion distribution (Balanced, Polarized, Unbalanced), and opinion framing (same, different).

Table 4 Percentage of opinion changes in Discussands due to logical fallacies.. Values indicate the percentage of opinion shifts in the *Discussant* agents exposed to each fallacy type by the *Opponent*'s statement. Results refer to Llama and Mistral agents discussing the *same boat* (S) and *different boat* (D) initial statement.