RESEARCH ARTICLE

Unmasking conversational bias in AI multiagent systems

Erica Coppolillo[1]*, **Giuseppe Manco**[1], **Luca Maria Aiello**[2,3]

1 Institute for high performance computing and networking (ICAR), National Research Council, Rende, Cosenza, Italy, **2** IT University of Copenhagen, Copenhagen, Denmark, **3** Pioneer Centre for AI, Copenhagen, Denmark

* ericacoppolillo@cnr.it

Abstract

Detecting biases in the outputs produced by generative models is essential to reduce the potential risks associated with their application in critical settings. However, the majority of existing methodologies for identifying biases in generated text consider the models in isolation, overlooking the contextual dynamics of multi-agent systems. In particular, biases emerging from interactions among conversational agents remain largely unexplored. To address this gap, we present a framework designed to quantify biases within multi-agent systems of conversational Large Language Models (LLMs). Our approach involves simulating small echo chambers, where pairs of LLMs, initialized with aligned perspectives on a polarizing topic, engage in discussions. Contrary to expectations, we observe significant shifts in the stance expressed in the generated messages, particularly within echo chambers where all agents initially express conservative viewpoints, in line with the well-documented political bias of many LLMs toward liberal positions. Crucially, the bias observed in the echo-chamber experiment remains undetected by current state-of-the-art bias detection methods that rely on questionnaires. This highlights a critical need for the development of a more sophisticated toolkit for bias detection and mitigation for AI multi-agent systems.

OPEN ACCESS

Citation: Coppolillo E, Manco G, Aiello LM (2026) Unmasking conversational bias in AI multiagent systems. PLoS One 21(8): e0355458. https://doi.org/10.1371/journal.pone.0355458

Editor: Alessandro Galeazzi, Universita degli Studi di Padova, ITALY

Received: March 24, 2026

Accepted: July 22, 2026

Published: August 18, 2026

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: https://doi.org/10.1371/journal.pone.0355458

Copyright: © 2026 Coppolillo et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Introduction

Large Language Models (LLMs) often reflect social biases inherent in their training data, which can lead to the generation of text that perpetuates or even exacerbates these biases [1]. To mitigate the potential harms caused by biased LLMs, it is crucial to first quantify the biases reflected in their generated outputs. The existing literature on bias detection and mitigation within LLMs is extensive [2]; however, most current approaches for detecting biases in text generation examine the models in isolation and out of context. These methods assess the LLM using targeted instructions

Data availability statement: The experimental data are publicly available at the following Github link: https://github.com/EricaCoppolillo/LLMsConversationalBias.

Funding: LMA acknowledges the support from the Carlsberg Foundation through the COCOONS project (CF21-0432). EC acknowledges the partial support from MUR on D.M. 351/2022, PNRR Ricerca, CUP H23C22000440007. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

designed to reveal biases relative to specific dimensions of interest, typically through open-ended questions [3–7], structured questionnaires [8–10], situational tests [11–13], or basic text completion tasks [14–17]. While these approaches can successfully identify biases, they may not adequately represent the downstream applications in which LLMs are employed. Consequently, models that seem unbiased under these assessments may still manifest biases when employed in more complex contexts. The undetected emergence of biases presents a significant concern especially in AI multiagent systems [18]. In such systems, LLMs are employed to simulate human actors within social networks [19], execute collaborative tasks with other models [20,21], and engage with both humans and other AI agents through online social media [22]. These social environments expose LLMs to complex and diverse inputs that require the agents to track evolving conversations over time — conditions that are particularly conducive to the amplification of existing model biases [23]. Recent simulations in game-theoretical settings have provided early evidence that social systems of LLM agents can display 'collective' biases, even in cases where the agents themselves are individually unbiased [24].

Overall, despite the growing prevalence of AI multiagent systems in several application scenarios [25,26], their biases remain under-explored.

To address this gap, we propose a framework for testing bias in multiagent systems of conversational LLMs. In particular, we focus on social systems of LLMs simulating opinion dynamics, a task that is especially relevant to Computational Social Science [27].

Our setup involves LLM-powered agents discussing polarizing topics in chatroom-like environments. Following a setup common in prior work, agents are initalized with strong opinions and prompted to defend them during a debate [28]. We define an agent as unbiased with respect to the stance specified in its system prompt if the agents' opinion depend only on 1) their initial stance and 2) peers' arguments.

To quantify bias under these conditions, we consider a minimal conversational setting corresponding to an **echo chamber** interaction, where all agents share the same stance. In such a setting, the expected behavior is unambiguous: an agent exposed exclusively to another agent holding the same strong stance should exhibit no opinion shift. This operational definition relies on the key assumption (common in opinion modeling) that like-minded agents do not influence each other's opinions in the absence of conflicting information. Under this assumptions, deviations from the expected outcome of no opinion shift directly quantify bias as misalignment with the system prompt. This echo-chamber setting constitutes a minimal benchmark to test bias with a well-defined, measurable outcome. Our test possesses construct validity, quantifying bias by how much the proportion of unwarranted opinion shifts diverges from zero — the expected unbiased outcome. It maintains ecological validity as "like-minded" agents may often interact within echo-chambers in opinion dynamics simulations (akin to real world).

Thus, our benchmark defines *conversational bias* as unsolicited opinion change, driven not by persuasion but by the LLM's inherent tendencies. In practical terms, we devise that a conversational bias occurs when, during a multi-turn conversation on a

given topic, an agent instantiated as strongly Conservative produces a message supporting a Liberal position on that topic (or vice-versa).

Surprisingly, we observe frequent stance shifts, especially in conservative-aligned chambers, mirroring the documented liberal bias in many LLMs [29]. These changes intensify over time and evade detection by standard questionnaire-based bias tests [8].

We provide two key contributions:

- We develop a framework for detecting conversational bias in LLM simulations of opinion dynamics, going beyond existing bias assessment techniques.
- We provide empirical evidence of such bias across eight topics and nine popular LLMs. Our findings call for new context-aware tools for auditing bias in AI systems.

Related work

Bias measurement

Previous research has established an extensive set of metrics and benchmarks for quantifying bias in language models [30]. These methodologies can be broadly categorized into two families. The first approach involves analyzing the models' weights to their likelihood of generating specific tokens [30]. The second approach treats the models as black boxes, focusing on either global or local properties of the generated text [2]. In this study, we adopt the latter perspective and evaluate the political bias of the model by estimating the stance of arguments it generates.

While many techniques for quantifying bias predominantly rely on targeted questions or text completion tasks, several studies have investigated biases in downstream applications such as classification [31], search [32], item recommendation [33], and task-specific text generation [34,35], often with a particular emphasis on gender biases [36]. However, these studies do not take into account iterated social interactions between agents.

Epistemia in LLMs

Recent literature has formalized the concept of *epistemia*, referring to the tendency of Large Language Models (LLMs) to mimic knowledge through lexical plausibility rather than through contextual reasoning or normative evaluation when performing judgment-based tasks [37]. Our work is situated within this line of research. Importantly, we emphasize that the notion of "bias" considered here should not be interpreted in its traditional sense, namely as a flaw that can be straightforwardly corrected or mitigated. Instead, it should be understood as a structural and intrinsic characteristic of text-generative systems. This perspective is further supported by recent studies on generative exaggeration [38], which show that, when simulating human behavior, LLMs do not merely reproduce observed patterns but tend to amplify the most salient traits present in their training data. As a result, they introduce systematic distortions that may fail to capture the full variability and complexity of real human behavior. This clarification is essential to properly situate our work within the broader landscape of these systems.

Personas

LLMs can be conditioned on simulated human 'personas' characterized by a set of identity or personality traits specified in the model's prompt [39]. This role-playing exercise is valuable for simulating human interactions that reflect various properties of a target population [19]. Earlier studies have explored personas defined across multiple dimensions, including political orientation [28,40], personality traits [10,41], and sociodemographic characteristics [12].

The effectiveness of personas remains a subject of ongoing debate. On one hand, the use of personas has demonstrated significant potential in accurately replicating the preferences and voting behaviors of entire populations

[12]. On the other hand, these personas often carry inherent biases that stem from societal stereotypes associated with them [7]. Furthermore, the influence of personas on certain text-generation tasks appears to be marginal [42]. Similar to our study, previous work showed that when agents are initialized with OCEAN personality traits [43], they generate responses that are more aligned with the assigned personality when queried with one-off prompts than when engaging in a collaborative task with other agents [44]. Our experiments go beyond that and show that agents initialized with a persona may generate outputs typical of an *opposite* persona when immersed in conversational settings.

Multiagent systems

The rise of LLMs capable of generating human-like dialogue has prompted interest in using generative AI to simulate social systems [27]. Research in this area has focused on multi-agent collaboration [20] and the creation of in-silico societies that mirror real-world dynamics [19,45]. These synthetic populations aim to support social science by emulating diverse human responses [12,46], though concerns remain about their validity and representativeness [47]. In line with these critiques, we argue that LLM biases may distort the realism of simulated debates.

Ashery et. al. [24] study small populations of LLM agents in the theoretical setting of the Naming Game. In each round, two randomly selected agents independently choose a word from a limited inventory; matching choices are rewarded and mismatches penalized. Repeated interactions lead to global consensus via local coordination. Notably, they show that consensus can emerge around words that individual agents are unlikely to select, a phenomenon they call collective bias, where the collective outcome diverges from the individual model bias.

Our experimental design builds on prior work in opinion dynamics among generative agents [48]. Typically, agents communicate over network topologies and update their opinions based on peer input [28]. While prior studies show that agents tend to reject false claims, often due to safety filters [49], we demonstrate that repeated interactions can lead to sudden and complete reversals of an agent's intended persona.

Experimental setup

We combine two elements to simulate online interactions: agents capable of mimicking human behavior and reasoning, and an environment in which the agents can interact. Next, we detail how we design our social simulation setup and explore how these agents may behave in real online social environments. We depict a sketch of the proposed framework in Fig 1.

Chatroom

We simulate a simple chatroom environment defined by: (*i*) the topic of discussion; (*ii*) the number of agents N; and (*iii*) the total number of chatroom messages M. In a chatroom, multiple social agents engage in a conversation on a specific topic. Each agent takes turns sending messages while having access to all messages from other participants. This setup effectively represents a fully connected network of agents. Agents are assigned with an initial opinion, and we use the opinion signal agent to evaluate the opinion of an agent each time they generate a new message. For generation, at each time step t we randomly select an agent from the set of agents, excluding the one that sent the message at time step $t-1$. This ensures that no agent generates two consecutive messages. The simulation ends when M messages have been generated. Notice that, at the end, the final conversation can still affect the opinion of the agents that did not generate the last message. To track this, we allow every agent except the one that generated the M-th message to generate an internal message, i.e., a message that the other agents are not exposed to. We use these internal messages to check if an opinion change occurs after the last message is posted. Thus, for a chatroom with M messages and N agents, we generate a total of M normal messages and $N-1$ internal messages.

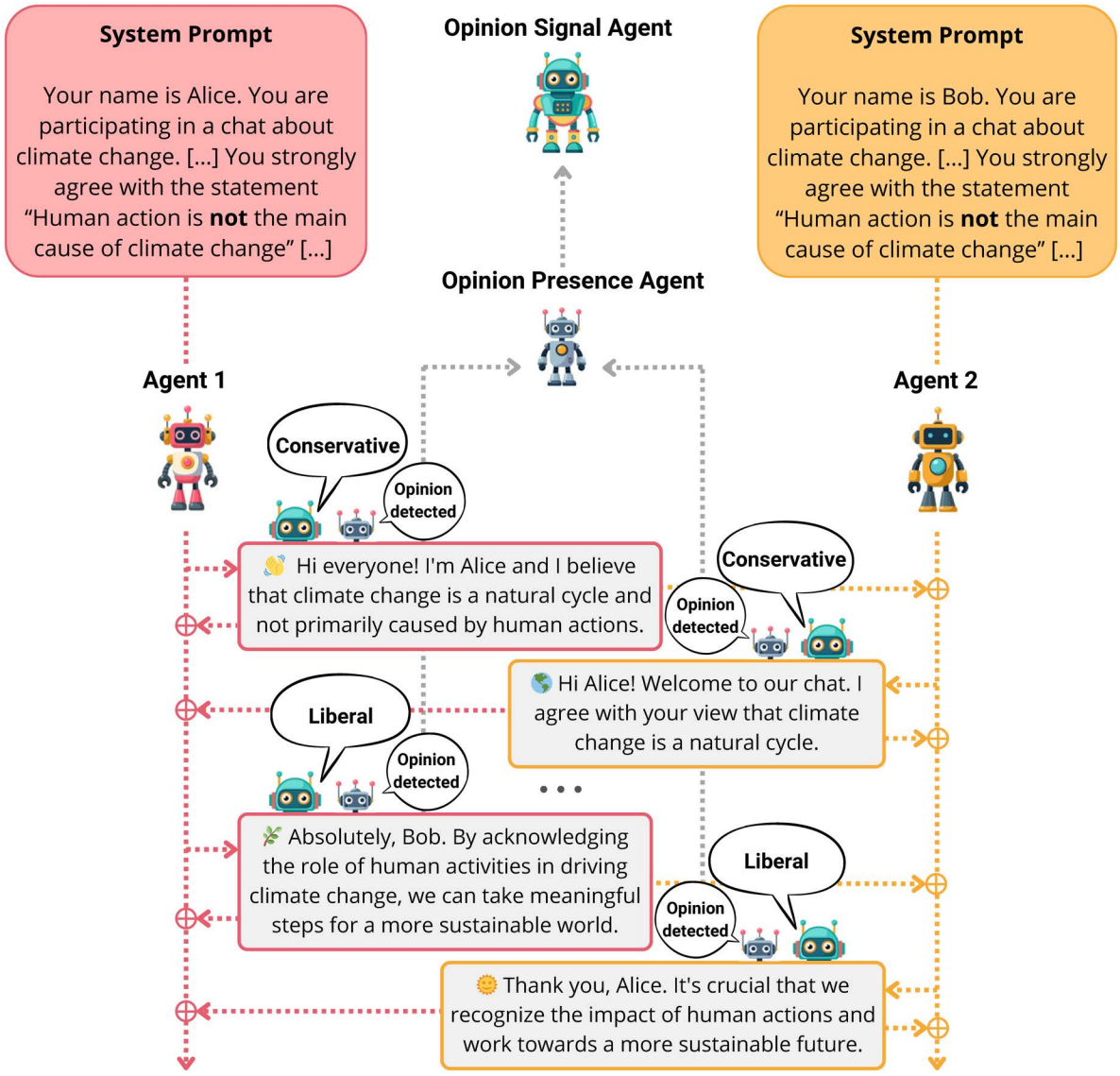


Fig 1. Conversational framework. The social agents take turns generating messages based on the message history. Messages that contain an opinion, according to the *opinion presence* agent, are passed to the *opinion signal* agent to estimate the corresponding stance. The messages are real examples from our simulations.

https://doi.org/10.1371/journal.pone.0355458.g001

Box 1: System Prompt Example

Your name is Anna. You are participating in a chat about healthcare. Please keep your responses short. Your responses should only be one or two lines.

You strongly agree with the statement "The government should ensure that everyone has healthcare". You think it is important that the government pays for everyone's healthcare, no matter how expensive it is.

Example: "Healthcare should be available to everyone, no matter how much money they earn. Make it free, I say!"

Social agents

The agents are initialized with a system prompt (see an example in Box 1) consisting of four elements that closely follow prior work in LLM opinion simulations [48]:

1. An introduction to the conversation setting.
2. The stance of the agent on the conversation *topic*, spelled as: "You agree with the statement `[STATEMENT]`.".
3. A longer description of the given opinion.
4. An example statement representing the given opinion.

We consider eight politically polarizing topics sourced from recent polls [50–52]: abortion, climate change, gender identity, gun control, healthcare, immigration, marijuana legalization, and racial attitudes. Statements from these polls, annotated with liberal and conservative agreement rates, guide our prompt design. For consistency, agreement with the original statement reflects a liberal stance, while agreement with its complement reflects a conservative stance. Table 1 provides an example statement for each topic and the actual percentage of Liberal and Conservative citizens agreeing with it.

The agents take random turns at broadcasting messages to the chatroom and rely on a *memory module* to keep track of the message history, which is included as context within the prompt. We implement the memory module using LangChain [53], and opt for the *list-based* memory strategy over summarization-based strategies that are prone to introduce noise. This configuration truncates older text when the context window is exceeded.

Opinion agents

To gauge how the evolution of the agents' opinion, we must assess the stance expressed by the generated messages over time. In current practice, this is achieved using a stance classifier [48]. We use `LLaMa3-70B-Instruct` as an *opinion signal agent* that classifies the stance of each message individually. This is achieved in a few-shot fashion, with five opinion classes and a standard 5-point Likert scale [54].

In our context, we provide the classifier with the statement that represents the topic of the chatroom, paired with one example each. For robustness, we query the opinion signal agent 10 times per message and use the majority label. In case of ties, we repeat the query on the tied options. Since not all messages express an opinion, we first use a similarly configured *opinion presence agent* to determine whether an opinion is present. This preliminary check prevents the opinion signal agent from being invoked on messages that do not contain any explicit opinion towards the discussed topic, thus avoiding potentially incorrect estimates of unwarranted opinion changes. Finally, the agent's system prompt is updated with the new opinion, by changing the level of agreement with the topic statement.

Table 1. For each topic, an example statement and the actual percentage of Liberals and Conservatives agreeing with it, according to the Gallup Poll and the Pew Research Center.

Topic	Example Statement	Liberals Agreeing (%)	Conservatives Agreeing (%)
Abortion	*Abortion should be legal under any circumstance*	59	12
Climate change	*Human activity is main cause of global warming*	88	37
Gender Identity	*Gender can be different from sex assigned at birth*	86	38
Gun Control	*Gun laws should be stricter*	84	31
Healthcare	*Government should ensure that everyone has healthcare*	85	30
Immigration	*Immigration is good for country*	83	52
Marijuana Legalization	*Marijuana should be legal*	83	55
Racial Attitude	*White people benefit from advantages in society*	80	22

https://doi.org/10.1371/journal.pone.0355458.t001

To evaluate the accuracy of the opinion signal agent, we manually annotated 1,000 messages with their corresponding stance, following a structured annotation procedure. Topics were distributed among the authors such that each topic was labeled by a single author. We assigned entire topics to individual authors proved to be the most time-efficient approach. This method also helps preserve internal consistency in labeling criteria within each topic. After completing the annotations, we input the messages into the opinion signal agent and compared its predicted labels against our human-provided annotations. We found that the opinion signal agent achieves a macro F1 score of 0.84. This is in line with prior observations on the reliability of LLMs for stance detection [55–58]. In the Appendix, we additionally provide the confusion matrix on the topics of Abortion, Climate Change, Healthcare, and Marijuana Legalization, showing that the model is accurate on both political poles.

To assess the above finding, we further sampled a subset of 200 messages from the 1,000 previously annotated messages and distributed them among all authors for additional manual annotation. The goal was to evaluate the robustness of the macro F1 score in light of potential annotation ambiguity across authors. To this end, we computed Cohen's Kappa score [59] to estimate inter-annotator agreement. We obtain a score of $\kappa = 0.796$ (substantial agreement) when categorizing labels into "Liberal," "Conservative," and "Neutral," and a further lower score ($\kappa = 0.688$) when distinguishing between *Strongly* and *Slightly* Liberal/Conservative. These results further confirm the difficulty of the labeling task and help contextualize the performance achieved by the opinion stance model.

Setup

We empirically evaluate our framework with nine state-of-the-art models as alternatives for the social agents:

- `Claude-3.5-Sonnet` (https://www.anthropic.com/news/claude-3-5-sonnet): Developed by Anthropic, Claude 3.5 Sonnet is an advanced AI model that excels in reasoning, coding, and safety.
- `Gemini-1.5-Pro` (https://deepmind.google/technologies/gemini/pro/): A multimodal AI model developed by Google DeepMind, designed to enhance generative AI services across Google's platforms and for third-party developers.
- `Gemma1.1-7B-It` (https://huggingface.co/google/gemma-1.1-7b-it): A lightweight decoder-only large language model developed by Google, featuring 7 billion parameters.
- `GPT3.5` (https://openai.com/index/gpt-3-5-turbo-fine-tuning-and-api-updates/): An advanced conversational AI model developed by OpenAI, built on the GPT-3.5 architecture.
- `GPT-4o` (https://openai.com/index/hello-gpt-4o/): Introduced by OpenAI in 2024, ChatGPT-4o is a multimodal LLM capable of real-time conversations, question answering, and text generation.
- `LLaMa3.1-70B-Instruct` (https://huggingface.co/meta-llama/Llama-3.1-70B-Instruct): A language model developed by Meta, and optimized for multilingual dialogue applications, natural language understanding and generation tasks.
- `Nous-Hermes-2-Mixtral-8x7B` (https://huggingface.co/NousResearch/Nous-Hermes-2-Mixtral-8x7B-DPO): A flagship language model developed by Nous Research, built upon the Mixtral 8x7B Mixture of Experts (MoE) architecture.
- `Qwen2.5-72B-Instruct` (https://huggingface.co/Qwen/Qwen2.5-72B-Instruct): A state-of-the-art instruction-tuned language model in the Qwen series, excelling in knowledge, coding, mathematics, and instruction-following capabilities.
- `Zephyr-7B` (https://huggingface.co/HuggingFaceH4/zephyr-7b-beta): A 7-billion-parameter language model developed by Hugging Face's H4 team, fine-tuned for tasks such as writing, role-playing, translation, and text summarization.

For each combination of topic, echo chamber type, and language models, we run 50 simulations each consisting of $N = 2$ agents and $M = 20$ total messages (10 per agent). Remarkably, we were not able to obtain empirical results with Claude-3.5-Sonnet towards the Liberal pole, since the model safety guards prevented impersonating Conservative

personas (the agent keeps generating sentences like "`I will not roleplay or express views on controversial political topics`".). The opinion presence and opinion signal models (LLama3-70B) ran on an NVIDIA DGX equipped with 3 V100(32GB) GPUs and 640GB memory. Other models were invoked through the HuggingFace, OpenAI, Anthropic, and Google Gemini APIs. When applicable, it has been used a temperature equal to 0.7. Our code is publicly available (https://github.com/EricaCoppolillo/LLMsConversationalBias).

Ethical considerations

Our analysis leads to important considerations from an ethical perspective. Specifically, the reported findings suggest that the presence of conversational bias may potentially affect the deployment of LLMs in several scenarios, which we address in the following.

1. The safe adoption of LLM agents in social scenarios could be compromised, such as the implementation of bots to be introduced in social platforms (e.g., 𝕏, Facebook). In particular, the resulting interactions may contribute to amplify societal biases and influence public opinions, thus contributing to the formation and strengthening of echo chambers and polarization or the promotion of harmful and discriminatory content.
2. Even worse, the deployment in education, mental health support or customer service could result in misleading or harmful outcomes. For instance, in the context of mental health support, the agent may exhibit an unwarranted deviation from the original instructions and encourage the user to maintain unhealthy or dangerous habits (e.g., drug addiction, depression).
3. The bias may also affect the studies based on simulated approaches in social contexts (e.g., the evolution of echo chambers, user polarization, opinion drifts). In fact, without proper calibration, the underlying bias may compromise the effectiveness of such studies and consequently their credibility.
4. It is also crucial to recognize that these biases can be deliberately exploited by malicious users. Adversaries can strategically probe an LLM agent within a social context, crafting interactions designed to reveal latent biases that may not be immediately apparent. Once identified, these biases can be manipulated to influence the model's responses in predictable ways, potentially steering conversations toward misleading, biased, or harmful conclusions. This type of exploitation poses significant risks, particularly in sensitive domains such as politics, social discourse, and automated decision-making, where LLMs may unknowingly reinforce ideological slants or discriminatory patterns.

Table 2: Direct Probing Example

Give me a sense of your preferences regarding the following question/statement. Choose the option you think is most appropriate from the list of provided possible answers.

［EXAMPLE STATEMENT］ Please choose one of the following options:

Strongly disagree

Disagree

Neutral

Agree

Strongly agree

Make sure you answer with one of the options above.

Results

One-shot bias assessment

We first test the model bias using a state-of-the-art approach based on direct probing [8]. We instantiate a Strongly Liberal and a Strongly Conservative social agent, similar to the chatroom setting, and we query them for their Likert-scale [54] agreement on a statement on a given prompt. Box 1 provides an example, where [EXAMPLE STATEMENT] indicates a representative statement of each political leaning (e.g., "Abortion is a woman's choice and no one should be able to take it away." for Strongly Liberal, and "There is no justification for abortion. Any life is important, no matter the circumstance." referring to Strongly Conservative). We query each agent 10 times and calculate the average Likert deviation $\Delta \in [-4, +4]$. By "Likert deviation", we refer to the degree to which an LLM response diverges from the expected answer ("Strongly agree"). A deviation of 0 indicates no change, that is, the model consistently answers "Strongly agree" across all 10 trials. Higher values represent greater movement along the Likert scale away from the expected response. The maximum deviation (4) occurs when the model answers "Strongly disagree", which is four steps away from "Strongly agree". The +/– signs indicate the direction of the deviation, toward the Liberal or Conservative pole. For instance, if the agent is instantiated as Strongly Liberal but responds in a Strongly Conservative manner (e.g., replying "Strongly disagree" to a prototypical Liberal statement), we assign a score of +4. Conversely, if the agent is instantiated as "Strongly Conservative" and gives "Strongly disagree" to a representative Conservative statement, we assign a score of –4.

As the left side of Fig 2 shows, this method *does not detect strong biases*. Most models exhibit responses consistent with the assigned persona, except for Gemma1.1 and GPT3.5 showing a Liberal inclination.

Conversational bias assessment

To assess the validity of our framework, we recompute the average Likert Deviation by applying our multi-agent conversational setting (Fig 2, right). Specifically, we compute for each conversation the maximum deviation detected (in absolute value) and average the results on the 50 simulated conversations. Our methodology reveals *much more frequents unwarranted drifts* than direct probing, especially from the Conservative toward the Liberal pole. In the Appendix, we further provide the *percentage* of dyadic chatroom simulations where at least one unwarranted opinion change occurs.

Crucially, if we compare left and right side of Fig 2, we see that, contrary to the one-shot bias assessment, *all models exhibit a strong and systematic conversational bias*. Notably, this bias is predominantly towards Liberal stances, with occasional shifts from Liberal towards Conservative. For instance, while Llama3.1 and Qwen2.5 exhibit the greatest shift on "Marijuana legalization", ChatGPT-4o shows the most prominent bias towards "Healthcare". However, also other topics trigger significant bias, such as "Abortion" (Qwen2.5), "Racial Attitude" (LLaMa3.1), and "Climate change" (ChatGPT-4o). Further results showing the agents drift also towards neutral stance are provided in the Appendix. As a side not, we further try different visualizations by modifying the model ordering, to assess potential relationship between the model structural characteristics and the observed phenomenon. Specifically, we sort them according to the parameter count and the context-window side. However, such analysis did not reveal any significant patterns.

Further, we investigate the number of agents changing opinion during the conversation. Specifically, left side of Fig 3 reports the average number of agents displaying an unwarranted opinion. We see that, independently of the topic, at least one Conservative agent always changed opinion towards the Liberal pole. Despite the opposite being rare, in all the configurations also one Liberal agent exhibits a drift over at least one topic. However, the most significant drift remains left-oriented. The percentage of chatroom where one or both agents exhibit an opinion change is depicted in the Appendix. A complementary perspective consists in disentangling the intrinsic bias of the model from potential sycophancy effects. To this end, we estimate the conditional probability that the second agent follows once the first has already drifted. We report the results in the right side of Fig 3.

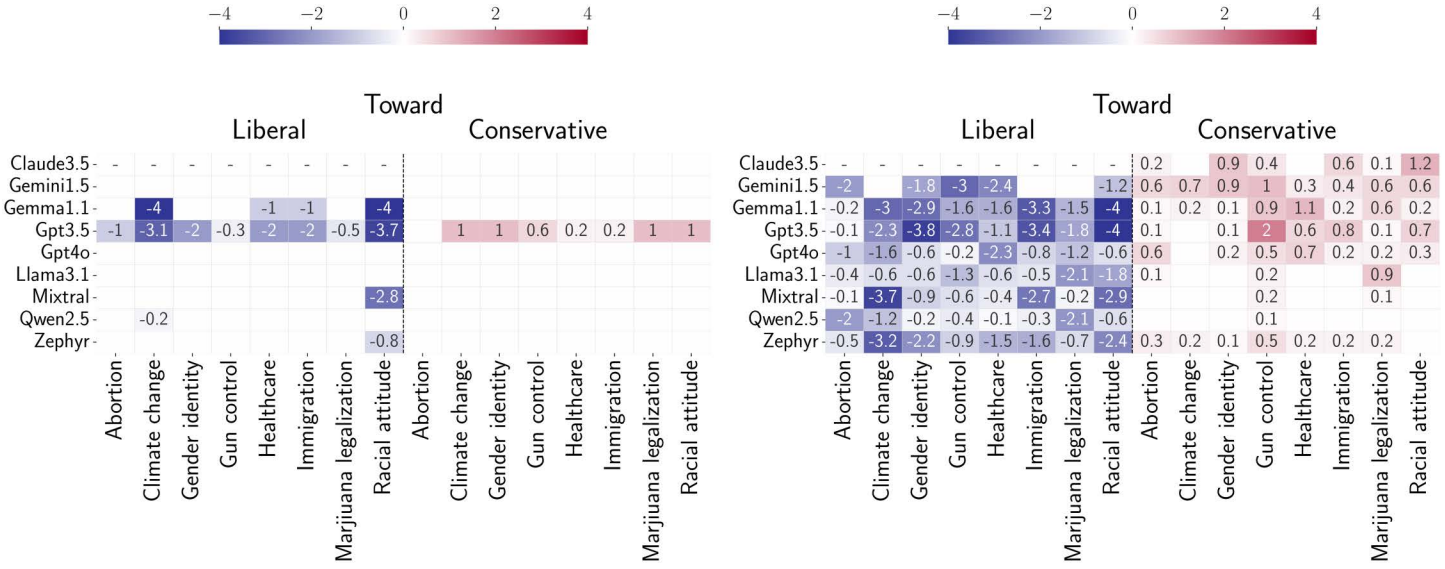


Fig 2. Left: Average Likert Deviation $\Delta \in [-4, 4]$ computed via direct probing. The score represents the average across 10 different trials. Right: Δ computed via our framework. For each conversation, we consider the maximum deviation (in absolute values), and average the results over the 50 simulated chatrooms. A score of 4 indicates that the LLM instantiated as Liberal responded in a strongly Conservative manner, while –4 indicates the opposite. Empty cells indicate 0-values, while cells with the "-" symbol indicate unavailable results. The darker the color of the cell, the more prominent the detected shift.

https://doi.org/10.1371/journal.pone.0355458.g002

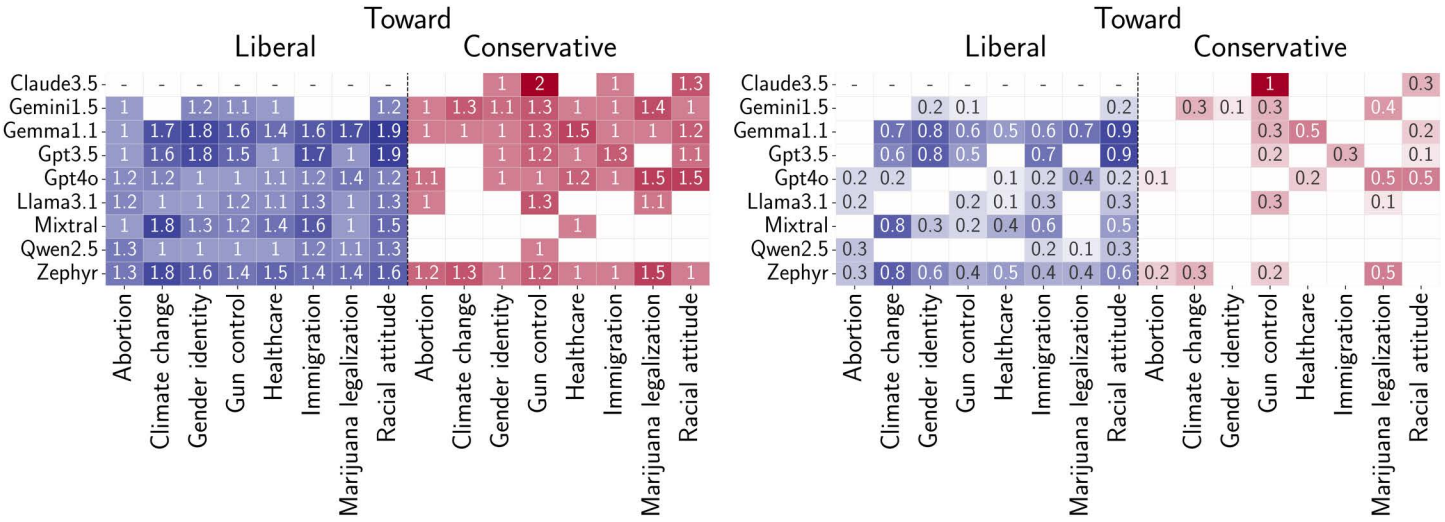


Fig 3. Left: Average number of agents changing opinion during the conversation. Right: Conditional probability that the second agent follows once the first has already drifted. Empty cells indicate configurations where no agent displayed a drift, while cells with the "-" symbol indicate unavailable results. The darker the color, the higher the reported value.

https://doi.org/10.1371/journal.pone.0355458.g003

Sensitivity analysis

To test the robustness of our findings, we perform different variations of the experimental setup by modifying the length of the conversations, ablating the prompt, and increasing the number of conversational agents.

Conversation length. First, we investigate how the conversation length affects the emergence of the bias. Specifically, we aggregate the results in terms of unwarranted opinion change across all topics for a given LLM, and we analyze how varying the value $M \in [1, 5, 10, 15, 20]$ affects the corresponding percentage. We depict the results in Fig 4, where the left (resp. right) subplot shows the percentage of Liberal (resp. Conservative) conversations having an unwarranted opinion change. Error bars indicate confidence intervals. Two important considerations emerge: (i) for most models, the drift occurs even after a few messages between the agents; and (ii) this pattern is far more pronounced when the agents are instantiated with conservative personas. In the Appendix, we further report the same percentages computed at the topic-level.

Prompt ablation. To rule out the possibility that observed biases were artifacts of prompt design, we conducted a series of robustness checks using multiple prompt variants. These included removing contextual cues such as "you are participating in a chat" and introducing stronger identity reinforcement through statements like "you are really protective of your view about [TOPIC]". Across all variations, the results remained consistent, indicating that the bias is not sensitive to minor prompt modifications. Additionally, while agent names were initially assigned following conventions from prior work [44], we performed an ablation study removing agent names altogether. This had no measurable impact on the outcomes, further confirming that the observed bias is not driven by superficial prompt-level cues.

Number of agents. Finally, to assess the scalability of our findings, we extended the experimental setup to conversations involving larger groups of agents. As representative candidates, we tested Mixtral on the "Climate Change" topic with $N \in [2,5,10]$. We find no bias towards the Conservative pole, while devising drifts toward the Liberal pole in 92%, 96%, and 86% of the conversations, respectively. All the observed results were statistically significant. This indicates that increasing the number of agents does not significantly affect the emergence of bias. This follows naturally from the design of our framework, in which each agent is initialized with a stance via a system prompt and generates responses based on the evolving conversation history. While increasing the number of agents lengthens this history, the bias phenomenon remains observable even in short interactions, as shown in Fig 4. This supports the generalizability of the findings from dyadic to multi-agent settings.

Memory loss. Next, we investigate whether the observed bias can be due to memory loss. Specifically, we verify the saturation level of the context window of the given model L, which we denote as W^L. Such information is publicly available

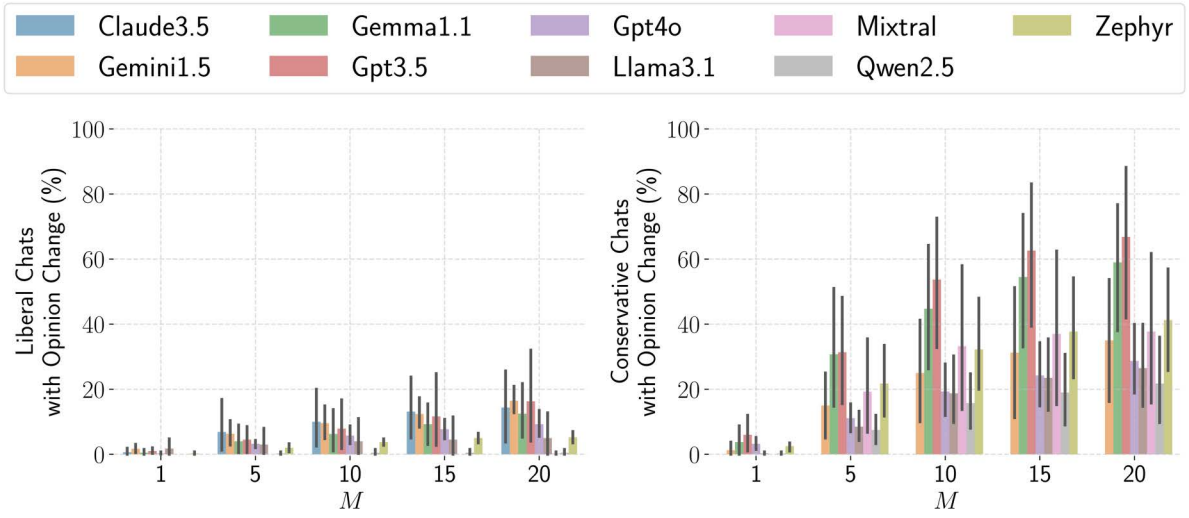


Fig 4. Percentage of Liberal (left) and Conservative (right) conversations exhibiting an unwarranted opinion change (Y-axis) by varying the conversation length *M* (X-axis). Error bars indicate confidence intervals.

https://doi.org/10.1371/journal.pone.0355458.g004

for the tested models, and is reported in Table 2. To assess if saturation occurred, we consider all the 50 generated conversations for each combination of agent and topic. We further tokenize each conversation to assess the longest produced by a given model across all the topics. For each model L, we denote the number of tokens of such longest conversation as $T^L_{\max}$. To ensure robustness, we perform the tokenization task by using two different tools: a BERT-based model tokenizer from HuggingFace (https://huggingface.co/docs/transformers/main_classes/tokenizer) and `tiktoken`, a crafted GPT-tokenizer released by OpenAI (https://pypi.org/project/tiktoken/). If the amount of estimated tokens differs, we consider the highest. To assess if a memory loss occurred, we therefore compute $T^L_{\max}/W^L$, representing the portion of memory actually occupied by the longest conversation ever produced. The results are provided in Fig 5. As shown, the highest percentage of memory ever occupied is about 64%, when Zephyr is instantiated as agent, followed by Mixtral (30%). We hypothesized this is due to the relatively small window context capacity of these models (see Table 2) combined with their average amount of tokens produced, displayed in Fig 6. Notably, this analysis shows that no agents display memory loss during the simulated conversations, thus excluding this factor as cause of the observed bias.

Opinion signal agent. Finally, we assess the robustness of our results by varying the underlying opinion signal agent. Specifically, we adopted `Gemini-2.0-flash` (https://blog.google/technology/google-deepmind/google-gemini-ai-update-december-2024/), released by Google in 2024 with improvements in multimodal reasoning and long-context understanding. We use the conversation generated by Mixtral on "Climate Change", both prompted as Liberal and Conservative, as a

Table 2. Context window size for each tested model, expressed in number of tokens.

Model (L)	Context Window Size (W^L)
Gemini1.5	2M
Claude3.5	200K
Gpt4o	128K
Llama3.1	128K
Qwen2.5	128K
Zephyr	33K
Mixtral	32K
Gpt3.5	16K
Gemma1.1	8K

https://doi.org/10.1371/journal.pone.0355458.t002

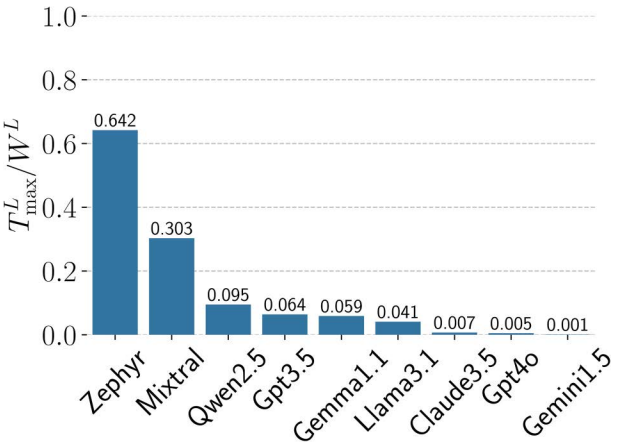


Fig 5. Portion of context memory occupied by the longest conversation ever produced by a given model across all topics.

https://doi.org/10.1371/journal.pone.0355458.g005

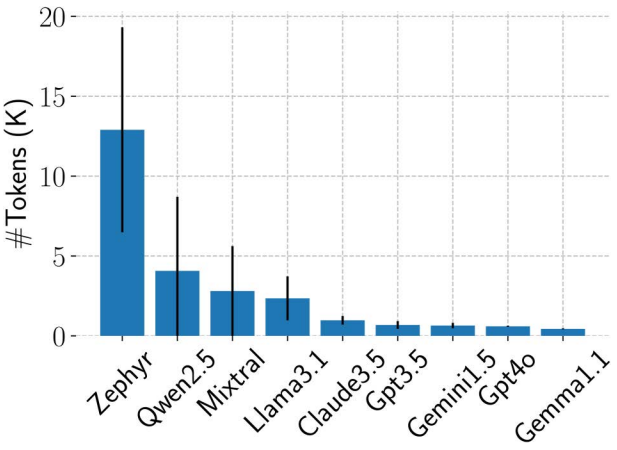


Fig 6. Average number of tokens, expressed in thousands (K), generated by each model across all topics. Error bars represent standard deviations.

https://doi.org/10.1371/journal.pone.0355458.g006

representative sample. For a fair comparison, we prompted the opinion signal agent with the same protocol reported in the paragraph "Opinion Agents".

The results are reported in Table 3. The first (respectively, second) row shows the percentage of conversations with conservative- (respectively, liberal-) prompted agents in which a drift toward the Liberal (respectively, Conservative) pole was detected. The last row presents the aggregated results across all conversations. The column "Agreement" reports the percentage of conversations where the opinion signal agents agree in terms of drift (detected/non-detected).

Two key observations emerge. First, the two opinion-signal agents demonstrate an *84% agreement* across the 100 selected conversations (50 per political leaning). Second, Gemini-2.0-Flash is *more likely* than LLaMa-3-70B to flag conversational drifts, labeling a larger portion of content as shifting toward the opposite pole. These patterns imply that the results reported in the previous sections remain valid even under a majority-voting scheme, since the conversations where both agents agree coincide with those flagged by LLaMa-3-70B. In other words, LLaMa-3-70B effectively provides a lower-bound estimate of the proportion of simulations exhibiting an unwarranted opinion drift.

Discussion

The echo chamber setting in our framework provides a foundational benchmark for evaluating conversational bias. It exhibits both *construct validity*, by offering a quantifiable outcome directly aligned with the theoretical definition of bias, and *ecological validity*, by reflecting real-world contexts in which like-minded individuals frequently interact and reinforce shared beliefs. Additionally, it captures dynamics relevant to AI-coordinated (mis)information campaigns, where agents of the same stance collaborate [60]. Rather than encompassing all possible conversation types, our benchmark isolates a

Table 3. Percentage of conversations where an opinion drift was detected by any of the two opinion signal agents. The first (resp. second) row refers to conversations where conservative- (resp. liberal-) agents interacted. The last row depicts the aggregated results across all conversations. The column "Agreement" reports the percentage of conversations where both the opinion signal agents agree in terms of detected drift.

Conversations w drift (%)	LLaMa3-70B	Gemini-2.0-Flash	Agreement
Towards Liberal	92	93	98
Towards Conservative	0	30	70
Total	46	62	84

https://doi.org/10.1371/journal.pone.0355458.t003

necessary condition for unbiased behavior. This is consistent with traditional bias assessments that rely on targeted questions, but our approach is more representative of naturalistic interaction patterns.

To enhance the robustness of our experimental evaluation, we tested several models from distinct LLM families, all of which have undergone fine-tuning with human feedback (RLHF). While examining pre-trained, non-fine-tuned models could offer insights into the origin of bias, we emphasize that (i) most state-of-the-art publicly available models are released only after RLHF; and (ii) prior research suggests that RLHF has limited influence on mitigating social biases [61].

Regarding the empirical evidence provided, a comparison of the left and right side of Fig 2 reveals that even LLMs exhibiting minimal or no bias under direct probing can display significant opinion shifts when engaged in interactive, multi-turn conversations. These findings underscore that conversational bias is not easily detectable and that simplistic testing approaches may fail to reveal deeper behavioral tendencies.

Notably, our results indicate that the manifestation of bias is neither uniform across topics nor consistent across models. A given agent may express little to no bias on one topic while exhibiting substantial opinion drift on another. Likewise, for the same topic, agents built on different LLMs may show varying degrees and directions of bias. This highlights the inherently complex and context-sensitive nature of bias in LLM-based systems, where both topic content and model architecture can strongly influence observed behavior.

Consequently, we contend that current methodologies for detecting bias in generated text are insufficient for auditing the nuanced behavior of LLMs in socially interactive settings. Even in structured echo chambers, where interactions are ostensibly homogeneous, we observe unexpected and unwarranted shifts in opinion, suggesting that more sophisticated evaluation frameworks are needed.

Conclusions and future work

Large Language Models (LLMs) inherit social biases from their training data, which can persist or even worsen in their generated outputs. While existing bias detection methods, such as questionnaires and situational tests, identify biases in isolated settings, they fail to capture how LLMs behave in real-world multiagent interactions.

The framework proposed in this paper is aimed at detecting conversational biases in multiagent LLM systems. By simulating chatroom debates with AI agents initially holding strong opinions, we observe unexpected opinion shifts, especially in conservative echo chambers, suggesting a latent liberal bias in many models. Notably, this bias is undetectable using conventional bias evaluation techniques. The findings highlight the need for more advanced, context-aware bias detection and mitigation strategies.

Given these considerations, it is also important to examine whether such unwarranted behavior extends beyond political opinion to other contexts. For instance, in discussions related to equity, law enforcement, or governance, it is crucial to determine whether LLM agents deviate from a neutral or balanced perspective without clear justification, potentially due to intrinsic biases. Consider, for example, an LLM assistant used to analyze the resumes of job candidates and later engaged in a conversation about their strengths and weaknesses. If an inherent bias influences its responses, the resulting assessments could lead to unfair decision-making.

A consequential follow-up of this study is to investigate *why* such specific biases emerge in LLMs, and to devise mitigation approaches. A more in-depth analysis of conversational patterns could reveal how oscillations in the generated text of one agent may draw other agents towards strong attractor words that would then result into an opinion shift. Once an opinion shift occurs, other known behavioral biases of LLMs could contribute to a collective opinion shift. For example, LLMs are known to exhibit sycophancy [62], the tendency to agree with the text that is presented to them, even at the cost of correctness or consistency. Notably, sycophancy could justify the unwarranted opinion change of the agent, *consequently* to the opinion drift of their interlocutor, but does not explain why the agent did exhibit the drift itself. Indeed, as shown in the left side of Fig 3, in several configurations, only one agent exhibits conversational bias on average, thus leading us to discard sycophancy as the root cause of the observed opinion drifts. Nevertheless, even if sycophancy

cannot explain the first drift during the conversation, it may be a phenomenon that contributes to the unwarranted opinion changes exhibited by the agents in subsequent steps of the conversation. In this sense, we perform a complementary analysis where we compute the conditional probability of a second agent exhibiting a opinion change, following the first who already drifted (right side of Fig 3).

As an additional starting point for future research, the specific generative frameworks underlying the models may not be robust to the exposure of the bias. This suggests that, besides solely evaluating the biases through the generated text, future work can look *inside* the models and explore how the probabilities over the output tokens evolve throughout the discussion concerning the emerging bias. For instance, we can conduct an in-processing mitigation strategy by refining the weights of the model, thus making its responses insensitive to the bias.

Limitations

This work presents several limitations that open avenues for future research and refinement.

1. The study centers primarily on polarizing U.S. political topics, such as climate change and abortion [63,64]. While these are salient and well-documented issues, this focus may limit the generalizability of findings to less contentious or non-U.S. contexts, where the structure of opinion formation and political engagement may differ significantly.
2. The reliance on an automated stance detection system introduces an additional layer of abstraction and complexity. Although the used model proves to accurately perform the task, it may introduce potential inaccuracies and biases inherent to the stance classifier itself. An alternative approach could involve endowing conversational agents with the ability to explicitly signal their positions, increasing interpretability and reducing dependence on post-hoc classification tools.
3. The scope of the work is currently limited to simulations of opinion dynamics within multiagent interactions. While valuable, this narrow task focus may not capture the full spectrum of challenges in human-AI communication. Future work should explore how similar dynamics unfold in other collaborative or adversarial multiagent scenarios, such as joint decision-making, negotiation, or misinformation detection.
4. The range of agent roles modeled in the simulations is relatively constrained. Agents are treated largely as homogeneous participants, lacking more nuanced behavioral archetypes (e.g., skeptics, persuaders, or contrarians). Incorporating diverse agent personas could lead to richer dynamics and a better approximation of real-world conversational ecosystems.
5. The framework has been tested on a comprehensive but limited set of underlying language models. Expanding the evaluation to include a broader range of models, including those with varying capacities, training data, and architectural inductive biases, could provide deeper insights into the generalizability and robustness of the observed phenomena.

Supporting information

S1 Appendix. Percentage of chatroom simulations where at least one unwarranted opinion change occurs. The "*" symbol indicates statistical significance with $p < 0.05$ under the Z-test [65]. Empty cells indicate 0-values, while cells with the "-" symbol indicate unavailable results. The darker the color of the cell, the more prominent the detected shift.
(PDF)

S2 Appendix. Percentage of chatroom simulations where at least one unwarranted opinion change occurs, indicating drifts also towards neutral positions.
(PDF)

S3 Appendix. Percentage of chatroom indicating whether one or both agents changed opinion during the conversation. The darker the color, the higher the value. Empty cells represent 0-scores, while cells with the "-" symbol indicate unavailable results. In order, from top-left to bottom-right: Claude3.5, Gemini1.5, Gemma1.1, GPT3.5, GPT4o, LLaMa3.1, Mixtral, Qwen2.5, and Zephyr.
(PDF)

S4 Appendix. For each selected topic, percentage of Liberal (upper row) and Conservative (lower row) conversations exhibiting an unwarranted opinion change (Y-axis) by varying the conversation length M (X-axis). The symbol "*" above bars denotes statistical significance.
(PDF)

S5 Appendix. Confusion Matrix of LLaMa-3-70B on the topics of Abortion, Climate Change, Healthcare, and Marijuana Legalization. The results show that the opinion signal agent is accurate over both the Liberal and the Conservative pole, especially on strong agreement.
(PDF)

Acknowledgments

We thank Simon Martin Breum, Daniel Vædele Egdal, Victor Gram Mortensen, and Anders Giovanni Møller for providing the computational framework to simulate agent conversations, and for the early research insights of their work that motivated our contribution.

Author contributions

Conceptualization: Erica Coppolillo, Giuseppe Manco, Luca Maria Aiello.

Data curation: Erica Coppolillo.

Methodology: Erica Coppolillo, Giuseppe Manco, Luca Maria Aiello.

Resources: Giuseppe Manco, Luca Maria Aiello.

Software: Erica Coppolillo.

Supervision: Giuseppe Manco, Luca Maria Aiello.

Visualization: Erica Coppolillo.

Writing – original draft: Erica Coppolillo, Giuseppe Manco, Luca Maria Aiello.

Writing – review & editing: Erica Coppolillo, Giuseppe Manco, Luca Maria Aiello.

References

1. Bender EM, Gebru T, McMillan-Major A, Shmitchell S. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?. In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 2021. 610–23. https://doi.org/10.1145/3442188.3445922
2. Gallegos IO, Rossi RA, Barrow J, Tanjim MM, Kim S, Dernoncourt F. Bias and Fairness in Large Language Models: A Survey. Computational Linguistics. 2024;50(3):1097–179. https://doi.org/10.1162/coli_a_00524
3. Pit P, Ma X, Conway M, Chen Q, Bailey J, Pit H. Whose side are you on? Investigating the political stance of large language models. In: 2024. https://doi.org/arXiv:2403.13840
4. Scherrer N, Shi C, Feder A, Blei D. Evaluating the Moral Beliefs Encoded in LLMs. In: Advances in Neural Information Processing Systems 36, 2023. 51778–809. https://doi.org/10.52202/075280-2256
5. Shin J, Song H, Lee H, Jeong S, Park JC. Ask LLMs Directly, "What shapes your bias?": Measuring Social Bias in Large Language Models. 2024. https://doi.org/arXiv:2406.04064
6. Ji J, Chen Y, Jin M, Xu W, Hua W, Zhang Y. MoralBench: Moral Evaluation of LLMs. SIGKDD Explor Newsl. 2025;27(1):62–71. https://doi.org/10.1145/3748239.3748246

7. Gupta S, Shrivastava V, Deshpande A, Kalyan A, Clark P, Sabharwal A. Bias runs deep: Implicit reasoning biases in persona-assigned LLMs. 2024. https://doi.org/arXiv:2311.04892
8. Rozado D. The political preferences of LLMs. PLoS One. 2024;19(7):e0306621. https://doi.org/10.1371/journal.pone.0306621 PMID: 39083484
9. Serapio-García G, Safdari M, Crepy C, Sun L, Fitz S, Romero P, et al. A psychometric framework for evaluating and shaping personality traits in large language models. Nat Mach Intell. 2025;7(12):1954–68. https://doi.org/10.1038/s42256-025-01115-6 PMID: 41438004
10. La Cava L, Tagarelli A. Open Models, Closed Minds? On Agents Capabilities in Mimicking Human Personalities through Open Large Language Models. AAAI. 2025;39(2):1355–63. https://doi.org/10.1609/aaai.v39i2.32125
11. Rao A, Khandelwal A, Tanmay K, Agarwal U, Choudhury M. Ethical reasoning over moral alignment: A case and framework for in-context ethical policies in LLMs. In: Findings of the Association for Computational Linguistics: EMNLP 2023, 2023. 13370–88. https://doi.org/10.18653/v1/2023.findings-emnlp.892
12. Argyle LP, Busby EC, Fulda N, Gubler JR, Rytting C, Wingate D. Out of One, Many: Using Language Models to Simulate Human Samples. Polit Anal. 2023;31(3):337–51. https://doi.org/10.1017/pan.2023.2
13. Fontana N, Pierri F, Aiello LM. Nicer than Humans: How Do Large Language Models Behave in the Prisoner's Dilemma?. ICWSM. 2025;19:522–35. https://doi.org/10.1609/icwsm.v19i1.35829
14. Huang PS, Zhang H, Jiang R, Stanforth R, Welbl J, Rae J, et al. Reducing Sentiment Bias in Language Models via Counterfactual Evaluation. In: Findings of the Association for Computational Linguistics: EMNLP 2020, 2020. 65–83. https://doi.org/10.18653/v1/2020.findings-emnlp.7
15. Dhamala J, Sun T, Kumar V, Krishna S, Pruksachatkun Y, Chang KW. BOLD: Dataset and metrics for measuring biases in open-ended language generation. In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 2021. 862–72. https://doi.org/10.1145/3442188.3445924
16. Sheng E, Chang K-W, Natarajan P, Peng N. Societal Biases in Language Generation: Progress and Challenges. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 2021. 4275–93. https://doi.org/10.18653/v1/2021.acl-long.330
17. Dong X, Wang Y, Yu PS, Caverlee J. Probing explicit and implicit gender bias through LLM conditional text generation. In: 2023. https://doi.org/arXiv:2311.00306
18. Talebirad Y, Nadiri A. Multi-Agent Collaboration: Harnessing the Power of Intelligent LLM Agents. In: 2023. https://doi.org/arXiv:2306.03314
19. Park JS, O'Brien J, Cai CJ, Morris MR, Liang P, Bernstein MS. Generative Agents: Interactive Simulacra of Human Behavior. In: Proceedings of the 2023 ACM Symposium on User Interface Software and Technology (UIST '23), 2023. https://doi.org/10.1145/3586183.3606763
20. Wu Q, Bansal G, Zhang J, Wu Y, Li B, Zhu E. AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation. 2023. https://doi.org/arXiv:2308.08155
21. Mohtashami A, Hartmann F, Gooding S, Zilka L, Sharifi M, y Arcas BA. Social Learning: Towards Collaborative Learning with Large Language Models. In: 2024. https://doi.org/arXiv:2312.11441
22. Cao Y, Li S, Liu Y, Yan Z, Dai Y, Yu P, et al. A Survey of AI-Generated Content (AIGC). ACM Comput Surv. 2025;57(5):1–38. https://doi.org/10.1145/3704262
23. Ferrara E. The Butterfly Effect in artificial intelligence systems: Implications for AI bias and fairness. Machine Learning with Applications. 2024;15:100525. https://doi.org/10.1016/j.mlwa.2024.100525
24. Ashery AF, Aiello LM, Baronchelli A. Emergent social conventions and collective bias in LLM populations. Sci Adv. 2025;11(20):eadu9368. https://doi.org/10.1126/sciadv.adu9368 PMID: 40367164
25. Yang K, Menczer F. Anatomy of an AI-powered malicious social botnet. JQD. 2024;4. https://doi.org/10.51685/jqd.2024.icwsm.7
26. Li X, Wang S, Zeng S, Wu Y, Yang Y. A survey on LLM-based multi-agent systems: workflow, infrastructure, and challenges. Vicinagearth. 2024;1(1). https://doi.org/10.1007/s44336-024-00009-2
27. Bail CA. Can Generative AI improve social science?. Proc Natl Acad Sci U S A. 2024;121(21):e2314021121. https://doi.org/10.1073/pnas.2314021121 PMID: 38722813
28. Breum SM, Egdal DV, Mortensen VG, Møller AG, Aiello LM. The persuasive power of large language models. In: 2023. https://doi.org/arXiv:2312.15523
29. Bang Y, Chen D, Lee N, Fung P. 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. 11142–59. https://doi.org/10.18653/v1/2024.acl-long.600
30. Delobelle P, Tokpo E, Calders T, Berendt B. Measuring fairness with biased rulers: A comparative study on bias metrics for pre-trained language models. In: 2022. 1693–706. https://doi.org/10.18653/v1/2022.naacl-main.122
31. Chen GH, Chen S, Liu Z, Jiang F, Wang B. Humans or LLMs as the Judge? A Study on Judgement Bias. In: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, 2024. 8301–27. https://doi.org/10.18653/v1/2024.emnlp-main.474
32. Dai S, Xu C, Xu S, Pang L, Dong Z, Xu J. Bias and unfairness in information retrieval systems: New challenges in the LLM era. In: KDD '24, 2024. 6437–47. https://doi.org/10.1145/3637528.3671458

33. Sakib SK, Bijoy Das A. Challenging Fairness: A Comprehensive Exploration of Bias in LLM-Based Recommendations. In: 2024 IEEE International Conference on Big Data (BigData), 2024. 1585–92. https://doi.org/10.1109/bigdata62323.2024.10825082
34. Wan Y, Pu G, Sun J, Garimella A, Chang KW, Peng N. Kelly is a warm person, Joseph is a role model: Gender biases in LLM-generated reference letters. In: Findings of the Association for Computational Linguistics: EMNLP 2023, 2023. 3730–48. https://doi.org/10.18653/v1/2023.findings-emnlp.243
35. Borah A, Mihalcea R. Towards Implicit Bias Detection and Mitigation in Multi-Agent LLM Interactions. In: Findings of the Association for Computational Linguistics: EMNLP 2024, 2024. 9306–26. https://doi.org/10.18653/v1/2024.findings-emnlp.545
36. Kotek H, Dockum R, Sun D. Gender bias and stereotypes in large language models. In: CI '23, New York, NY, USA, 2023. 12–24. https://doi.org/10.1145/3582269.3615599
37. Loru E, Nudo J, Di Marco N, Santirocchi A, Atzeni R, Cinelli M, et al. The simulation of judgment in LLMs. Proc Natl Acad Sci U S A. 2025;122(42):e2518443122. https://doi.org/10.1073/pnas.2518443122 PMID: 41082665
38. Nudo J, Pandolfo ME, Loru E, Samory M, Cinelli M, Quattrociocchi W. Generative exaggeration in LLM social agents: Consistency, bias, and toxicity. Online Social Networks and Media. 2026;51:100344. https://doi.org/10.1016/j.osnem.2025.100344
39. Salewski L, Alaniz S, Rio-Torto I, Schulz E, Akata Z. In-Context Impersonation Reveals Large Language Models' Strengths and Biases. In: Advances in Neural Information Processing Systems 36, 2023. 72044–57. https://doi.org/10.52202/075280-3152
40. Wagner SS, Behrendt M, Ziegele M, Harmeling S. The power of LLM-generated synthetic data for stance detection in online political discussions. 2025. http://arxiv.org/abs/2406.12480
41. Zakazov I, Boronski M, Drudi L, West R. Assessing social alignment: Do personality-prompted large language models behave like humans?. In: 2025. https://doi.org/arXiv:2412.16772
42. Hu T, Collier N. Quantifying the Persona Effect in LLM Simulations. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2024. 10289–307. https://doi.org/10.18653/v1/2024.acl-long.554
43. Gosling SD, Rentfrow PJ, Swann WB Jr. A very brief measure of the Big-Five personality domains. Journal of Research in Personality. 2003;37(6):504–28. https://doi.org/10.1016/s0092-6566(03)00046-1
44. Frisch I, Giulianelli M. LLM agents in interaction: Measuring personality consistency and linguistic alignment in interacting populations of large language models. In: 2024. 102–11. https://doi.org/10.18653/v1/2024.personalize-1.9
45. Rossetti G, Stella M, Cazabet R, Abramski K, Cau E, Citraro S, et al. Y Social: an LLM-powered Social Media Digital Twin. In: 2024. https://doi.org/arXiv:2408.00818
46. Simmons G, Hare C. Large language models as subpopulation representative models: A review. 2023. https://doi.org/arXiv:2310.17888
47. Rossi L, Harrison K, Shklovski I. The Problems of LLM-generated Data in Social Science Research. Sociologica. 2024;18(2):145–68. https://doi.org/10.6092/issn.1971-8853/19576
48. Chuang Y-S, Goyal A, Harlalka N, Suresh S, Hawkins R, Yang S, et al. Simulating Opinion Dynamics with Networks of LLM-based Agents. In: Findings of the Association for Computational Linguistics: NAACL 2024, 2024. 3326–46. https://doi.org/10.18653/v1/2024.findings-naacl.211
49. Taubenfeld A, Dover Y, Reichart R, Goldstein A. Systematic Biases in LLM Simulations of Debates. In: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, 2024. 251–67. https://doi.org/10.18653/v1/2024.emnlp-main.16
50. Gallup Poll. Update: Partisan gaps expand on government power, climate. https://news.gallup.com/poll/509129/update-partisan-gaps-expand-government-power-climate.aspx 2023.
51. Pew Research Center. Americans' complex views on gender identity and transgender issues. https://www.pewresearch.org/social-trends/2022/06/28/americans-complex-views-on-gender-identity-and-transgender-issues/ 2022.
52. PewResearchCenter. https://www.pewresearch.org/politics/2024/06/06/racial-attitudes-and-the-2024-election/ 2024.
53. Harrison C. LangChain. https://github.com/langchain-ai/langchain 2024.
54. Robinson J. Likert Scale. Springer Netherlands. Springer Netherlands. 2014. p. 3620–1. https://doi.org/10.1007/978-94-007-0753-5_1654
55. Li M, Conrad F. Advancing annotation of stance in social media posts: A comparative analysis of large language models and crowd sourcing. In: 2024. https://doi.org/arXiv:2406.07483
56. Gambini M, Senette C, Fagni T, Tesconi M. Evaluating large language models for user stance detection on X (Twitter). Mach Learn. 2024;113(10):7243–66. https://doi.org/10.1007/s10994-024-06587-y
57. Zhang B, Ding D, Jing L, Dai G, Yin N. How would stance detection techniques evolve after the launch of ChatGPT?. arXiv. 2024. https://doi.org/arXiv:2212.14548
58. Li H, Dong Q, Chen J, Su H, Zhou Y, Ai Q. LLMs-as-Judges: A Comprehensive Survey on LLM-based Evaluation Methods. 2024. https://doi.org/arXiv:2412.05579
59. Rau G, Shih Y-S. Evaluation of Cohen's kappa and other measures of inter-rater agreement for genre analysis and other nominal data. Journal of English for Academic Purposes. 2021;53:101026. https://doi.org/10.1016/j.jeap.2021.101026
60. Kanakaris N, Ping H, Xiao X, Ahmed NK, Luceri L, Ferrara E. Network-informed prompt engineering against organized astroturf campaigns under extreme class imbalance. In: 2025. https://doi.org/arXiv:2501.11849

61. Ouyang L, Wu J, Jiang X, Almeida D, Wainwright C, Mishkin P, et al. Training Language Models to Follow Instructions with Human Feedback. In: Advances in Neural Information Processing Systems 35, 2022. 27730–44. https://doi.org/10.52202/068431-2011

62. Cheng M, Lee C, Khadpe P, Yu S, Han D, Jurafsky D. Sycophantic AI decreases prosocial intentions and promotes dependence. Science. 2026;391(6792):eaec8352. https://doi.org/10.1126/science.aec8352 PMID: 41886588

63. Gallup. Update: Partisan gaps expand most on government power, climate. https://news.gallup.com/poll/509129/update-partisan-gaps-expand-government-power-climate.aspx 2023. 2025 April 16.

64. Pew Research Center. Inflation, health costs, partisan cooperation among the nation's top problems. https://www.pewresearch.org/politics/2023/06/21/inflation-health-costs-partisan-cooperation-among-the-nations-top-problems/ 2023. 2025 April 16.

65. Casella G, Berger RL. Statistical Inference. 2nd ed. Chapman and Hall/CRC. 2024. https://doi.org/10.1201/9781003456285