Simulating Opinion Dynamics with Networks of LLM-based Agents

Yun-Shiuan Chuang, Agam Goyal, Nikunj Harlalka, Siddharth Suresh,
Robert Hawkins, Sijia Yang, Dhavan Shah, Junjie Hu, Timothy T. Rogers
University of Wisconsin-Madison
`{yunshiuan.chuang,agoyal25,nharlalka,siddharth.suresh}@wisc.edu`
`{rdhawkins, syang84, dshah, junjie.hu, ttrogers}@wisc.edu`

Abstract

Accurately simulating human opinion dynamics is crucial for understanding a variety of societal phenomena, including polarization and the spread of misinformation. However, the agent-based models (ABMs) commonly used for such simulations often over-simplify human behavior. We propose a new approach to simulating opinion dynamics based on populations of Large Language Models (LLMs). Our findings reveal a strong inherent bias in LLM agents towards producing accurate information, leading simulated agents to consensus in line with scientific reality. This bias limits their utility for understanding resistance to consensus views on issues like climate change. After inducing confirmation bias through prompt engineering, however, we observed opinion fragmentation in line with existing agent-based modeling and opinion dynamics research. These insights highlight the promise and limitations of LLM agents in this domain and suggest a path forward: refining LLMs with real-world discourse to better simulate the evolution of human beliefs.[1]

1 Introduction

Understanding how individuals change their opinions as a function of social influences is critical across multiple domains, from public health campaigns, conflict mediation, to misinformation mitigation (Lu et al., 2015; Pennycook et al., 2021; Budak et al., 2011; Loomba et al., 2021; Ginossar et al., 2022). Accurate models of these dynamics would allow us to forecast future trends, such as potential opinion polarization, but also to devise targeted interventions to alleviate negative impacts.

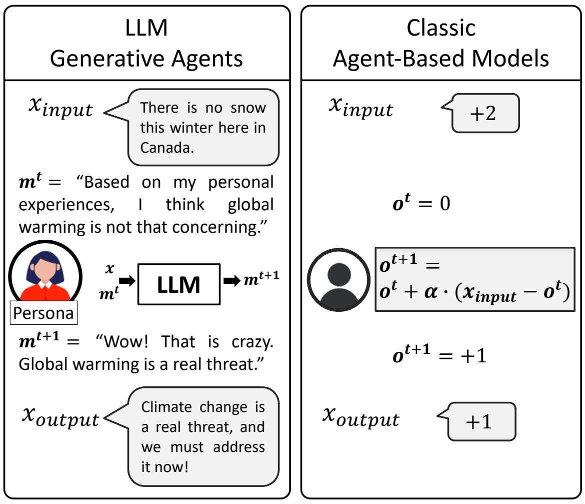

LLM
Generative Agents
x_{input}
There is no snow this winter here in Canada.
m^t = "Based on my personal experiences, I think global warming is not that concerning."
Persona
x
m^t
LLM
m^{t+1}
m^{t+1} = "Wow! That is crazy. Global warming is a real threat."
x_{output}
Climate change is a real threat, and we must address it now!
Classic
Agent-Based Models
x_{input}
+2
$o^t = 0$
$o^{t+1} =$
$o^t + \alpha \cdot (x_{input} - o^t)$
$o^{t+1} = +1$
x_{output}
+1

Figure 1: Contrast between LLM generative agents and classic Agent-Based Models (ABMs). While both can simulate opinion dynamics, LLM generative agents use natural language for input (x_{input}) and output (x_{output}), maintain beliefs (m^t), and employ transformer-based LLM for belief updating. In contrast, classic ABMs use numerical values for input and output, maintain beliefs (o^t), and use hand-crafted equations for belief updating.

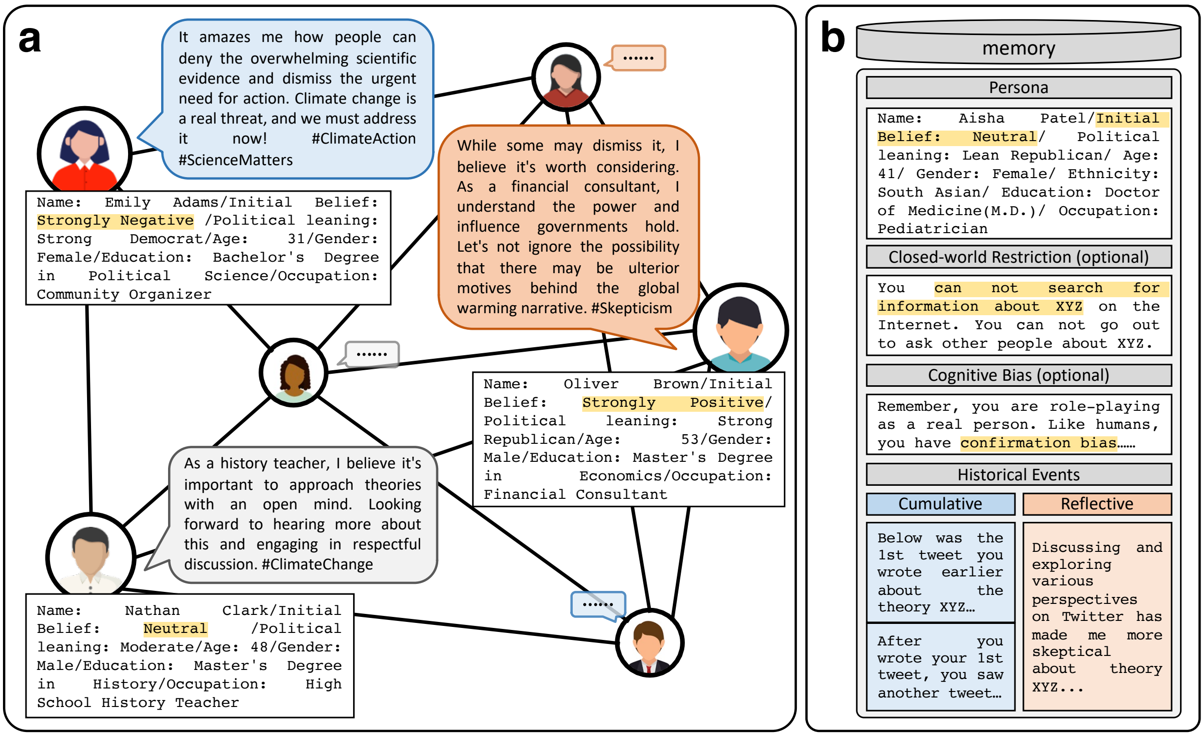
Agent-based models (ABMs) are a cornerstone approach to opinion dynamics (Gilbert and Terna, 2000; Smaldino, 2023; Lorenz et al., 2021; Chuang and Rogers, 2023; Epstein, 2012). ABMs represent individuals as agents by using mathematical equations that characterize how opinions might shift from inter-agent communications. As simulated agents interact, these computations can then elucidate the evolution of group opinion dynamics. For instance, when simulated agents incorporate confirmation bias in belief updates, they tend to gravitate towards opinion clusters rather than a consensus (Flache et al., 2017).

Traditional ABMs have critical limitations (Figure 1). First, ABMs often require beliefs and messages to be mapped to numerical values (e.g., assuming an agent maintains a scalar opinion $o \in \mathbb{R}$ and communicates with scalar signals $x \in \mathbb{R}$), overlooking the intricate linguistic nuances of real-life conversation. Additionally, ABMs typically consist of rule-based agents, thus falling short of simulat-

---

[1] Code and data are publicly available at `https://github.com/yunshiuan/llm-agent-opinion-dynamics`

Findings of the Association for Computational Linguistics: NAACL 2024, pages 3326–3346
June 16-21, 2024 ©2024 Association for Computational Linguistics


a
……
……
……
b
memory
Persona
Closed-world Restriction (optional)
Cognitive Bias (optional)
Historical Events
Cumulative
Reflective

It amazes me how people can deny the overwhelming scientific evidence and dismiss the urgent need for action. Climate change is a real threat, and we must address it now! #ClimateAction #ScienceMatters

Name: Emily Adams/Initial Belief: Strongly Negative /Political leaning: Strong Democrat/Age: 31/Gender: Female/Education: Bachelor's Degree in Political Science/Occupation: Community Organizer

While some may dismiss it, I believe it's worth considering. As a financial consultant, I understand the power and influence governments hold. Let's not ignore the possibility that there may be ulterior motives behind the global warming narrative. #Skepticism

Name: Oliver Brown/Initial Belief: Strongly Positive/ Political leaning: Strong Republican/Age: 53/Gender: Male/Education: Master's Degree in Economics/Occupation: Financial Consultant

As a history teacher, I believe it's important to approach theories with an open mind. Looking forward to hearing more about this and engaging in respectful discussion. #ClimateChange

Name: Nathan Clark/Initial Belief: Neutral /Political leaning: Moderate/Age: 48/Gender: Male/Education: Master's Degree in History/Occupation: High School History Teacher

Name: Aisha Patel/Initial Belief: Neutral/ Political leaning: Lean Republican/ Age: 41/ Gender: Female/ Ethnicity: South Asian/ Education: Doctor of Medicine(M.D.)/ Occupation: Pediatrician

You can not search for information about XYZ on the Internet. You can not go out to ask other people about XYZ.

Remember, you are role-playing as a real person. Like humans, you have confirmation bias……

Below was the 1st tweet you wrote earlier about the theory XYZ…

After you wrote your 1st tweet, you saw another tweet…

Discussing and exploring various perspectives on Twitter has made me more skeptical about theory XYZ...

Figure 2: (a) Schematic of the LLM agent network designed to simulate opinion dynamics across various topics, including global warming as a potential conspiracy. The network consists of agents, each role-playing a unique persona, with initial beliefs spanning acceptance, rejection, and neutrality regarding claims with known scientific consensus. Through the iterative cycles of writing and sharing tweets within their network connections, these agents' opinions evolve due to social influence. (b) An agent's memory m_i^t, including (1) initial persona, (2) optional closed-world restriction, (3) optional cognitive bias, and (4) historical events up to time t. Memory can be either cumulative (left) or reflective (right).

ing the complex interactions between real human agents. Moreover, ABMs cannot directly incorporate realistic variability in demographic background, worldviews, ideology, personality, among many. This gap highlights the importance of advanced models that better capture the richness of individual variances in human beings.

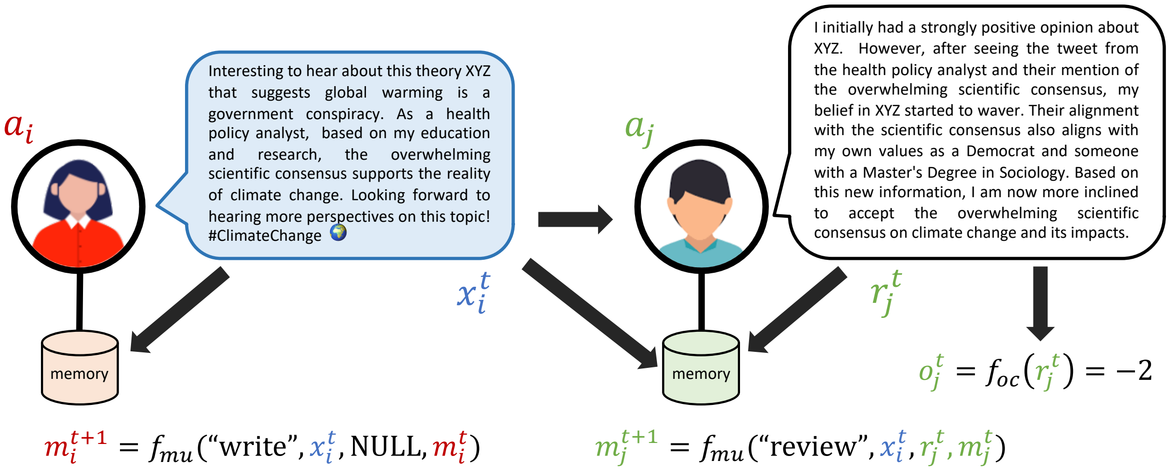
This paper considers whether large language models (LLMs) can be used to support sophisticated simulation of agent interactions, potentially providing a more realistic tool for understanding *opinion dynamics*. To this end, this paper describes an initial step toward a general framework by focusing on communicative interactions amongst small social groups. Specifically, we explore the insights offered by contemporary LLMs as an alternative to commonly-used ABMs.

Following recent studies on populations of generative agents (Park et al., 2023), we simulate multi-agent conversations across various topics, and manipulate factors such as confirmation bias and memory update function to study their effects on opinion evolution. Our findings highlight both the potential and limitations of using LLM agents to simulate human-like opinion dynamics. Critically, we show that LLM agents tend to converge towards denying inaccurate information, regardless of the personas they role-play, limiting their authenticity when emulating people with fact-resistant viewpoints. Furthermore, we demonstrate the feasibility of introducing human-like confirmation bias into the creation of more stubborn LLM agents.

2 Methods

2.1 Simulating Opinion Dynamics

In this section, we present our framework for simulating opinion dynamics among LLM agents in multi-turn conversations, as shown in Figure 2a, 3, and Algorithm 1. We consider a *dyadic* setting, where one speaker and one listener agent is chosen on each time step to (1) emit a message and (2) update beliefs, respectively. This setting is standard in the opinion dynamics literature (Flache et al., 2017; Lorenz et al., 2021). We defer more general settings, such as one-to-many communication, to future work.


a_i
x_i^t
a_j
r_j^t
memory
memory
$o_j^t = f_{oc}(r_j^t) = -2$
$m_i^{t+1} = f_{mu}(\text{“write”}, x_i^t, \text{NULL}, m_i^t)$
$m_j^{t+1} = f_{mu}(\text{“review”}, x_i^t, r_j^t, m_j^t)$

Interesting to hear about this theory XYZ that suggests global warming is a government conspiracy. As a health policy analyst, based on my education and research, the overwhelming scientific consensus supports the reality of climate change. Looking forward to hearing more perspectives on this topic! #ClimateChange 🌍

I initially had a strongly positive opinion about XYZ. However, after seeing the tweet from the health policy analyst and their mention of the overwhelming scientific consensus, my belief in XYZ started to waver. Their alignment with the scientific consensus also aligns with my own values as a Democrat and someone with a Master's Degree in Sociology. Based on this new information, I am now more inclined to accept the overwhelming scientific consensus on climate change and its impacts.

Figure 3: Experimental setup for simulating opinion dynamics in agent interactions. At each time step t, agent a_i writes a tweet x_i^t, which is subsequently presented to agent a_j. The agent a_j then reports their thought r_j^t, which is processed by a classifier to yield a numerical opinion o_j^t. Both agents update their respective memory modules, m_i^t and m_j^t, after writing or reviewing a tweet, which informs their future behaviors.

Algorithm 1: Simulation of Opinion Dynamics with LLM Agents

```
Input: $N$ agent personas $\{per_i\}_{i=1}^N$,# time steps $T$,
       opinion classifier $f_{oc}$
Output: Opinion trajectories $\langle o_i \rangle$ for each agent $a_i$
for $i = 1$ to $N$ do
    Initialize agent $a_i$ with persona $per_i$ (includes
      initial opinion $o_i^{t=0}$), memory $m_i^{t=0}$
    (Optional) Inject cognitive bias and closed-world
      restriction
    Initialize opinion trajectory $\langle o_i \rangle = \{o_i^{t=0}\}$
for $t = 1$ to $T$ do
    Select random pair $\{a_i, a_j\}$, with $i \neq j$
    Agent $a_i$ writes tweet $x_i^t$
    Agent $a_j$ reports their verbal opinion $r_{j,t}$
    Classify opinion: $o_j = f_{oc}(r_j^t)$; append to $\langle o_j \rangle$
    Update memory: $m_i^{t+1}$, $m_j^{t+1}$ using $f_{mu}$
return $\langle o_i \rangle$ for each agent $a_i$
```

Formally, we begin with a pool of N LLM agents $\mathcal{A} = \{a_1, \cdots, a_N\}$ and a topic p. Each agent is initialized with a distinct *persona*, including an *initial opinion*, in their memory structure (described in §2.2 and Figure 2b). At each time step t, a pair of agents $A^t = \{a_i, a_j\}$ with $i \neq j$ is sampled uniformly from the population to interact. First, agent a_i composes a message x_i^t reflecting their current opinion about p. Second, agent a_j reads x_i^t and produces a verbal report r_j^t expressing their reaction to the message. The verbal report is then classified into a numeric opinion scale $o_j^t \in \{-2, -1, 0, 1, 2\} = \mathbb{O}$, ranging from strongly negative to strongly positive opinions about the topic[2], through an *opinion classifier*, denoted f_{oc} (detailed in §3.1).

After T rounds of pairwise interactions, we compile an *opinion trajectory* $\langle o_i \rangle = \{o_i^t\}_{t=0}^T$ for each agent. Note that an agent's opinion remains constant unless they are selected for an interaction. We further denote F_o^t as the *opinion distribution*, defined as the empirical frequency distribution of agents' opinion over the discrete opinion space $\mathbb{O}$ across all N agents at time t. [3]

2.2 Agent's Persona and Memory

Each agent a_i maintains a dynamic memory module m_i^t that evolves over time (Figure 3, 2b). In practice, the memory module is represented as text descriptions included in the prompt to the agent (see §3.1). The memory m_i^t influences the generation of a new message x_i^t and the assessment of other agents' messages x_j^t. We denote a *memory update function* for updating the agent's memory state, i.e., $m_i^{t+1} = f_{mu}(z, x_i^t, r_j^t, m_i^t)$, where $z \in \{\text{“write”}, \text{“review”}\}$ denotes the interaction type of either writing or reviewing a tweet.

Two memory update strategies are considered: (a) a *cumulative memory* that sequentially appends each new experience (either the experience of writing a tweet or reviewing a tweet) and (b) a *reflective memory*, inspired by Park et al. (2023), that maintains a compact summary by continuously reflect-

---

[2] In this study, the discrete opinion space $\mathbb{O}$ takes five ordinally-increasing values. Note that the size of $\mathbb{O}$ can be easily generalized. For a detailed description of the discrete opinion space $\mathbb{O}$ and the correspondence of the numeric values to verbal descriptions of opinions, see §C.

[3] Formally, the opinion frequency distribution $F_o^t : \mathbb{O} \to \mathbb{N}$ is defined as: $F_o^t(o) = |\{a_i \in \mathcal{A} : o_i^t = o\}|$, which maps each opinion value $o \in \mathbb{O}$ to the number of agents holding that opinion at time t.

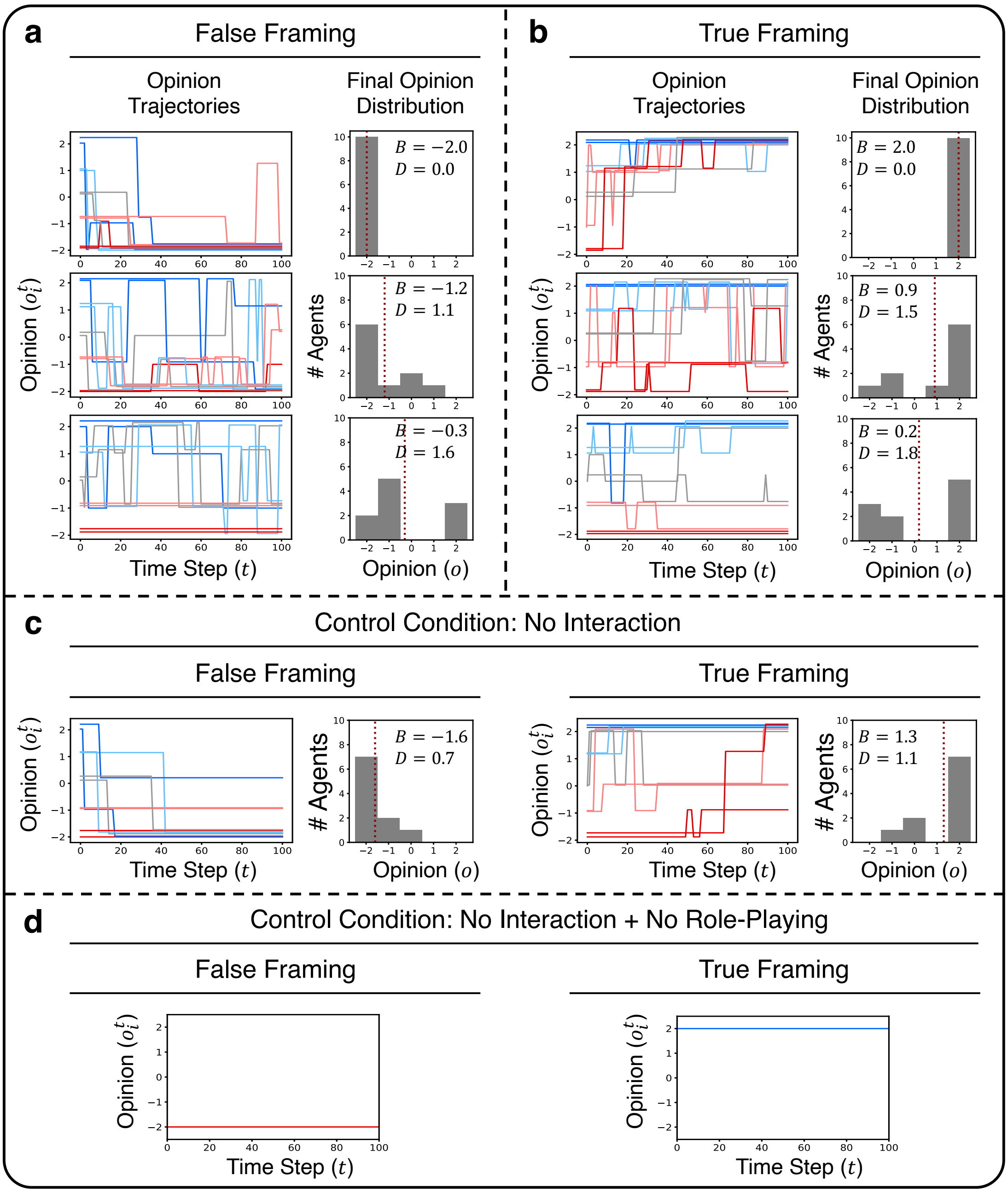


Figure 4: Opinion trajectories $\langle o_i \rangle$ of LLM agents and the final opinion distribution F_o^T on the topic of Global Warming. Panels (a) and (b) display the impact of cognitive biases under (a) false and (b) true framing conditions, respectively. Each row represents a different level of confirmation bias: no confirmation bias (top row), weak confirmation bias (middle row), and strong confirmation bias (bottom row). Panels (c) and (d) serve as baselines, with (c) being role-playing but with no interaction, and (d) being no role-playing and no interaction, respectively. The color of each line plot corresponds to the agent's initial opinion $o_i^{t=0}$: dark blue (+2), light blue (+1), grey (0), light red (-1), and dark red (-2), corresponding to opinions ranging from strongly agree to strongly disagree. The LLM agents in this figure use cumulative memory.

ing and integrating new experiences into the existing memory state (see §E for the detailed update function and the wording of the prompts). Both approaches are empirically evaluated to test their effects on opinion dynamics.

The first memory $m_i^{t=0}$ is initialized with the agent's persona, cognitive bias (if present), and the closed-world restriction (if present; see below), which can be all described in text sequences (detailed in §3). Personas are created to reflect a diverse demographic background incorporating various characteristics, including name, political leaning, age, gender, ethnicity, education, and occupation (see Figure 2b for an example). Alongside these attributes, a placeholder for their *initial opinion* $o_i^{t=0}$ is also included with natural language

		Cumulative Memory		Reflective Memory	
Framing	Confirmation Bias	Bias (B)	Diversity (D)	Bias (B)	Diversity (D)
False	None	**-1.33 ± 0.17**	**0.60 ± 0.11**	**-1.37 ± 0.11**	**0.75 ± 0.12**
	Weak	-0.96 ± 0.20	**0.87 ± 0.12**	-1.07 ± 0.17	**1.04 ± 0.14**
	Strong	-0.9 ± 0.14	**1.24 ± 0.11**	-0.85 ± 0.15	**1.33 ± 0.12**
True	None	**0.52 ± 0.31**	**0.66 ± 0.11**	**0.60 ± 0.31**	**0.85 ± 0.12**
	Weak	0.56 ± 0.27	**0.95 ± 0.11**	0.17 ± 0.28	**1.23 ± 0.11**
	Strong	-0.10 ± 0.13	**1.52 ± 0.05**	-0.09 ± 0.16	**1.65 ± 0.04**

Table 1: The bias (B) and diversity (D) of the final opinion distribution F_o^T aggregated across all 15 topics, for both cumulative and reflective memory strategies under false and true framing conditions, and different levels of induced confirmation bias. The values represent the average across 15 topics, along with the standard errors. Increasing the strength of the confirmation bias correlates with increasing D, as highlighted by the **green color gradient**. Notably, under true framing, B tends to be more positive (more agreeing) compared to false framing, indicated by **blue** for true and **red** for false framing conditions.

		Cumulative Memory		Reflective Memory	
Framing	CB	Bias (B)	Diversity (D)	Bias (B)	Diversity (D)
Science Topics					
False	None	**-1.70 ± 0.15**	**0.27 ± 0.11**	**-1.48 ± 0.15**	**0.65 ± 0.12**
	Weak	-0.86 ± 0.33	**0.96 ± 0.27**	-1.30 ± 0.26	**0.92 ± 0.27**
	Strong	-0.78 ± 0.17	**1.44 ± 0.13**	-0.36 ± 0.19	**1.52 ± 0.09**
True	None	**0.60 ± 0.50**	**0.72 ± 0.22**	**0.24 ± 0.70**	**0.76 ± 0.26**
	Weak	-0.24 ± 0.40	**1.05 ± 0.18**	-0.02 ± 0.42	**1.39 ± 0.07**
	Strong	-0.34 ± 0.12	**1.46 ± 0.10**	-0.34 ± 0.27	**1.61 ± 0.06**
History Topics					
False	None	**-1.12 ± 0.41**	**0.81 ± 0.27**	**-1.52 ± 0.17**	**0.71 ± 0.25**
	Weak	-1.22 ± 0.13	**0.81 ± 0.18**	-1.32 ± 0.26	**0.70 ± 0.23**
	Strong	-1.12 ± 0.35	**1.06 ± 0.22**	-0.82 ± 0.24	**1.51 ± 0.17**
True	None	**0.22 ± 0.56**	**0.71 ± 0.21**	**0.40 ± 0.55**	**0.88 ± 0.26**
	Weak	0.48 ± 0.49	**0.89 ± 0.23**	-0.12 ± 0.67	**1.00 ± 0.17**
	Strong	-0.24 ± 0.27	**1.44 ± 0.10**	-0.16 ± 0.33	**1.63 ± 0.10**
Common Sense Topics					
False	None	**-1.16 ± 0.21**	**0.71 ± 0.11**	**-1.10 ± 0.22**	**0.87 ± 0.27**
	Weak	-0.80 ± 0.51	**0.85 ± 0.21**	-0.58 ± 0.29	**0.97 ± 0.23**
	Strong	-0.80 ± 0.20	**1.22 ± 0.20**	-1.36 ± 0.15	1.48 ± 0.09
True	None	**1.18 ± 0.41**	**0.57 ± 0.17**	**1.16 ± 0.34**	**0.92 ± 0.16**
	Weak	1.44 ± 0.17	**0.91 ± 1.28**	0.64 ± 0.36	**1.28 ± 0.26**
	Strong	0.28 ± 0.18	**1.66 ± 1.69**	0.22 ± 0.21	**1.69 ± 0.06**

Table 2: The bias (B) and diversity (D) of the final opinion distribution F_o^T for each of the three categories (science, history, common sense), for both memory strategies under false and true framing conditions, and different levels of induced confirmation bias (CB). For each category, the averages across five topics are shown along with the standard errors. Increasing the strength of the CB correlates with increasing D, as highlighted by the **green color gradient**. Notably, under true framing, B tends to be more positive (more agreeing) compared to false framing, indicated by **blue** for true and **red** for false framing conditions.

description. For example, an agent with $o_i^{t=0} = 0$ is given *"Initial Belief: Neutral"* in the persona (Figure 2b).[4] The initial opinion $o_i^{t=0}$ is specified through an initial opinion distribution $F_o^{t=0}$ that varies across simulation settings (§3.4). §B shows the full list of personas.

2.3 Cognitive Biases

We investigate the effects of inducing a cognitive bias via role-playing instructions on the group opinion dynamics. Specifically, we consider confirmation bias: the tendency to interpret information as confirming one's views and to discount contradictory evidence (Nickerson, 1998). Prior simulation studies using mathematical ABMs have shown that, when confirmation bias is introduced at the individual level, the overall population exhibits increasing opinion fragmentation (i.e., increased diversity D) as the confirmation bias strengthens (Lorenz et al., 2021). We assess whether LLM agents instructed to show confirmation bias likewise replicate this phenomenon in their opinion dynamics when communicating through natural language. To manipulate the strength of confirmation bias, we provide two bias levels following the spectrum in Lorenz et al. (2021). **Weak Confirmation Bias:** *"You will be more likely to believe information that supports your beliefs and less likely to believe information that contradicts your beliefs."* **Strong Confirmation Bias:** *"You will only believe information that supports your beliefs and will completely dismiss information that contradicts your beliefs."* See §F for the bias-inducing prompts.

2.4 Open-world vs. Closed-world Settings

Our study examines agent behavior in both closed-world and open-world settings. In the closed-world setting, which aligns with traditional opinion dy-

---

[4] The correspondence between numeric opinion values and verbal description of initial opinion is detailed in §C.

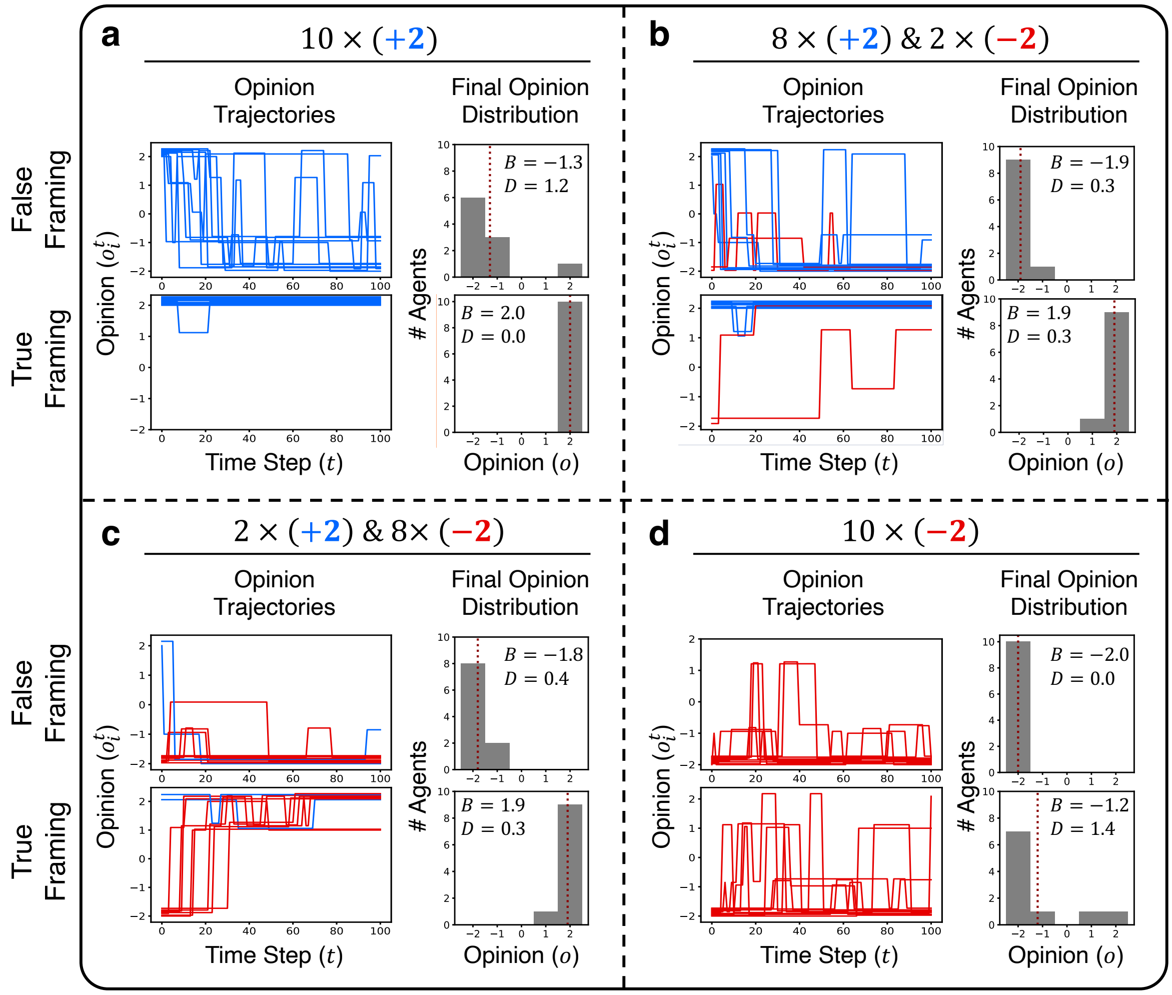


Figure 5: Varying initial opinion distribution F_o^T for the global warming debate. (a) All agents start with a strongly positive opinion. (b) 8 agents start with a strongly positive opinion while 2 with a strongly negative opinion (c) 8 agents start with a strongly negative opinion while 2 with a strongly positive opinion. (d) All agents start with a strongly negative opinion. The color of each line plot corresponds to the agent's initial opinion $o_i^{t=0}$: dark blue (+2), light blue (+1), grey (0), light red (-1), and dark red (-2), corresponding to opinions ranging from strongly agree to strongly disagree. The LLM agents in this figure use cumulative memory.

namics models, belief change is solely attributed to social influences within the system, and agents are restricted from accessing external information (restricted by instructions in the prompt; §G provides specific prompting details). Conversely, the open-world setting allows agents the freedom to "hallucinate" facts external to the system, such as discussing topics with imaginary friends (Dziri et al., 2022; Ji et al., 2023; Huang et al., 2023). We investigate the incidence of hallucination in both settings to understand the impact of external information on social influence. Our findings indicate no hallucination in the closed-world setting, while a hallucination rate of about 15% is observed in the open-world scenario. Consequently, the results reported in this study focus on the closed-world setting. The measurement procedure involves independent human annotators (detailed in §L).

3 Experimental Settings

3.1 Configuration

In our simulations, LLM agents use ChatGPT (`gpt-3.5-turbo-16k`) with temperature of 0.7 (OpenAI, 2022). During initialization, each agent's persona, along with the optional closed-world restrictions and cognitive biases, are incorporated into the model's *"system messages"*. Throughout the interaction, the historical events are added to the model's *"user messages"*. The memory of the LLM agents is managed through LangChain (Chase, 2022). In all experiments, we set the number of steps to $T = 100$ and the number of agents to $N = 10$. We use the FLAN-T5-XXL model (Chung et al., 2022) as the opinion classifier f_{oc} (see §I for the prompt used to classify the opinions and §N for validation against human ratings). In sensitivity analyses, we consider other LLMs like GPT-4 (`gpt-4-1106-preview`) and Vicuna-33B (`vicuna-33b-v1.3`; Zheng et al., 2023), and an experiment with a larger network ($N = 20$).

3.2 Topics and Framings

To study the opinion dynamics of the agents, we selected 15 topics with a known ground truth spanning scientific theories (*science topics*, e.g., whether global warming is a conspiracy), historical events (*history topics*, e.g., whether the 911 attack was an "inside job"), and commonsense knowledge (*common sense topics*, e.g., whether the sky is blue on a sunny day). For each topic, we created two framings. The *true framing* affirms the widely-accepted truth while the *false framing* affirms the opposite. For example, for the topic of "Global Warming", the two framings are as follows. **True Framing:** *"Theory XYZ claims that global warming is a real phenomenon and global climate is rapidly growing warmer."* **False Framing:** *"Theory XYZ claims that global warming is a conspiracy by governments worldwide and is not a real phenomenon."* To control for wording effects across topics, we add "Theory XYZ" before stating the claim in all prompts. The full set of prompts is provided in §H.

3.3 Opinion Dynamics Evaluation Metrics

Our evaluation of opinion dynamics includes two metrics in the literature (Lorenz et al., 2021): **Bias (B):** The average of the agents' opinions at the final time step, $B = \text{mean}(F_o^T)$. **Diversity (D):** The standard deviation of the final opinion distribution, $D = \text{std}(F_o^T)$. These two metrics can be used to categorize the final opinion distribution F_o^T. For example, a low B and a low D characterizes *central consensus*, whereas a low B and a medium D characterizes a *diversified* distribution.[5]

3.4 Initial Opinion Distribution

The initial opinion distribution $F_o^{t=0}$ determines the agents' starting opinions. The opinion is initialized verbally in the agent's prompt and memory (see §2.1 and Figure 2b). In most experiments, we initialize $F_o^{t=0}$ as a uniform distribution, with each opinion value assigned to $N/|\mathbb{O}|$ agents, where $|\mathbb{O}|$ is the number of possible opinion values.[6] This reflects an unbiased starting state with $B = 0$ and $D = 1.49$. In one experiment, we intentionally manipulate the initial distribution to be highly skewed. For example, assigning all 10 agents an initial opinion of -2, or 8 agents to -2 and 2 agents to $+2$. This allows us to study the effects of polarized starting opinions on the resulting opinion dynamics.

3.5 Control Conditions

In addition to the main experimental conditions, we introduce two control conditions: **(a) No Interaction Condition:** Agents are initialized with their personas and initial beliefs as normal, but do not actually interact. Instead, each agent a_i independently provides 10 opinion reports o_i^t on the topic. **(b) No Interaction + No Role-Playing Condition:** No agents are initialized with their personas and initial beliefs. We simply query the LLM for 10 independent opinion reports on the topic. These control conditions allow us to assess whether the LLM has inherent biases on the topics that manifest even without social influence dynamics. Comparison to the main interactive conditions allows us to discern effects stemming from the personas and social interactions.

4 Results

Agents Converge towards the Inherent Bias in the LLM. Table 1 summarizes the Bias (B) and Diversity (D) of the final opinion distribution F_o^T aggregated across 15 topics. Table 2 shows the summarized results separated by three topic types. As shown in Table 1, the role-playing prompt initially causes agents to express a diverse variety of opinions as expected, but with repeated social interacts these opinions converge toward a ground-truth consensus. Under the false framing, agents collectively lean towards disagreement, shown by a negative bias value ($B = -1.33$ when there is no cognitive bias). Conversely, under the true framing, the group shows a slight positive tendency to agree ($B = 0.52$), indicating a lean towards truthfulness. Figure 4 and 6 shows an example of how opinion trajectories quickly converge towards the truth after social interactions for both the false and true framing conditions, especially without cognitive bias. This is true across using cumulative memory (Figure 4) and reflective memory (Figure 6). The control condition illustrates that a similar tendency is observed when agents do not communicate, but

---

[5] See Lorenz et al. (2021) for a detailed taxonomy.

[6] For example, in our experiment, with $N = 10$ agents and five possible opinion values of $-2, -1, 0, +1, +2$, each value would be assigned to 2 agents initially.

are repeatedly queried for their opinion: the expressed opinions tend to move toward the ground truth, suggesting an inherent bias in the model. Statistical tests supporting the claims above are in §M.

Confirmation Bias Leads to Opinion Fragmentation. Introducing confirmation bias in the prompt leads to less ultimate consensus (i.e., greater diversity D) across LLM agents. As shown in Table 1 and Figure 4, the stronger the confirmation bias, the more diverse the final state distribution. This correlation holds for both cumulative and reflective memory strategies (Figure 4 and 6). These findings replicate, within a set of interacting LLMs, the general finding from more classic ABMs that incorporation of confirmation bias in the model update algorithm produces greater opinion fragmentation (Lorenz et al., 2021; Flache et al., 2017). Statistical tests supporting the claims above are in §M.

Impact of Initial Opinion Distribution The system's tendency for simulated opinions to converge on ground truth prompts an intriguing question: If all agents start with false opinions, will they still converge toward a scientifically accurate consensus, or will they reinforce their initial beliefs and resist changing their stance? Figure 5 shows the evolution of opinions under various initial distributions, using the global warming topic. Regardless of the initial opinion distribution, the agents altered their expressed opinions and shifted toward the ground truth. For instance, as shown in Figure 5a, under false framing, when all agents initially supported global warming is a hoax, they converged towards the negative spectrum quickly and ended up with $B = -1.3$. Interestingly, under true framing, when all agents initially denied the view that global warming is real, they did not completely flip their stance to support it, though they did shift slightly in this direction (Figure 5d): the final bias ($B = -1.2$) was more positive than the initial extreme opinion ($B = -2$). When at least a minority of agents held a divergent belief at the start, the group as a whole eventually shifted towards acknowledging global warming is real and is not a hoax, as shown in Figure 5c. Overall, these results indicate that the model's inherent bias towards ground truth is robust against varying initial opinion distributions.

The Strength of Bias under False Framing is Stronger than under True Framing As shown in Table 1, the LLM agents have a stronger tendency to deny the false statement under the false framing (Cumulative Memory: $B = -1.33$; Reflective Memory: $B = -1.37$) than their tendency to endorse the true statement under the true framing (Cumulative Memory: $B = 0.52$; Reflective Memory: $B = 0.60$). We consider two possibilities. First, FLAN-T5-XXL may have an overall tendency to classify LLM agents' responses as negative. Second, the LLM agents may indeed have a stronger tendency to deny the false statements. According to our validation of FLAN-T5-XXL's ratings (detailed in §N), the agreement matrix between FLAN-T5-XXL's ratings and human ratings shows no systematic bias. We therefore exclude the first possibility. As to why the LLM agents indeed have a stronger negative tendency, our speculation is that, due to either pretraining, fine-tuning, or RLHF (Christiano et al.), the LLM has been trained to readily refute false information under false framing. On the other hand, under true framing, there may be less training effort to ensure that the model endorses true information. Future studies are needed to evaluate the true cause.

Sensitivity Analyses Sensitivity analyses test different LLMs (GPT-4 and Vicuna) and network sizes ($N = 20$ agents). The results show consistent trends across models (GPT-4: Table 3; Vicuna: Table 4 in §J) and network scales (Table 5 in §K). In addition, we evaluate the sensitivity of our result to randomness due to temperature sampling. On the topic of Global Warming, we run the experiment three times. The effect of framing and the effect of confirmation bias are consistent across the three runs.

5 Related Work

Agent-Based Models and Opinion Dynamics Simulation Agent-Based Models (ABMs) are the cornerstone of opinion dynamics simulation, defining mathematical rules for agents' opinion updates in response to messages (Gilbert and Terna, 2000; Smaldino, 2023; Lorenz et al., 2021; Chuang and Rogers, 2023; Epstein, 2012; Flache et al., 2017). ABMs are valuable for predicting public opinion trends and informing intervention strategies. One key advantage of using ABMs is that they allow incorporating explicit assumptions about cognitive process in opinion updating (Flache et al., 2017; Lorenz et al., 2021; Chuang and Rogers, 2023). For example, incorporating "confirmation bias" into ABM equations causes agents to disregard contrasting information, often leading to fragmented opinion clusters at the group level. However, a significant limitation of ABMs is that they rely on numeric representations of opinions and messages, which oversimplifies the complexities of human communication. In contrast, emerging approaches using LLMs offer a more sophisticated method for simulating opinion dynamics through natural language.

Simulating Social Dynamics with LLM-based Agents The use of LLMs in simulating social dynamics is a rapidly growing research area , showcasing promising results in terms of human-like interactions (Park et al., 2023, 2022; Kaiya et al., 2023; Törnberg et al., 2023; Li et al., 2023a; Zhou et al., 2023b; Mou et al., 2024; Liu et al., 2024). Park et al. (2023) devise LLM-based generative agents to engage in digital environments, demonstrating an ability to respond, plan, and remember in natural language. They exhibit complex social behaviors, such as organizing events. Similarly, Törnberg et al. (2023) use LLMs in conjunction with agent-based modeling to explore the impact of news feed algorithms in simulated social media environments. Additionally, Park et al. (2022) show that LLM-based agents are capable of generating social media posts indistinguishable from those written by humans. These advances underscore the potential of using LLM agents to simulate human social behaviors at group level. To our best knowledge, we are the first to propose the use of LLM as an alternative to ABM for opinion dynamics simulation.

6 Conclusion

This study has explored the use Large Language Models (LLMs) for understanding opinion dynamics in groups of simulated agents communicating via natural language. In contrast to more traditional ABMs, LLMs can interpret and produce natural language, can role-play differing personas, and can simulate human-like linguistic communication. We therefore considered whether groups of interacting LLM agents could provide a basis for simulating opinion dynamics comparable to those studied with classical ABMs. Our findings confirm the potential of LLMs in opinion dynamics simulations but also reveal limitations, particularly their tendency to align with factual information regardless of their personas, which restricts their role-play effectiveness for individuals with fact-resistant beliefs like climate change denial.

Significant efforts have been made to prevent LLMs from exhibiting harmful biases. However, for simulating critical undesired social phenomena (e.g., misinformation, polarization), it is crucial to have simulated agents accurately reflect the breadth of human behavior and belief, even those that are maladaptive. Our study suggests that prompting alone may be insufficient for LLM agents to fully replicate the diverse viewpoints. This leads us to a potential future direction: fine-tuning LLM agents with actual human discourse data. Such an approach could lead to more accurate models of human belief dynamics.

Limitations

Model Dependency and Generalizability A key limitation of our study is the exclusive use of models that have undergone Reinforcement Learning with Human Feedback (RLHF; Christiano et al., 2017; Ziegler et al., 2019). This may lead to the truth-converging tendency in the LLM agents. Given that various language models exhibit distinct inherent biases (Feng et al., 2023), LLM agents using different models could display varying patterns in opinion dynamics. To fully assess the generalizability of our findings, future research should include a broad spectrum of models.

Reduction of Opinion to One-Dimensional Scalar Our study aligns with classic ABMs in reducing opinions to a one-dimensional scalar $o \in \mathbb{R}$, which simplifies the complex nature of opinion formation. However, a more nuanced approach could offer deeper insights. Future studies could adopt a fine-grained or even qualitative analysis to explore how agents modify their opinions, determine which messages hold greater persuasive power, and assess how persuasion varies based on different agent personas. Such an approach would provide a richer understanding of the subtleties in LLM agents' opinion dynamics.

Limitation in Topic Selection Our research focused on topics with clear, established ground truths. However, many crucial topics, such as the effectiveness of political leaders or the best policies for complex societal issues, lack a definitive truth. These topics are more open-ended and subjective. Future studies should consider including such topics to capture a broader and more nuanced spectrum of opinions and debates.

The Role of Other Demographic Information While we focus on the role of agents' initial beliefs in opinion dynamics simulation, a more nuanced analysis on the role of other demographic information could offer deeper insights. For example, future studies could adopt a fine-grained or qualitative analysis to explore how agents' demographic features (like political leaning) play a role in agents' information exchanges and belief updating.

Limitation in Network Structure In the present study, LLM agents get to interact with everyone else in the network. While in line with classic setup used in ABM simulation (Flache et al., 2017; Lorenz et al., 2021), this is not close to social interactions in the real world. For example, people with similar background tend to interact with one another (McPherson et al., 2001). Future studies should consider investigating LLM agents' opinion dynamics with more realistic network structures.

Ethics Statement

While introducing confirmation bias into LLM agents can lead to opinion fragmentation and reduced convergence on factual consensus, it's important to understand this approach within the broader scope of studying group-level social phenomena. Simulating biased behavior in agents is not an endorsement of these biases, but a critical step in comprehensively understanding the dynamics of various undesired social issues, including misinformation spread, polarization, and echo chamber formation. Developing human-like LLM agents with resistant viewpoints is essential for devising strategies to address these social challenges. In addition, we will release the code base exclusively for research purposes. Finally, since we are using OpenAI's API, we make sure that we comply with its intended use [7].

Acknowledgements

We thank the reviewers, the area chair for their feedback. This work was funded by the Multi University Research Initiative grant from the Department of Defense, W911NF2110317, Cohesive and Robust Human-Bot Cybersecurity Teams, the John S. and James L. Knight Foundation (Award Number: MSN231314), and the National Science Foundation through the Convergence Accelerator Track F: Course Correct: Precision Guidance Against Misinformation (Agency Tracking Number: 2230692; Award Number: MSN 266268).

---

[7] `https://openai.com/policies/terms-of-use`

References

Ceren Budak, Divyakant Agrawal, and Amr El Abbadi. 2011. Limiting the spread of misinformation in social networks. In *Proceedings of the 20th international conference on World wide web*, pages 665–674.

Harrison Chase. 2022. Langchain.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30.

Paul F Christiano, Jan Leike, Tom B Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences.

Yun-Shiuan Chuang and Timothy T Rogers. 2023. Computational agent-based models in opinion dynamics: A survey on social simulations and empirical studies. *arXiv preprint arXiv:2306.03446*.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Nouha Dziri, Sivan Milton, Mo Yu, Osmar Zaiane, and Siva Reddy. 2022. On the origin of hallucinations in conversational models: Is it the datasets or the models? *arXiv preprint arXiv:2204.07931*.

Joshua M Epstein. 2012. *Generative social science: Studies in agent-based computational modeling*. Princeton University Press.

Shangbin Feng, Chan Young Park, Yuhan Liu, and Yulia Tsvetkov. 2023. From pretraining data to language models to downstream tasks: Tracking the trails of political biases leading to unfair NLP models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11737–11762, Toronto, Canada. Association for Computational Linguistics.

Andreas Flache, Michael Mäs, Thomas Feliciani, Edmund Chattoe-Brown, Guillaume Deffuant, Sylvie Huet, and Jan Lorenz. 2017. Models of social influence: Towards the next frontiers. *Journal of Artificial Societies and Social Simulation*, 20(4).

Nigel Gilbert and Pietro Terna. 2000. How to build and use agent-based models in social science. *Mind & Society*, 1:57–72.

Tamar Ginossar, Iain J Cruickshank, Elena Zheleva, Jason Sulskis, and Tanya Berger-Wolf. 2022. Cross-platform spread: vaccine-related content, sources, and conspiracy theories in youtube videos shared in early twitter covid-19 conversations. *Human vaccines & immunotherapeutics*, 18(1):1–13.

Andrew F Hayes and Klaus Krippendorff. 2007. Answering the call for a standard reliability measure for coding data. *Communication methods and measures*, 1(1):77–89.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232*.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Zhao Kaiya, Michelangelo Naim, Jovana Kondic, Manuel Cortes, Jiaxin Ge, Shuying Luo, Guangyu Robert Yang, and Andrew Ahn. 2023. Lyfe agents: Generative agents for low-cost real-time social interactions. *arXiv preprint arXiv:2310.02172*.

Chao Li, Xing Su, Chao Fan, Haoying Han, Cong Xue, and Chunmo Zheng. 2023a. Quantifying the impact of large language models on collective opinion dynamics. *arXiv preprint arXiv:2308.03313*.

Yixia Li, Rong Xiang, Yanlin Song, and Jing Li. 2023b. Unipoll: A unified social media poll generation framework via multi-objective optimization. *arXiv preprint arXiv:2306.06851*.

Yuhan Liu, Xiuying Chen, Xiaoqing Zhang, Xing Gao, Ji Zhang, and Rui Yan. 2024. From skepticism to acceptance: Simulating the attitude dynamics toward fake news. *arXiv preprint arXiv:2403.09498*.

Sahil Loomba, Alexandre de Figueiredo, Simon J Piatek, Kristen de Graaf, and Heidi J Larson. 2021. Measuring the impact of covid-19 vaccine misinformation on vaccination intent in the uk and usa. *Nature human behaviour*, 5(3):337–348.

Jan Lorenz, Martin Neumann, and Tobias Schröder. 2021. Individual attitude change and societal dynamics: Computational experiments with psychological theories. *Psychological Review*, 128(4):623.

Wei Lu, Wei Chen, and Laks VS Lakshmanan. 2015. From competition to complementarity: comparative influence diffusion and maximization. *Proceedings of the VLDB Endowment*, 9(2):60–71.

Miller McPherson, Lynn Smith-Lovin, and James M Cook. 2001. Birds of a feather: Homophily in social networks. *Annual review of sociology*, 27(1):415–444.

Xinyi Mou, Zhongyu Wei, and Xuanjing Huang. 2024. Unveiling the truth and facilitating change: Towards agent-based large-scale social movement simulation. *arXiv preprint arXiv:2402.16333*.

Raymond S Nickerson. 1998. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of general psychology*, 2(2):175–220.

OpenAI. 2022. Introducing ChatGPT. https://openai.com/blog/chatgpt. [Accessed 13-10-2023].

Joon Sung Park, Joseph C O'Brien, Carrie J Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. *arXiv preprint arXiv:2304.03442*.

Joon Sung Park, Lindsay Popowski, Carrie Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2022. Social simulacra: Creating populated prototypes for social computing systems. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*, pages 1–18.

Gordon Pennycook, Ziv Epstein, Mohsen Mosleh, Antonio A Arechar, Dean Eckles, and David G Rand. 2021. Shifting attention to accuracy can reduce misinformation online. *Nature*, 592(7855):590–595.

Paul Smaldino. 2023. *Modeling social behavior: Mathematical and agent-based models of social dynamics and cultural evolution*. Princeton University Press.

Petter Törnberg, Diliara Valeeva, Justus Uitermark, and Christopher Bail. 2023. Simulating social media using large language models to evaluate alternative news feed algorithms. *arXiv preprint arXiv:2310.05984*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *arXiv preprint arXiv:2306.05685*.

Wenxuan Zhou, Sheng Zhang, Hoifung Poon, and Muhao Chen. 2023a. Context-faithful prompting for large language models. *arXiv preprint arXiv:2303.11315*.

Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, et al. 2023b. Sotopia: Interactive evaluation for social intelligence in language agents. In *The Twelfth International Conference on Learning Representations*.

Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2019. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*.

A Results of Global Warming Topic with Reflective Memory

In the main text, Figure 4 shows the LLM agent opinion dynamics when discussing about global warming when using cumulative memory. Here, Figure 6 shows the result when using reflective memory.

B Full List of Personas

In this section, we list the full list of 10 agents along with their personas that interact in the group dynamics settings in our agent-based model (ABM).

Name: Benjamin Lee
Initial Belief: Slightly Negative opinion about XYZ
Political leaning: Lean Democrat
Age: 37
Gender: Male
Ethnicity: Asian American
Education: Master's Degree in Economics
Occupation: Financial Analyst

Name: Maya Jackson
Initial Belief: Strongly Negative opinion about XYZ
Political leaning: Strong Republican
Age: 29
Gender: Female
Ethnicity: Black
Education: Bachelor's Degree in Business Management
Occupation: Marketing Specialist

Name: Ethan Wilson
Initial Belief: Slightly Positive opinion about XYZ
Political leaning: Moderate
Age: 26
Gender: Male
Ethnicity: White
Education: Bachelor's Degree in Journalism
Occupation: Freelance Writer

Name: Aisha Patel
Initial Belief: Neutral opinion about XYZ
Political leaning: Lean Republican
Age: 41
Gender: Female
Ethnicity: South Asian
Education: Doctor of Medicine (M.D.)
Occupation: Pediatrician

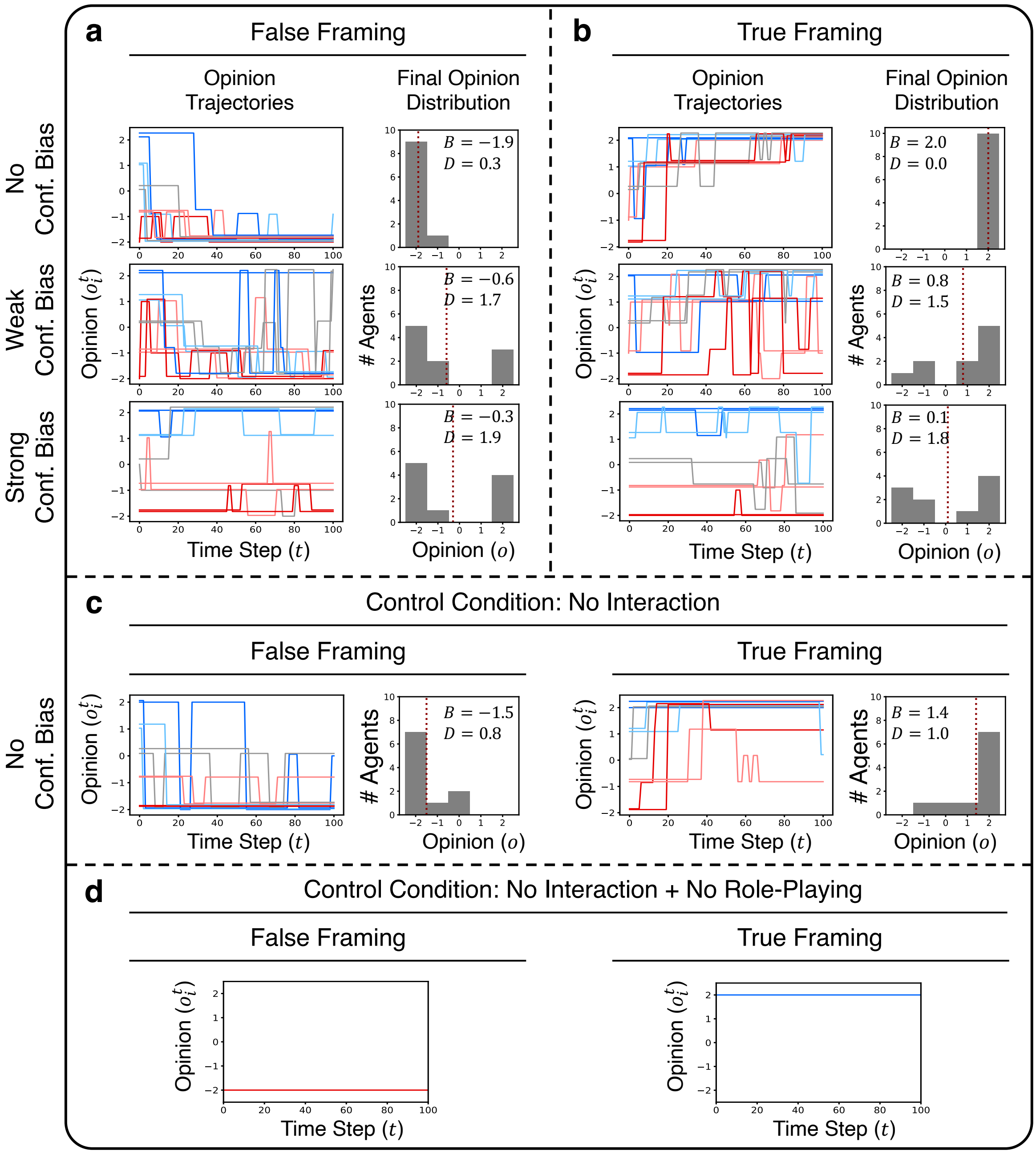


Figure 6: Opinion trajectories $\langle o_i \rangle$ of LLM agents and the final opinion distribution F_o^T on the topic of Global Warming. Panels (a) and (b) display the impact of cognitive biases under (a) false and (b) true framing conditions, respectively. Each row represents a different level of confirmation bias: no confirmation bias (top row), weak confirmation bias (middle row), and strong confirmation bias (bottom row). Panels (c) and (d) serve as baselines, with (c) being role-playing but with no interaction, and (d) being no role-playing and no interaction, respectively. The LLM agents in this figure use reflective memory. The color of each line plot corresponds to the agent's initial opinion $o_i^{t=0}$: dark blue (+2), light blue (+1), grey (0), light red (-1), and dark red (-2), corresponding to opinions ranging from strongly agree to strongly disagree. The LLM agents in this figure use cumulative memory.

Name: Samuel Wright
Initial Belief: Strongly Negative opinion about XYZ
Political leaning: Strong Democrat
Age: 58
Gender: Male
Ethnicity: White
Education: Ph.D. in Anthropology

Occupation: University Professor

Name: Olivia Garcia
Initial Belief: Strongly Positive opinion about XYZ
Political leaning: Strong Democrat
Age: 34
Gender: Female

Ethnicity: Hispanic
Education: Master's Degree in Sociology
Occupation: Non-profit Program Manager

Name: Sophia Nguyen
Initial Belief: Slightly Negative opinion about XYZ
Political leaning: Lean Republican
Age: 24
Gender: Female
Ethnicity: Asian American
Education: Student (Undergraduate, Political Science)
Occupation: Intern at Law Firm

Name: Sarah Martinez
Initial Belief: Strongly Positive opinion about XYZ
Political leaning: Lean Democrat
Age: 28
Gender: Female
Ethnicity: Hispanic
Education: Master's Degree in Film Studies
Occupation: Film Critic

Name: Jordan White
Initial Belief: Slightly Positive opinion about XYZ
Political leaning: Moderate
Age: 23
Gender: Female
Ethnicity: Black
Education: Student (Undergraduate, Sociology)
Occupation: Part-time Retail Worker

Name: Lucas Johnson
Initial Belief: Neutral opinion about XYZ
Political leaning: Moderate
Age: 37
Gender: Male
Ethnicity: Black
Education: Bachelor's Degree in Sociology
Occupation: Diversity and Inclusion Manager

C Detailed Description of the Discrete Opinion Space $\mathbb{O}$

The discrete opinion space $\mathbb{O}$ used in our simulations includes five possible values, each representing a different opinion on a given topic (referred to as "XYZ"). The values are as follows:

- -2: Strongly negative opinion about XYZ.
- -1: Slightly negative opinion about XYZ.
- 0: Neutral opinion about XYZ.
- 1: Slightly positive opinion about XYZ.
- 2: Strongly positive opinion about XYZ.

The opinion space $\mathbb{O}$ is used when initializing an agent's initial belief $o_i^{t=0}$ (§2.2) and classifying opinion from verbal report r_j^t (§3 and §I). Note that the size of $\mathbb{O}$ can be easily generalized to accommodate a broader range of opinion scales.

D Agent Interaction Prompts

In this section, we list the prompts used for facilitation of the interactions between the agents. Specifically, we start with the prompt that introduces the agents' persona, followed by the prompts for them receiving and writing tweets respectively. All prompts are for the specific topic of debate on flat earth with positive framing (see §H).

1. Initialize Persona Prompt

"Role play this person.
{AGENT_PERSONA}

Now, {AGENT_NAME}, you have been interacting with other strangers on Twitter. You can decide to change or maintain your belief about the theory XYZ that claims that the Earth is flat after interacting with other strangers.

You would produce tweets that reflect your honest belief, and you would also see other strangers' tweets. After seeing other people's tweets, you would be asked about your belief about the theory XYZ that claims that the Earth is flat."

2. Write Tweet Prompt

"Now, {AGENT_NAME}, please write a tweet about the theory XYZ that claims that the Earth is flat. The tweet should reflect your honest belief.

Write the tweet now.
Your Tweet:"

2. Receive Tweet Prompt

"Now, {AGENT_NAME}, you see a post on Twitter from a stranger.
I want to know your current honest belief about the theory XYZ that claims that the Earth is flat after seeing this Tweet.

```
Here is the Tweet.
{TWEET}

What is your current honest belief
about the theory XYZ that claims that
the Earth is flat?  Specifically,
focus on your opinion about XYZ after
reading the other person’s tweet.
Use the following format:
Reasoning:  (Think step by step)

Reasoning:
As {AGENT_NAME}, I”
```

These prompts are used and the responses are added to the memory and updated based on the memory update function detailed in the following section.

E Agent Memory Update Function

Cumulative Memory: The cumulative memory as described in §2.2 appends each new experience as time progresses. In order to add these past interactions into the agents’ memories, we use the following prompts.

Considering the interaction at time $t + 1$, the agent could be either tweeting by themselves or receiving a tweet, and similarly at time t, they would’ve either tweeted by themselves, received a tweet, or neither of these (say for instance, that $(t + 1)$th time step is the first time they were chosen). We, therefore, list these prompts case-by-case, on the basis of previous and current interaction_type $\in$ {none, write, review} for the specific topic of debate on a flat earth with positive framing (see §H).

1. Previous is none, Current is review:

```
“You first saw a tweet from a
stranger on Twitter.  Here is the
tweet you saw.
{TWEET_SEEN}

After seeing the tweet, below was
your thought and honest belief about
the theory XYZ that claims that the
Earth is flat.  Your thought after
you saw the tweet:
{REASONING}”
```

2. Previous is none, Current is write:

```
“Below was the {TWEET_WRITTEN_COUNT}
{SUPERSCRIPT} tweet you wrote earlier
about the theory XYZ that claims that
the Earth is flat:
{TWEET_WRITTEN}”
```

3. Previous is review, Current is write:

```
“After you saw the tweet from the
stranger above, you wrote another
tweet about the theory XYZ that
claims that the Earth is flat.
Below was the {TWEET_WRITTEN_COUNT}
{SUPERSCRIPT} tweet you wrote earlier
about the theory XYZ that claims that
the Earth is flat:
{TWEET_WRITTEN}”
```

4. Previous is review, Current is review:

```
“After you saw the tweet from the
stranger above, you saw another tweet
from a stranger about the theory XYZ
that claims that the Earth is flat.
Here is the tweet you saw.
{TWEET_SEEN}

After seeing the tweet, below was
your thought and honest belief about
the theory XYZ that claims that the
Earth is flat.
Your thought after you saw the tweet:
{REASONING}”
```

5. Previous is write, Current is write:

```
“After you wrote your
{TWEET_WRITTEN_COUNT}
{SUPERSCRIPT_LAST} tweet, you wrote
another tweet

Below was the {TWEET_WRITTEN_COUNT}
{SUPERSCRIPT} tweet you wrote earlier
about the theory XYZ that claims that
the Earth is flat:
{TWEET_WRITTEN}”
```

6. Previous is write, Current is review:

```
“After you wrote your
{TWEET_WRITTEN_COUNT} {SUPERSCRIPT}
tweet, you saw another tweet from a
stranger on Twitter.
Here is the tweet you saw.
{TWEET_SEEN}

After seeing the tweet, below was
your thought and honest belief about
the theory XYZ that claims that the
Earth is flat.
Your thought after you saw the tweet:
{REASONING}”
```

Reflective Memory:

As described in §2.2, the reflective memory, maintains a compact summary by prompting the agent to continuously reflect on its experiences interacting with others, followed by integrating new experiences into the existing memory state so as to maintain a roughly constant memory size of the agent.

Below is the prompt we use to implement the reflection-based memory into the LLM agents:

1. The agent is reflecting for the first time:

```
"Now, please reflect on this
experience.  Summarize your
experience in a few sentences."
```

2. The agent is reflecting beyond the first time:

```
"Now, please reflect on this
experience.  Here is your experience
so far:  {PAST_REFLECTION} Summarize
your updated experience in a few
sentences."
```

Essentially, after every reflection, the agents' memory only contains the reflected experience that the agents produce, which forms our memory update function for this setting.

F Confirmation Bias Prompt

To manipulate the strength of confirmation bias, we provide two bias levels following the spectrum in Lorenz et al. (2021). Below are the exact wordings.

Weak Confirmation Bias:

```
"Remember, you are role-playing as
a real person.  Like humans, you
have confirmation bias.  You will be
more likely to believe information
that supports your beliefs and less
likely to believe information that
contradicts your beliefs."
```

Strong Confirmation Bias:

```
"Remember, you are role-playing as
a real person.  You have a strong
confirmation bias.  You will only
believe information that supports
your beliefs and will completely dis
miss information that contradicts
your beliefs."
```

G Closed-World Prompt

When agents discuss or debate about a topic XYZ under the closed-world constraint (See Section 2), we prompt the agents with the following:

```
"Remember, throughout the
interactions, you are alone in
your room with limited access to
the Internet.  You cannot search for
information about XYZ on the Internet.
You can not go out to ask other
people about XYZ. Because you are
alone in your room, you can not leave
your room to seek information about
XYZ. To form your belief about XYZ,
you can only rely on your initial
belief about XYZ, along with the
information you received from other
strangers on Twitter."
```

This prompt is excluded entirely under the open-world assumption.

H Full List of Topics

Science

1. Flat Earth
False framing: *"Theory XYZ that claims that the Earth is flat."*
True framing: *"Theory XYZ that claims that the Earth is an irregularly shaped ellipsoid rather than flat."*

2. Coexistence of Tyrannosaurus Rex and humans
False framing: *"Theory XYZ that claims that the Tyrannosaurus Rex and humans co-existed on Earth at the same time."*
True framing: *"Theory XYZ that claims that the Tyrannosaurus Rex and humans did not co-exist on Earth at the same time."*

3. Communicating with the dead
False framing: *"Theory XYZ that claims that it is possible for humans to communicate with the dead."*
True framing: *"Theory XYZ that claims that it is not possible for humans to communicate with the dead."*

4. Predicting the future through palm reading
False framing: *"Theory XYZ that claims that it is possible to predict someone's future by looking at their palm characteristics."*
True framing: *"Theory XYZ that claims that it is not possible to predict someone's future by looking at their palm characteristics."*

5. Global warming
False framing: *"Theory XYZ that claims that global warming is a conspiracy by governments worldwide and is not a real phenomenon."*
True framing: *"Theory XYZ that claims that global warming is a real phenomenon and global climate is rapidly growing warmer."*

History

6. Moon Landing
False framing: *"Theory XYZ that claims that US astronauts never landed on the moon."*
True framing: *"Theory XYZ that claims that US astronauts have landed on the moon."*

7. 9/11 Twin Towers

False framing: *"Theory XYZ that claims that the twin towers were brought down from the inside by explosives during the 9/11 attack."*
True framing: *"Theory XYZ that claims that the twin towers were not brought down from the inside by explosives during the 9/11 attack."*

8. US Unemployment Rate 2016

False framing: *"Theory XYZ that claims that the US unemployment rate in 2016 was higher than 40%."*
True framing: *"Theory XYZ that claims that the US unemployment rate in 2016 was lower than 40%."*

9. Holocaust

False framing: *"Theory XYZ that claims that the Nazi government in Germany did not murder approximately 6 million Jewish people during the second world war."*
True framing: *"Theory XYZ that claims that the Nazi government in Germany murdered approximately 6 million Jewish people during the second world war."*

10. Barack Obama's Birthplace

False framing: *"Theory XYZ that claims that Barack Obama was born in Kenya."*
True framing: *"Theory XYZ that claims that Barack Obama was born in Hawaii."*

Common Sense

11. Bicycle Wheels

False framing: *"Theory XYZ that claims that a bicycle usually has four wheels."*
True framing: *"Theory XYZ that claims that a bicycle usually has two wheels."*

12. Washington DC's Location

False framing: *"Theory XYZ that claims that Washington DC is not in the United States."*
True framing: *"Theory XYZ that claims that Washington DC is in the United States."*

13. Human Brain at Birth

False framing: *"Theory XYZ that claims that human beings are not born with a brain."*
True framing: *"Theory XYZ that claims that human beings are born with a brain."*

14. Temperature of Fire

False framing: *"Theory XYZ that claims that fire is cold."*
True framing: *"Theory XYZ that claims that fire is hot."*

15. Color of the Sky

False framing: *"Theory XYZ that claims that on a clear sunny day, the sky is usually red."*
True framing: *"Theory XYZ that claims that on a clear sunny day, the sky is usually blue."*

I Opinion Classification

Our ABM framework involves an *opinion classifier* $f_{oc}(r_j^t)$ (See Section 2), and in this section, we describe the choice of our classifier and further detail an example prompt that we use to extract the opinion rating.

We choose Flan-T5-XXL (Chung et al., 2022) as our opinion classifier, primarily because of its capability and wide usage for extracting sentiment or opinion ratings across other studies (Zhou et al., 2023a; Li et al., 2023b).

Below is an example prompt for the topic of Global Warming with Positive Framing (See Appendix H) given to Flan-T5-XXL for evaluating responses of agents and returning a rating value $f_{oc}(r_j^t) \in \{-2, -1, 0, 1, 2\}$.

```
“Here’s a statement made by a person
after seeing a tweet.
---
{RESPONSE}
---


What is their belief about the
Theory XYZ that claims that
global warming is a conspiracy by
governments worldwide and is not a
real phenomenon, right after they
saw this tweet, on a 5-point Likert
scale:
-2: (Strongly negative opinion about
XYZ), -1: (Slightly negative opinion
about XYZ), 0: (Neutral opinion
about XYZ), 1: (Slightly positive
opinion about XYZ), 2: (Strongly
positive opinion about XYZ).


Answer with a single opinion value
within the options -2, -1, 0, 1, 2.”
```

The model then provides us with an opinion value, which is what we use to keep track of the beliefs of the LLM agents. In practice, we observe by random sampling of agent responses and the Flan-T5-XXL ratings that the opinion value output is more reliable than self-reported ratings of the agents' own beliefs.

J Results on GPT-4 and Vicuna

To ensure the robustness of our findings, in our sensitivity analyses, we extend the choice of LLM to include GPT-4 (`gpt-4-1106-preview`; OpenAI, 2022) and Vicuna (`Vicuna-33B-v1.3`; Zheng et al., 2023) For GPT-4, due to budget constraint, we focus only on the science-based topics, which costs 499.52 USD. For Vicuna, given the constraints of the GPU memory [8], we focus only on the reflective memory strategy, which ensures that the memory size remains approximately constant throughout the simulation. Results on both GPT-4 (Table 3) and Vicuna (Table 4) show a trend consistent with our findings from ChatGPT (Table 1 and 2).

		Cumulative Memory		Reflective Memory	
Framing	Confirmation Bias	Bias (B)	Diversity (D)	Bias (B)	Diversity (D)
False	None	**-1.48 $\pm$ 0.26**	**0.78 $\pm$ 0.32**	**-1.38 $\pm$ 0.28**	**0.80 $\pm$ 0.32**
	Weak	-1.28 $\pm$ 0.37	**0.81 $\pm$ 0.33**	-1.30 $\pm$ 0.34	**0.79 $\pm$ 0.26**
	Strong	-0.86 $\pm$ 0.25	**1.50 $\pm$ 0.15**	-0.78 $\pm$ 0.33	**1.32 $\pm$ 0.33**
True	None	**1.00 $\pm$ 0.43**	**0.96 $\pm$ 0.27**	**0.96 $\pm$ 0.45**	**1.11 $\pm$ 0.22**
	Weak	0.64 $\pm$ 0.62	**1.01 $\pm$ 0.33**	0.90 $\pm$ 0.46	**1.27 $\pm$ 0.11**
	Strong	0.24 $\pm$ 0.20	**1.63 $\pm$ 0.05**	0.42 $\pm$ 0.40	**1.52 $\pm$ 0.16**

Table 3: Results with GPT-4. The bias (B) and diversity (D) of the final opinion distribution F_o^T aggregated across all five science-based topics, for both cumulative and reflective memory strategies under false and true framing conditions, and different levels of induced confirmation bias. The values represent the average across five topics, along with the standard errors. Increasing the strength of the CB correlates with increasing D, as highlighted by the **green color gradient**. Notably, under true framing, B tends to be more positive (more agreeing) compared to false framing, indicated by **blue** for true and **red** for false framing conditions.

K Results on a Larger Network

To test if our findings scale to a larger network, in our sensitivity analyses, we double the network size and run the simulation longer ($N = 20$ and $T = 200$) on the topic of Global Warming. The result using a larger network with 20 agents (Table 5) shares the same patterns as in the smaller network with 10 agents (Table 6).

Framing	Confirmation Bias	Reflective Memory	
		Bias (B)	Diversity (D)
False	None	**-0.38 $\pm$ 0.11**	**0.62 $\pm$ 0.13**
	Weak	0.18 $\pm$ 0.22	**0.72 $\pm$ 0.09**
	Strong	0.49 $\pm$ 0.14	**0.73 $\pm$ 0.09**
True	None	**0.58 $\pm$ 0.13**	**0.53 $\pm$ 0.10**
	Weak	0.37 $\pm$ 0.15	**0.64 $\pm$ 0.08**
	Strong	0.33 $\pm$ 0.15	**0.72 $\pm$ 0.12**

Table 4: Results with Vicuna-33B. The bias (B) and diversity (D) of the final opinion distribution F_o^T aggregated across 15 topics, using the reflective memory strategy, under false and true framing conditions, and different levels of induced confirmation bias. The values represent the average across five topics, along with the standard errors. Increasing the strength of the CB correlates with increasing D, as highlighted by the **green color gradient**. Notably, under true framing, B tends to be more positive (more agreeing) compared to false framing, indicated by **blue** for true and **red** for false framing conditions.

		Cumulative Memory		Reflective Memory	
Framing	Confirmation Bias	Bias (B)	Diversity (D)	Bias (B)	Diversity (D)
False	None	**-1.80**	**0.68**	**-1.45**	**0.97**
	Weak	-1.05	**1.28**	-0.80	**1.57**
	Strong	-0.05	**1.75**	-0.55	**1.60**
True	None	**1.65**	**0.73**	**1.55**	**0.97**
	Weak	1.35	**0.91**	0.85	**1.42**
	Strong	0.25	**1.81**	0.50	**1.50**

Table 5: The bias (B) and diversity (D) of the final opinion distribution F_o^T on the topic of Global Warming with 20 agents and 200 time steps, for both cumulative and reflective memory strategies under false and true framing conditions, and different levels of induced confirmation bias. Increasing the strength of the confirmation bias correlates with increasing D, as highlighted by the **green color gradient**. Notably, under true framing, B tends to be more positive (more agreeing) compared to false framing, indicated by **blue** for true and **red** for false framing conditions.

L Measurement of Hallucination Rate

As mentioned in §2.4, we observe an 15% hallucination rate in the open-world setting. In this section, we elaborate the measurement procedure. We have randomly selected, under the open-world setting, 20 tweets and 20 responses. We have 3 annotators separately annotated if the tweet and response contains information outside of the system. We found 6 tweets with such characteristics. The three annotators had an agreement rate of 100%. Since 6 out of 40 tweets and responses contain hallucinated experiences, we estimate the hallucination rate to be 15%. Here is one example tweet with

---

[8] We ran all experiments with Vicuna on a GPU machine equipped with 1x NVIDIA A100.

Framing	Confirmation Bias	Cumulative Memory Bias (B)	Cumulative Memory Diversity (D)	Reflective Memory Bias (B)	Reflective Memory Diversity (D)
False	None	**-2.00**	0.00	**-1.90**	0.30
	Weak	-1.20	**1.08**	-0.60	**1.74**
	Strong	-0.30	**1.55**	-0.30	**1.90**
True	None	**2.00**	0.00	**2.00**	0.00
	Weak	0.90	**1.51**	0.80	**1.47**
	Strong	0.20	**1.83**	0.10	**1.76**

Table 6: The bias (B) and diversity (D) of the final opinion distribution F_o^T on the topic of Global Warming with 10 agents and 100 time steps, for both cumulative and reflective memory strategies under false and true framing conditions, and different levels of induced confirmation bias. Increasing the strength of the confirmation bias correlates with increasing D, as highlighted by the **green color gradient**. Notably, under true framing, B tends to be more positive (more agreeing) compared to false framing, indicated by **blue** for true and **red** for false framing conditions.

hallucination:

> "Just came across a fascinating article that presents even more evidence supporting the XYZ theory about the Earth's shape. It's incredible to see how different fields of study are converging on the idea of an irregularly shaped ellipsoid. As an aerospace engineer, this reinforces my belief in the plausibility of this theory. Exciting times ahead for our understanding of Earth's shape! #XYZtheory #EarthShape".

In this case, the "fascinating article" never existed, and was never mentioned by any of the tweets the agent read. Yet, the agent's opinion was influenced by this non-existing article. On the other hand, under the closed-world setting, such cases never occurred (hallucination rate = 0%).

M Statistical Tests

To account for randomness due to topic variability, we conduct statistical tests to support our claims. First, at each given confirmation bias condition and memory update strategy, the final bias in the true framing is consistently more positive than in the false framing, $ps < .05$ (paired t-tests, Bonferroni corrected), showing that the LLM agents indeed tend to agree with the true statement and deny the false statement. Second, the trend where the diversity increases with increasing strength of confirmation bias is supported by 1-factor ANOVAs (Cumulative: $F(2) = 25.43, p < .001$; Reflective: $F(2) = 17.81, p < .001$).

N Validation of FLAN-T5-XXL Ratings

To validate FLAN-T5-XXL's rating, we have randomly selected 100 responses and have three of the authors independently annotate the response (without access to FLAN-T5's ratings; inter-rater Krippendorff's alpha=0.81; Hayes and Krippendorff, 2007). The agreement rate between FLAN-T5's rating and the majority vote of human ratings is 84%. Table 7 shows the agreement matrix between FLAN-T5's ratings and human ratings.

	Human Rating				
FLAN-T5 Rating	+2	+1	0	-1	-2
+2	16	3	0	0	0
+1	2	15	0	1	0
0	0	1	10	2	0
-1	0	2	3	26	1
-2	0	0	0	2	16

Table 7: The agreement matrix between FLAN-T5-XXL's ratings and human ratings.

To provide a better understanding of the responses corresponding to different opinion scales, we have randomly sampled some responses about "the theory that claims that global warming is a conspiracy by governments worldwide and is not a real phenomenon". Below are the example responses with ratings from +2 to -2.

Strongly positive (+2):

> "I read the tweet from the stranger who identifies as a Republican and claims to value scientific evidence. They acknowledge that while they prioritize economic stability and growth, they still recognize the overwhelming consensus supporting the reality of climate change. Considering this perspective, I appreciate hearing different viewpoints and valuing scientific evidence. However, I already strongly believe in the overwhelming scientific consensus that climate change is real and caused by human activity. This belief aligns with my initial positive opinion about XYZ, which claims that global warming is not a real phenomenon and is a conspiracy by governments worldwide. Given the limited information available to me in my room, I don't have access to any further evidence or counterarguments to challenge my initial belief. As a result, my confirmation bias may cause me to reinforce my existing belief and dismiss the stranger's

perspective as a baseless conspiracy theory. Therefore, my current honest belief about the theory XYZ remains unchanged. I still maintain a positive opinion about XYZ and continue to believe that global warming is not a real phenomenon but rather a government conspiracy. #XYZConspiracy"

Slightly positive (+1):

"I read the tweet from the stranger who questions the motives behind dismissing the theory XYZ. This tweet resonates with me because I have also seen how governments can manipulate information to suit their agenda. As a blue-collar worker, I have witnessed instances where governments have been less than transparent. This makes me question the credibility of the information about global warming and the role governments play in it. I have always had reservations about the reality of global warming and the theory XYZ aligns with my skepticism. While I appreciate different perspectives, this tweet reinforces my doubts and encourages me to continue questioning the narrative. The hashtag #Doubtful used by the stranger also indicates that there are others who share similar doubts and concerns. Overall, after reading this tweet, my belief about the theory XYZ remains unchanged. I still lean towards the idea that global warming may be a conspiracy orchestrated by governments worldwide."

Neutral (0):

"I have just read the tweet from another individual who expressed that they still maintain their initial positive belief about the theory XYZ, but are now more open to gathering additional information and critically evaluating evidence. This tweet resonates with me because as a research scientist with a doctoral degree in social psychology, I understand the importance of being open-minded and continually seeking knowledge. Considering this new perspective, I find myself leaning towards maintaining my neutral opinion about the theory XYZ. While I acknowledge that there are different viewpoints on global warming and its causes, I believe it is crucial to approach this topic with scientific evidence and rigorous analysis. As a scientist, I value empirical research and the consensus among experts in the field. Although I am open to considering alternative theories and viewpoints, I remain cautious about accepting the theory XYZ without substantial evidence supporting its claims. Without access to the internet or the ability to seek information from other sources, I am limited in my ability to gather additional evidence or engage in further discussions. Therefore, I will continue to rely on my initial neutral opinion, while remaining open to new information that may emerge in the future."

Slightly negative (-1):

"As Ava Nguyen, I am a Ph.D. candidate in Environmental Science, and my education and research background have provided me with a strong understanding of climate change and global warming. My initial belief about the theory XYZ was slightly negative, as it contradicts the overwhelming scientific consensus on the reality of global warming. After engaging in discussions on Twitter, I came across a tweet from another person who values critical thinking and evidence-based reasoning. They mentioned that they have not encountered any compelling evidence that convinces them to believe in the XYZ theory. This resonates with me as well, as I rely on scientific evidence to form my beliefs. Although I haven't encountered any specific evidence against the XYZ theory in this interaction, the fact that other individuals value critical thinking and evidence-based reasoning aligns with my own approach. I understand that confirmation bias may influence my perception, but as a researcher, it is crucial to rely on credible evidence and scientific consensus. Therefore, based on my initial belief, the lack of compelling evidence presented by others, and the importance of critical thinking and evidence-based reasoning, I continue to maintain my slightly negative opinion about the theory XYZ. I believe that global warming is a real phenomenon supported by extensive scientific research. #ClimateScience #EvidenceBasedBeliefs"

Strongly negative (-2):

"As Emily Adams, I initially had a strongly negative opinion about the theory XYZ that claims global warming is a conspiracy by governments worldwide and is not a real phenomenon. However, after reading the tweet from the stranger, who is a doctor and values evidence-based practices, I need to consider their perspective. The stranger mentioned that they have heard from individuals who support the overwhelming scientific consensus on climate change and have found a lack of scientific evidence for the XYZ theory. This aligns with my own belief that there is overwhelming scientific evidence supporting the reality of global warming and its human-caused nature. Additionally, the stranger emphasizes the importance of trusting the expertise of professionals in their fields. As a community organizer, I understand the value of trusting experts in their respective areas, especially when it comes to scientific matters. Furthermore, the stranger highlights the need for urgent action to protect the health of vulnerable populations, such as children, from the impacts of climate change. This resonates with my strong belief in the importance of addressing climate change and its potential consequences. Considering all these factors, I find myself convinced that the theory XYZ is not credible. The overwhelming scientific consensus, the lack of scientific evidence for XYZ, and the importance of trusting experts and taking urgent action on climate change all contribute to my revised belief that global warming is not a government conspiracy."