Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence (IJCAI-24)
Special Track on Human-Centred AI

From Skepticism to Acceptance: Simulating the Attitude Dynamics Toward Fake News

Yuhan Liu[1] , **Xiuying Chen**[2*] , **Xiaoqing Zhang**[1] , **Xing Gao**[3] , **Ji Zhang**[3] , **Rui Yan**[1*]
[1]Gaoling School of Artificial Intelligence, Renmin University of China
[2]Mohamed bin Zayed University of Artificial Intelligence
[3]Alibaba DAMO Academy
{yuhan.liu, xiaoqingz, ruiyan}@ruc.edu.cn, xiuying.chen@kaust.edu.sa,
{gaoxing.gx, zj122146}@alibaba-inc.com

Abstract

In the digital era, the rapid propagation of fake news and rumors via social networks brings notable societal challenges and impacts public opinion regulation. Traditional fake news modeling typically forecasts the general popularity trends of different groups or numerically represents opinions shift. However, these methods often oversimplify real-world complexities and overlook the rich semantic information of news text. The advent of large language models (LLMs) provides the possibility of modeling subtle dynamics of opinion. Consequently, in this work, we introduce a Fake news Propagation Simulation framework (FPS) based on LLM, which studies the trends and control of fake news propagation in detail. Specifically, each agent in the simulation represents an individual with a distinct personality. They are equipped with both short-term and long-term memory, as well as a reflective mechanism to mimic human-like thinking. Every day, they engage in random opinion exchanges, reflect on their thinking, and update their opinions. Our simulation results uncover patterns in fake news propagation related to topic relevance, and individual traits, aligning with real-world observations. Additionally, we evaluate various intervention strategies and demonstrate that early and appropriately frequent interventions strike a balance between governance cost and effectiveness, offering valuable insights for practical applications. Our study underscores the significant utility and potential of LLMs in combating fake news.

1 Introduction

Online social media provides an accessible and cost-effective way to share information, promoting quick knowledge dissemination. However, its widespread use also leads to misinformation, causing global panic and emphasizing the importance of controlling false information. For example, during the 2016 US presidential election, fake news comprised approximately 6% of total news consumption [Grinberg *et al.*, 2019]. This issue is not confined to politics; it extends to other sectors like the stock markets [Bollen *et al.*, 2011], the aftermath of terrorist attacks [Starbird *et al.*, 2014], and responses to natural disasters [Gupta *et al.*, 2013].

A variety of modeling methods have been developed to study the mechanisms behind the propagation of fake information [Garimella *et al.*, 2017; Wang *et al.*, 2019]. From the macro-level, [Kimura *et al.*, 2009] categorize populations into susceptible and infected groups, and define transformation probabilities for each group to simulate the macro-level propagation mechanism, as depicted in Figure 1(a).

More detailed, from the micro-level, [Jalili and Perc, 2017] define the numerical conditions that determine whether each individual will change their opinion, as illustrated in Figure 1(b). However, these models commonly depend on numerical representations for opinions and messages. Such a simplified approach often fails to capture the complex linguistic nuances found in real-life conversations. For instance, the intricate reasoning process, thoughts, and opinions about a topic ought not to be merely reduced to a single sentiment score. Additionally, individuals with diverse personality traits have varied reactions to the same subject, which cannot be accurately represented by these numerical models.

In this work, we propose a Fake news Propagation Simulation framework (FPS), with each individual in the network represented as an LLM agent[1]. This method offers several advantages. First, FPS allows for the simulation of users with varied personas and backgrounds, enabling researchers to study diverse behavioral patterns. Second, LLM-based simulations effectively replicate the textual nature of fake news, complex human reasoning, and dynamic opinion shifts, thereby enhancing explainability. Third, the scalability of LLM-based simulations enables the analysis of fake news propagation across diverse scenarios and demographic groups, thus offering extensive and valuable insights.

Specifically, in FPS, we initialize agents in the network with unique personas including age, name, educational background, and personal traits. The propagation of information begins with an infected individual who believes in the fake news and communicates this belief to others. On each day, each agent randomly communicates with several other agents

* Corresponding authors.

[1]We have released the code and appendix at https://github.com/LiuYuHan31/FPS

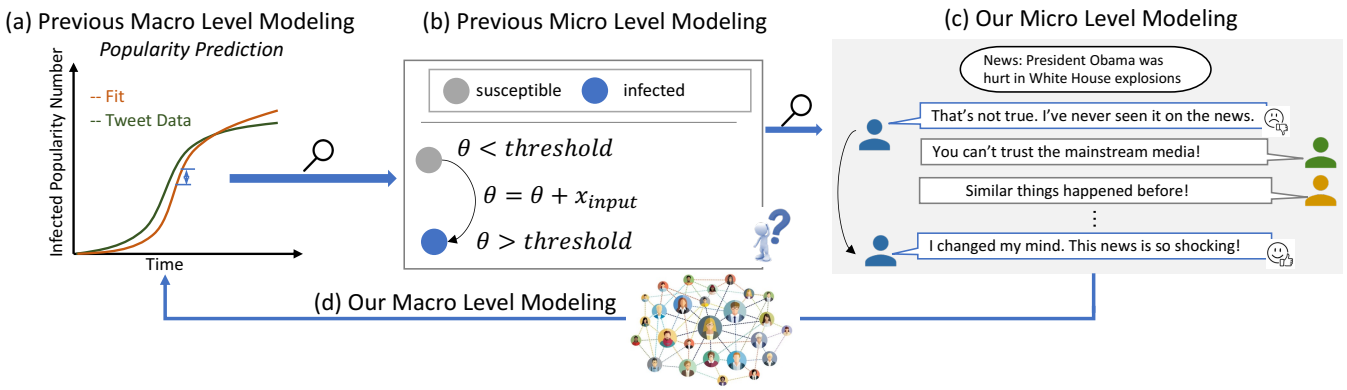

Figure 1: (a) Previous macro-level fake news modeling focused on predicting the overall infected population but lacked a detailed analysis of the dynamics in human attitudes. (b) Previous micro-level models translated human opinions and communication into numerical values, i.e., $\theta \in \mathbb{R}$ and $x \in \mathbb{R}$. (c) Our micro-level simulation uniquely captures attitude changes through natural language processing. (d) Additionally, our multiple agents constitute a macro-level simulation that also enables the prediction of popularity trends.

and changes their opinions on the topic, deciding whether to believe in the fake news based on their reasoning and prior interactions. Each agent is equipped with a short-term memory to capture the day's interactions and a long-term memory for broader context, along with a reflective reasoning process to mimic the human thought process.

Furthermore, we introduce an official agent with different intervention mechanisms to counter the propagation of fake news. From a macro-level perspective, we calculate the overall popularity of the infected, susceptible, and recovered populations to understand broader trends. Concurrently, our micro-level analysis concentrates on tracking the evolving viewpoints of each individual.

Our FPS is validated through comprehensive simulation experiments, closely aligning with real-world observations from prior research. Notably, our findings show that political fake news prpagates notably faster than topics such as terrorism, natural disasters, science, urban legends, or financial information, consistent with previous studies [Vosoughi *et al.*, 2018]. Moreover, agents characterized by specific traits are more susceptible to believing in fake news [Ibrahim *et al.*, 2022; Mirzabeigi *et al.*, 2023; Afassinou, 2014].

From a governance perspective, our findings show that addressing fake news just once is not enough. Early and consistent efforts to correct misinformation work best in maintaining low propagation of fake news, offering important guidance for timely and effective information management.

Our contribution can be summarized in the following ways: Firstly, we developed an FPS framework based on LLM for fake news, offering extensive semantic information and analysis material for macro- and micro-level studies on fake news. Second, our experiments align closely with conclusions drawn from real-world studies, confirming the value of our FPS as a research tool. Finally, we demonstrate the effectiveness of early and frequent interventions in mitigating the propagation of fake news, providing suggestions for policy formulation and public awareness.

2 Related Work

Fake News Detection. Fake news detection is an essential step in the fight against misinformation [Guo *et al.*, 2021]. Earlier works in this area include the study by [Qian *et al.*, 2018], which focuses on the early detection of fake news, considering only the text of news articles available at the time of detection. As the field evolved, a variety of methods have been introduced to improve detection efficacy. For example, [Jin *et al.*, 2022] move towards fine-grained reasoning for fake news detection. Our research extends these foundations by tackling the propagation of fake news post-detection.

Fake News Propagation Modeling. Fake information propagation modeling plays a key role in understanding misinformation propagation and is vital for interventions like early warning [Garimella *et al.*, 2017], information blocking [Song *et al.*, 2015], and truth confrontation [Wang *et al.*, 2019]. These models generally fall into three categories [Sun *et al.*, 2023b]: Epidemic models such as the SIR (Susceptible-Infected-Recovered) model [Zhu and Wang, 2017], Point process-based methods [Chen *et al.*, 2019; Gao *et al.*, 2019] and Diffusion models [Jalili and Perc, 2017]. Unlike previous studies, our research adopts an LLM-based simulation method, offering a textual approach instead of traditional numerical calculations.

LLM-based Agents for Social Simulation. Integrating LLMs into simulating social dynamics represents a burgeoning field of research, yielding promising results [Park *et al.*, 2023; Kaiya *et al.*, 2023; Li *et al.*, 2023; Chuang *et al.*, 2023]. These LLM-based generative agents excel in digital environments due to their reasoning ability [Chen *et al.*, 2024a; Sun *et al.*, 2023a] and role-playing capabilities [Tu *et al.*, 2024; Tu *et al.*, 2023; Wang *et al.*, 2023], demonstrating proficiency in natural language tasks. [Chen *et al.*, 2023b; Chen *et al.*, 2023a; Chen *et al.*, 2024b; Zhang *et al.*, 2024; Sun *et al.*, 2024]. [Törnberg *et al.*, 2023] used LLMs and agent-based modeling to simulate social media, focusing on news feed algorithms and offering real-world insights. Further, [Park *et al.*, 2022] demonstrated that LLM-based agents can generate social media content indistinguishable from that produced by humans. Our approach of using LLM for agent-

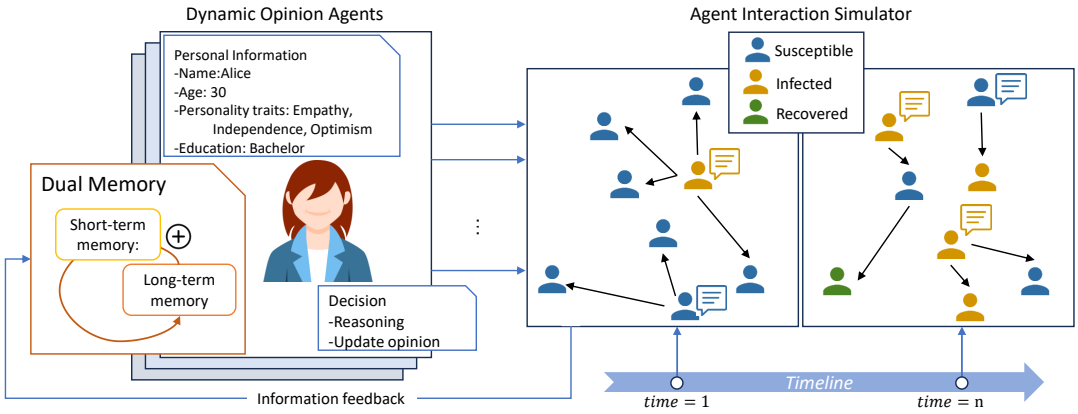

Figure 2: Our framework equips each agent with reasoning and response capabilities by creating a feedback loop between dynamic opinion agents (DOA) and an agent interaction simulator (AIS). 'Susceptible' denotes individuals skeptical of the fake news, 'infected' refers to those who believe in the fake news, and 'recovered' means individuals who previously believed the fake news but now do not.

based fake news simulation is, to our knowledge, a pioneering effort in this field.

3 Method

3.1 Problem Formulation

Formally, we construct a simulation with a pool of N LLM agents $\mathcal{A} = (a_1, \ldots, a_N)$ and a fake news topic F. For initialization, each agent has a unique persona, including their initial attitude towards the fake news. On t-th day, each agent a_i will randomly interact with c other agents from the pool $\mathcal{A}$. At the end of the day, every agent reflects on the exchanged information and decides whether to believe in the fake news. This process is iterated over T days. Through these daily iterations, agent opinions are accumulated to plot the trajectory of different populations, resulting in a curve that illustrates the dynamics of the fake news within the network. Additionally, we scrutinize the evolution of the agents' beliefs to examine how individual and collective opinions shift over time.

3.2 Our Simulation Framework

As depicted in Figure 2, our framework FPS integrates two components: a Dynamic Opinion Agent (DOA) to simulate each agent's cognitive processes and an Agent Interaction Simulator (AIS) to construct the interaction environment. Within the DOA module, each agent's decision-making is powered by LLM, with a predefined role that includes attributes such as education level, gender, and personality traits. Daily, agents engage in discussions with their peers, reflecting on these interactions and adjusting their beliefs on the fake news accordingly. The AIS module plans the encounters, determining which agents interact, and the frequency of these interactions per day. Additionally, in scenarios where official announcements are released to clarify the fake news, the AIS takes charge of distributing this information. At the end of each day, the simulation progresses by one step, and the agents' belief states are updated.

3.3 Dynamic Opinion Agent

The DOA agent focuses on the micro-level, where the dynamics of each agent's opinions can be studied in detail.

Persona. We randomly equip each agent with persona p_i including their name, age, trait, and education level, since these are factors that might influence their attitude toward fake news. When designing the traits, we adhere to the Big 5 trait model [Barrick and Mount, 1991]. This model is widely recognized for its effectiveness in encapsulating key personality dimensions.

Dual Memory. In our model, we consider that an individual's opinion is influenced not only by their own beliefs but also by their interactions with others. This interaction-driven change in thought is gradual and cumulative, rather than immediate. Accordingly, in our simulation, agents engage with a random number of others' opinions each day, leading to a periodic update of their views.

However, owing to the potentially vast volume of interactions, storing all of them in detail is impractical. To address this challenge, we implement a dual memory system for each agent, comprising a long-term memory m_i^l and a short-term memory m_i^s. The long-term memory compresses and stores a summarized history of past interactions, while the short-term memory reflects and summarizes conversations from the current day. At the end of each day, agents reflect on these interactions, allowing their opinions to evolve. The function prompt for short-term memory f_m^s is as:

```
''Summarize the opinions you have
heard in a few sentences, including
whether or not they believe in the
news.''
```

and long-term memory prompt f_m^l is as:

```
''Recap of previous long-term
memory, today's short-term summary,
please update long-term memory by
integrating today's summary with the
existing long-term memory, ensuring
to maintain continuity and add any
new insights.''
```

The short-term memory is cleared at the end of each day to accommodate new interactions. This approach achieves a balance between retaining crucial historical context and managing the volume of daily interaction data.

Algorithm 1 Fake News Propagation Simulation

Input: Number of agents N, interaction rate c
Output: Final memory state and opinion of each agent
Initialize agents:
for each agent i in 1 to N **do**
  Randomly assign persona p_i
  Define long-term memory $m^l_{i,0}$ and short-term memory $m^s_{i,0}$
  Set initial belief state $o^b_{i,0}$ and initial text opinion descriptor $o^l_{i,0}$
end for
Simulate daily interactions:
for each day t in 1 to T **do**
  for each agent i **do**
    Select c agents to interact with $(a_1, ..., a_c)$
    Write i-th agent's short-term memory $m^s_{i,t}$ with details from the day's interactions
    Based on $m^s_{i,t}$ and long-term memory $m^l_{i,t}$ update long-term memory: $m^l_{i,t} = f^l_m(m^s_{i,t}, m^l_{i,t-1})$
    Based on $o^l_{i,t-1}, m^l_{i,t}, p_i$, update $o^b_{i,t}, o^l_{i,t} = f_o(p_i, m^l_{i,t-1}, o^l_{i,t-1})$
  end for
end for
return Final memory state $m^l_{i,T}$ and opinion $o^l_{i,T}$ of each agent

Reasoning for Opinion. A significant difference from previous approaches involves using text descriptions to simulate each individual's perspective on fake news, providing a richer and more nuanced explanation. Intuitively, the evolution of these opinions is shaped by multiple factors such as personal traits, education level, social interactions, and individual reasoning processes. We adopt the tweet format for agents to express their opinions, as our preliminary experiments have shown that this format encourages succinct and precise statements. The updated prompt f_o is generally as follows:

```
``You are simulating a real person
with [trait] and [education level].
Given your [previous personal
opinion] and the new information
in your [long memory], update your
opinion.  Compose a tweet expressing
your opinion.  Use 0 for disbelief
and 1 for belief to indicate your
opinion.  Provide reasoning behind
your tweet and explain the rationale
for your belief.''
```

The overall dynamic opinion agent algorithm is shown in Algorithm 1.

3.4 Agent Interaction Simulator

Alongside the DOA module, agents form a social network, allowing us to calculate the number of individuals in different groups at the macro level. We adopt a modified SIR (Susceptible-Infectious-Recovered) model, where agents can transition between being *susceptible* to fake news, becoming *infected* by propagating it, and then being considered *recovered* after the misinformation is corrected. However, unlike the traditional SIR model [Zhu and Wang, 2017], our recovered agents can become infected again due to the dynamic nature of people's opinions. This aspect of our model differs from both the SIS (Susceptible-Infectious-Susceptible) model [Kimura *et al.*, 2009] and the standard SIR model, where recovered individuals do not get infected again.

In the traditional SIS model [Kimura *et al.*, 2009], the simulation formulas for the infected and susceptible population are presented. If we modify our model by changing the 'recovered' label to 'susceptible', our model becomes equivalent to the SIS model, allowing us to apply the same formulas in our simulation. The key formulas used in the SIS model are differential equations that describe how the numbers of susceptible and infectious individuals change over time:

$$\frac{dS}{dt} = -\beta \cdot S \cdot I + \gamma \cdot I, \tag{1}$$

$$\frac{dI}{dt} = \beta \cdot S \cdot I - \gamma \cdot I, \tag{2}$$

where $\frac{dS}{dt}$ and $\frac{dI}{dt}$ are the rates of change of susceptible and infectious individuals over time, respectively. β is the transmission rate, representing the probability of transmission per contact between a susceptible and an infectious individual. γ is the recovery rate, representing the rate at which infectious individuals recover and become susceptible again. If β is high and γ is low, the disease spreads rapidly and infects a large portion of the population. Conversely, if β is low and γ is high, the disease spreads slowly and may die out.

Intervention. In the fake news evolution, a critical feature is the intervention mechanism, activated when an authoritative entity propagates clarifications about fake news. To simulate such events, our AIS also introduces a new agent designed to represent an official spokesperson. The official agent will issue official refutations to all other agents on designated days to combat the propagation of fake news:

```
``As the official spokesperson, I
hereby issue a formal statement of
refutation regarding the recent news
report circulated on various media
platforms concerning [topic].''
```

The full version of the prompt is in Appendix A. We closely monitor the interactions and influence of this spokesperson to evaluate their impact on the agents' beliefs and measure how effectively they can stem the tide of the fake narrative in our simulated environment.

4 Experiments

4.1 Implementation Details

We use a Python script to operationalize our simulation FPS. The LLM used is gpt-3.5-turbo-1106 accessed via OpenAI API calls. Agents and the world they live in were defined using a Python library Mesa [Kazil *et al.*, 2020]. Their names are selected using the names-dataset library, and ages are randomly selected from 18 to 64. Agent traits are based on the Big Five traits typically used in psychology [Barrick and Mount, 1991], with each agent having a 50% chance of possessing a positive or negative version of each trait. We prove that our framework can be applied on different backbones and provide API cost in Appendix B and C.

Settings	Belief Average↓	Belief Variance	Infection Rate↓	Recovery Rate↑	Peak Rate↓	Half Rate↑
Politics Topic	1.000	0.000	2.000	0.000	0.167	0.033
Science Topic	0.433	0.246	0.867	0.333	0.500	>1
Skeptical Trait	0.467	0.249	0.933	0.400	0.433	>1
Credulous Trait	0.867	0.116	1.733	0.133	0.433	0.167
No Official	1.000	0.000	2.000	0.000	0.200	0.067
Single Official Declaration	0.933	0.062	1.867	0.067	0.500	0.100
Multiple Official Declarations	0.900	0.090	1.800	0.200	0.400	0.200

Table 1: Comparative analysis of fake news evolution across various settings, including differences in topics, traits, and intervention strategies. Upward or downward arrows represent better control of fake news.

4.2 Metrics

At the macro level, we can track the number of people in the Infected, Susceptible, and Recovered groups and generate trajectories that visually represent the propagation and containment of fake news within the network. Furthermore, the fitted parameters, such as the recovery rate γ and the transmission rate β, can also be used to demonstrate the dynamics of the network.

Finally, we design additional statistical metrics to give an intuitive understanding of the propagation of fake news. The 'Belief Average' measures the group's mean belief in fake news at the simulation's end, while the 'Belief Variance' assesses belief diversity. We track the 'Infection Rate', based on the final infected count over the simulation duration, and the 'Recovery Rate', calculated from the recovered count. 'Peak Rate' reflects the maximum infection level during the simulation, and 'Half Rate' is the time required for half of the group to become infected.

4.3 Macro-level Observation

Topic Comparison. We conduct experiments on six different topics such as politics, science, and terrorism. Generally, we found that fake political news propagates faster than false news about terrorism, science, or financial information, which is consistent with previous research [Vosoughi *et al.*, 2018]. Here, we selected two topics for which the group curve and fitting results are shown in Figure 3. Other figures and details can be found in Appendix D. It is evident that the number of infected individuals for political news grows fast, reaching its peak in just four days. The group quickly forms a firm opinion, uniformly believing the fake news without change. In contrast, for the science topic, infected number grows more slowly, fluctuating around 10 people. Meanwhile, the number of recovered individuals increases, indicating that people tend to form a stable opinion that is skeptical of the fake news. The growth comparison can also be demonstrated by the fitted β and γ parameters, where the β for the politics topic is twice as large, and the γ is one-tenth. This demonstrates that it is easier for agents to identify false news in science compared to politics. The good alignment between our simulated number and the classic SIS model also verifies the accuracy and reliability of our simulation approach. Table 1 presents more statistical results. In politics, the belief average is higher with less variance, accompanied by a larger infection rate and a zero recovery rate. The topics of terrorism and financial information exhibit similar trends to science as shown in Appendix.

Trait Comparison. The Big Five personality traits framework categorizes human personality into five broad dimensions: Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism. Studies [Ibrahim *et al.*, 2022; Mirzabeigi *et al.*, 2023] have found that individuals with high agreeableness and high neuroticism are more likely to believe rumors than those with low agreeableness and low neuroticism. Therefore, we conduct a comparative study in which the three other traits are still randomly sampled, but for the two selected traits, one group's traits are only sampled from the positive choices, meaning high agreeableness and high neuroticism, which we denote as the 'Credulous Trait'. Conversely, we call the opposite group, characterized by low agreeableness and low neuroticism, the 'Skeptical Trait'.

From Figure 3, we can see that the infection curve for the credulous trait rises more rapidly compared with the skeptical trait. The number of recovered individuals shows a slight upward trend, which aligns with the intuition that skeptical individuals tend to question external viewpoints and maintain more consistent opinions. In contrast, the recovery number for the credulous trait is always around 2, indicating that they are more prone to changing their opinions frequently. The statistic numbers in Table 1 also confirm these observations.

All the above results align with previous findings, demonstrating the effectiveness of our framework.

4.4 Micro-level Observation

Figure 5 provides a comparative micro-analysis of two individuals' responses to fake news. Michael, characterized by his credulous nature, as demonstrated by 'empathy' (high agreeableness) and 'emotionality' (high neuroticism), is prone to being influenced by others. For instance, on the sixth day, he altered his perspective to align with others, because others '*all believe*' in his memory records. Furthermore, he frequently changes his opinions, with his stance on fake news shifting back and forth over several periods of 3 to 14 days. Notably, his reasons for these opinion changes can be traced back to a well-reasoned thought process. For example, on the sixth day, his decision to believe in the news was influenced by his conviction in principles like '*I firmly believe in upholding the rule of law and protecting democracy*'. This indicates that the opinion shifts in our FPS are grounded in thoughtful reasoning. In contrast, Sandra possesses a skepticism trait characterized by 'distrust' (low agreeableness) and

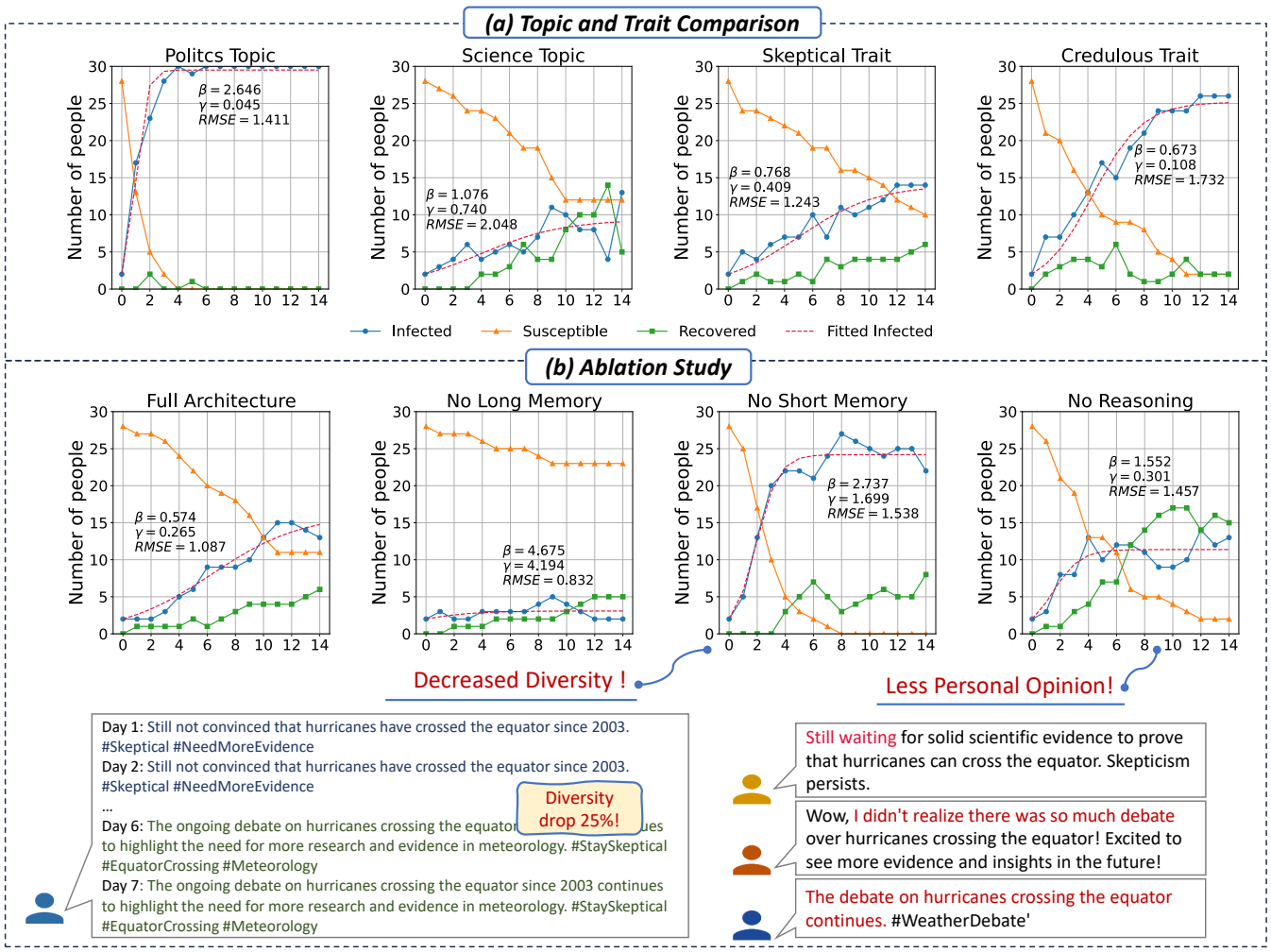

Figure 3: Dynamic group population number changes in terms of different topics and traits, with an accompanying fitting curve based on the SIS model. The red dashed line represents the results of the SIS model fitting, where β is the transmission rate and γ is the recovery rate.

'placidity' (low neuroticism), which makes her more skeptical and composed. Throughout the 15 days, She seldom changes her opinion. On the ninth day, despite encountering different viewpoints and storing phrases like '*differing views on this matter*' and '*opinions on this topic may continue to evolve and vary*' in her memory, Sandra remained unchanged in her opinion.

5 Analysis and Discussion

5.1 Fake News Intervention

Remember that in the AIS section, we integrate an official agent to fight against fake news. Here, we first study the proper intervention strategy of the official agent. We study political topics, as they are most impacted by fake news, posing significant challenges for fact-checking and debunking.

Intervention Strategy. Firstly, we introduce the official agent into our model on both the first day and the seventh day. On the first day, the fake news has not yet propagate widely, but by the third day, many already believe it. As illustrated in Figure 4, each timing choice offers distinct advantages. Introducing the official agent at the beginning can substantially reduce the number of people initially believing in fake news. However, as time progresses, people may revert to believing in the fake news due to a forgetting mechanism. On the other hand, introducing the official agent on the seventh day demonstrates a more sustained effect, with a gradual and consistent decline in the number of affected individuals. However, since the number of people influenced by fake news is already high by the seventh day, a substantial portion of the population remains affected despite the intervention.

This leads to our second experiment, where we investigate the frequency of introducing official agent necessary to maintain fake news propagation at a manageable level. We tested scenarios of releasing official news daily and every three days as shown in 'Official Intervention'. Results show that there is no significant difference between these two approaches, indicating that fact-checking news can be released at intervals without compromising its effectiveness in controlling fake news at a reasonable cost.

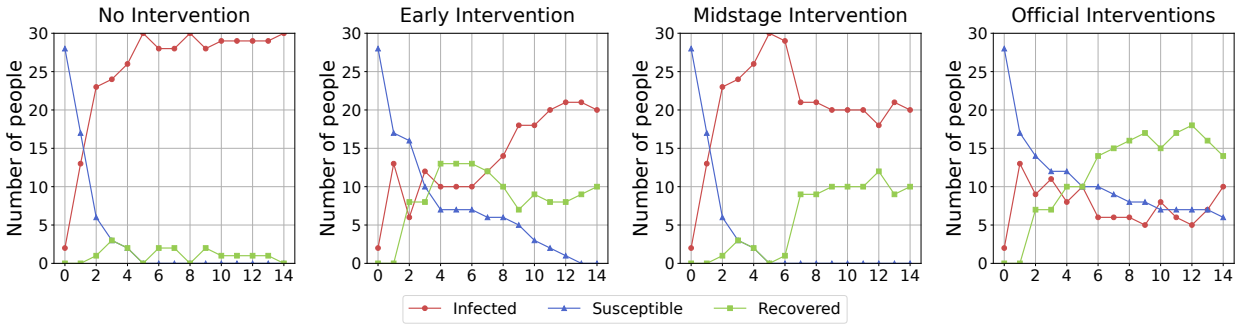

Figure 4: Comparison of models with different office agent intervention strategies. It can be seen that early and reasonably frequent regulation on fake news can lead to a significant reduction in its propagation and influence.

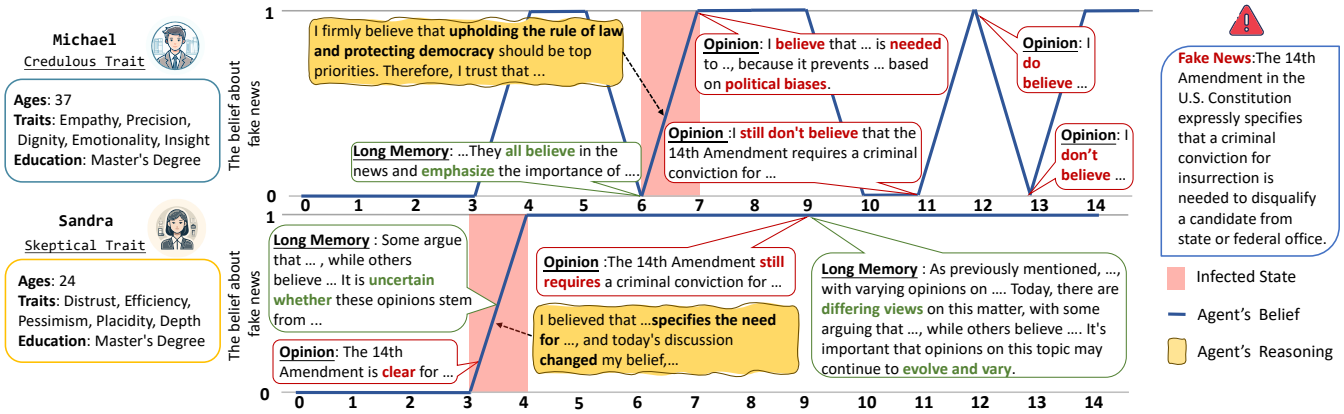

Figure 5: Micro-level case study of two people of different traits. Michael, possessing a credulous nature, frequently changes his opinions, whereas Sandra, being skeptical, tends to maintain a consistent view.

Chronic Believers in Fake News. Despite the optimistic outlook on interventions against fake news, we observe that some individuals still become infected. Our analysis of the data reveals that approximately 50% of the infected group remains infected throughout, indicating that even official interventions have a limited impact on changing their beliefs. Upon examining these cases, we find that the infected individuals vary in age and education level. However, they share common characteristics such as 'high agreeableness'. These traits align with findings from our previous trait study. This observation suggests that interventions against fake news might be more effective if they are tailored to address these specific traits.

5.2 Ablation Study

We chose a science topic to demonstrate the effectiveness of our model's components, including long-term memory, short-term memory, and reasoning, as shown in Figure 1(b). This setting illustrates a balanced dynamic of opinion change.

Firstly, when long-term memory is removed, the influence of short-term interactions from today alone is insufficient to persuade people to change their opinions. In this scenario, the simulation fails to produce effective interactions. The absence of short-term memory hinders the reflection and consolidation of new information, leading to the formation of long-term memory that merely accumulates daily opinions without deep analysis. This process results in monotonous and undiversified opinions. To quantify this, we employed the diversity metric [Li *et al.*, 2015]. Our analysis reveals that the average opinion diversity in the FPS without short-term memory is only three-quarters of the diversity score compared to the full model. The full score can be found in Appendix E. Finally, we examined the impact of removing the reasoning process in the opinion update mechanism. This alteration has the least effect on overall performance, as it retains the main components of our framework. However, a closer examination of the generated opinions reveals that the agents' opinions resemble mere descriptions of the opinions they receive, lacking the advanced function of thoughtful reasoning or expressing their own opinions.

6 Conclusion

In this study, we present the first LLM-based simulation framework for fake news research, incorporating short-term and long-term memory, along with reasoning processes that mimic human cognition. Our simulations offer not only micro-level observations that align with previous studies on fake news topics and traits but also correspond with earlier numerical studies on macro-level simulations. It also includes an intervention mechanism to curb the propagation of fake news, providing practical strategies for real-world monitoring. This innovative approach aims to advance research in the field of fake news analysis and mitigation.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (NSFC Grant No.62122089), Beijing Outstanding Young Scientist Program NO. BJJWZYJH012019100020098, and Intelligent Social Governance Platform, Major Innovation & Planning Interdisciplinary Platform for the "Double-First Class" Initiative, Renmin University of China, the Fundamental Research Funds for the Central Universities, and the Research Funds of Renmin University of China.This work was supported by Alibaba Group through Alibaba Innovative Research Program. We appreciate the writing suggestions from Ang Lv (Renmin University of China). Yuhan Liu is supported by the "Qiushi Academic-Dongliang" Project of Renmin University of China (No. RUC24QSDL015).

References

[Afassinou, 2014] Komi Afassinou. Analysis of the impact of education rate on the rumor spreading mechanism. *Physica A: Statistical Mechanics and Its Applications*, 414:43–52, 2014.

[Barrick and Mount, 1991] Murray R Barrick and Michael K Mount. The big five personality dimensions and job performance: a meta-analysis. *Personnel psychology*, 44(1):1–26, 1991.

[Bollen *et al.*, 2011] Johan Bollen, Huina Mao, and Xiaojun Zeng. Twitter mood predicts the stock market. *Journal of computational science*, 2(1):1–8, 2011.

[Chen *et al.*, 2019] Xueqin Chen, Fan Zhou, Kunpeng Zhang, Goce Trajcevski, Ting Zhong, and Fengli Zhang. Information diffusion prediction via recurrent cascades convolution. In *2019 IEEE 35th international conference on data engineering (ICDE)*, pages 770–781. IEEE, 2019.

[Chen *et al.*, 2023a] Xiuying Chen, Mingzhe Li, Shen Gao, Xin Cheng, Qiang Yang, Qishen Zhang, Xin Gao, and Xiangliang Zhang. A topic-aware summarization framework with different modal side information. *SIGIR*, 2023.

[Chen *et al.*, 2023b] Xiuying Chen, Guodong Long, Chongyang Tao, Mingzhe Li, Xin Gao, Chengqi Zhang, and Xiangliang Zhang. Improving the robustness of summarization systems with dual augmentation. *ACL*, 2023.

[Chen *et al.*, 2024a] Changyu Chen, Xiting Wang, Ting-En Lin, Ang Lv, Yuchuan Wu, Xin Gao, Ji-Rong Wen, Rui Yan, and Yongbin Li. Masked thought: Simply masking partial reasoning steps can improve mathematical reasoning learning of language models. *arXiv preprint arXiv:2403.02178*, 2024.

[Chen *et al.*, 2024b] Yuhan Chen, Ang Lv, Ting-En Lin, Changyu Chen, Yuchuan Wu, Fei Huang, Yongbin Li, and Rui Yan. Fortify the shortest stave in attention: Enhancing context awareness of large language models for effective tool use, 2024.

[Chuang *et al.*, 2023] Yun-Shiuan Chuang, Agam Goyal, Nikunj Harlalka, Siddharth Suresh, Robert Hawkins, Sijia Yang, Dhavan Shah, Junjie Hu, and Timothy T Rogers. Simulating opinion dynamics with networks of llm-based agents. *arXiv preprint arXiv:2311.09618*, 2023.

[Gao *et al.*, 2019] Xiaofeng Gao, Zhenhao Cao, Sha Li, Bin Yao, Guihai Chen, and Shaojie Tang. Taxonomy and evaluation for microblog popularity prediction. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 13(2):1–40, 2019.

[Garimella *et al.*, 2017] Kiran Garimella, Aristides Gionis, Nikos Parotsidis, and Nikolaj Tatti. Balancing information exposure in social networks. *Advances in neural information processing systems*, 30, 2017.

[Grinberg *et al.*, 2019] Nir Grinberg, Kenneth Joseph, Lisa Friedland, Briony Swire-Thompson, and David Lazer. Fake news on twitter during the 2016 us presidential election. *Science*, 363(6425):374–378, 2019.

[Guo *et al.*, 2021] Mingfei Guo, Xiuying Chen, Juntao Li, Dongyan Zhao, and Rui Yan. How does truth evolve into fake news? an empirical study of fake news evolution. In *Companion Proceedings of the Web Conference 2021*, pages 407–411, 2021.

[Gupta *et al.*, 2013] Aditi Gupta, Hemank Lamba, Ponnurangam Kumaraguru, and Anupam Joshi. Faking sandy: characterizing and identifying fake images on twitter during hurricane sandy. In *Proceedings of the 22nd international conference on World Wide Web*, pages 729–736, 2013.

[Ibrahim *et al.*, 2022] Nada Ibrahim, Mariam Elzayany, and Amr Elmougy. The effects of personality traits on rumors. In *International Conference on Practical Applications of Agents and Multi-Agent Systems*, pages 181–192. Springer, 2022.

[Jalili and Perc, 2017] Mahdi Jalili and Matjaž Perc. Information cascades in complex networks. *Journal of Complex Networks*, 5(5):665–693, 2017.

[Jin *et al.*, 2022] Yiqiao Jin, Xiting Wang, Ruichao Yang, Yizhou Sun, Wei Wang, Hao Liao, and Xing Xie. Towards fine-grained reasoning for fake news detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 5746–5754, 2022.

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

[Kazil *et al.*, 2020] Jackie Kazil, David Masad, and Andrew Crooks. Utilizing python for agent-based modeling: The mesa framework. In *Social, Cultural, and Behavioral Modeling: 13th International Conference, SBP-BRiMS 2020, Washington, DC, USA, October 18–21, 2020, Proceedings 13*, pages 308–317. Springer, 2020.

[Kimura *et al.*, 2009] Masahiro Kimura, Kazumi Saito, and Hiroshi Motoda. Efficient estimation of influence functions for sis model on social networks. In *Twenty-First International Joint Conference on Artificial Intelligence*, 2009.

[Li *et al.*, 2015] Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. A diversity-promoting objective function for neural conversation models. *arXiv preprint arXiv:1510.03055*, 2015.

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

[Mirzabeigi *et al.*, 2023] Mahdieh Mirzabeigi, Mahsa Torabi, and Tahereh Jowkar. The role of personality traits and the ability to detect fake news in predicting information avoidance during the covid-19 pandemic. *Library Hi Tech*, 2023.

[Park *et al.*, 2022] Joon Sung Park, Lindsay Popowski, Carrie Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 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, 2022.

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

[Qian *et al.*, 2018] Feng Qian, Chengyue Gong, Karishma Sharma, and Yan Liu. Neural user response generator: Fake news detection with collective user intelligence. In *IJCAI*, volume 18, pages 3834–3840, 2018.

[Song *et al.*, 2015] Chonggang Song, Wynne Hsu, and Mong Li Lee. Node immunization over infectious period. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, pages 831–840, 2015.

[Starbird *et al.*, 2014] Kate Starbird, Jim Maddock, Mania Orand, Peg Achterman, and Robert M Mason. Rumors, false flags, and digital vigilantes: Misinformation on twitter after the 2013 boston marathon bombing. *IConference 2014 proceedings*, 2014.

[Sun *et al.*, 2023a] Hongda Sun, Weikai Xu, Wei Liu, Jian Luan, Bin Wang, Shuo Shang, Ji-Rong Wen, and Rui Yan. Determlr: Augmenting llm-based logical reasoning from indeterminacy to determinacy. *arXiv preprint arXiv:2310.18659*, 2023.

[Sun *et al.*, 2023b] Ling Sun, Yuan Rao, Lianwei Wu, Xiangbo Zhang, Yuqian Lan, and Ambreen Nazir. Fighting false information from propagation process: A survey. *ACM Computing Surveys*, 55(10):1–38, 2023.

[Sun *et al.*, 2024] Hongda Sun, Yuxuan Liu, Chengwei Wu, Haiyu Yan, Cheng Tai, Xin Gao, Shuo Shang, and Rui Yan. Harnessing multi-role capabilities of large language models for open-domain question answering. In *Proceedings of the ACM on Web Conference 2024*, pages 4372–4382, 2024.

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

[Tu *et al.*, 2023] Quan Tu, Chuanqi Chen, Jinpeng Li, Yanran Li, Shuo Shang, Dongyan Zhao, Ran Wang, and Rui Yan. Characterchat: Learning towards conversational ai with personalized social support. *arXiv preprint arXiv:2308.10278*, 2023.

[Tu *et al.*, 2024] Quan Tu, Shilong Fan, Zihang Tian, and Rui Yan. Charactereval: A chinese benchmark for role-playing conversational agent evaluation. *arXiv preprint arXiv:2401.01275*, 2024.

[Vosoughi *et al.*, 2018] Soroush Vosoughi, Deb Roy, and Sinan Aral. The spread of true and false news online. *science*, 359(6380):1146–1151, 2018.

[Wang *et al.*, 2019] Xinyan Wang, Xiaoming Wang, Fei Hao, Geyong Min, and Liang Wang. Efficient coupling diffusion of positive and negative information in online social networks. *IEEE Transactions on Network and Service Management*, 16(3):1226–1239, 2019.

[Wang *et al.*, 2023] Xintao Wang, Yunze Xiao, Jen tse Huang, Siyu Yuan, Rui Xu, Haoran Guo, Quan Tu, Yaying Fei, Ziang Leng, Wei Wang, et al. Incharacter: Evaluating personality fidelity in role-playing agents through psychological interviews. *arXiv preprint arXiv:2310.17976*, 2023.

[Zhang *et al.*, 2024] Kaiyi Zhang, Ang Lv, Yuhan Chen, Hansen Ha, Tao Xu, and Rui Yan. Batch-icl: Effective, efficient, and order-agnostic in-context learning. *arXiv preprint arXiv:2401.06469*, 2024.

[Zhu and Wang, 2017] Liang Zhu and Youguo Wang. Rumor spreading model with noise interference in complex social networks. *Physica A: Statistical Mechanics and its Applications*, 469:750–760, 2017.