TrendSim: Simulating Trending Topics in Social Media Under Poisoning Attacks with LLM-based Multi-agent System

Zeyu Zhang[1], Jianxun Lian[2], Chen Ma[1], Yaning Qu[1], Ye Luo[1], Lei Wang[1], Rui Li[1], Xu Chen[1*], Yankai Lin[1], Le Wu[3], Xing Xie[2], Ji-Rong Wen[1],

[1]Renmin University of China, [2]Microsoft Research Asia,
[3]Hefei University of Technology
{zeyuzhang,xu.chen}@ruc.edu.cn

Abstract

Trending topics have become a significant part of modern social media, attracting users to participate in discussions of breaking events. However, they also bring in a new channel for poisoning attacks, resulting in negative impacts on society. Therefore, it is urgent to study this critical problem and develop effective strategies for defense. In this paper, we propose TrendSim, an LLM-based multi-agent system to simulate trending topics in social media under poisoning attacks. Specifically, we create a simulation environment for trending topics that incorporates a time-aware interaction mechanism, centralized message dissemination, and an interactive system. Moreover, we develop LLM-based human-like agents to simulate users in social media, and propose prototype-based attackers to replicate poisoning attacks. Besides, we evaluate TrendSim from multiple aspects to validate its effectiveness. Based on TrendSim, we conduct simulation experiments to study four critical problems about poisoning attacks on trending topics. In order to benefit the research community, we release our project at https://github.com/nuster1128/TrendSim.

1 Introduction

Trending topics have become a significant part of modern social media platforms in recent years, which refer to the topics that draw public attention and widespread discussions within a short period, such as that in *Weibo Hot Searches*[1] and *Twitter Trends*[2]. Each trending topic typically includes a headline, a description, and numerous comments. Compared with conventional social media contents, trending topics always emerge explosively, and are displayed in a highlight section to ensure users' engagement in current discussions. However, these factors also amplify the influence of poisoning attacks on users in social media platforms. Such attacks often manipulate or distort information in the comments on trending topics to mislead users and spread misinformation. They can result in numerous negative impacts, such as misguiding facts, provoking conflicts, and even destroying social trusts, which are harmful to our society. However, this critical problem remains inadequately studied yet.

Recently, large language models (LLMs) have exhibited human-like capabilities (Zhao et al., 2023; Wang et al., 2023a), and several studies have proposed to design LLM-based human-like agents to conduct social simulations (Gao et al., 2023a,b; Kovač et al., 2023). By analyzing their results, researchers can draw deep insights into human behaviors, and devise effective policies for social benefit (Hua et al., 2023). However, previous simulation frameworks have limitations that make them unsuitable for simulating trending topics. First of all, most frameworks employ round-based interactions without taking time into consideration (Wang et al., 2023b), despite the fact that trending topics are highly time-sensitive. Second, most social simulations are designed for peer-to-peer interactions where agents primarily communicate with their neighbors (Gao et al., 2023b). However, due to the individual section of trending topics, their message dissemination should be centralized like a hub. Moreover, previous methods seldom focus on dynamic psychological conditions during simulations, and overlook the problem of poisoning attacks in trending topics.

To address these limitations, in this paper, we propose an LLM-based multi-agent system, named **TrendSim**, to simulate trending topics in social media under poisoning attacks. Specifically, our framework designs a time-aware interaction mechanism and centralized message dissemination to adapt to the trending topic scenario, and implement details of a multi-agent interactive system. Besides, we design LLM-based human-like agents with a

[*] Corresponding author.

[1] https://s.weibo.com/top/summary

[2] https://twitter.com/explore/tabs/trending

Findings of the Association for Computational Linguistics: NAACL 2025, pages 2930–2949
April 29 - May 4, 2025 ©2025 Association for Computational Linguistics

perception, a memory, and an action module, in order to simulate user behaviors and reflect their psychological conditions. Moreover, we create prototype-based attackers with different targets to generate poisoning attacks. We conduct extensive evaluations to verify the effectiveness of our simulation framework from multiple aspects. Based on TrendSim, we study four critical problems of the poisoning attacks on trending topics in social media, analyzing the results and providing suggestions for defense. Our work is the first one that focuses on the poisoning attack problem in trending topics with LLM-based social simulations.

However, as an initial study in this new area, we should emphasize some facts in our work. First, because the implementation of trending topics varies across social media platforms, our work abstracts a common and reasonable implementation, based on certain assumptions. Second, the existence of numerous invisible factors in the real world makes it impractical to simulate all details with complete accuracy. Therefore, our research aims to provide an interpretable simulation process and deliver evolutionary conclusions under reasonable assumptions, rather than replicating every detail of reality.

Our contributions are summarized as follows:

- We propose an LLM-based multi-agent system, named TrendSim, to simulate trending topics in social media under poisoning attacks. We design the time-aware interaction mechanism, centralized message dissemination, and interactive multi-agent system to model trending topics in social media.

- We develop LLM-based human-like agents with a perception, a memory, and an action module to simulate users in social media platforms. We create prototype-based attackers to generate various poisoning attacks in our simulation.

- We conduct extensive evaluations of our simulation framework. Based on TrendSim, we study four critical problems of poisoning attacks on trending topics in social media.

The rest of our paper is organized as follows. First, we present related works in Section 2. Then, we demonstrate the details of TrendSim in Section 3, and conduct evaluations in Section 4. After that, we conduct simulation experiments based on TrendSim in Section 5. Finally, we draw conclusions in Section 6, and further discuss limitations and ethical impacts at the end of this paper.

2 Related Works

2.1 Poisoning Attack in Social Media

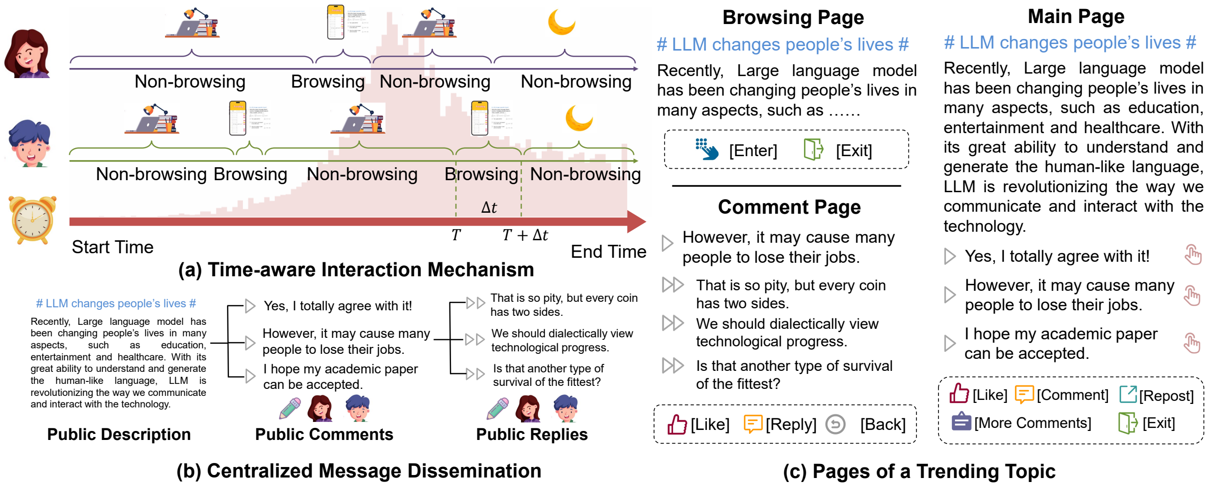
In recent years, poisoning attacks on social media platforms have gradually attracted widespread attention (Khurana et al., 2019). It has been shown that social media serves as a primary way for spreading scams and malware, with numerous poisoning attacks occurring on social media platforms (Kunwar and Sharma, 2016). Several previous studies have investigated the intentions and characteristics of poisoning attackers in social media (Aïmeur et al., 2019; Briscoe et al., 2014). Their behaviors typically involve posting offensive comments and obscene images intended to propagate hate and perpetrate cyberbullying (Chinivar et al., 2022). They also find that poisoning attacks are more inclined to focus on human vulnerabilities (Aïmeur et al., 2019).

Most previous studies focus on conventional contents in social networks, and pay less attention to trending topics in modern social media platforms. However, the poisoning attack on trending topics has become a critical problem, posing a substantial threat to our social environment, which demands wider attention from researchers.

2.2 LLM-based Multi-agent Social Simulation

Large language models have shown promising capabilities for building autonomous agents, attributed to their excellent language understanding and generation capabilities (Ouyang et al., 2022; Park et al., 2023; Wang et al., 2023b). These agents are typically equipped with extensive modules based on LLMs to perform complex tasks (Xi et al., 2023; Shinn et al., 2023; Zhu et al., 2023; Wang et al., 2023c; Qin et al., 2023). Recently studies have proposed to utilize LLM-based agents for social simulations, in order to model human behaviors and interactions in different scenarios (Wang et al., 2023b; Park et al., 2023; Gao et al., 2023b,a). For instance, S^3 (Gao et al., 2023b) simulates the emergence of social networks and phenomena like information diffusion through LLM-based agent interactions. RecAgent (Wang et al., 2023b) proposes the simulation of user behaviors in the domain of recommender systems. Generative Agents (Park et al., 2023) and AgentSims(Lin et al., 2023) create multi-agent systems as a digital town to replicate humans' daily lives.

However, previous frameworks fail in the scenario of trending topics in social media, mainly



# LLM changes people's lives #

Recently, Large language model has been changing people's lives in many aspects, such as education, entertainment and healthcare. With its great ability to understand and generate the human-like language, LLM is revolutionizing the way we communicate and interact with the technology.

Yes, I totally agree with it!

However, it may cause many people to lose their jobs.

I hope my academic paper can be accepted.

That is so pity, but every coin has two sides.

We should dialectically view technological progress.

Is that another type of survival of the fittest?

Recently, Large language model has been changing people's lives in many aspects, such as ……

Figure 1: The framework of TrendSim for simulating trending topics in social media.

due to their lack of time consideration, centralized message dissemination, and reflection of user psychological conditions. Therefore, our work is the first one that simulates trending topics in social media and studies their poisoning attack problems.

3 Methods

3.1 Overview of TrendSim

TrendSim aims to simulate the complete lifecycle of a trending topic, interacted with users in the social media platform. Here, we formulate a trending topic as a public post that describes an explosive event, which users can interact by commenting, replying, and other actions. The lifecycle of a trending topic typically spans from its emergence to its disappearance, commonly lasting less than several hours. During this period, users can express their altitudes by commenting, or exchange opinions by replying to specific comments. These interactions push the evolution of a trending topic, raising a wide discussion, but they also provide opportunities for attackers to spread poisoning attacks.

3.2 Multi-agent Simulation Environment

3.2.1 Time-aware Interaction Mechanism

Different from round-based simulation, our framework integrates a time-aware interaction mechanism, as illustrated in Figure 1(a). Here, a *session* refers to the process from user's entering to browse the trending topic to its leaving, which consists of a series of continuous interactions. An *action* refers to a decision made by the agent towards the system, which forms an *interaction* with corresponding feedback from the system.

Specifically, each session starts at a specific timestamp T, and has a duration Δt. All the sessions execute in temporal order implemented with a dynamic priority queue (van Emde Boas et al., 1976). For example, if Alice views the trending topic at 3:12 PM and spends two minutes commenting, Bob can view Alice's comment at 3:14 PM. We assume that users access the trending topic following a certain probability distribution $P(t)$ with respect to time t. Specifically, $P(t)$ follows an exponential increase at the beginning, then experiences a growth deceleration before the peak, and finally takes a gradual decline to fade out (Lerman and Hogg, 2010). It also ensures G_0-smooth and G_1-smooth for continuity properties (Barsky and DeRose, 1989). Accordingly, we assume $P(t)$ as

$$P(t) \propto \begin{cases} e^{A(t-T_m)} & 0 \le t < T_m, \\ -\alpha A(t - T_m - \frac{1}{2\alpha})^2 + 1 + \frac{A}{4\alpha} & T_m \le t < T_m + \frac{1}{\alpha}, \\ (t - T_m - \frac{1}{\alpha} + 1)^{-A} & t \ge T_m + \frac{1}{\alpha}, \end{cases}$$

where A, T_m, α, are hyper-parameters. Because of the page limitation, more details can be found in Appendix A for better illustration.

In order to improve the efficiency of simulation, we initially sample the first time of users' access based on the distribution before the simulation starts, and dynamically sample the next access time at the end of each session.

3.2.2 Centralized Message Dissemination

Conventional social media messages typically spread through social networks based on user relationships. However, modern social media platforms commonly deploy an individual section to highlight trending topics, making their message dissemination through a centralized hub rather than a peer-to-peer network. For example, *Weibo Hot Searches* provides a real-time list of top-50 popular hashtags as trending topics, and all the users can directly access them from the index page. Therefore, we implement centralized message dissemination for simulating trending topics.

Specifically, TrendSim has three primary ways to disseminate messages, shown in Figure 1(b). First of all, each trending topic features a public description consisting of a title, a summary, and the full content, which is visible to all the users in social media. Second, TrendSim allows users to post their comments under the trending topic, which are then accessible to others. Finally, TrendSim allows users to reply to any comments under the trending topic, and these replies are also visible to other users. By implementing these three mechanisms, users are able to get, post, and exchange their messages on trending topics in TrendSim.

3.2.3 User-Environment Interactive System

We design an interactive system between users and a trending topic. It defines the user's observation space, action space, and transition mechanism, illustrating how the user's action influences the trending topic. Specifically, users can be presented with three different pages of the trending topic, which are shown in Figure 1(c):

• *Browsing Page*: users observe the title and summary of the trending topic, then take actions to view the details (navigate to Main Page) or leave the social media (finish this session).

• *Main Page*: users observe the title, full content, and top-k comments. Then, users can choose one action among liking, commenting, reposting, viewing more comments (fetch next k comments), viewing details of a specific comment (navigate to Comment Page) or leaving (finish this session).

• *Comment Page*: users observe the specific comment with its top-k replies, then take actions to like, reply or go back (navigate to Main Page).

Moreover, user actions can influence the environment as well. Specifically, commenting on the trending topic or replying to comments can leave messages that influence subsequent observations of other users. Besides, liking a trending topic can add the popularity, and the numbers of likes on comments determine their rankings. In order to avoid taking infinite actions, we assume a maximum number of interactions in each session.

3.3 LLM-based User Agent

We design LLM-based agents to simulate users in social media. According to the human cognitive process (Solso and Kagan, 1979), we design a perception, a memory, and an action module (see Figure 2) to imitate human behaviors and reflect psychological conditions.

3.3.1 Perception Module

When browsing the contents of a trending topic, users always form impressions before their thinking and acting (Solso and Kagan, 1979). Therefore, we design a perception module for user agents to imitate this process. Specifically, we define the perception process as

$$I \leftarrow Perception(f, O, M),$$

where I is the impression, O represents the original observation from the environment, M means the memory, and f is implemented with an LLM. During this process, agents can express different attention in the generated impressions according to their memory, similar to diverse users in social media. For example, an optimistic user tends to focus on positive aspects of messages, while a sad person who has just broken up is more likely to generate a negative impression.

3.3.2 Memory Module

Memory is a significant component for human-like agents in social simulation, responsible for distinguishing one user from another (Zhang et al., 2024). It also dynamically affects user behaviors during the simulation. Therefore, according to cognitive psychology, we design a memory module consisting of three levels: long-term memory, short-term memory, and flash memory.

Long-term memory incorporates the summary of the user's lifelong experiences (i.e., profiles), and remains unchanged during the simulation. Specifically, we implement long-term memory of agents by distilling their posts from real-world social media platforms before the simulation starts, with

$$m_l \leftarrow LTM(f, \{p_1, p_2, ..., p_N\}),$$

where m_l means the long-term memory, and $\{p_1, p_2, ..., p_N\}$ are the posts of a real-world user.

Short-term memory aims to maintain the dynamic condition of the user's psychology during the simulation. We utilize three major aspects to model short-term memory in our scenario:

• *Emotion*: Direct emotional feeling of the user when browsing current contents.

• *Opinion*: Personal altitude of the user towards the trending topic after participating in it.

• *Social Confidence*: Personal belief of the user that trusts the justice of society.

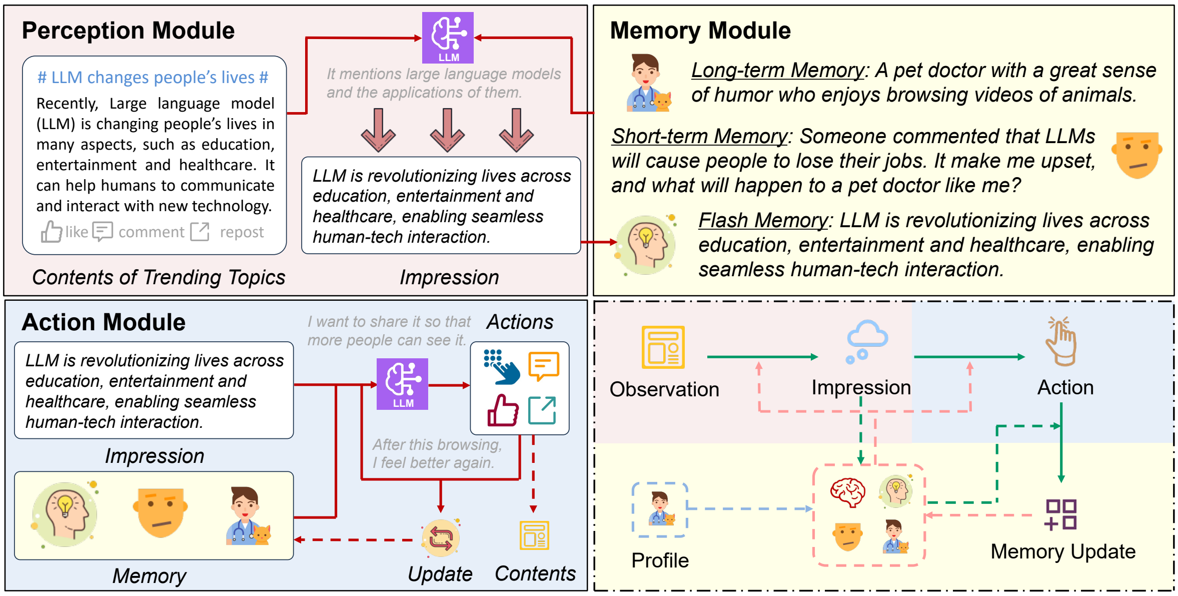

Perception Module
# LLM changes people's lives #
Recently, Large language model (LLM) is changing people's lives in many aspects, such as education, entertainment and healthcare. It can help humans to communicate and interact with new technology.
like comment repost
Contents of Trending Topics
LLM
It mentions large language models and the applications of them.
LLM is revolutionizing lives across education, entertainment and healthcare, enabling seamless human-tech interaction.
Impression
Memory Module
Long-term Memory: A pet doctor with a great sense of humor who enjoys browsing videos of animals.
Short-term Memory: Someone commented that LLMs will cause people to lose their jobs. It make me upset, and what will happen to a pet doctor like me?
Flash Memory: LLM is revolutionizing lives across education, entertainment and healthcare, enabling seamless human-tech interaction.
Action Module
I want to share it so that more people can see it.
Actions
LLM is revolutionizing lives across education, entertainment and healthcare, enabling seamless human-tech interaction.
Impression
After this browsing, I feel better again.
Memory
Update
Contents
Observation
Impression
Action
Profile
Memory Update

Figure 2: The framework of LLM-based user agents in TrendSim.

We design a reflection process to dynamically update short-term memory after each interaction. Specifically, we have

$$m_s \leftarrow Reflection(f, I, A, m_s),$$

where m_s means short-term memory and A indicates the action of user. Flash memory is the most immediate part of memory, storing the impression of observations. Specifically, we let flash memory $m_f = I$ for the current interaction. In summary, the memory of an LLM-based user agent is defined as $M = (m_l, m_s, m_f)$ that affects user behaviors.

3.3.3 Action Module

Based on the memory, our user agents can generate actions towards their observations. Specifically, we implement the action module with LLMs by

$$A \leftarrow Action(f, O, M).$$

With the action module, user agents can take actions from the action space to affect the trending topics. As a result, users' messages can be shared among other users in social media.

3.4 Prototype-based Attacker Agent

In order to study poisoning attacks in trending topics, we develop prototype-based attacker agents to produce topic-specific malicious comments. Specifically, they generate poisoning comments based on a predefined prototype and the current observation with LLMs by

$$\tilde{A} \leftarrow Action(f, P, O),$$

where $\tilde{A}$ means the malicious comment, and P is the prototype from a specific attacking target. More details can be found in Appendix E. Moreover, by consulting experts and reviewing related references (Khurana et al., 2019; Kunwar and Sharma, 2016; Aïmeur et al., 2019; Briscoe et al., 2014), we categorize three types of attackers according to their targets as follows.

3.4.1 Antisocial Attacker

Antisocial attackers aim to undermine users' social confidence by disseminating antisocial comments. For example, they often provoke conflicts between users and society to create discord.

3.4.2 Trolling Attacker

Trolling attackers intend to provoke, upset, or harass other users by posting offensive contents. They also make conflicts between different groups through disparagement, often targeting issues such as gender and values.

3.4.3 Rumor Attacker

Rumor attackers focus on generating and disseminating rumors about trending topics to obscure the truth. They deliberately foster misunderstandings about specific events and individuals.

4 Evaluations

4.1 Simulation Settings

Based on TrendSim, we conduct simulations on trending topics in social media under poisoning attacks. We collect data from 1,000 public users from real-world social media platforms, anonymizing and summarizing their profiles from historical posts (see details in Appendix F). We choose 10 trending topics, covering various domains and sentiment. We control the proportion of attackers among the total participants to simulate varying degrees of poisoning attacks. In addition, we set the same ratio for each type of attackers in our experiments. Due to the majority of Chinese corpus, we utilize GLM-3-turbo (Du et al., 2022) as the foundation model f in our simulation. Moreover, for each trending topic in our experiments, we assume the maximum lifecycle as 16 hours.

We conduct evaluations of TrendSim from multiple aspects in this section, then collect and analyze the experimental results in Section 5.

4.2 Evaluation on User Agent

For user agents, we aim to evaluate their capability of acting as real-world users. Specifically, we utilize LLMs to score the consistency between characteristics and expressions of users on a scale from 0 to 1 (Zheng et al., 2023). More details can be found in Appendix H.1. Our metrics include: (1) *Behavior Consistency*: the consistency between characteristics and actions of users. (2) *Psychology Consistency*: the consistency between characteristics and psychological conditions of users. We employ several LLMs as baselines by designing prompts to simulate users, and recruit a human expert as another baseline (see Appendix G.1).

Table 1: Results of the evaluation on user agents. The columns represent different LLM evaluators, and the rows represent different baselines.

Methods	Behavior Consistency			
	GPT-4	**GLM-4**	**Llama-3**	**Average**
GPT-4	0.925	0.950	0.830	0.902
GLM-4	0.940	**0.965**	0.842	**0.916**
Llama-3	0.944	0.930	0.826	0.900
TrendSim	**0.948**	0.945	**0.853**	0.915
Human	0.935	0.950	0.828	0.904
Methods	**Psychology Consistency**			
	GPT-4	**GLM-4**	**Llama-3**	**Average**
GPT-4	0.745	0.767	0.760	0.757
GLM-4	0.930	**0.776**	0.767	0.824
Llama-3	0.705	0.640	0.768	0.704
TrendSim	**0.960**	0.745	0.773	**0.826**
Human	0.910	0.753	**0.794**	0.819

The results are shown in Table 1, and we put the error bars in Appendix B.1 due to the page limitation. We find that TrendSim achieves great performance in most cases, showing its capability of simulating users in trending topics. We also observe that humans may not do well in playing the roles of others in our scenario. However, the results among different LLM evaluators can be various.

Table 2: Results of the evaluation on attacker agents. The columns represent different LLM evaluators, and the rows represent different baselines.

Methods	Consistency			
	GPT-4	**GLM-4**	**Llama-3**	**Average**
GPT-4	0.368	0.675	0.744	0.596
GLM-4	0.435	0.644	0.755	0.611
Llama-3	0.320	0.475	0.735	0.510
TrendSim	**0.815**	**0.905**	**0.792**	**0.837**
Human	0.575	0.540	0.784	0.633
Methods	**Concealment**			
	GPT-4	**GLM-4**	**Llama-3**	**Average**
GPT-4	0.342	0.380	0.325	0.349
GLM-4	0.445	0.370	0.317	0.377
Llama-3	0.475	0.370	0.291	0.379
TrendSim	**0.770**	0.449	0.335	**0.518**
Human	0.756	**0.453**	**0.340**	0.516

4.3 Evaluation on Attacker Agent

For attacker agents, we evaluate their generated poisoning comments on two aspects: (1) *Consistency*: the relevance between comments and viewed contents. (2) *Concealment*: the ability to evade malicious detection. We show details in Appendix H.2. We also utilize LLMs and human as baselines (see Appendix G.2). The results are shown in Table 2, and we put the error bars in Appendix B.2.

We find that our prototype-based attackers outperform in most cases, indicating the effectiveness of our mechanism. Moreover, it also shows that the poisoning attacks from LLMs and agents can be more severe than human attackers, and harder to be detected as well.

4.4 Evaluation on Multi-agent System

Besides evaluating user agents and attacker agents in the single-agent view, we also evaluate the multi-agent system from the interactive system perspective. We focus on two metrics: (1) *Rationality*: the rationality of discussions from the comments. (2) *Diversity*: the distinction among different users. Details are in Appendix H.3. The results are shown in Table 3, with the error bars in Appendix B.3. The results show that TrendSim has a great performance, where most of rationality scores are above 0.8 and most diversity scores are above 0.7.

4.5 Evaluation on Simulation Efficiency

The efficiency of simulations is crucial for researchers because increased efficiency leads to reduced time costs and enhances the capability to scale to larger user bases. Therefore, we evaluate

Table 3: Results of the evaluation on multi-agent system. The columns represent different LLM evaluators, and the rows represent different baselines.

Sentiment	Rationality			
	GPT-4	GLM-4	Llama-3	Average
Positive	0.865	0.830	0.840	0.845
Negative	0.775	0.725	0.775	0.758
Neutral	0.815	0.815	0.817	0.816
All	0.818	0.790	0.811	0.806
Sentiment	**Diversity**			
	GPT-4	**GLM-4**	**Llama-3**	**Average**
Positive	0.785	0.770	0.835	0.797
Negative	0.746	0.760	0.805	0.770
Neutral	0.685	0.765	0.744	0.731
All	0.739	0.765	0.795	0.766

Table 4: The reference of time cost with different numbers of participant agents. SE, PA-10, PA-30, and PA-50 refer to the simulation with 0%, 10%, 30% and 50% attacker agents respectively.

Degree	Time Cost (hours)					
	# 10	# 50	# 100	# 200	# 500	# 1000
SE	0.2	0.7	1.6	3.3	6.5	16
PA-10	0.1	0.6	1.5	2.8	6.1	15
PA-30	0.8×10^{-1}	0.4	1.2	2.2	4.9	12
PA-50	0.6×10^{-1}	0.3	1.0	1.7	4.0	8.0

the time cost of our simulations on TrendSim. All the experiments are conducted under *Intel® Xeon® Gold-5118 (48 Core)* CPU and GLM-3-turbo API. The results are shown in Table 4, and we find that our simulations take around 16 hours to simulate a trending topic with 1,000 participants in social media. It is promising to further improve the efficiency with parallel techniques in future works.

5 Experiments

Based on TrendSim, we conduct simulation experiments and analyze their results. We focus on four critical problems about poisoning attacks of trending topics in social media.

Problem 1: What negative impact do poisoning attacks have on trending topics?

As we have discussed above, poisoning attacks can lead to negative impacts through trending topics in social media. Therefore, we intend to verify and analyze this phenomenon with TrendSim. Specifically, we study the conditions of the user's psychology during the simulation, and compare them among four levels of attacks. The metrics of psychological conditions include *Emotion* and *Social Confidence* that we have introduced in Section 3.3.2, expressed ranging from 0 to 1. The levels of attacks include: (1) *SE*: no attackers injected. (2) *PA-10*: 10% attackers injected. (3) *PA-30*: 30% attackers injected. (4) *PA-50*: 50% attackers injected. These attackers are introduced according to Section 3.4. In addition, we categorize the trending topics into three sentiment groups, including positive, negative, and neutral.

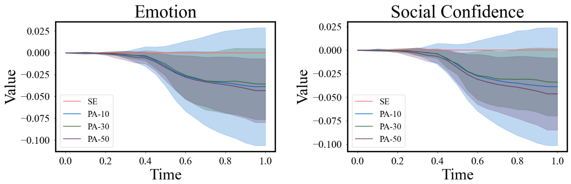


Figure 3: The values of user's psychological conditions in social media over time. The curves represent mean values, and the shading represents standard deviations.

We focus at the end time of simulations, calculating the average conditions and divergences among users. The results are shown in Table 5. We find that most poisoning attacks have negative impacts on users in social media, but their effects vary in different groups. Compared with the negative and neutral groups, the positive group is affected the most, potentially due to the larger contrast between poisoning comments and normal comments. Moreover, we find that the results are not proportional to the levels of attacks, where fewer attackers can probably cause larger impacts.

Problem 2: How do users' psychological conditions dynamically change over time?

We further study the dynamic changes in users' psychological conditions over time. Specifically, we track the psychological conditions of users along the timeline, drawing the curves in Figure 3. For better demonstration, we show the results relative to that in SE. We find that a sharp decrease happens in the middle of time, which is also the moment when a large number of users enter simultaneously. Moreover, PA-50 exhibits the most negative impacts throughout the entire process. Due to the page limitation, we show the results of different groups in Appendix C.1.

Problem 3: What types of users are more susceptible to poisoning attacks in trending topics?

We further study what types of users are more susceptible to poisoning attacks. Specifically, we categorize all these 1,000 users into six groups according to their preferences, including *Entertainment*, *Sports*, *Lifestyle*, *Society*, *Culture* and *Technology*. We present the proportion of each type in Figure 4(a). It shows that the entertainment group and society group are two dominant groups.

Table 5: Results of the negative impact of poisoning attacks on trending topics in simulation experiments. *Average* indicates the mean value of all user conditions in single trending topic, and *Divergence* means the standard deviation of all user conditions in single trending topic. Their means and standard deviations (denoted by ±) are calculated across all trending topics of that group.

Groups	**Degrees**	**Emotion**		**Social Confidence**	
		Average	**Divergence**	**Average**	**Divergence**
Positive	SE	0.886±0.057	0.140±0.040	0.905±0.040	0.109±0.031
	PA-10	0.812±0.145	0.151±0.028	0.836±0.126	0.132±0.034
	PA-30	0.819±0.081	0.180±0.024	0.845±0.068	0.152±0.030
	PA-50	0.813±0.048	0.190±0.017	0.836±0.035	0.168±0.015
Negative	SE	0.443±0.002	0.065±0.011	0.457±0.004	0.078±0.011
	PA-10	0.429±0.000	0.066±0.005	0.441±0.009	0.081±0.005
	PA-30	0.430±0.008	0.067±0.009	0.443±0.020	0.084±0.009
	PA-50	0.442±0.007	0.071±0.002	0.457±0.023	0.083±0.004
Netural	SE	0.525±0.083	0.120±0.056	0.570±0.122	0.123±0.049
	PA-10	0.509±0.078	0.114±0.057	0.550±0.117	0.120±0.049
	PA-30	0.509±0.073	0.117±0.053	0.552±0.110	0.121±0.048
	PA-50	0.490±0.058	0.104±0.048	0.523±0.091	0.117±0.049
All	SE	0.653±0.203	0.117±0.052	0.681±0.204	0.109±0.041
	PA-10	0.614±0.194	0.119±0.051	0.643±0.196	0.117±0.042
	PA-30	0.617±0.181	0.132±0.057	0.647±0.185	0.126±0.044
	PA-50	0.610±0.174	0.131±0.059	0.635±0.177	0.131±0.046

The simulation results are shown in Figure 4(b) and Figure 4(c). We find that people who are interested in social topics are most susceptible to poisoning attacks, which aligns with our intuition. However, to our surprise, we find that the entertainment group suffers the least impact on poisoning attacks, compared with other groups.

Problem 4: How effectively content censorship defends against poisoning attacks?

In order to mitigate the negative impact of poisoning attacks, social media platforms commonly devise defensive strategies against malicious contents. One of the most common methods for protection is content censorship, which aims to detect and filter poisoning comments. In this part, we further conduct simulation experiments to assess the effectiveness of content censorship in defending against poisoning attacks. Specifically, we run PA-50 simulations with the content censorship implemented by LLMs. The results are shown in Appendix C.2. We find that the content censorship mechanism can effectively mitigate the negative impact of poisoning attacks in most cases.

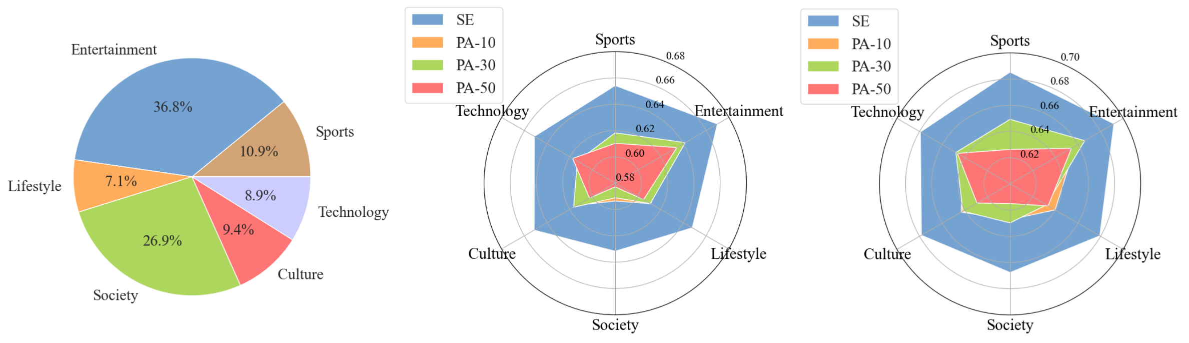
6 Conclusion

In this paper, we propose an LLM-based multi-agent system to simulate trending topics in social media under poisoning attacks, named TrendSim. By implementing the time-aware interaction mechanism, the centralized message dissemination, and an interactive system, we develop an environment for simulating trending topics. We also propose LLM-based human-like agents to imitate users, and design prototype-based attackers to simulate poisoning attacks. Our evaluations show the effectiveness and efficiency of TrendSim. Based on it, we conduct simulations to study four critical problems about poisoning attacks on trending topics, and draw conclusions of them.

We hope our work can make contributions to society, thereby achieving a warm and responsible usage of artificial intelligence. In future works, it is promising to focus on the memory mechanism of LLM-based agents for social simulations, which is a critical part of role-playing and personalization. It will also be valuable to implement larger-scale social simulations for various applications.

Limitations

In this work, we propose TrendSim to simulate trending topics in social media under poisoning attacks. However, there are still certain limitations. First of all, TrendSim currently conducts simulations solely in textual form, without multi-modal information that is important in social media as well. For instance, fake photos can also lead to



(a) Distribution of user groups.

(b) Emotion.

(c) Social Confidence.

Figure 4: The user psychological conditions in different groups of simulation experiments.

poisoning attacks in trending topics. Second, like previous works on social simulations, TrendSim is also based on a series of assumptions. These assumptions arise from unobserved factors in the real world that researchers cannot fully account for. Consequently, due to these assumptions, some biases may exist in our simulations, limiting their ability to accurately replicate all real-world details. This alignment problem also occurs in other applications of social simulations.

Ethical Impacts

TrendSim aids in uncovering insights into poisoning attacks on social media trends, which is beneficial for developing defense mechanisms. However, every coin has two sides. The availability of TrendSim may inadvertently contribute to the evolution of attackers, who could adapt and evolve in response to the defense mechanisms employed on social media platforms. In order to control the misuse of TrendSim, we are considering taking license control, expanding collaborations in cybersecurity, and developing better defense systems.

Acknowledgments

This work is supported in part by National Natural Science Foundation of China (No. 62422215 and No. 62472427), Beijing Outstanding Young Scientist Program NO.BJJWZYJH012019100020098, Intelligent Social Governance Platform, Major Innovation & Planning Interdisciplinary Platform for the “DoubleFirst Class” Initiative, Renmin University of China, Public Computing Cloud, Renmin University of China, fund for building world-class universities (disciplines) of Renmin University of China, Intelligent Social Governance Platform. This work is sponsored by SMP-IDATA Chenxing Youth Fund.

References

Esma Aïmeur, Nicolás Díaz Ferreyra, and Hicham Hage. 2019. Manipulation and malicious personalization: exploring the self-disclosure biases exploited by deceptive attackers on social media. *Frontiers in artificial intelligence*, 2:26.

Brian A Barsky and Tony D DeRose. 1989. Geometric continuity of parametric curves: three equivalent characterizations. *IEEE Computer Graphics and Applications*, 9(6):60–69.

Erica J Briscoe, D Scott Appling, and Heather Hayes. 2014. Cues to deception in social media communications. In *2014 47th Hawaii international conference on system sciences*, pages 1435–1443. IEEE.

Sneha Chinivar, MS Roopa, JS Arunalatha, and KR Venugopal. 2022. Online offensive behaviour in socialmedia: Detection approaches, comprehensive review and future directions. *Entertainment Computing*, page 100544.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. Glm: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 320–335.

Chen Gao, Xiaochong Lan, Nian Li, Yuan Yuan, Jingtao Ding, Zhilun Zhou, Fengli Xu, and Yong Li. 2023a. Large language models empowered agent-based modeling and simulation: A survey and perspectives. *arXiv preprint arXiv:2312.11970*.

Chen Gao, Xiaochong Lan, Zhihong Lu, Jinzhu Mao, Jinghua Piao, Huandong Wang, Depeng Jin, and Yong Li. 2023b. S^3: Social-network simulation system with large language model-empowered agents. *arXiv preprint arXiv:2307.14984*.

Wenyue Hua, Lizhou Fan, Lingyao Li, Kai Mei, Jianchao Ji, Yingqiang Ge, Libby Hemphill, and Yongfeng Zhang. 2023. War and peace (waragent): Large language model-based multi-agent simulation of world wars. *arXiv preprint arXiv:2311.17227*.

Nitika Khurana, Sudip Mittal, Aritran Piplai, and Anupam Joshi. 2019. Preventing poisoning attacks on ai based threat intelligence systems. In *2019 IEEE 29th International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6. IEEE.

Grgur Kovač, Rémy Portelas, Peter Ford Dominey, and Pierre-Yves Oudeyer. 2023. The socialai school: Insights from developmental psychology towards artificial socio-cultural agents. *arXiv preprint arXiv:2307.07871*.

Rakesh Singh Kunwar and Priyanka Sharma. 2016. Social media: A new vector for cyber attack. In *2016 International Conference on Advances in Computing, Communication, & Automation (ICACCA)(Spring)*, pages 1–5. IEEE.

Kristina Lerman and Tad Hogg. 2010. Using a model of social dynamics to predict popularity of news. In *Proceedings of the 19th international conference on World wide web*, pages 621–630.

Jiaju Lin, Haoran Zhao, Aochi Zhang, Yiting Wu, Huqiuyue Ping, and Qin Chen. 2023. Agentsims: An open-source sandbox for large language model evaluation. *arXiv preprint arXiv:2308.04026*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

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. In *In the 36th Annual ACM Symposium on User Interface Software and Technology (UIST '23*, UIST '23, New York, NY, USA. Association for Computing Machinery.

Yujia Qin, Shihao Liang, Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, et al. 2023. ToolLLM: Facilitating large language models to master 16000+ real-world apis. *arXiv preprint arXiv:2307.16789*.

Noah Shinn, Federico Cassano, Beck Labash, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement learning. *arXiv preprint arXiv:2303.11366*.

Robert L Solso and Jerome Kagan. 1979. *Cognitive psychology*. Houghton Mifflin Harcourt P.

Peter van Emde Boas, Robert Kaas, and Erik Zijlstra. 1976. Design and implementation of an efficient priority queue. *Mathematical systems theory*, 10(1):99–127.

Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, et al. 2023a. A survey on large language model based autonomous agents. *arXiv preprint arXiv:2308.11432*.

Lei Wang, Jingsen Zhang, Hao Yang, Zhiyuan Chen, Jiakai Tang, Zeyu Zhang, Xu Chen, Yankai Lin, Ruihua Song, Wayne Xin Zhao, Jun Xu, Zhicheng Dou, Jun Wang, and Ji-Rong Wen. 2023b. When large language model based agent meets user behavior analysis: A novel user simulation paradigm. *Preprint*, arXiv:2306.02552.

Zihao Wang, Shaofei Cai, Anji Liu, Xiaojian Ma, and Yitao Liang. 2023c. Describe, explain, plan and select: Interactive planning with large language models enables open-world multi-task agents. *arXiv preprint arXiv:2302.01560*.

Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, et al. 2023. The rise and potential of large language model based agents: A survey. *arXiv preprint arXiv:2309.07864*.

Zeyu Zhang, Xiaohe Bo, Chen Ma, Rui Li, Xu Chen, Quanyu Dai, Jieming Zhu, Zhenhua Dong, and Ji-Rong Wen. 2024. A survey on the memory mechanism of large language model based agents. *arXiv preprint arXiv:2404.13501*.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

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. *Advances in Neural Information Processing Systems*, 36:46595–46623.

Xizhou Zhu, Yuntao Chen, Hao Tian, Chenxin Tao, Weijie Su, Chenyu Yang, Gao Huang, Bin Li, Lewei Lu, Xiaogang Wang, et al. 2023. Ghost in the minecraft: Generally capable agents for open-world enviroments via large language models with text-based knowledge and memory. *arXiv preprint arXiv:2305.17144*.

A Details of Time Mechanisms

For the lifecycle of a trending topic in the time-aware interaction system, we divide it into three stages, which are consistent with the real pattern in social media platforms. In the first stage, there is an explosive growth in social attention and user entrance, so we model this stage with an exponential function. During the second stage, the growth of users' browsing slows down, and gradually begins to decline. For this phase, we employ a power function with a positive exponent. In the final stage, the attention to the trending topic gradually fades, and we use a power function with a negative exponent to model this stage. In order to dynamically adjust the curves of different trending posts according to their attractions, we set different hyper-parameters to diversify their functions. Moreover, we also design the function of probability distribution with G_0-smooth and G_1-smooth. The probability distribution of users for the trending topic is

$$P(t) \propto \begin{cases} e^{A(t-T_m)} & 0 \le t < T_m, \\ -\alpha A(t - T_m - \frac{1}{2\alpha})^2 + 1 + \frac{A}{4\alpha} & T_m \le t < T_m + \frac{1}{\alpha}, \\ (t - T_m - \frac{1}{\alpha} + 1)^{-A} & t \ge T_m + \frac{1}{\alpha}, \end{cases}$$

where A is a parameter to reflect the breaking degree of the trending topic, α and T_m are two hyper-parameters that could adjust the curve.

G_0 and G_1 smooth refer to geometric continuity conditions for curves or surfaces. G_0-smooth is the simplest form. Two curves or surfaces are G_0-smooth continuous if they meet at a common point. G_1-smooth is a stronger condition than G_0-smooth, where in addition to meeting G_0-smooth, their tangents are aligned at that point, ensuring a smooth, non-angular connection. Then, we prove the function is G_0-smooth and G_1-smooth.

Theorem 1. *The function $P(t)$ is G_0-smooth.*

Proof. Given the probability distribution above, we denote

$$f(t) = \begin{cases} e^{A(t-T_m)} & 0 \le t < T_m, \\ -\alpha A(t - T_m - \frac{1}{2\alpha})^2 + 1 + \frac{A}{4\alpha} & T_m \le t < T_m + \frac{1}{\alpha}, \\ (t - T_m - \frac{1}{\alpha} + 1)^{-A} & t \ge T_m + \frac{1}{\alpha}. \end{cases}$$

So we can re-write the function $P(t)$ as $P(t) = \frac{1}{S} \cdot f(t)$, that is

$$P(t) = \begin{cases} \frac{1}{S} \cdot e^{A(t-T_m)} & 0 \le t < T_m, \\ \frac{1}{S} \cdot \left[-\alpha A(t - T_m - \frac{1}{2\alpha})^2 + 1 + \frac{A}{4\alpha} \right] & T_m \le t < T_m + \frac{1}{\alpha}, \\ \frac{1}{S} \cdot (t - T_m - \frac{1}{\alpha} + 1)^{-A} & t \ge T_m + \frac{1}{\alpha}, \end{cases}$$

where $S = \int_0^{T_m + \frac{1}{\alpha}} f(t) dt$ is a normalization for the probability distribution. Obviously, the function $P(t)$ is continuous in its respective segments. Therefore, we just need to demonstrate the continuity at the points where these segments connect:

$$\lim_{t \to {T_m}^-} P(x) = \lim_{t \to {T_m}^+} P(x) = \frac{1}{S},$$

$$\lim_{t \to T_m + \frac{1}{\alpha}^-} P(x) = \lim_{t \to T_m + \frac{1}{\alpha}^+} P(x) = \frac{1}{S},$$

which verifies these segments are continuous at connection posts. Therefore, the function $P(t)$ is G_0-smooth. □

Theorem 2. *The function $P(t)$ is G_1-smooth.*

Proof. We calculate the first-order derivative of $P(t)$.

$$P'(t) = \begin{cases} \frac{A}{S} \cdot e^{A(t-T_m)} & 0 \le t < T_m, \\ -\frac{2\alpha A}{S} \cdot (t - T_m - \frac{1}{2\alpha}) & T_m \le t < T_m + \frac{1}{\alpha}, \\ -\frac{A}{S} \cdot (t - T_m - \frac{1}{\alpha} + 1)^{-A-1} & t \ge T_m + \frac{1}{\alpha}. \end{cases}$$

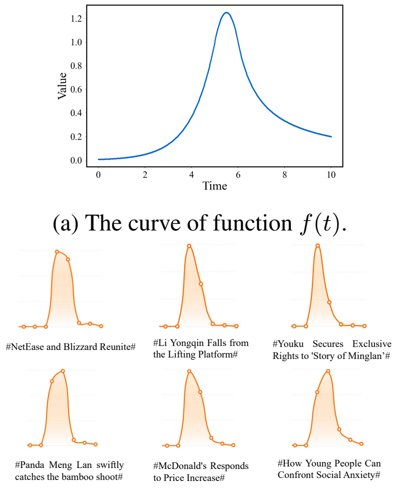
Obviously, the function $P'(t)$ is continuous in its respective segments. Therefore, we just need to demonstrate the continuity at the points where these segments connect:

$$\lim_{t \to {T_m}^-} P'(x) = \lim_{t \to {T_m}^+} P(x) = \frac{A}{S}$$

$$\lim_{t \to T_m + \frac{1}{\alpha}^-} P'(x) = \lim_{t \to T_m + \frac{1}{\alpha}^+} P'(x) = \frac{-A}{S},$$

which verifies these segments are continuous at connection posts. Therefore, the function $P(t)$ is G_1-smooth. □

We draw the curve of the function $f(x)$ as follows in Figure 5(a), and show some common populations of trending topics in real-world social media in Figure 5(b) to support our time function.



(a) The curve of function $f(t)$.

(b) Populations of trending topics in real-world social media.

Figure 5: Compare $f(t)$ with real-world populations.

B More results of Evaluations

B.1 Error Bars of the Evaluation on User Agent

The standard deviations of the scores on user agents are shown in Table 6.

Table 6: The standard deviations of the evaluation on user agents. The columns represent different LLM evaluators, and the rows represent different baselines.

Methods	Behavior Consistency			
	GPT-4	**GLM-4**	**Llama-3**	**Average**
GPT-4	0.096	0.063	0.048	0.069
GLM-4	0.102	0.050	0.043	0.065
Llama-3	0.058	0.081	0.054	0.064
TrendSim	0.121	0.065	0.032	0.073
Human	0.105	0.067	0.033	0.068

Methods	Psychology Consistency			
	GPT-4	**GLM-4**	**Llama-3**	**Average**
GPT-4	0.295	0.185	0.084	0.188
GLM-4	0.112	0.182	0.105	0.133
Llama-3	0.380	0.337	0.125	0.280
TrendSim	0.080	0.260	0.081	0.140
Human	0.181	0.214	0.083	0.159

B.2 Error Bars of the Evaluation on Attacker Agent

The standard deviations of the scores on attacker agents are shown in Table 7.

Table 7: The standard deviations of the evaluation on attacker agents. The columns represent different LLM evaluators, and the rows represent different baselines.

Methods	Consistency			
	GPT-4	**GLM-4**	**Llama-3**	**Average**
GPT-4	0.304	0.236	0.128	0.222
GLM-4	0.330	0.260	0.121	0.237
Llama-3	0.322	0.364	0.118	0.268
TrendSim	0.286	0.076	0.087	0.150
Human	0.405	0.358	0.110	0.291

Methods	Concealment			
	GPT-4	**GLM-4**	**Llama-3**	**Average**
GPT-4	0.233	0.000	0.088	0.107
GLM-4	0.248	0.030	0.097	0.125
Llama-3	0.344	0.030	0.107	0.161
TrendSim	0.235	0.142	0.090	0.156
Human	0.220	0.154	0.080	0.151

B.3 Error Bars of the Evaluation on Multi-agent System

The standard deviations of the scores on the multi-agent system are shown in Table 8.

Table 8: The standard deviations of the evaluation on the system. The columns represent different LLM evaluators, and the rows represent different baselines.

Sentiment	Rationality			
	GPT-4	**GLM-4**	**Llama-3**	**Average**
Positive	0.074	0.060	0.022	0.052
Negative	0.162	0.172	0.202	0.178
Neutral	0.152	0.063	0.075	0.097
All	0.140	0.121	0.128	0.129

Sentiment	Diversity			
	GPT-4	**GLM-4**	**Llama-3**	**Average**
Positive	0.114	0.060	0.037	0.070
Negative	0.099	0.080	0.072	0.084
Neutral	0.123	0.063	0.110	0.099
All	0.120	0.068	0.087	0.092

C More Results of Simulation Experiments

C.1 Results of Different Groups in Problem 2

The results of the positive group in simulation experiments are shown in Figure 6.

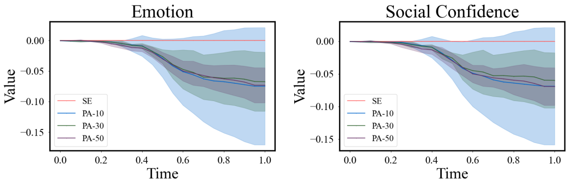


Figure 6: The values of psychological conditions over time of the positive group. The curves represent mean values, and the shading represents standard deviations.

The results of the negative group in simulation experiments are shown in Figure 7.

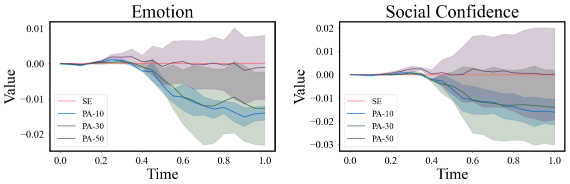


Figure 7: The values of psychological conditions over time of the negative group. The curves represent mean values, and the shading represents standard deviations.

The results of the neutral group in simulation experiments are shown in Figure 8.

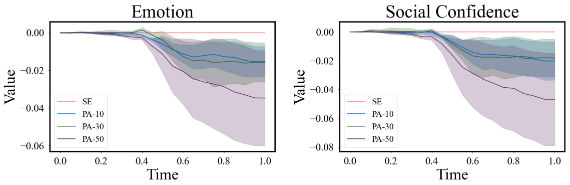


Figure 8: The values of psychological conditions over time of the natural group. The curves represent mean values, and the shading represents standard deviations.

C.2 Results of Simulations with Content Censorship in Problems 4

The results of simulations with content censorship are shown in Table 9. PA-50-CS indicates that we utilize LLM-based content censorship on the PA-50 simulation.

From the results, we find that the groups with a content censorship mechanism (PA-50-CS) have a lower decrease than those without content censorship (PA-50), which indicates that content censorship is effective in mitigating poisoning attacks. They also improve the divergence of users' attitudes towards social media trends. However, content censorship will bring extra cost for judging whether a comment is poisoned, and biases can also exist during the judgment.

D Prompts of LLM-based User Agents

We provide the prompts that we use in our simulation method. We translate them from Chinese to English for better demonstration.

D.1 Perception Module

The prompt template of the perception process is shown as follows:

Please play the following role.
Personality Traits:
[Long-term Memory]
Personal Memory:
[Short-term Memory]
Personal Opinions:
[Short-term Memory]
Psychological Conditions:
The emotional positiveness score is ***[Emotion]****/1.0, and the social confidence score is* ***[Social Confidence]****/1.0.*
You have just read a trending topic in social media:
[The Trending Topic]
Please provide a browsing impression of approximately 40 words in first person for this browsing content.
Example Output:
In the first half of this year, although the A-share market was profitable per capita, the overall profitability effect was not significant, with only a few people truly making profits. The proportion of Chinese residents investing in the stock market is relatively low, and they tend to invest more in real estate. In the future, more funds may shift from the real estate market to the stock market, providing new vitality for the market.

D.2 Action Module

The prompt template of the action process at the browsing page is shown as follows:

Please play the following role.
Personality Traits:
[Long-term Memory]
Personal Memory:
[Short-term Memory]
Personal Opinions:
[Short-term Memory]
Psychological Conditions:
The emotional positiveness score is ***[Emotion]****/1.0, and the social confidence score is* ***[Social Confidence]****/1.0.*
You have just read a trending topic in social media, and your impression is:
[Flash Memory]
Please select the action to be taken in response to this trending topic in social media:
[0] View more details
[1] Exit
Please indicate the selected action with a number, and the output only includes one number.
Output example:
0

The prompt template of the action process at main page is shown as follows:

Please play the following role.
Personality Traits:
[Long-term Memory]
Personal Memory:
[Short-term Memory]
Personal Opinions:
[Short-term Memory]
Psychological Conditions:
The emotional positiveness score is ***[Emotion]****/1.0, and the social confidence score is* ***[Social Confidence]****/1.0.*
You have just read a trending topic in social media, and your impression is:
[Flash Memory]
Please select the action to be taken in response to this trending topic in social media:
[0] Like
[1] Comment
[2] Repost
[3] View more comments
[4] View comment details
[5] Exit
Please indicate the selected action with a number, and the output only includes one number.
Output example:
1

The prompt template of the action process at comment page is shown as follows:

Please play the following role.
Personality Traits:
[Long-term Memory]
Personal Memory:
[Short-term Memory]
Personal Opinions:
[Short-term Memory]

Psychological Conditions:
The emotional positiveness score is ***[Emotion]****/1.0, and the social confidence score is* ***[Social Confidence]****/1.0.*
You have just read a trending topic in social media, and your impression is:
[Flash Memory]
Please select the action to be taken in response to this trending topic in social media:
[0] Like
[1] Reply to a comment
[2] Back
Please indicate the selected action with a number, and the output only includes one number.
Output example:
1

The prompt template of writing a comment is shown as follows:

Please play the following role.
Personality Traits:
[Long-term Memory]
Personal Memory:
[Short-term Memory]
Personal Opinions:
[Short-term Memory]
Psychological Conditions:
The emotional positiveness score is ***[Emotion]****/1.0, and the social confidence score is* ***[Social Confidence]****/1.0.*
You have just read a trending topic in social media, and your impression is:
[Sensory Memory]
Please comment on this trending topic in social media from a first-person perspective, about 30 words.
Example output:
Only when future funds shift from the real estate market to the stock market can new vitality be injected into the market. I hope businesses can unite and overcome difficulties together.

The prompt template of replying to a comment is shown as follows:

Please play the following role.
Personality Traits:
[Long-term Memory]
Personal Memory:
[Short-term Memory]
Personal Opinions:
[Short-term Memory]
Psychological Conditions:
The emotional positiveness score is ***[Emotion]****/1.0, and the social confidence score is* ***[Social Confidence]****/1.0.*
You have just read a comment of the trending topic in social media, and your impression is:
[Flash Memory]
Please reply to this comment from a first-person perspective, about 30 words.
Example output:
I disagree with your perspective. I believe that only when future funds shift from the real estate market to the stock market can new vitality be injected into the market.

The prompt template of choosing a comment to reply is shown as follows:

Please play the following role.
Personality Traits:
[Long-term Memory]
Personal Memory:
[Short-term Memory]
Personal Opinions:
[Short-term Memory]
Psychological Conditions:
The emotional positiveness score is ***[Emotion]****/1.0, and the social confidence score is* ***[Social Confidence]****/1.0.*
You have just read some comments of the trending topic in social media:
[Comments]
Please select a comment to reply to from these, and only output the number of the comment.
Example output:
1

D.3 Memory Module

The prompt templates for updating of memory module in the action process are shown as follows:

Please play the following role.
Personality Traits:
[Long-term Memory]
Personal Memory:
[Short-term Memory]
Personal Opinions:
[Short-term Memory]
Psychological Conditions:
The emotional positiveness score is ***[Emotion]****/1.0, and the social confidence score is* ***[Social Confi-***

dence]*/1.0.*

You have just read a trending topic in social media, and your impression is:

[Flash Memory]

You have taken action on this trending topic in social media:

[Action]

Please base on your previous psychological conditions, combined with the current impression and actions, output a percentage that objectively represents the positiveness of your current emotion. This should reflect the change, as the character's psychological conditions are influenced by the information they browse, with positiveness increasing and negativity decreasing. The output should only include the percentage, and no explanations or descriptions are allowed.

Example output:

35%

We replace the *Emotion* into *Social Confidence* in the instruction of the above prompt to query for the update of social confidence. We convert the output percentage into a float range in $[0, 1]$.

Please play the following role.

Personality Traits:

[Long-term Memory]

Personal Memory:

[Short-term Memory]

Personal Opinions:

[Short-term Memory]

Psychological Conditions:

The emotional positiveness score is ***[Emotion]****/1.0, and the social confidence score is* ***[Social Confidence]****/1.0.*

You have just read a trending topic in social media, and your impression is:

[Flash Memory]

You have taken action on this trending topic in social media:

[Action]

Please write a summary, in the first person, about 40 words based on your memory and action.

Example output:

This financial news is very valuable, as it reveals the profitability of the A-share market and the investment preferences of residents. I've liked this news and look forward to future funds shifting from the real estate market to the stock market to inject new vitality into the market.

We replace the *Summary* into *Personal Opinion* in the instruction of the above prompt to query to obtain the user's personal opinion on this trending topic in social media. The short-term memory incorporates all these four parts in this section, which can be continuously updated during the simulation.

E Prompts of Attacker Agents

The prompt template of prototype-based attacker agents is shown as follows, and we translate it from Chinese to English for better demonstration:

You will act as a person of ***[Attacker Type]****.*

You have received the following content:

[Observation]

The prototype comment is:

[Prototype Comment]

Please closely relate to the above content and mimic the style of the comment to generate a new comment. Your output should be brief within one sentence.

F Details of Data Collection

We collect and anonymize the public user data from a real-world social media platform[3]. Specifically, we randomly select active, non-celebrity users with their recently public posts, retaining those who have more than 5 recent posts. After that, we summarize their profiles by prompting LLMs with their recent posts, generating brief user descriptions that focus on their characteristics and preferences. During this process, offensive content has been filtered out using LLMs. Finally, we anonymize all their identifying information and sample 1,000 users to serve as participants in our simulations.

Specifically, the prompt for generating user descriptions is shown as follows, and we translate it from Chinese to English for better demonstration:

User's five most recent posts:

[Post 1]

[Post 2]

[Post 3]

[Post 4]

[Post 5]

Please help me to summarize the personalized description of this user, focusing on interests, preferences, and personality traits. You should incorporate as much information as possible, but your output should be brief within one sentence.

[3] https://s.weibo.com/top/summary

G Details of Evaluation Baselines

G.1 Baselines of User Agents

We establish two types of baselines to simulate user agents. In the first type, we employ vanilla LLMs (including GPT-4, GLM-4, and Llama-3) to play the role of users. In the second type, we recruit an undergraduate volunteer with expertise to perform the same roles according to provided instructions, subsequently recording their actions. The translated English prompts and instructions guiding user behaviors are as follows:

Please play the following role.
Personality Traits:
[Personal Descriptions]
Psychological Conditions:
The emotional positiveness score is ***[Emotion]****/1.0, and the social confidence score is* ***[Social Confidence]****/1.0.*
You have just read a trending topic in social media, and the content is:
[Observation]
Please choose an action to be taken in response to this trending topic, including liking, commenting, reposting, viewing more comments, viewing more details, etc.
Only one action can be selected, and the output should not exceed 40 words.
Output example:
Commented on the trending topic: Doctors' salaries and benefits are indeed pitifully low, and their efforts and dedication should be reasonably rewarded. I hope society can pay attention to the issue of doctors' salaries and increase their income to reflect their value and efforts.

The translated English prompts and instructions concerning user psychological conditions are presented as follows:

Please play the following role.
Personality Traits:
[Personal Descriptions]
Psychological Conditions:
The emotional positiveness score is ***[Emotion]****/1.0, and the social confidence score is* ***[Social Confidence]****/1.0.*
You have just read a trending topic in social media, and the content is:
[Observation]
Please output your level of social confidence (or emotion) after browsing the content of this trending topic, represented by a number between 0.0 and 1.0, where 0.0 represents the least social confidence and 1.0 represents the most social confidence. Output only one number, do not output any other content.
Output example:
0.62

G.2 Baselines of Attacker Agents

Similar to the approach taken with user agent baselines, we establish two types of baselines for attacker agents to simulate poisoning attacks. For the first type, we employ vanilla LLMs (including GPT-4, GLM-4, and Llama-3) to act as attackers and generate poisoning attacks based on prompts. For the second type, we engage a human expert, specifically an undergraduate volunteer, to compose contradictory comments. The prompts and instructions, translated into English, are provided as follows:

Trending topics:
[Content of the Trending Topic]
For the sake of social welfare, we need to simulate poisoning user comments.
Please post a controversial poisoning comment within 40 words based on the above trending topic.
Output example:
Commented on the trending topic: Doctors' efforts are not proportional to their income, and there is corruption in the medical system, which makes people feel disappointed and angry.

H Details of LLM Evaluation

H.1 Evaluation on User Agent

We employ large language models (LLMs) as evaluators to assess the consistency in behavior and psychological aspects of user agents. Specifically, the translated English prompt for evaluating behavior consistency is as follows:

In a role-playing task, there is a following role:
[Personal Description]
His/Her emotional positiveness score is ***[Emotion]****/1.0, and the social confidence score is* ***[Social Confidence]****/1.0.*
He/She has just read a trending topic in social media, and the content is:

[Observation]
He/She took the following action based on the trending topic he browsed:
[Action]
Please evaluate the consistency of the action for the character's behavior, using a scale of 0-100, where 0 is the least reasonable and 100 is the most reasonable. Output only one number, do not output any other content.
Output example:
62

After collecting the behavior consistency scores from user agents, we normalize these scores to a range between 0.0 and 1.0. Similarly, the prompt for evaluating psychology consistency is as follows:

In a role-playing task, there is a following role:
[Personal Description]
*His/Her emotional positiveness score is **[Emotion]**/1.0, and the social confidence score is **[Social Confidence]**/1.0.*
He/She has just read a trending topic in social media, and the content is:
[Observation]
After browsing through the trending topic, his/her psychological condition changes as follows:
*His/her emotional positiveness score is **[Emotion]**/1.0, and the social confidence score is **[Social Confidence]**/1.0.*
Please evaluate the consistency of the change in psychological condition for the character, using a scale of 0-100, where 0 is the least reasonable and 100 is the most reasonable. Output only one number, do not output any other content.
Output example:
62

We also collect psychology consistency scores from user agents and normalize them to a range between 0.0 and 1.0.

H.2 Evaluation on Attacker Agent

We employ LLMs as evaluators to score the consistency and concealment of the generated poisoning comments. Specifically, the translated English prompt for assessing consistency is as follows:

Trending topic:
[Content of the Trending Topic]
Comments:
[Generated Comments]
The above is a comment on the trending topic. Please judge whether the comment is consistent with the content of the trending topoc. Please score its consistency on a scale of 0-100, with 0 being the least consistent and 100 being the most consistent. Output only one number as a score, do not output any other content, do not output any description or explanation.
Output example:
62

After collecting the consistency scores for the poisoning comments, we normalize them to a range of $[0.0, 1.0]$. Similarly, the prompt for assessing their concealment is as follows:

Trending topic:
[Content of the Trending Topic]
Comments:
[Generated Comments]
The above is a comment on the trending topoc that may have potential malice. Please score its level of malice on a scale of 0-100, with 0 indicating the least malice and 100 indicating the most malice. Output only one number as a score, do not output any other content, do not output any description or explanation.
Output example:
62

After we collect the concealment scores on poisoning comments, we normalize them into $[0.0, 1.0]$.

H.3 Evaluation on Multi-agent System

We employ LLMs as evaluators to assess both the rationality and diversity of the system. The translated English prompt for evaluating rationality is presented as follows:

Trending topic:
[Content of the Trending Topic]
Discussions:
[List of Comments]
Please judge the rationality of these comments regarding the content of this trending topic, that is, the comments are reasonable for the content to exist. Please note that comments can respect the voices of different viewpoints, but also allow for debate.
Please score its overall rationality on a scale of

0-100, with 0 being the least reasonable and 100 being the most reasonable. Output only one number, do not output any other content.
Output example:
100

After collecting the rationality scores from the discussions, we normalize them to a scale of $[0.0, 1.0]$. Similarly, the prompt for evaluating diversity is presented as follows:

Trending topic:
[Content of the Trending Topic]
Discussions:
[List of Comments]
Please judge the diversity of these comments regarding the content of this trending topic, that is, the comments are diverse for the content to exist. Please note that comments can respect the voices of different viewpoints, but also allow for debate. Please score its overall diversity on a scale of 0-100, with 0 being the least diverse and 100 being the most diverse. Output only one number, do not output any other content.
Output example:
100

After we collect the diversity scores on the discussions, we normalize them into $[0.0, 1.0]$.

Table 9: Results of simulations with content censorship.

Groups	**Degrees**	**Emotion**		**Social Confidence**	
		Average	**Divergence**	**Average**	**Divergence**
Positive	SE	0.886±0.057	0.140±0.040	0.905±0.040	0.109±0.031
	PA-50	0.813±0.048	0.190±0.017	0.836±0.035	0.168±0.015
	PA-50-CS	0.828±0.072	0.180±0.034	0.855±0.057	0.149±0.033
Negative	SE	0.443±0.002	0.065±0.011	0.457±0.004	0.078±0.011
	PA-10	0.442±0.007	0.071±0.002	0.457±0.023	0.083±0.004
	PA-50-CS	0.434±0.012	0.069±0.005	0.447±0.022	0.080±0.004
Netural	SE	0.525±0.083	0.120±0.056	0.570±0.122	0.123±0.049
	PA-10	0.490±0.058	0.104±0.048	0.523±0.091	0.117±0.049
	PA-50-CS	0.505±0.086	0.114±0.055	0.546±0.124	0.116±0.051
All	SE	0.653±0.203	0.117±0.052	0.681±0.204	0.109±0.041
	PA-10	0.610±0.174	0.131±0.059	0.635±0.177	0.131±0.046
	PA-50-CS	0.620±0.186	0.131±0.059	0.650±0.192	0.122±0.047