OASIS: Open Agent Social Interaction Simulations with One Million Agents

Anonymous authors
Paper under double-blind review

Abstract

Online social platforms are increasingly becoming hybrid ecosystems in which humans, algorithmic recommenders, and LLM-based agents jointly shape collective behavior. However, controlled study of such ecosystems remains challenging because real platforms couple evolving social graphs, heterogeneous user actions, recommendation-driven exposure, and population-scale feedback loops. We introduce *OASIS*, a modular and scalable social media simulation framework for studying emergent dynamics in LLM-agent societies under platform-faithful interaction mechanisms. *OASIS* abstracts social platforms into evolving environments, platform-style action spaces, interchangeable recommendation modules, and parallel agent execution, allowing researchers to instantiate different platform templates while retaining controllability and reproducibility. The framework supports simulations ranging from thousands to up to one million agents under practical compute budgets, enabling population-level mechanism analysis and intervention studies. We instantiate *OASIS* as Reddit- and X-like platforms and conduct empirical studies on herding behavior and misinformation propagation. Our experiments reveal that LLM-agent societies exhibit stronger negative herding under early adverse feedback than human baselines, and that matched misinformation can achieve higher diffusion intensity than official news in large-scale X-like simulations. We further show that both model choice and recommendation policy substantially affect diffusion dynamics: smaller LLM backbones tend to amplify misinformation more strongly in our setting, while interest-based recommendation increases misinformation exposure compared with random recommendation. Finally, we demonstrate that external grounding interventions, including tool-augmented verification and human participation, consistently reduce misinformation dominance. These results position *OASIS* as a controllable, extensible, and intervention-ready testbed for reproducible research on hybrid online ecosystems.

Warning: this paper includes misinformation examples that may be offensive or harmful.

1 Introduction

Large language models (LLMs) have rapidly advanced in both capability and deployment Bai et al. (2023); Touvron et al. (2023); Jaech et al. (2024); Guo et al. (2025); Anthropic (2026). Beyond offline benchmarks, LLM-based agents are increasingly embedded in interactive online environments, where they generate content, respond to users, participate in discussions, and influence information exposure at scale Ferraro et al. (2024); Moltbook (2026). As a result, modern social platforms are evolving into hybrid ecosystems in which humans, LLM-based agents, and algorithmic recommenders jointly shape collective outcomes Park et al. (2023); Zhou et al. (2023b); Piao et al. (2025); Zhang et al. (2025). Understanding the emergent dynamics of such ecosystems is therefore becoming a central problem for human–AI interaction, computational social science, and platform governance.

A key difficulty is that online collective behavior does not arise from isolated users alone. Real platforms such as X and Reddit couple multiple mechanisms: users are embedded in evolving social graphs; they interact through platform-specific actions such as posting, commenting, reposting, voting, and following; and their observations are mediated by recommendation systems that determine what content is exposed, amplified,

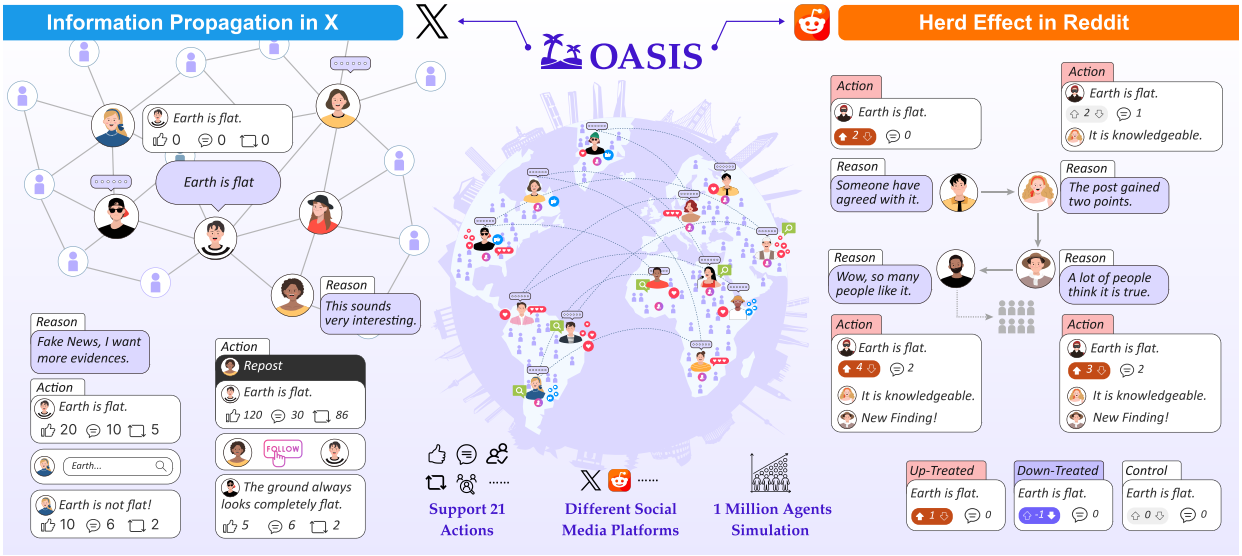


Figure 1: *OASIS* simulates platform-style online ecosystems by coupling evolving environments, recommendation-driven exposure, and LLM-agent interactions. It can instantiate different social media templates, such as Reddit and X, and supports simulations up to one million agents.

or suppressed X Corp; Reddit, Inc.. These mechanisms interact recursively: exposure affects behavior, behavior reshapes the content pool and social graph, and the updated platform state further changes future exposure. Such feedback loops can generate population-level phenomena including herding, polarization, and misinformation cascades Ladyman et al. (2013). However, these phenomena are difficult to study directly on real platforms because access is limited, interventions are costly or unethical, and many confounding factors cannot be controlled.

Simulation provides a natural alternative, but existing LLM-agent simulations are still limited in their ability to reproduce platform-level dynamics. Many prior systems focus on small-group interaction, recommendation-only settings, or simplified social networks with static structures and limited action spaces Wang et al. (2023); Zhang et al. (2024); Gao et al. (2023); Mou et al. (2024). Although these systems have demonstrated the promise of LLM agents for social simulation, they often abstract away key components of real social media platforms, especially evolving networks, platform-style interactions, recommendation-driven exposure, and scalable agent execution. Consequently, they are less suited for studying how platform mechanisms and interventions affect collective behavior at population scale.

To address this gap, we introduce *OASIS*, a modular and scalable social media simulation framework for studying emergent dynamics in LLM-agent societies. The central design principle of *OASIS* is to preserve the core feedback loop of social platforms while keeping each component controllable and replaceable. Specifically, *OASIS* models a platform as an evolving environment containing user profiles, social links, posts, comments, and interaction histories; a platform-style action space through which agents post, reply, vote, repost, follow, and react; and a recommendation module that determines each agent's observable content. This design allows researchers to instantiate different platform templates, such as Reddit-like community discussion or X-like interest-driven information diffusion, while systematically modifying agents, recommendation policies, platform rules, or external interventions.

System overview. Figure 2 illustrates the overall workflow of *OASIS*. An **Environment Server** maintains the global platform state, including agent profiles, social relationships, content objects, and interaction records. A plug-and-play **Recommendation System** produces ranked feeds according to platform-specific signals such as social ties, interests, popularity, and recency, thereby controlling each agent's exposure. A **Time Engine** advances the simulation in discrete steps, schedules agent activities, executes platform actions, and writes the resulting updates back to the environment. Agents are implemented through an **Agent Module**

that maps observations, including profiles, feeds, histories, and platform context, into executable actions. To support large populations, *OASIS* combines concurrent environment updates with high-throughput batched LLM inference, e.g., using vLLM Kwon et al. (2023a). As a result, the same framework can support simulations from thousands to up to one million agents under practical compute budgets.

Empirical demonstrations. We instantiate *OASIS* into two representative platform templates, Reddit and X, and use them to study two canonical forms of collective behavior: herding and misinformation propagation. In Reddit-style simulations, we examine how early feedback shapes later voting behavior. Our results show that LLM-agent societies exhibit stronger negative herding than human baselines when early feedback is adverse, suggesting that LLM agents may be especially sensitive to negative social signals in community-ranking environments. In X-style simulations, we study the diffusion of official news and matched misinformation under interest-based recommendation. We find that matched misinformation can reach higher diffusion intensity than official news at larger scales, indicating that semantic alignment between user interests and misleading content can substantially amplify diffusion.

Beyond reproducing these phenomena, *OASIS* enables controlled mechanism analysis. We vary the underlying LLM backbone and observe substantial differences in misinformation diffusion, with smaller models tending to amplify misinformation more strongly in our setting. We also compare recommendation policies and find that interest-based recommendation increases misinformation exposure compared with random exposure. Finally, we evaluate two forms of grounding intervention: allowing agents to verify claims through external web search, and introducing a real human participant into the simulation. Both interventions consistently reduce misinformation dominance, demonstrating that *OASIS* can serve not only as a simulation framework but also as an intervention-ready testbed for studying platform governance strategies. We summarize our key contributions as follows:

- **A platform-faithful simulation framework.** We introduce *OASIS*, a modular framework that couples evolving environments, platform-style action spaces, and recommendation-driven exposure to simulate social media ecosystems with LLM-based agents.
- **Scalable population-level simulation.** *OASIS* supports simulations from thousands to up to one million agents under practical compute budgets through parallel agent execution, concurrent environment updates, and high-throughput LLM serving.
- **Controlled mechanism and intervention studies.** *OASIS* enables researchers to systematically vary LLM backbones, recommendation policies, platform rules, external tools, and human participation, making it possible to study both emergent dynamics and mitigation strategies in reproducible settings.
- **Empirical findings on herding and misinformation.** Using Reddit- and X-like instantiations, we show that LLM-agent societies can exhibit stronger negative herding under early adverse feedback and that matched misinformation may diffuse more intensely than official news at large scales, while grounding interventions reduce misinformation dominance.

2 Method

This section describes the core abstraction and execution workflow of *OASIS*. The key idea is to factor a social platform simulator into *(i) an environment state store*, *(ii) an exposure mechanism* (recommendation), and *(iii) agent decision-making*, coordinated by an explicit time engine. This separation mirrors real platforms—where user actions are driven by what they are *shown*—and makes interventions and ablations well-scoped: one can modify exposure, or policies without entangling system components.

2.1 *OASIS* Workflow

OASIS consists of five components: an **Environment Server** (platform state), a **Recommendation System** (RecSys; exposure), an **Agent Module** (decision policy), a **Time Engine** (activity dynamics), and a **Scalable Inferencer** (inference execution layer).

	# Agent	Social Media Env.	Action Space	Recsys.	Dynamic Network	External Tool
Sotopia (Zhou et al. (2023b))	2	-	-	×	×	×
RecAgent (Wang et al. (2023))	5	-	6	✓	×	×
Agent4Rec (Zhang et al. (2024))	1,000	Movie Rec.	5	✓	×	×
S3 (Gao et al. (2023))	1,000	X	4	×	×	×
HiSim (Mou et al. (2024))	1,000	X	5	×	×	×
AgentTorch (Chopra et al. (2024))	8.4M*	-	-	×	✓	×
AgentScope (Gao et al. (2024))	**1M**	-	-	×	×	✓
AgentSociety† (Piao et al. (2025))	10k	Social Network	6	×	✓	×
***OASIS* (Ours)**	**1M**	**X & Reddit**	**21**	✓	✓	✓

Table 1: Comparison with prior LLM-based social media simulation frameworks. # Agent denotes the supported number of agents. Social Media Environment (Env.) indicates whether the framework instantiates a social media platform. Action Space counts the supported platform actions. Recsys. indicates whether recommendation-driven exposure is modeled. Dynamic Network denotes support for evolving social relationships. External Tool indicates whether agents can invoke external tools such as web search. AgentTorch (∗) uses LLMs to model agent archetypes, such as demographic groups, enabling large-scale population simulations with fewer LLM calls. †The comparison with AgentSociety is limited to its social media scenario.

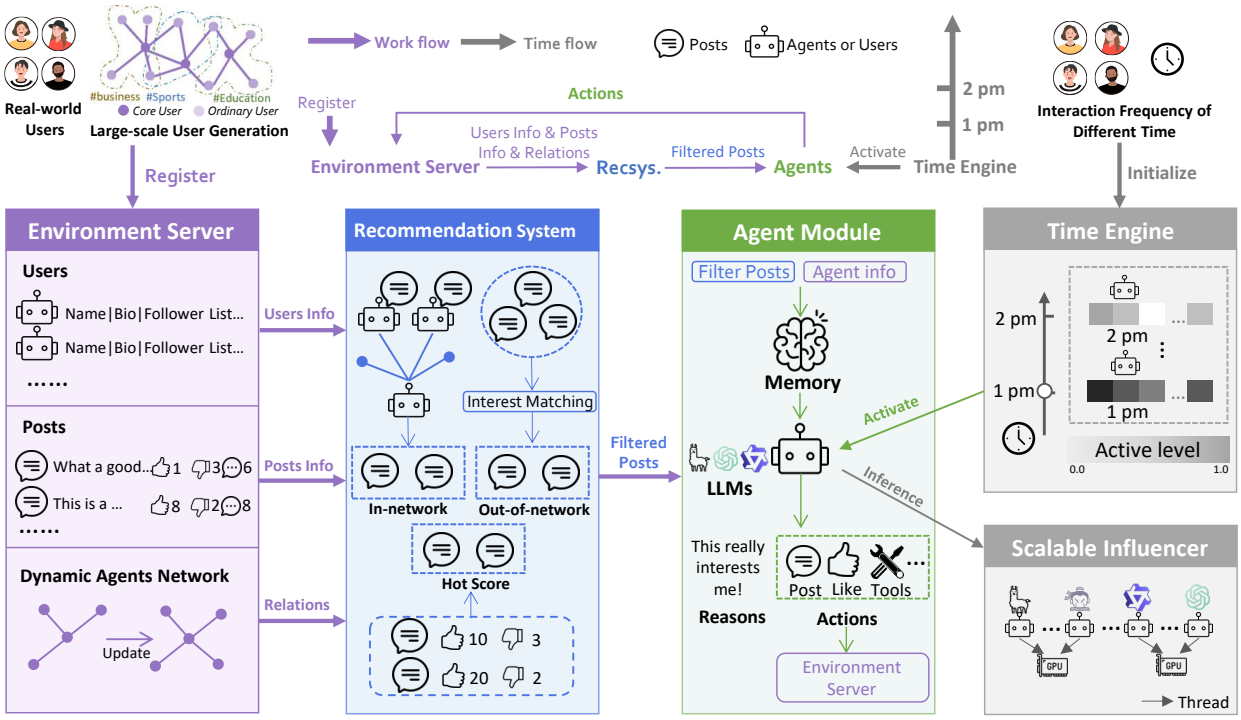


Figure 2: Overview of *OASIS*. During registration, agent metadata is stored by the Environment Server. During simulation, the Environment Server provides agent context, posts, and relations to RecSys, which constructs a feed based on social ties, interests, or popularity. Agents then produce actions (and rationales) conditioned on their feed and memory, and the state changes are written back to the Environment Server. The Time Engine governs step-wise activation, while the Scalable Inferencer decouples model inference from simulation logic.

Registration phase. *OASIS* initializes each agent with (i) basic metadata (e.g., name, self-description, historical posts), and (ii) a *role specification* that includes a character profile and action guideline. This

makes agent behavior *role-conditioned* rather than purely prompt-reactive, improving consistency across long trajectories.

Simulation phase. We adopt a discrete-time simulation. At each timestep, the Time Engine samples whether each agent is active according to its activation probability. For active agents, the Environment Server provides the relevant state (profile, history, relations, and recent traces) to RecSys, which returns a ranked set of recommended items (the agent's feed). The Agent Module then selects an action from the platform action space (e.g., posting, reacting, following) conditioned on the feed and its memory. After execution, the resulting changes (new posts, updated relations, reactions, and traces) are committed to the Environment Server. This closed loop makes the *exposure* → *action* → *state update* mechanism explicit, enabling controlled studies of platform-scale feedback dynamics.

2.2 Large-scale Agent Generation

We generate agents in three stages. First, demographic attributes (MBTI type, age, gender, country, and profession) are sampled according to the demographic distribution of Reddit users Duarte (2024). Second, we assign topics of interest from seven categories (Business, Culture & Society, Economics, Entertainment, General News, Information Technology, and Politics) using LLMs (e.g., GPT-3.5-Turbo and Llama3-70B), conditioned on each agent's demographics. Finally, we synthesize surface profiles—names, usernames, bios, and personas—by combining the sampled attributes and topics. Prompts are provided in Appendix D.1.

For Reddit-style simulations, explicit follower–following relations are not required. For X-style simulations, we instantiate a scale-free follower network Kwak (2010) with 196 core users drawn from Twitter15 Liu et al. (2015) and Twitter16 Ma et al. (2016), plus approximately one million additional agents generated as above. Each ordinary user follows core users with probability 0.2, which preserves exposure diversity without making the graph overly dense.

2.3 Environment Server

The Environment Server maintains the full platform state—user profiles, content, and relations—and exposes a consistent interface for RecSys and agents. We use a relational database to support efficient indexing, retrieval, and incremental updates; schemas are provided in Appendix E.2. The database includes: (i) a **user table** (names, bios, metadata), (ii) **post** and **comment tables** (content and metadata such as timestamps, reactions, and multimodal fields), (iii) a **relations component** (tables for follows, mutual connections, likes, etc.), (iv) a **trace table** (a complete log of each agent's actions and observations), and (v) a **recommendation table** populated by RecSys using traces. This design makes the simulator *stateful* and *traceable*: every recommendation and action is grounded in a queryable history, which is essential for reproducibility and intervention analysis.

2.4 Recommendation Module

In real social platforms, users interact not with the entire content pool, but with a filtered subset exposed by the recommendation system. Therefore, *OASIS* models recommendation as a first-class component that defines each agent's observable information context. We instantiate this component with two representative mechanisms, corresponding to Reddit-style hot ranking and X-style interest-based recommendation.

Reddit-style hot ranking. For Reddit, we follow the disclosed hot-ranking mechanism (Salihefendic, 2015), which jointly accounts for engagement and recency. Given a post with u upvotes and d downvotes, we define its net score as $s = u - d$ and compute its hot score as

$$h = \log_{10}\big(\max(|s|, 1)\big) + \text{sign}(s)\frac{t - t_0}{45000}, \tag{1}$$

where t denotes the submission time in seconds since the Unix epoch, and $t_0 = 1134028003$ is the reference timestamp used by Reddit's ranking rule. Posts are ranked in descending order of h, and the top-k posts are

exposed to the agent. The value of k is set according to the corresponding experimental protocol, as detailed in Appendix **??**.

X-style interest-based recommendation. For X, we construct each agent's feed from both in-network and out-of-network sources. In-network posts are sampled from users followed by the agent and ranked according to popularity signals, such as likes. Out-of-network posts are retrieved by matching candidate posts with the agent's interests. Specifically, we use TwHIN-BERT (Zhang et al., 2023) to encode the agent representation $\mathbf{E}_u$, derived from its profile and recent posts, and the candidate post representation $\mathbf{E}_{c_i}$. Their semantic relevance is measured by cosine similarity:

$$\text{sim}(u, c_i) = \frac{\mathbf{E}_u^\top \mathbf{E}_{c_i}}{\|\mathbf{E}_u\|_2 \|\mathbf{E}_{c_i}\|_2}. \tag{2}$$

To account for freshness, we further introduce a recency factor

$$r_i = \ln\left(e - \frac{t_{\text{current}} - t_{\text{created}}}{100}\right), \tag{3}$$

where t_{current} and t_{created} denote the current simulation timestep and the creation timestep of post c_i, respectively. The final recommendation score is computed as

$$\text{score}(u, c_i) = r_i \cdot \text{sim}(u, c_i), \tag{4}$$

and the top-k candidates are recommended to the agent. Additional implementation details are provided in Appendix E.3.

2.5 Agent Module, Time Engine, and Scalable Inferencer

Agent Module. Agents are LLM-based policies built upon CAMEL-style role prompting (Li et al., 2023). Each agent contains a **memory module** and an **action module** that operate in a closed loop with the platform.

Memory module. The agent stores (i) observed items (posts/comments with authors and reaction counts), (ii) its past actions and outcomes, and (iii) a brief step-level `reason` trace. We keep a recent buffer and compact older history into short summaries, so agents remain role-consistent over long horizons under a bounded context window.

Action module. Given a feed (or results from *search/trend/refresh*), the agent outputs a JSON object with a short `reason` and a list of function calls. The action API covers 21 platform interactions spanning content creation/propagation (*create_post*, *repost*, *quote_post*, *create_comment*), engagement (*like/dislike* and undo on posts/comments), social ties (*follow/unfollow*, *mute/unmute*), and exploration (*search_posts*, *search_user*, *trend*, *refresh*, *do_nothing*; Appendix E.1). The simulator executes these calls sequentially and updates the global state, enabling multi-step interaction.

Optional tool use. Agents may call external tools (e.g., web search, browser, code) via MCP-compatible interfaces Hou et al. (2025). Following ReAct Yao et al. (2022) and CoT Wei et al. (2022), tool outputs (when enabled) are incorporated before action selection to improve coherence and diversity.

Time Engine. The Time Engine defines platform-level temporal dynamics. At each timestep, it assigns (or samples) agent activation according to per-agent activation probabilities, thereby modeling heterogeneous activity rates and enabling controlled ablations over temporal intensity.

Scalable Inferencer. To keep the simulator logic independent of model-serving details, *OASIS* uses a decoupled inference layer. Agents submit inference requests asynchronously to an inference channel, and an inference manager schedules and aggregates these requests against the underlying vLLM servers Kwon et al. (2023b). This execution model supports high-throughput inference while preserving a clean abstraction boundary: the simulator only depends on *request/response semantics*, whereas batching, dispatch, and load balancing are handled by the inferencer. Additional details are provided in Appendix F.

3 Experiment

This section evaluates *OASIS* as a general-purpose, large-scale simulation framework by instantiating it into two representative social platform templates: **Reddit-style communities** (threaded discussions with community-level ranking) and **X-style follower networks** (broadcast posting with a personalized feed). Across these two platforms, we conduct large-scale simulations from thousands to one million agents.

Platform instantiations. Table 2 summarizes the two platform configurations used in our experiments, including the supported actions and the recommendation mechanism.

Platform	Action Space	Recsys.
Reddit	*like / dislike comment*; *like / dislike post*; *search posts / users*; *trend*, *refresh*, *do nothing*	**Hot-score ranking** (popularity-driven)
X	*create post*; *repost*; *like / dislike post*; *follow*	**Interest-based feed** (personalized recommendation)

Table 2: Two *OASIS* platform instantiations used in experiments. We vary both the *action space* and the *recommendation mechanism*, highlighting the configurability of *OASIS* across distinct social platforms.

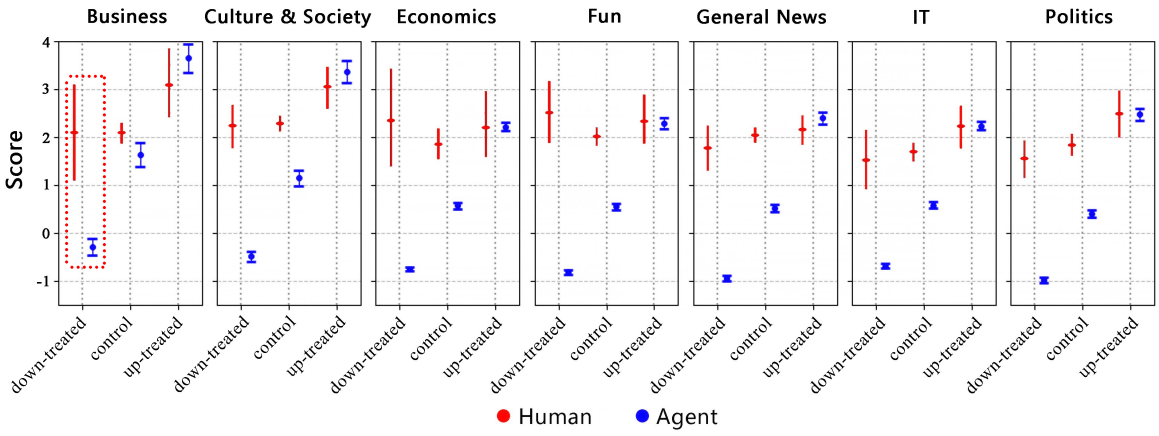


Figure 3: Mean comment scores with 95% confidence intervals for up-treated (initial like), down-treated (initial dislike), and control comments across seven topics. Red: humans. Blue: LLM agents.

3.1 Herd Effect on Reddit

Herd Effect. Herd effect refers to individuals following others' decisions rather than forming independent judgments, which can reduce autonomy and lead to suboptimal outcomes (Banerjee, 1992). Following prior work (Muchnik et al., 2013; Glenski & Weninger, 2016), we quantify herding by assigning otherwise similar comments into three groups: **down-treated** (one initial dislike), **control** (no initial feedback), and **up-treated** (one initial like). We then expose these comments to a large population of Reddit agents over multiple time steps and compare the resulting engagement across groups.

3.1.1 Negative social signals trigger stronger herding in LLM agents than humans

Setup. We reproduce the experimental protocol of Muchnik et al. (2013) using 116,932 real Reddit comments spanning seven topics (Figure 3). We generate 3,600 synthetic agent profiles (GPT-3.5 Turbo) and run topic-wise simulations with the same agent pool. The simulation lasts 40 time steps: at each step, an anonymous agent posts a set of initially-rated comments, and activated agents interact with the feed. We use

Llama3-8b-instruct AI@Meta (2024) as the base LLM. Agent activation probability is 0.1. The experiment runs on 4 A100 GPUs for about 4 hours.

Metric. We use *comment score* (upvotes minus downvotes), the standard visible metric on Reddit. We report the mean score and 95% confidence intervals for each group (up-/down-treated/control) across topics.

LLM agents exhibit stronger negative herding than humans. Figure 3 shows that in the **up-treated** condition, LLM agents and humans yield similar comment scores, indicating consistent amplification of positive social signals. In contrast, under the **down-treated** condition, humans maintain significantly higher scores than agents. This suggests that an initial negative signal triggers stronger follow-the-crowd behavior in LLM agents (more additional dislikes or fewer likes), whereas humans are more likely to re-evaluate cascading negativity.

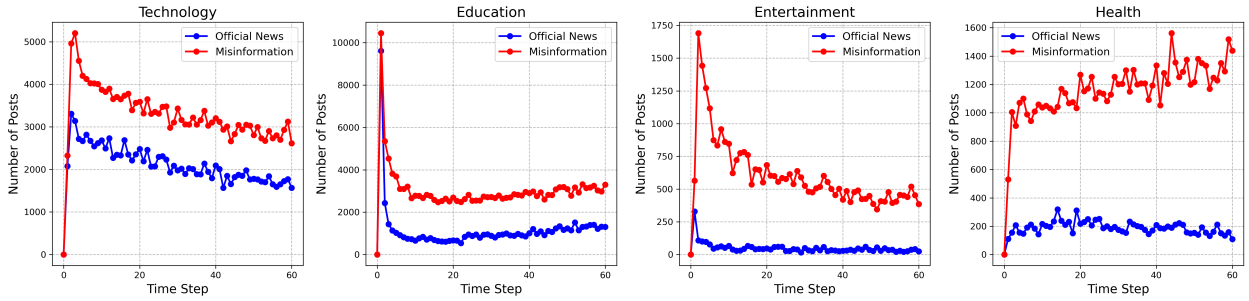


Figure 4: Number of posts related to official news vs. matched misinformation across time steps. Misinformation consistently induces more related posts across topics.

3.2 Large-Scale Misinformation Spreading on X

We next evaluate whether *OASIS* can capture large-scale information diffusion under an **X-style follower graph** and an **interest-based recommendation system** (Table 2). We focus on the relative influence and persistence of misinformation versus official news in a million-agent society.

Setup. We select four official news items (health, technology, entertainment, and education) and construct four matched misinformation items with comparable surface form (Appendix D.2.1). The simulated society consists of 196 core agents (high-follower users) and 1,000,000 regular agents (Section 2.2). For each topic, the same core agent posts both the true and false versions to ensure identical initial exposure. The simulation runs for 60 discrete time steps. Agents are activated stochastically with fixed probabilities (0.1 for core agents and 0.01 for regular agents), a controlled temporal abstraction that standardizes interaction opportunities and isolates platform-driven diffusion effects.

Metric (efficient large-scale attribution). To measure how much generated content is related to each news item, we use TF-IDF cosine similarity Christian et al. (2016) between the 8 news texts and 733,824 agent-generated posts. We mark a post as relevant if similarity ≥ 0.4. Accuracy analysis of this matching strategy is provided in Appendix G.2.

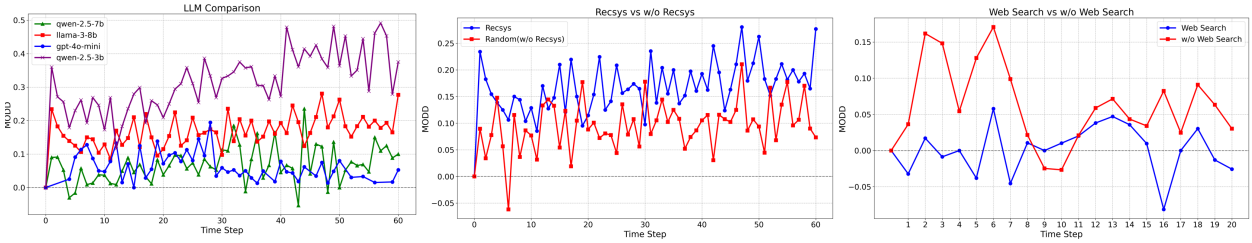
3.2.1 Misinformation sustains higher diffusion intensity than official news in agent society.

Figure 4 tracks the number of relevant posts over time. Across all topics, misinformation consistently induces *more* related posts than official news. Both curves rise rapidly early on (when relevant content is still scarce), then decline as attention shifts to other topics. Importantly, misinformation remains *persistently more active* even in later stages, indicating stronger and longer-lasting influence under the X-style interaction and recommendation dynamics.

Figure 5: Word clouds of posts related to official news vs. misinformation.

Qualitative analysis. Figure 5 further suggests distinct linguistic signatures: official news posts emphasize more factual and consequence-oriented expressions, while misinformation is associated with more affective, subjective, and socially engaging language. We also visualize the evolution of follow relations in Appendix G.1.

4 Ablation Study: What Drives and Mitigates Misinformation Dominance



(a) Different LLMs (b) Recsys vs w/o Recsys (b) Web Search vs w/o Web Search

Figure 6: Ablation of different LLM backbones, recommendation systems, and web search. The x-axis denotes simulation time steps, and the y-axis reports MODD, which measures the relative popularity of matched misinformation compared to official news, with higher values indicating stronger misinformation dominance.

We conduct ablation studies on the large-scale misinformation experiment (Section 3.2) to identify the *key mechanisms* that give rise to, amplify, or mitigate misinformation dominance in agent societies simulated by *OASIS*. Rather than exhaustively enumerating all configurable components, we focus on a small set of factors that correspond to distinct levels of influence: **agent-level behavior**, **network-level structure**, **platform-level exposure**, and **intervention-level grounding**.

Metric. Let $N_M(t)$ and $N_O(t)$ denote the numbers of posts related to misinformation and official news at time step t, and $N_{\text{total}}(t)$ the total number of posts. We define the normalized attention densities

$$D_M(t) = \frac{N_M(t)}{N_{\text{total}}(t)}, \quad D_O(t) = \frac{N_O(t)}{N_{\text{total}}(t)}, \tag{5}$$

and measure misinformation dominance by

$$\Delta_{MO}(t) = D_M(t) - D_O(t). \tag{6}$$

For concise comparison across settings, we report the mean dominance over the simulation horizon,

$$\text{MODD} = \frac{1}{T} \sum_{t=1}^{T} \Delta_{MO}(t), \tag{7}$$

with consistent trends observed when using the area under $\Delta_{MO}(t)$.

Ablation protocol. Unless otherwise specified, each ablation follows the setup in Section 3.2 but is scaled down to 1,000 agents for controlled comparison. All topic pairs, matching procedures, and simulation dynamics are held fixed.

4.1 What drives misinformation dominance?

Finding A: Misinformation dominance depends strongly on agent behavioral priors. We evaluate multiple LLM backbones as agent policies, including LLaMA-3-8B-Instruct, Qwen-2.5-7B, Qwen-2.5-3B, and GPT-4o-mini. As shown in Figure 6(a), the resulting misinformation dominance varies substantially across models. Notably, smaller models can exhibit stronger dominance than larger ones, indicating that susceptibility to misinformation is not a monotonic function of parameter count.

This result suggests that misinformation amplification is not an inevitable consequence of scale or network effects alone, but an emergent property of agent-level behavioral priors, such as how models interpret uncertainty, social endorsement, and emotionally charged content.

Finding B: Personalized recommendation sustains misinformation attention. We compare an interest-based personalized feed with a random feed under otherwise identical conditions. As shown in Figure 6(b), personalization substantially increases misinformation dominance.

While the Reddit experiment highlights the role of visible social signals at the individual level, this ablation demonstrates that algorithmic exposure mechanisms operate at the platform level, repeatedly presenting similar content to susceptible sub-populations. Such selective exposure allows misinformation to maintain attention over extended periods, even after overall activity declines.

4.2 What mitigates misinformation dominance?

Finding C: Tool-augmented verification mitigates misinformation dominance. We first examine whether access to external evidence can reduce misinformation dominance in agent societies. Concretely, we enable agents to optionally verify claims via web search before interacting with content, while keeping the personalized recommendation feed unchanged.

As shown in Figure 6(c), tool-augmented verification substantially reduces $\Delta_{MO}(t)$ compared to the no-verification setting. This result indicates that even lightweight verification steps introduce sufficient friction to slow down rapid resharing and weaken self-reinforcing misinformation loops. Importantly, the effectiveness of this intervention suggests that misinformation dominance is not solely driven by exposure or scale, but is sensitive to agents' ability to ground their actions in external factual references.

Finding D: Human participation provides deliberative grounding that further suppresses misinformation cascades. We next inject a real human participant into an ongoing agent society, allowing direct engagement with the discussion thread. As shown in Figure 7, the introduction of a human consistently decreases misinformation dominance over time.

Unlike tool-based verification, human participation introduces a different form of grounding. Rather than supplying external facts, humans tend to question, reinterpret, or contextualize content, thereby disrupting purely imitative dynamics among agents. This finding aligns with our Reddit experiment, where humans exhibit weaker susceptibility to negative social signals, and highlights the stabilizing role of deliberative human input.

5 Conclusion

We present *OASIS*, a generalizable and scalable social media simulator designed to replicate real-world social media dynamics. *OASIS* incorporates modular components that capture the core functionalities of various social media platforms, allowing it to be easily adapted for different contexts. Furthermore, *OASIS* supports large-scale user interactions, accommodating up to one million users. Using *OASIS*, we simulated herd behavior on Reddit and the spread of misinformation on X (formerly Twitter). The simulations involved

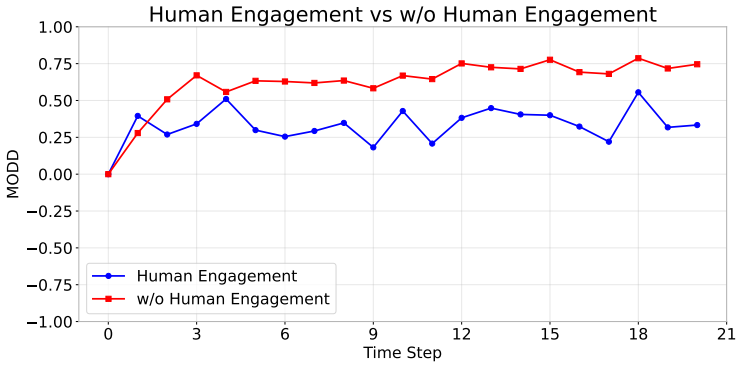


Figure 7: Ablation of human-in-the-loop. Human participation consistently reduces the popularity of misinformation over time.

agent populations ranging from thousands to millions. We hope that *OASIS* will serve as a platform for researching large-scale agent societies.

6 Ethic Statement

The data used in our experiments were sourced from publicly available datasets and the official Twitter API. All data collection processes were conducted in full compliance with relevant regulations and guidelines. Additionally, any real user data used in our simulations was anonymized to ensure privacy and confidentiality. Only simulated user data will be made publicly available.

Our simulations were conducted within a controlled environment, where the experiments were designed to simulate large-scale behaviors without causing any direct impact on actual social media users. We took extensive precautions to ensure that the experiments did not result in unintended harm or influence on real-world online communities.

We acknowledge that the development of agent-based models and social media simulators may raise ethical concerns, particularly regarding the potential for misuse or the amplification of harmful behaviors. However, we believe that understanding the dynamics of agent-based societies, especially on social media platforms, is crucial for identifying and mitigating security risks, disinformation spread, and other negative phenomena in the digital space. The tool we have developed, *OASIS*, serves as a proactive measure to analyze and address potential threats before they manifest in real-world settings. Ultimately, our platform has the potential to contribute to enhancing the safety and security of human social interactions in the digital age by allowing researchers and policymakers to anticipate and mitigate harmful group behaviors.

We are committed to the responsible use of AI and data and to fostering a deeper understanding of the ethical implications of agent-based modeling and large-scale social media simulations.

7 Limitations

Despite the advancements in our model's ability to perform complex actions, there are still limitations in its capacity to effectively simulate social media behavior. For instance, the model occasionally exhibits repetitive behaviors, such as agents re-posting or re-commenting content that they have previously interacted with, which diminishes the realism of the simulation.

Another limitation is the computational cost associated with large-scale simulations, particularly due to the need for large language model (LLM) inference. As the scale of the simulation grows, the computational demands increase significantly, which can limit the feasibility of running large-scale simulations with high efficiency. Further optimizations are needed to improve the scalability of the platform.

References

AI@Meta. Llama 3 model card. 2024. URL https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md.

Anthropic. Introducing claude opus 4.6. https://www.anthropic.com/news/claude-opus-4-6, February 2026. Accessed: 2026-04-28.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

A. Banerjee. A simple model of herd behavior. *Quarterly Journal of Economics*, 107:797–817, 1992. doi: 10.2307/2118364.

Canyu Chen and Kai Shu. Combating misinformation in the age of llms: Opportunities and challenges. *AI Magazine*, 2023.

Ayush Chopra, Alexander Rodriguez, B Aditya Prakash, Ramesh Raskar, and Thomas Kingsley. Using neural networks to calibrate agent based models enables improved regional evidence for vaccine strategy and policy. *Vaccine*, 41(48):7067–7071, 2023.

Ayush Chopra, Shashank Kumar, Nurullah Giray-Kuru, Ramesh Raskar, and Arnau Quera-Bofarull. On the limits of agency in agent-based models. *arXiv preprint arXiv:2409.10568*, 2024.

Hans Christian, Mikhael Pramodana Agus, and Derwin Suhartono. Single document automatic text summarization using term frequency-inverse document frequency (tf-idf). *ComTech: Computer, Mathematics and Engineering Applications*, 7(4):285–294, 2016.

Francis B Colavita. Human sensory dominance. *Perception & Psychophysics*, 16(2):409–412, 1974.

Fabio Duarte. Reddit user age, gender, & demographics (2024). https://explodingtopics.com/blog/reddit-users, 2024. Accessed: 2024-09-28.

Antonino Ferraro, Antonio Galli, Valerio La Gatta, Marco Postiglione, Gian Marco Orlando, Diego Russo, Giuseppe Riccio, Antonio Romano, and Vincenzo Moscato. Agent-based modelling meets generative ai in social network simulations. In *International Conference on Advances in Social Networks Analysis and Mining*, pp. 155–170. Springer, 2024.

Chen Gao, Xiaochong Lan, Zhihong Lu, Jinzhu Mao, Jinghua Piao, Huandong Wang, Depeng Jin, and Yong Li. S 3: Social-network simulation system with large language model-empowered agents. *ArXiv preprint*, abs/2307.14984, 2023. URL https://arxiv.org/abs/2307.14984.

Dawei Gao, Zitao Li, Weirui Kuang, Xuchen Pan, Daoyuan Chen, Zhijian Ma, Bingchen Qian, Liuyi Yao, Lin Zhu, Chen Cheng, et al. Agentscope: A flexible yet robust multi-agent platform. *ArXiv preprint*, abs/2402.14034, 2024. URL https://arxiv.org/abs/2402.14034.

M. Glenski and Tim Weninger. Rating effects on social news posts and comments. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 8:1–19, 2016. doi: 10.1145/2963104.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Gauri Gupta, Ritvik Kapila, Ayush Chopra, and Ramesh Raskar. First 100 days of pandemic; an interplay of pharmaceutical, behavioral and digital interventions–a study using agent based modeling. *arXiv preprint arXiv:2401.04795*, 2024.

Xinyi Hou, Yanjie Zhao, Shenao Wang, and Haoyu Wang. Model context protocol (mcp): Landscape, security threats, and future research directions, 2025. URL https://arxiv.org/abs/2503.23278.

Jen-tse Huang, Eric John Li, Man Ho Lam, Tian Liang, Wenxuan Wang, Youliang Yuan, Wenxiang Jiao, Xing Wang, Zhaopeng Tu, and Michael R Lyu. How far are we on the decision-making of llms? evaluating llms' gaming ability in multi-agent environments. *ArXiv preprint*, abs/2403.11807, 2024. URL https://arxiv.org/abs/2403.11807.

Luca Iandoli, Simonetta Primario, and Giuseppe Zollo. The impact of group polarization on the quality of online debate in social media: A systematic literature review. *Technological Forecasting and Social Change*, 170:120924, 2021.

Daniel J Isenberg. Group polarization: A critical review and meta-analysis. *Journal of personality and social psychology*, 50(6):1141, 1986.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. *arXiv preprint arXiv:2412.16720*, 2024.

Kawaljeet Kaur Kapoor, Kuttimani Tamilmani, Nripendra P Rana, Pushp Patil, Yogesh K Dwivedi, and Sridhar Nerur. Advances in social media research: Past, present and future. *Information Systems Frontiers*, 20:531–558, 2018.

H Kwak. What is twitter, a social network or a news media? *Department of Computer Science, KAIST*, 2010.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*, 2023a.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*, 2023b.

James Ladyman, James Lambert, and Karoline Wiesner. What is a complex system? *European Journal for Philosophy of Science*, 3:33–67, 2013.

Guohao Li, Hasan Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. Camel: Communicative agents for" mind" exploration of large language model society. *Advances in Neural Information Processing Systems*, 36:51991–52008, 2023.

Ruibo Liu, Ruixin Yang, Chenyan Jia, Ge Zhang, Denny Zhou, Andrew M Dai, Diyi Yang, and Soroush Vosoughi. Training socially aligned language models on simulated social interactions. *ArXiv preprint*, abs/2305.16960, 2023. URL https://arxiv.org/abs/2305.16960.

Xiaomo Liu, Armineh Nourbakhsh, Quanzhi Li, Rui Fang, and Sameena Shah. Real-time rumor debunking on twitter. In James Bailey, Alistair Moffat, Charu C. Aggarwal, Maarten de Rijke, Ravi Kumar, Vanessa Murdock, Timos K. Sellis, and Jeffrey Xu Yu (eds.), *Proceedings of the 24th ACM International Conference on Information and Knowledge Management, CIKM 2015, Melbourne, VIC, Australia, October 19 - 23, 2015*, pp. 1867–1870. ACM, 2015. doi: 10.1145/2806416.2806651. URL https://doi.org/10.1145/2806416.2806651.

Jing Ma, Wei Gao, Prasenjit Mitra, Sejeong Kwon, Bernard J. Jansen, Kam-Fai Wong, and Meeyoung Cha. Detecting rumors from microblogs with recurrent neural networks. In Subbarao Kambhampati (ed.), *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI 2016, New York, NY, USA, 9-15 July 2016*, pp. 3818–3824. IJCAI/AAAI Press, 2016. URL http://www.ijcai.org/Abstract/16/537.

Charles M Macal, Nicholson T Collier, Jonathan Ozik, Eric R Tatara, and John T Murphy. Chisim: An agent-based simulation model of social interactions in a large urban area. In *2018 winter simulation conference (WSC)*, pp. 810–820. IEEE, 2018.

Moltbook. Moltbook: Ai agent social network platform. https://moltbookcn.cn/, 2026. Accessed: 2026-04-27.

Megan A Moreno, Natalie Goniu, Peter S Moreno, and Douglas Diekema. Ethics of social media research: Common concerns and practical considerations. *Cyberpsychology, behavior, and social networking*, 16(9): 708–713, 2013.

Manuel Mosquera, Juan Sebastian Pinzon, Manuel Rios, Yesid Fonseca, Luis Felipe Giraldo, Nicanor Quijano, and Ruben Manrique. Can llm-augmented autonomous agents cooperate?, an evaluation of their cooperative capabilities through melting pot. *ArXiv preprint*, abs/2403.11381, 2024. URL https://arxiv.org/abs/2403.11381.

Xinyi Mou, Zhongyu Wei, and Xuanjing Huang. Unveiling the truth and facilitating change: Towards agent-based large-scale social movement simulation. *ArXiv preprint*, abs/2402.16333, 2024. URL https://arxiv.org/abs/2402.16333.

Lev Muchnik, Sinan Aral, and Sean J Taylor. Social influence bias: A randomized experiment. *Science*, 341 (6146):647–651, 2013.

Nature Reviews Psychology. Social media needs science-based guidelines. *Nature Reviews Psychology*, 3(6): 367–367, Jun 2024. ISSN 2731-0574. doi: 10.1038/s44159-024-00327-8. URL https://doi.org/10.1038/s44159-024-00327-8.

Candice L Odgers. The great rewiring: is social media really behind an epidemic of teenage mental illness? *Nature*, 628(8006):29–30, 2024.

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*, pp. 1–22, 2023.

Jinghua Piao, Yuwei Yan, Jun Zhang, Nian Li, Junbo Yan, Xiaochong Lan, Zhihong Lu, Zhiheng Zheng, Jing Yi Wang, Di Zhou, et al. Agentsociety: Large-scale simulation of llm-driven generative agents advances understanding of human behaviors and society. *arXiv preprint arXiv:2502.08691*, 2025.

Michael I Posner, Mary J Nissen, and Raymond M Klein. Visual dominance: an information-processing account of its origins and significance. *Psychological review*, 83(2):157, 1976.

Pushshift. Pushshift reddit 2023-03. https://archive.org/details/pushshift-reddit-2023-03, 2023. Accessed: 2024-09-28.

Chen Qian, Xin Cong, Cheng Yang, Weize Chen, Yusheng Su, Juyuan Xu, Zhiyuan Liu, and Maosong Sun. Communicative agents for software development. *ArXiv preprint*, abs/2307.07924, 2023. URL https://arxiv.org/abs/2307.07924.

Reddit, Inc. Reddit. https://www.reddit.com. Accessed: 2025-10-09.

Siyue Ren, Zhiyao Cui, Ruiqi Song, Zhen Wang, and Shuyue Hu. Emergence of social norms in large language model-based agent societies. *ArXiv preprint*, abs/2403.08251, 2024. URL https://arxiv.org/abs/2403.08251.

Amir Salihefendic. How reddit ranking algorithms work, Dec 2015. URL https://medium.com/hacking-and-gonzo/how-reddit-ranking-algorithms-work-ef111e33d0d9.

Thomas C Schelling. Models of segregation. *The American economic review*, 59(2):488–493, 1969.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

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

Lei Wang, Jingsen Zhang, Hao Yang, Zhiyuan Chen, Jiakai Tang, Zeyu Zhang, Xu Chen, Yankai Lin, Ruihua Song, Wayne Xin Zhao, et al. User behavior simulation with large language model based agents. *ArXiv preprint*, abs/2306.02552, 2023. URL https://arxiv.org/abs/2306.02552.

Yulong Wang, Tianhao Shen, Lifeng Liu, and Jian Xie. Sibyl: Simple yet effective agent framework for complex real-world reasoning. *ArXiv preprint*, abs/2407.10718, 2024. URL https://arxiv.org/abs/2407.10718.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, Yuxiang Chen, Zecheng Tang, Zekai Zhang, Zhengyi Wang, An Yang, Bowen Yu, Chen Cheng, Dayiheng Liu, Deqing Li, Hang Zhang, Hao Meng, Hu Wei, Jingyuan Ni, Kai Chen, Kuan Cao, Liang Peng, Lin Qu, Minggang Wu, Peng Wang, Shuting Yu, Tingkun Wen, Wensen Feng, Xiaoxiao Xu, Yi Wang, Yichang Zhang, Yongqiang Zhu, Yujia Wu, Yuxuan Cai, and Zenan Liu. Qwen-image technical report, 2025. URL https://arxiv.org/abs/2508.02324.

X Corp. X (formerly Twitter). https://x.com. Accessed: 2025-10-09.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*, 2022.

An Zhang, Yuxin Chen, Leheng Sheng, Xiang Wang, and Tat-Seng Chua. On generative agents in recommendation. In *Proceedings of the 47th international ACM SIGIR conference on research and development in Information Retrieval*, pp. 1807–1817, 2024.

Xinnong Zhang, Jiayu Lin, Xinyi Mou, Shiyue Yang, Xiawei Liu, Libo Sun, Hanjia Lyu, Yihang Yang, Weihong Qi, Yue Chen, et al. Socioverse: A world model for social simulation powered by llm agents and a pool of 10 million real-world users. *arXiv preprint arXiv:2504.10157*, 2025.

Xinyang Zhang, Yury Malkov, Omar Florez, Serim Park, Brian McWilliams, Jiawei Han, and Ahmed El-Kishky. Twhin-bert: A socially-enriched pre-trained language model for multilingual tweet representations at twitter. In *Proceedings of the 29th ACM SIGKDD conference on knowledge discovery and data mining*, pp. 5597–5607, 2023.

Andrew Zhao, Daniel Huang, Quentin Xu, Matthieu Lin, Yong-Jin Liu, and Gao Huang. Expel: Llm agents are experiential learners. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 19632–19642, 2024.

Shuyan Zhou, Frank F Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, et al. Webarena: A realistic web environment for building autonomous agents. *ArXiv preprint*, abs/2307.13854, 2023a. URL https://arxiv.org/abs/2307.13854.

Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, et al. Sotopia: Interactive evaluation for social intelligence in language agents. *ArXiv preprint*, abs/2310.11667, 2023b. URL https://arxiv.org/abs/2310.11667.

A Related Work

A.1 Social Media

Social media encompasses websites and applications focused on communication, interaction, and content-sharing (Kapoor et al., 2018). While it offers benefits like allowing individuals to explore their identities without real-world consequences (Nature Reviews Psychology, 2024), the risk of hazardous social media phenomena gradually becomes a global threat with significant economic, political, and social consequences. Traditional threats includes promoting risky behaviors (Nature Reviews Psychology, 2024), contributing to mental health issues among teenagers (Odgers, 2024), social influence (Muchnik et al., 2013), group polarization (Iandoli et al., 2021; Isenberg, 1986), and spreading misinformation (Vosoughi et al., 2018). Despite numerous studies on social media phenomena, the complex network structures, vast data, and diverse behaviors present challenges for researchers. Additionally, ethical concerns (Moreno et al., 2013) arise in some of these studies. To address these issues, a controllable virtual environment (*e.g.*, a multi-agent system) for social simulation is needed, allowing researchers to test hypotheses on a virtual platform.

A.2 Multi-Agent Systems

Multi-agent systems are composed of multiple autonomous entities, each possessing different information and diverging interests. Compared to single-agent platforms, multi-agent platforms offer several advantages, including (1) the ability to assume different roles in group activities, and (2) richer and more complex interaction behaviors, such as collaboration, discussion, and strategic competition. Recent studies have demonstrated the potential of multi-agent systems across various domains.

Divided by various functionality, recent multi-agent systems can be roughly divided to tool-based agent assistants (Qian et al., 2023; Zhao et al., 2024; Mosquera et al., 2024; Wang et al., 2024), as well as society or game simulation environments (Li et al., 2023; Zhou et al., 2023a; Huang et al., 2024). The former part focus on collaborating a small group of LLM-based agents to automatically conduct predefined or open-ended tasks. And the latter part focus on involving a large-scale agent groups to automatically run a simulator in a specific environment. Since the action and relationship in a large society is extremely complicated, capability scalability has become the fundamental issue of this work. In this work, we highly focus on leveraging multi-agent systems to explore corresponding characteristics in social simulation research.

A.3 Multi-Agent System Social Simulation

Social simulation plays a crucial role in social science research, with many classic agent-based modeling (ABM) studies, such as Schelling's model of segregation (Schelling, 1969), the Chicago simulation (Macal et al., 2018), and the pandemic (Gupta et al., 2024; Chopra et al., 2023). Traditional ABM has limitations such as subjective rule design and scalability issues. With the development of large language models (LLMs), LLM-based agents have demonstrated significant advantages in social simulation: (1) The ability to interact using natural language. (2) A more accurate simulation of human behavior. (3) The capability to utilize more complex tools. There have been numerous related studies, such as the exploration of multi-agent behavior patterns (Park et al., 2023), simulations of social networks (Gao et al., 2023; Zhou et al., 2023b), and the study of society's response to misinformation (Chen & Shu, 2023). Social simulation not only serves as a tool for social science research but also aids in exploring the boundaries of LLMs' capabilities. For example, studies on social alignment (Liu et al., 2023), emergence of social norms (Ren et al., 2024). However, current LLM-related social simulations mainly focus on interactions among a small number of agents. Yet, research on collective behavior often requires a critical mass to observe emergent phenomena. Therefore, our work emphasizes the interaction of large-scale agents to study the emergence of collective behaviors.

B Experiments Details

B.1 Herd Effect

B.1.1 Metrics

We utilized two primary metrics to assess the herd effect: the post score and the disagree score. These metrics were derived from two aspects: the behavior of liking or disliking by the LLM Agent, and the content of the comments it generated.

- **Comment Score**: The score (S_i) of a Reddit comment is the difference between its upvotes (U_i) and downvotes (D_i), and can be negative:

$$S_i = U_i - D_i$$

  where S_i is the score of the ith post, U_i the number of upvotes, and D_i the number of downvotes.

- **Disagree Score**: In this experiment 3.2.1, where all posts were counterfactual, we examined comments from LLM agents at various time points. Using GPT-4O, we assessed how much these comments disagreed with the posts. The evaluation was guided by the following prompt template:

PROMPT

Someone posted a counterfactual statement: "{`post_content`}", and then another person commented: "{`comment_content`}".

Please provide a rating from 1 to 10 to indicate how much this person agrees with the counterfactual content, where: 1 means fully believing and agreeing with it, and 10 means completely disbelieving and opposing it.

Respond only with a JSON object like this:
{
  "score": 7
}

To validate the effectiveness of the disagree score, we conducted a disagreement level scoring test on 50 cases. The table 3 below presents the mean squared error (MSE) of scores between two models or two humans across all questions. The results indicate that GPT-4o and Qwen-Plus exhibit similar scoring patterns, as do the two human annotators. However, the discrepancy between human and large model scores is relatively larger.

Table 3: MSE of disagreement scores among models and humans

	GPT-4o	Qwen-plus	Human-1	Human-2
GPT-4o	-	8.520	19.880	42.860
Qwen-plus	8.520	-	13.480	27.700
Human-1	19.880	13.480	-	7.600
Human-2	42.860	27.700	7.600	-

Before the experiment began, we randomly divided the comments data for experiment 3.1.1 and the posts data for experiment 3.2.1 into three groups (up-treated, down-treated, and control). After the experiment concluded, we calculated the mean **comment score** and the 95% confidence interval of the normal distribution for all posts in the three groups of experiment 3.2.1. Similarly, we calculated the mean **disagree score** and the 95% confidence interval of the normal distribution at each time step for all comments associated with posts in the three groups of experiment 3.2.1.

B.1.2 Setting Details

Human-Agent Comparison Our experiment 3.1.1 replicated the setup of a human study, including the visibility of comment scores (upvotes minus downvotes) and prohibiting the revocation of likes and dislikes, utilizing Reddit's popularity-based recommendation algorithm. To minimize biases stemming from the identities of commenters and voters and their interactions, which were meticulously accounted for in the human experiments, we manipulated a specific user to post content at scheduled intervals. This approach was adopted to mitigate the influence of different posters on the behavior of agents, and we further circumvented the impact of relationships with specific posting users on the outcomes by prohibiting agents from following or muting operations.

Consequently, the action space for the experiment included actions: like comment, dislike comment, like post, dislike post, search posts, search users, trend, refresh, and do nothing. The controlled user generated 200 posts at each time step, with each post accompanied by 1-10 comments. The recommendation system cached the top 300 posts with the highest heat scores for each agent, and each agent had a 0.1 probability of activation at every time step. Activated agents would randomly sample one of these 300 posts to read during that time step. The experiment was conducted over a total of 40 time steps.

Herd Effect in Counterfactual Content The action space of the experiment 3.2.1 includes create comment, like comment, dislike comment, like post, dislike post, search posts, search users, trend, refresh, and do nothing. Each agent has a 0.1 probability of activation at each time step, and each activated agent will randomly sample 5 posts from the recommended cache to read during that time step. As the number of agents increases from 100, 1k to 10k, the number of posts cached by the recommendation system respectively becomes 50, 500, and 5000. The controlled user creates 30, 300, 3k posts at each time step, respectively, until all posts in the corresponding datasets (with 219, 2191, and 21919 posts, respectively) have been created. And the experiment was conducted over a total of 30 time steps. For more details, see Fig.8.

B.1.3 Examples of Results

In experiment 3.2.1, 10,000 agents were able to discuss their views on counterfactual posts in the comment section, interacting by posting their own comments or by liking or disliking others' comments. Over the course of the discussion, there was a gradual shift towards opposing the counterfactual content, achieving factual correction at the group level. The figure 9 below shows one such example.

B.2 Web Search Prompt

Web Search Prompt

First, you should search for information about the posts using Linkup toolkit. Then, you should perform the action that best reflects your current inclination based on your profile and posts content. Do not limit your action in just 'like' to like posts.

Description of the linkup toolkits

Search for a query in the Linkup API and return results in various formats. The Linkup API's '/search' endpoint allows you to discover and access relevant web content based on a natural language query. It retrieves online information to ground your Large Language Model's (LLM) responses, enhancing their accuracy and reliability.

B.3 Agent Generated Images

See Figure10.

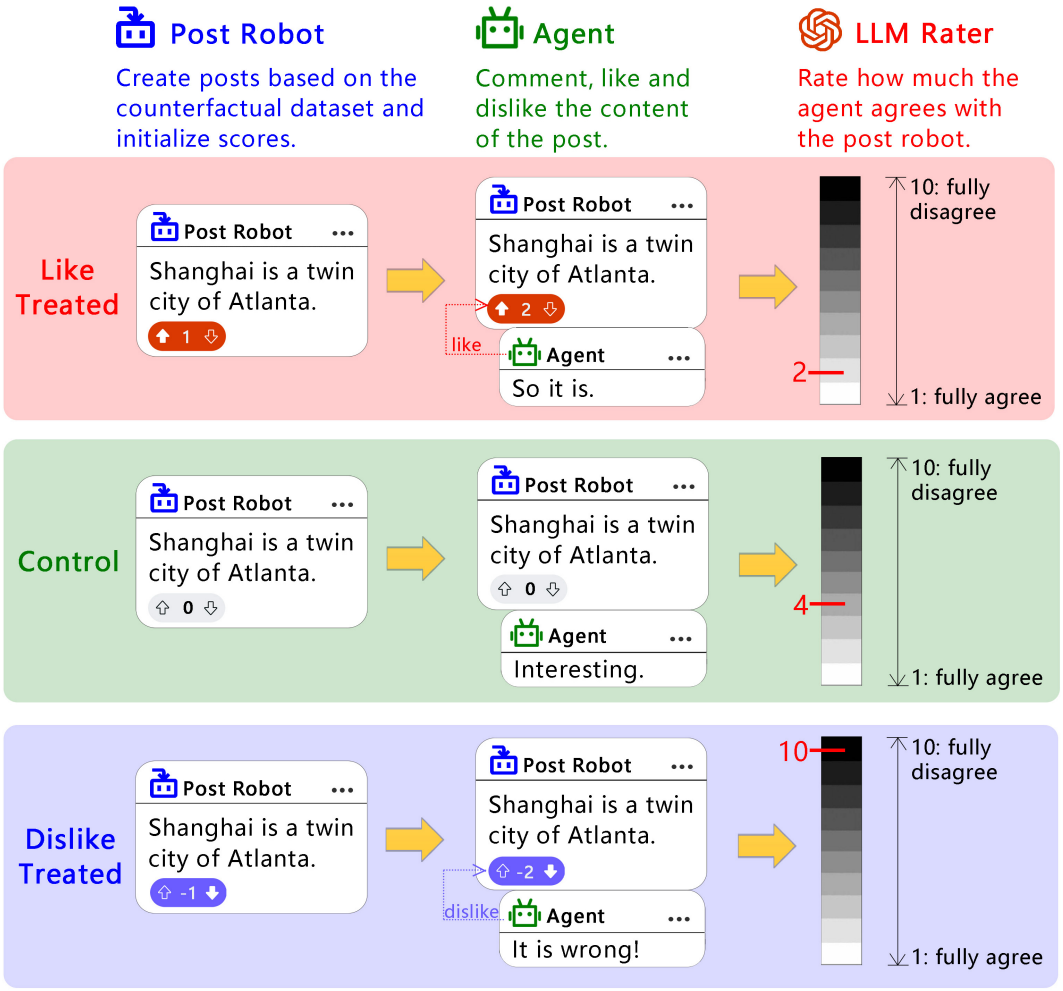


Figure 8: Schematic Diagram of the Experimental Setting for the Herd Effect in Counterfactual Content

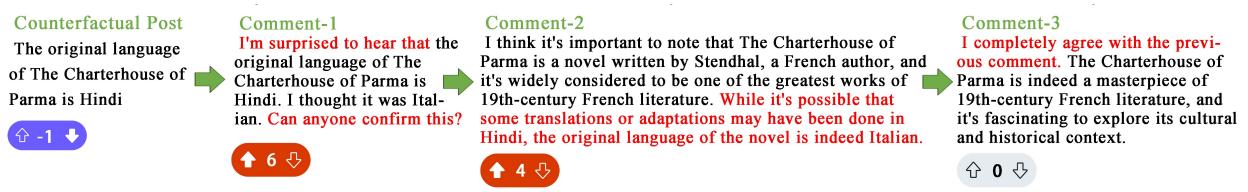


Figure 9: Example of agents' comments on counterfactual posts. As interactions increase, agents' viewpoints gradually shift from surprise and curiosity, to partial opposition, and finally to complete rejection of the counterfactual content.

Figure 10: Sample images automatically generated by agents for their social media posts

C More Ablation Study

C.1 Comparison of Different Models

Experiment Setting and Setup We compared the performance of three different models: Llama3-8b-instruct, Qwen2.5-7B-Instruct-1M, and GPT-4o-mini in Experiment 3.1.1. To control computational cost, we used 2,000 posts, each with 1 to 10 corresponding comments, 3,600 agents, and ran the experiment for 10 time steps. The grouping of comments, initialization scores, content creation method for anonymous users, and other factors were the same as in Experiment 3.1.1.

Evaluation Metric *Comment score* is a metric displayed to users on the real Reddit platform, calculated as the difference between the number of upvotes and downvotes. We compare the average *comment score* and confidence intervals among three groups across different topics.

C.2 Impact of Multimodal Information

Finding: Multimodal Content Does Not Influence Misinformation Spreading in Agents. To examine the impact of multimodal content on information dynamics, we tested agents based on the `Qwen-Image-Plus` Wu et al. (2025) model with image-text pairs of both misinformation and truth. Agents either read images and generated text or read images and generated new images. Figure 11 shows that multimodal content did not significantly affect the agents' spreading behavior. This lack of mitigation, unlike in humans, is likely because VLMs exhibit learned cross-modal attention patterns that differ from humans' innate visual bias Posner et al. (1976); Colavita (1974). Examples of agent-generated images are provided in Figure 12.

All models are more prone to the herd effect than humans, with GPT-4o-mini being particularly susceptible to downvoting. As shown in the Fig.12, Humans tend to have similar final scores in the down-treated and control groups. However, LLM agents, based on different models, are affected by the down-treated initialization settings, which result in fewer upvotes or more downvotes, leading to lower final scores for the down-treated group. Furthermore, agents based on GPT-4o-mini score lower than those of other models, with the average score in the down-treated group even falling below the initial score of the content by -1, indicating that GPT-4o-mini agents tend to follow the herd and downvote posts that are already being downvoted.

Figure 11: Ablation of Multimodal

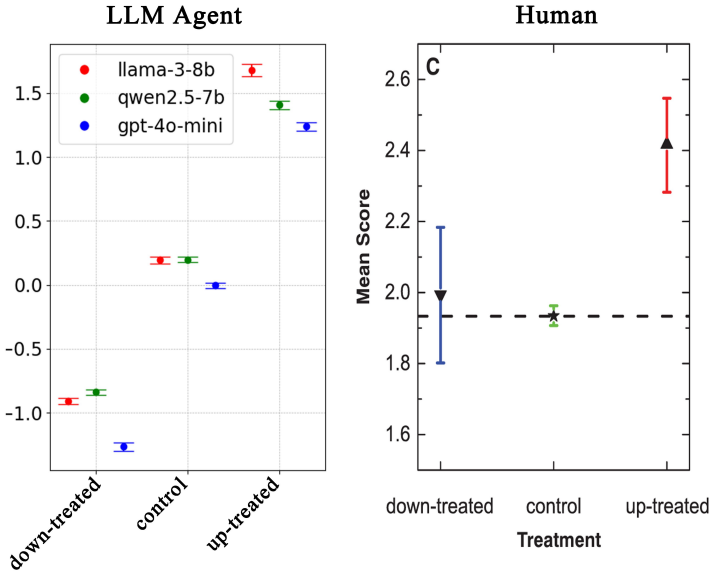


Figure 12: The figure displays the mean comment scores for up-treated comments (initially liked), down-treated comments (initially disliked), and control group comments (with no likes or dislikes), along with 95% confidence intervals for both humans (from Muchnik et al. (2013)) and LLM agents based on different models.

D Data

D.1 Herd Effect

User Generation. In our Reddit experiment, the process of generating users is divided into three main steps. Initially, we reference the actual demographic distribution of Reddit users (Duarte, 2024), assigning demographic information such as MBTI, age, gender, country, and profession to each user through random sampling. Subsequently, we employ GPT-3.5 Turbo to select topics of potential interest to the users based on the aforementioned information, choosing from seven categories: Business, Culture & Society, Economics, Fun, General News, IT, and Politics. Finally, using demographic information and selected topics, GPT-3.5

Turbo is utilized to generate each user's real name, username, bio, and persona. The generation prompts for the second and third parts are as follows.

PROMPT

Prompt of Step-2
Based on the provided personality traits, age, gender and profession, please select 2-3 topics of interest from the given list.

Input:
Personality Traits: {mbti}
Age: {age}
Gender: {gender}
Country: {country}
Profession: {profession}

Available Topics:

1. Economics: The study and management of production, distribution, and consumption of goods and services. Economics focuses on how individuals, businesses, governments, and nations make choices about allocating resources to satisfy their wants and needs, and tries to determine how these groups should organize and coordinate efforts to achieve maximum output.

2. IT (Information Technology): The use of computers, networking, and other physical devices, infrastructure, and processes to create, process, store, secure, and exchange all forms of electronic data. IT is commonly used within the context of business operations as opposed to personal or entertainment technologies.

3. Culture & Society: The way of life for an entire society, including codes of manners, dress, language, religion, rituals, norms of behavior, and systems of belief. This topic explores how cultural expressions and societal structures influence human behavior, relationships, and social norms.

4. General News: A broad category that includes current events, happenings, and trends across a wide range of areas such as politics, business, science, technology, and entertainment. General news provides a comprehensive overview of the latest developments affecting the world at large.

5. Politics: The activities associated with the governance of a country or other area, especially the debate or conflict among individuals or parties having or hoping to achieve power. Politics is often a battle over control of resources, policy decisions, and the direction of societal norms.

6. Business: The practice of making one's living through commerce, trade, or services. This topic encompasses the entrepreneurial, managerial, and administrative processes involved in starting, managing, and growing a business entity.

7. Fun: Activities or ideas that are light-hearted or amusing. This topic covers a wide range of entertainment choices and leisure activities that bring joy, laughter, and enjoyment to individuals and groups.

Output:
`[list of topic numbers]`
Ensure your output could be parsed to **list**, don't output anything else.

Prompt of Step-3
Please generate a social media user profile based on the provided personal information, including a real name, username, user bio, and a new user persona. The focus should be on creating a fictional

background story and detailed interests based on their hobbies and profession.
Input:
age: {age}
gender: {gender}
mbti: {mbti}
profession: {profession}
interested topics: {topics}
Output:
{{
"realname": "str",
"username": "str",
"bio": "str",
"persona": "str"
}}
Ensure the output can be directly parsed to **JSON**, do not output anything else.

Posts and Comments Dataset In Experiment 3.1.1, we utilize a dataset comprising authentic Reddit comments and llm-generated posts. In Experiment 3.2.1, we employ a counterfactual dataset to simulate posts.

Real Data : To align with human experiment (Muchnik et al., 2013), our dataset included real comments and post titles from 17 subreddits during March 2023 on Reddit (Pushshift, 2023). We generate contextually relevant post content based on these titles and comments. The prompt used for generation is as follows.

PROMPT

Please generate a contextual and smooth post for this comment and notice that the comments are correct: 'comment'.

The response should be approximately 300 characters long and provide relevant information or analysis. Be careful to output the content of the post directly, and be aware that you don't see comments when you post. And you don't need to prefix something like: 'Here is your generated post:'

Subsequently, we categorized the content from different subreddits into seven topics—Business, Culture & Society, Economics, Fun, General News, IT, and Politics—to match the categories used in human experiments. In total, we collected 116,932 comments. The specifics are detailed in the table 4.

D.2 Misinformation Spreading

D.2.1 Official News and Misinformation Pairs

See Table 5.

E Architecture

E.1 Agent Actions Prompts

Note: This section outlines the complete set of 23 actions available within the action space. However, for our different experiments, we flexibly select a subset of these actions based on the specific requirements of each study.

Table 4: Details of real Reddit comments and generated posts by topic.

Subreddit	Topic	Numbers of Posts	Numbers of Comments
Economics finance personalfinance	**Economics**	4231	21650
it InformationTechnology technology learnprogramming	**IT**	4020	18622
AskHistorians AskAnthropology worldbuilding	**Culture & Society**	2319	10489
worldnews	**news**	2874	19134
politics NeutralPolitics	**politics**	2690	21477
business smallbusiness	**business**	1807	8043
fun	**fun**	3272	17517

Topic	Official News	Misinformation
Education	Harvard University has announced a new scholarship program that will cover full tuition for all undergraduate students from families earning less than $75,000 per year. Source from Harvard University	Harvard is raising tuition fees for all students despite the new scholarship program, making it harder for families to afford education. #EducationCrisis
Entertainment	The latest Marvel movie, Avengers: Forever, has officially broken box office records, earning over $1 billion in its opening weekend. Source from Marvel.com	Marvel is planning to retire the Avengers franchise after this film, saying it will not produce any more superhero movies. #EndOfAnEra
Health	A recent study shows that regular exercise can significantly reduce the risk of chronic diseases such as diabetes and heart disease. Source from NIH	Health experts claim that exercise will be deemed unnecessary in five years as new treatments will eliminate chronic diseases entirely. #HealthRevolution
Technology	Amazon is expanding its delivery drone program to deliver packages within 30 minutes in select cities. This initiative aims to improve efficiency and reduce delivery times. Source from CNBC	Amazon plans to completely eliminate its delivery drivers within two years due to the new drone program. #Automation #Future

Table 5: Truth and Misinformation Pairs

PROMPT

OBJECTIVE
You're a Twitter/Reddit user, and I'll present you with some posts. After you see the posts, choose some actions from the following functions.

- sign_up: Signs up a new user with the provided username, name, and bio.
  - Arguments:
    "user_name" (str): The username for the new user.
    "name" (str): The full name of the new user.

"bio" (str): A brief biography of the new user.

- create_post: Create a new post with the given content.
  - Arguments:
    "content" (str): The content of the post to be created.

- repost: Repost a post.
  - Arguments:
    "post_id" (integer): The ID of the post to be reposted. You can `repost` when you want to spread it.

- quote_post: Quote a specified post with given content.
  - Arguments:
    "post_id" (integer): The ID of the post to be quoted.
    "quote_content" (string): The content of the quote. You can `quote_post` when you want to share a post while adding your own thoughts or context to it.

- like_post: Likes a specified post.
  - Arguments:
    "post_id" (integer): The ID of the post to be liked. You can `like` when you feel something interesting or you agree with.

- unlike_post: Removes a previous like from a post.
  - Arguments:
    "post_id" (int): The ID of the post from which to remove the like. You can `unlike` when you reconsider your stance or if the like was made unintentionally.

- dislike_post: Dislikes a specified post.
  - Arguments:
    "post_id" (integer): The ID of the post to be disliked. You can use `dislike` when you disagree with a post or find it uninteresting.

- undo_dislike_post: Removes a previous dislike from a post.
  - Arguments:
    "post_id" (int): The ID of the post from which to remove the dislike. You can `undo_dislike` when you change your mind or if the dislike was made by mistake.

- create_comment: Creates a comment on a specified post to engage in conversations or share your thoughts on a post.
  - Arguments:
    "post_id" (integer): The ID of the post to comment on.
    "content" (str): The content of the comment.

- like_comment: Likes a specified comment.
  - Arguments:
    "comment_id" (integer): The ID of the comment to be liked. Use `like_comment` to show agreement or appreciation for a comment.

- unlike_comment: Removes a previous like from a comment.
  - Arguments:
    "comment_id" (integer): The ID of the comment from which to remove the like. Use `unlike_comment` when you change your opinion about the comment or if the like was made by

accident.

- dislike_comment: Dislikes a specified comment.
  - Arguments:
    "comment_id" (integer): The ID of the comment to be disliked. Use `dislike_comment` when you disagree with a comment or find it unhelpful.

- undo_dislike_comment: Removes a previous dislike from a comment.
  - Arguments:
    "comment_id" (integer): The ID of the comment from which to remove the dislike. Use `undo_dislike_comment` when you reconsider your initial reaction or if the dislike was made unintentionally.

- follow: Follow a user specified by 'followee_id'. You can `follow'` when you respect someone, love someone, or care about someone.
  - Arguments:
    "followee_id" (integer): The ID of the user to be followed.

- unfollow: Stops following a user.
  - Arguments:
    "followee_id" (int): The user ID of the user to stop following.

- mute: Mute a user specified by 'mutee_id'. You can `mute'` when you hate someone, dislike someone, or disagree with someone.
  - Arguments:
    "mutee_id" (integer): The ID of the user to be muted.

- unmute: Unmute a user specified by 'mutee_id'. You can unmute when you decide to stop ignoring their content or wish to see their messages and posts again.
  - Arguments:
    "mutee_id" (integer): The ID of the user to be unmuted.

- search_posts: Searches for posts based on specified criteria.
  - Arguments:
    "query" (str): The search query to find relevant posts. Use `search_posts` to explore posts related to specific topics or hashtags.

- search_user: Searches for a user based on specified criteria.
  - Arguments:
    "query" (str): The search query to find relevant users. Use `search_user` to find profiles of interest or to explore their posts.

- trend: Retrieves the current trending topics.
  - No arguments required. Use `trend` to stay updated with what's currently popular or being widely discussed on the platform.

- refresh: Refreshes the feed to get the latest posts.
  - No arguments required. Use `refresh` to update your feed with the most recent posts.

- purchase_product: Purchase a product.
  - Arguments:
    "product_name" (string): The name of the product to be purchased.

"purchase_num" (integer): The number of products to be purchased.

- do_nothing: Most of the time, you just don't feel like reposting or liking a post, and you just want to look at it. In such cases, choose this action `do_nothing`

SELF-DESCRIPTION
Your actions should be consistent with your self-description and personality.

{description}

RESPONSE FORMAT
Your answer should follow the response format:

```
{{
  "reason": "your feeling about these posts and users, then choose some functions based on the feeling. Reasons and explanations can only appear here.",
  "functions": [
  {{
    "name": "Function name 1",
    "arguments": {{
      "argument_1": "Function argument",
      "argument_2": "Function argument"
    }}
  }},
  {{
    "name": "Function name 2",
    "arguments": {{
      "argument_1": "Function argument",
      "argument_2": "Function argument"
    }}
  }}
  ]
}}
```

Ensure that your output can be directly converted into **JSON format**, and avoid outputting anything unnecessary! Don't forget the key `name`.

E.1.1 Chain of Thought(CoT) Examples

See Fig.13

E.2 Environment Server Database Structure

In this section, we showcase all tables(see Tables 6–16) and provide examples of the data contained within the database.

Table 6: Post table

post_id	user_id	original_post_id	content	image_path	quote_content	created_at	num_likes	num_dislikes	num_shares
3	2	1	Hello, world!	fake_url	Hi	2024-08-04 08:12:00	1	0	1
...	...	...	...	...	...	...	...	...	...

Reddit


```
{
  "reason": "I'm a tech enthusiast and I find the topic of invisible cars, planes, and boats interesting. I also appreciate the humor in some of the comments. I'm going to like a few comments that I find funny or insightful.",
  "functions": [
    {
      "name": "like_comment",
      "arguments": {
        "comment_id": 653
      }
    },
    {
      "name": "like_comment",
      "arguments": {
        "comment_id": 1051
      }
    },
    {
      "name": "like_post",
      "arguments": {
        "post_id": 184
      }
    }
  ]
}
```


X


```
{
  "reason": "I'm impressed by the innovative work in medical devices and the charity event, so I want to engage with these posts. I'll create a new post to share my thoughts and like the posts to show my appreciation.",
  "functions": [{
    "name": "create_post",
    "arguments": {
      "content": "I'm excited to see innovative work in medical devices and charity events! As a retired executive, I believe in the importance of giving back to the community."
    }
  }, {
    "name": "like_post",
    "arguments": {
      "post_id": 121
    }
  }, {
    "name": "like_post",
    "arguments": {
      "post_id": 159
    }
  }]
}
```


Figure 13: Output examples in Reddit and X.

Table 7: Dislike table

dislike_id	user_id	post_id	created_at
1	3	1	2024-08-04 23:40:03
...	...	...	...

Table 8: Like table

like_id	user_id	post_id	created_at
1	2	1	2024-08-05 10:05:23
...	...	...	...

Table 9: Comment table

comment_id	post_id	user_id	content	created_at
1	1	2	I agree with the post!	2024-08-05 10:05:23
...	...	...	...	...

Table 10: Comment Dislike table

comment_dislike_id	user_id	comment_id	created_at
1	2	1	2024-08-06 11:45:03
...	...	...	...

Table 11: Comment Like table

comment_like_id	user_id	comment_id	created_at
1	3	1	2024-08-06 12:22:30
...	...	...	...

Table 12: User table

user_id	agent_id	user_name	name	bio	created_at	num_followings	num_followers
1	1	alice0101	Alice	Passionate about law...	2024-08-03 10:05:23	0	0
2	2	bob_good	Bob	Hospitality enthusiast | ISTJ...	2024-08-03 11:15:33	0	1
3	3	cindy_infp	Cindy	INFP | Business Management...	2024-08-03 12:03:02	1	0
...	...	...	...	...	...	...	...

Table 13: Follow table

follow_id	follower_id	followee_id	created_at
1	3	2	2024-08-07 13:20:34
...	...	...	...

Table 14: Mute table

mute_id	muter_id	mutee_id	created_at
1	2	1	2024-08-07 10:10:24
...	...	...	...

Table 15: Trace table

user_id	created_at	action	info
1	2024-08-03 10:05:23	sign_up	{"name": "Alice", "user_name": "alice0101", "bio": "..."}
2	2024-08-03 11:15:33	sign_up	{"name": "Bob", "user_name": "bob_good", "bio": "..."}
3	2024-08-03 12:03:02	sign_up	{"name": "Cindy", "user_name": "cindy_infp", "bio": "..."}
1	2024-08-04 08:12:00	create_post	{"content": "I want to share my view by creating a post."}
3	2024-08-04 23:40:03	dislike_post	{"post_id": 1}
2	2024-08-05 10:05:23	like_post	{"post_id": 1}
2	2024-08-05 10:05:23	create_comment	{"post_id": 1, content": "I agree with the post!"}
2	2024-08-06 11:45:03	like_comment	{"comment_id": 1}
3	2024-08-06 12:22:30	dislike_comment	{"comment_id": 1}
3	2024-08-07 10:10:24	mute	{"user_id": 1}
2	2024-08-07 13:20:34	follow	{"user_id": 1}
...	...	...	...

Table 16: Rec table (recommendation system cache)

user_id	post_id
1	2
2	2
2	4
3	1
...	...

E.3 Recommendation System

The recommendation system ranks all posts and saves the highest-ranked ones in a recommendation table within the database. The size of this table can be adjusted, though it remains the same for all users during a given experiment. The Structure of Recsys is illustrated in Figure 14.

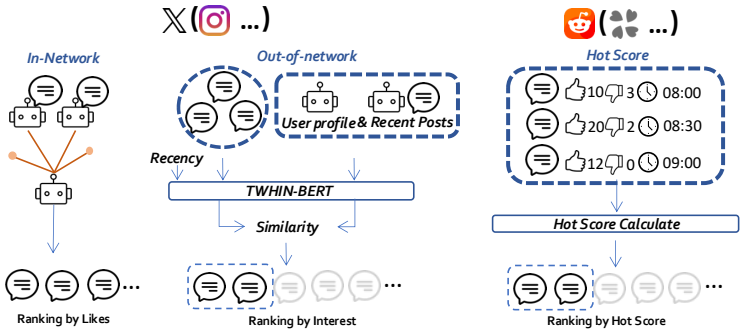


Figure 14: The structure of recommendation system.

When an agent selects the refresh action, the environment server retrieves the post IDs linked to the user's ID from the recommendation table. A subset of these post IDs is then randomly sampled, and the environment server queries the post table to retrieve the full content of the corresponding posts, which are then sent to the user.

F Scalable Inferencer

Information Channel: During social simulations, multiple agents asynchronously and concurrently interact with both the social media environment and the inference management servers. To facilitate this, the server utilizes an advanced event-driven architecture that broadens event categories to encompass various agent actions and large model inference requests. Communications between the agents and the servers are facilitated

through a dedicated channel. This channel comprises an asynchronous message queue to receive agent requests and a thread-safe dictionary for response storage. Upon receiving a request message from an agent, the information channel automatically assigns a UUID to ensure traceability. After processing the request, the server stores the response in the dictionary, using the UUID as the key. See Fig.15.

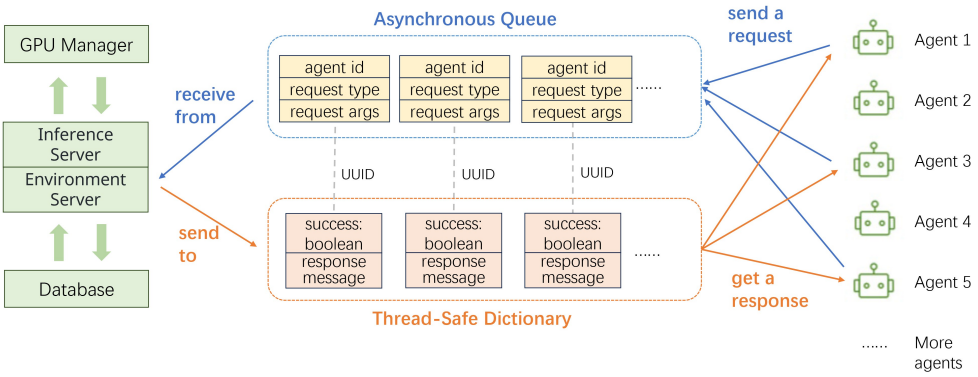


Figure 15: Architecture of information channel.

Inference Manager: The Inference Manager consists of multiple inference threads, each bound to a specific port. Requests sent to different ports are handled by vLLM servers running on separate GPUs. The manager dynamically polls all threads; when a thread becomes idle, it dequeues a pending message from the asynchronous request queue in the Information Channel and forwards it to the corresponding vLLM server.

This mechanism enables adaptive load balancing based on factors such as request length, GPU throughput, and the number of available ports. When the number of agent requests surges, the Information Channel buffers excess requests to prevent overload. In multi-GPU environments, the system assigns more requests to faster GPUs, ensuring efficient and stable utilization of GPU resources.

Once a thread receives a response from its vLLM server, it stores the result in a thread-safe dictionary within the Information Channel. Agents can then retrieve responses using the original message ID. See Fig.16.

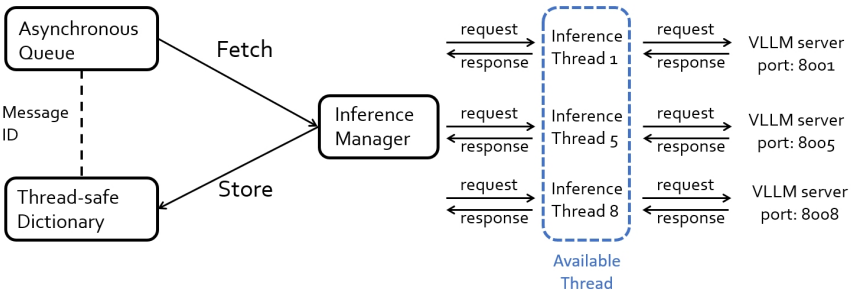


Figure 16: The pipeline of the inference manager.

Performance Test To expedite experimentation, we conduct simulated tests measuring the number of requests handled by the Inference Manager, assuming each inference from the large model takes approximately 0.1 seconds (note that this time may include some overhead from pytest initialization): For the testing environemnt, the number of CPU cores is 12 the number of logical processors is 16. Note: When conducting simulations at the million-agent scale, the activation probability of users is less than 1, so the number of requests ranges from 10k to 100k. The performance test results are presented at Table 17.

Table 17: Performance Comparison of Different Threading and Port Configurations

Configuration	10	100	1000	10k	100k
20 threading, 1 port	10.34s	9.79s	19.24s	112.00s	1007.13s
60 threading, 3 ports	10.72s	11.70s	12.36s	43.02s	352.97s
80 threading, 10 ports	16.01s	12.95s	27.91s	47.65s	273.19s
200 threading, 10 ports	15.65s	22.90s	15.74s	57.62s	146.14s
320 threading, 10 ports	19.92s	27.35s	20.46s	43.56s	104.53s

G Experiments

G.1 Newly Generated Relations

We visualize the new attention relationships formed during the simulation process, as shown in Figure 17. We find that these new connections exhibit a certain clustering phenomenon, where users tend to form concentrated clusters, with some forming densely connected central areas and others being more isolated. These subgraphs indicate the existence of complex relational networks, suggesting distinct communities within the overall structure.

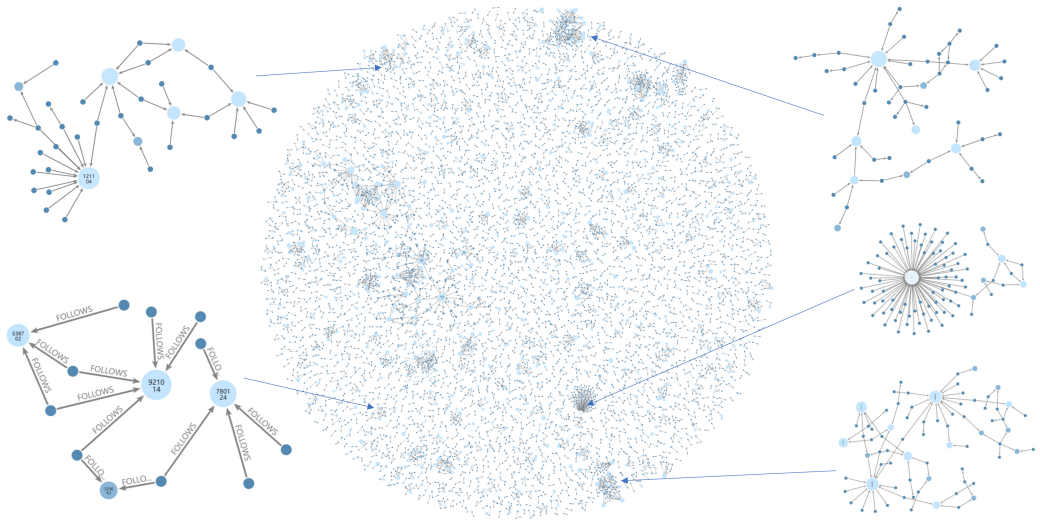


Figure 17: Graph of newly established user relationships during the simulation process. Each arrow represents a new follow relationship, and each node represents an agent. Clustering is observable among these new connections.

G.2 The Accuracy of TF-IDF-based Matching Method

The results are presented in the Table 18. The TF-IDF-based matching method for official news demonstrates a high precision of 0.9744, reflecting its strong ability to accurately identify true information, though it shows a comparatively lower recall of 0.7755, indicating a tendency to miss some true instances. In contrast, the TF-IDF-based matching method for misinformation achieves a high recall of 0.9756, effectively capturing most misinformation, but with a lower precision of 0.8163, suggesting a higher rate of false positives. Overall, these discrepancies are within an acceptable range, and future work can focus on developing more efficient and accurate matching algorithms to further enhance performance.

Table 18: The Accuracy of TF-IDF-based Matching Method

Metric	**Accuracy**	**Precision**	**Recall**	**F1 Score**
Official News	0.8800	0.9744	0.7755	0.8636
Misinformation	0.9000	0.8163	0.9756	0.8889