



Towards efficient optimization of multi-agent social simulation via large language models

Kun Zhang¹ · Xiaoyan Yu² · Hao Peng¹ · Zhe Yang¹ · Ye Tian³ · Hao Jin⁴ · Tuoyu Feng⁴ · Hui Lin⁴

Received: 11 October 2025 / Accepted: 18 December 2025 / Published online: 12 January 2026
© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2026

Abstract

Public opinion on social media during major societal events is highly vulnerable to manipulation by misinformation and extreme emotions, posing severe threats to mainstream values and public trust. While most existing studies focus on passive detection of negative content, systematic approaches for proactively optimizing opinion guidance remain scarce. To address this issue, we propose a Social Influence Text Generation Network, namely SITGNet, a novel multi-agent social simulation framework powered by large language models (LLMs). SITGNet populates a digital social sandbox with cognitively sophisticated agents, each possessing distinct profiles, memories, and behavioral patterns grounded in empirical data. This enables the high-fidelity reproduction of opinion dynamics, capturing emergent phenomena from complex user interactions and psychological transitions. Within this platform, we introduce Guiding Agents, a specialized class of agents designed to proactively steer discourse. These agents leverage a multi-stage retrieval-augmented generation (RAG) pipeline to synthesize information from external knowledge bases, grounding their outputs in factual evidence to generate nuanced and influential content. Through extensive, large-scale simulations, we systematically evaluate various intervention strategies. Our results demonstrate that a hybrid deployment of Guiding Agents can effectively suppress negative sentiment propagation, alleviate group polarization, and guide opinion trajectories toward rational outcomes. In addition to building the simulation framework, this study treats Social Influence Text Generation as a key modeling objective, enabling systematic evaluation of text-based opinion guidance strategies.

Keywords Large language models · Retrieval-augmented generation · Multi-agent social simulation · Social influence text generation

1 Introduction

With the deep penetration of social media into everyday life, online platforms have become the primary arenas for the formation, contestation, and dissemination of public opinion during major societal events, such as public health crises, security incidents, natural disasters, and debates

on technology ethics [1, 2]. The decentralized structure, algorithmic amplification, and rapid diffusion of information in social networks enable emotions and viewpoints to converge and spread at unprecedented speed. On the one hand, such dynamics can promote public awareness, speed up information sharing, and help gather useful resources during emergencies [3]. On the other hand, the same fast information-spreading mechanisms also bring serious challenges: during the early and peak stages of major events, online spaces are often filled with unconfirmed information, strongly emotional comments, one-sided discussions, and even organized harmful activities [4]. These patterns can distort public discussions, strengthen extreme opinions, and reinforce existing biases. Over time, such problems may not only mislead people's understanding but also reduce trust in public institutions, increase conflicts between groups, and gradually lead to deeper divisions in public opinion [5, 6].

✉ Hao Peng
penghao@buaa.edu.cn

¹ School of Cyber Science and Technology, Beihang University, Beijing 100191, China

² Beijing Institute of Technology, Beijing 100081, China

³ Laboratory for Advanced Computing and Intelligence Engineering, Wuxi 214000, China

⁴ China Academy of Electronics and Information Technology, Beijing 100041, China

Understanding these patterns is therefore a key challenge for computational social science and an urgent social issue.

These risks raise a key scientific question: *Can we accurately simulate and analyze the evolutionary trajectory of online public opinion to better understand the mechanisms of information diffusion and collective emotion dynamics?* While numerous studies have focused on defensive tasks such as rumor detection [7], sentiment analysis [8], and bot or troll detection [9], most efforts remain limited to passively identifying and filtering harmful content. Such reactive approaches, although useful, do not provide a systematic way to reproduce and study the underlying processes of opinion evolution. In contrast, simulation-based frameworks that model user interactions, emotional feedback, and network dynamics remain underexplored. Building an effective multi-agent social simulation environment requires integrating multiple complex capabilities, including user emotion modeling [10], social influence prediction [11], network-aware diffusion modeling [12], and goal-driven agent design [13].

The recent convergence of LLMs and multi-agent simulation offers unprecedented opportunities to address these challenges in a principled and systematic manner. On the one hand, LLMs exhibit strong capabilities in contextual reasoning, adaptive dialogue, and controllable text generation, enabling the simulation of diverse roles, tones, and interaction patterns in online discourse [14, 15]. These properties allow LLMs to mimic realistic user behaviors more faithfully than traditional rule-based or template-driven systems, thus providing a richer foundation for studying emergent opinion dynamics. On the other hand, agent-based modeling has long been recognized as a powerful paradigm for exploring how micro-level interactions among individuals give rise to macro-level collective phenomena such as consensus formation, fragmentation, or polarization [16–18]. When combined, LLMs and agent-based simulation create a controllable “digital social sandbox” that allows researchers to reproduce and manipulate complex opinion ecosystems, examine counterfactual scenarios, and systematically evaluate the outcomes of different intervention strategies. This paper demonstrates that the synergy of LLM-driven agents enhances the accuracy and scalability of simulating complex social dynamics in online environments, and based on this synergy, we design and implement a multi-agent simulation environment to reproduce the full trajectory of opinion evolution on social networks during crisis-driven events.

To systematically investigate opinion dynamics and intervention strategies, we develop SITGNet, a multi-agent social simulation framework powered by LLMs. SITGNet integrates heterogeneous agents with distinct roles and psychological profiles, enabling us to reproduce realistic trajectories of opinion formation and diffusion during

crisis-driven events. In particular, we construct LLM-based agents that simulate user behaviors with high fidelity by combining the generative and reasoning capabilities of LLMs with real-world user data. This design allows each agent to capture diverse attributes such as demographic features, personality, emotional tendencies, and occupation, which are parameterized and grounded in empirical data collected from social platforms. By anchoring agent behaviors in both large-scale models and observed user characteristics, SITGNet achieves a closer alignment with actual online ecosystems and provides a more reliable foundation for studying the mechanisms of opinion evolution.

The framework supports both baseline simulations without intervention and experimental settings with proactive guidance, thereby offering a controlled environment to evaluate how different factors—such as content framing, messenger identity, and network structure—jointly shape collective outcomes. In baseline simulations without intervention, SITGNet reproduces patterns consistent with recent findings in computational social science. For instance, messages originating from highly influential actors—commonly referred to as key opinion leaders (KOLs)—disproportionately drive shifts in public sentiment and trigger large-scale cascades across the network [19, 20]. By contrast, even factually accurate and informative content often exhibits limited impact when expressed in emotionally neutral or cognitively distant terms, whereas emotionally charged narratives spread more rapidly and achieve wider reach. This “emotional prioritization” effect aligns with recent evidence on online emotional contagion and the differential diffusion of discrete emotions in digital platforms [21, 22]. These observations highlight key levers for effective opinion intervention: beyond factual correctness, success depends critically on the credibility and perceived authority of the messenger, the timing and structural position of dissemination, and the emotional resonance of the content.

Building on these insights, SITGNet incorporates specialized *guiding agents* that proactively intervene in opinion dynamics. To address the critical challenge that early malicious content can trigger avalanche-like cascades of negativity and misinformation, guiding agents employ a retrieval-augmented generation (RAG) mechanism. By dynamically retrieving knowledge from authoritative sources—such as trusted institutions, verified media, or domain-specific repositories—and combining it with persuasive rhetorical and emotional framing, these agents generate counter-narratives that are both credible and engaging. Extensive experiments within SITGNet demonstrate that strategically deploying guiding agents effectively shifts opinion trajectories toward constructive directions, suppresses misinformation cascades, mitigates long-term polarization, and amplifies the diffusion of rational and

cooperative discourse. Together, these results show that proactive, generative interventions grounded in both social theory and LLM-based capabilities hold strong promise for fostering healthier online ecosystems and offer a new methodological lens for the study of opinion governance in the digital age.

The main contributions of this study are summarized as follows:

- *Simulation Framework*: We develop SITGNet, an LLM-driven multi-agent simulation framework for modeling the dynamic evolution of online public opinion. The system captures the interplay among user viewpoints, emotions, and behaviors under diverse event scenarios.
- *LLM-based Agent Construction*: We design cognitively grounded LLM-based agents equipped with perception, memory, and action modules to simulate human-like behaviors in social media environments. These agents can autonomously perceive, reason, and interact, forming the foundation of large-scale social simulations.
- *Guidance Mechanism*: We design a retrieval-augmented guidance mechanism that allows guiding agents to generate factual and context-relevant content. Through goal-driven reasoning and short- to long-term planning, the agents provide timely interventions to reduce negative sentiment and stabilize public opinion.
- *Evaluation Framework*: We build a set of metrics and conduct systematic experiments to evaluate different guidance strategies. The framework measures the effectiveness and social impact of interventions, providing a practical basis for future opinion-guidance research.

2 Preliminaries

In this section, we briefly introduce the foundational concepts and standard processes underlying the three key components of our study: LLM-based agents, retrieval-augmented generation, and social simulation systems. These concepts provide the theoretical and technical basis upon which SITGNet is constructed.

LLM-based Agent Design LLM-based agents are autonomous entities capable of perceiving context, reasoning, and generating human-like responses through natural language models. Typically, an LLM agent operates in a perception–decision–action cycle. At each time step, the agent receives textual or structured input from the environment, using LLM produces a contextually appropriate action such as a response, summary, or judgment. The internal structure usually includes three main components:

- (1) profile module describing static attributes (e.g., identity, role, personality),
- (2) memory module that retains past interactions and supports long-term consistency, and
- (3) reasoning module that employs the LLM to infer intentions or generate actions. This design enables agents to simulate cognitive behavior patterns observed in human communication, and has been widely used in interactive simulations, decision-making studies, and cognitive modeling. *Retrieval-Augmented Generation* RAG is a framework that combines information retrieval with language generation to enhance factual accuracy and contextual relevance. In this approach, the system first transforms the user input into a query and retrieves a small collection of related documents from an external knowledge base. These retrieved materials are then incorporated into the input context and provided to the language model for response generation. By grounding the generation process in real, authoritative information, RAG enables the model to access up-to-date or domain-specific knowledge without additional training. This design effectively reduces hallucinations, improves factual consistency, and increases the interpretability of the model's reasoning process.

Social Simulation Systems Social simulation systems model how individuals or groups interact within a social environment to study emergent collective behaviors. A typical social simulation consists of three elements:

- (1) an environment defining global rules of interaction and information flow,
- (2) a population of agents that act according to personal attributes, goals, and perception of the environment, and
- (3) a scheduler that controls temporal evolution, allowing events to unfold across discrete or continuous time steps. At each step, agents perceive the current state, make decisions based on internal models or external stimuli, and update both their states and the environment.

3 Method

Our simulation framework is composed of a comprehensive environment model that dictates the rules of information propagation and an agent model that governs individual decision-making processes. At its core, the simulation loop operates as follows: (1) The environment evolves based on event heat dynamics and calculates a global Negative Sentiment Index (NSI). (2) Each agent perceives information from the environment. (3) The agent's cognitive module

(memory, reflection, planning) processes this information to decide on an action. (4) Guiding agents, equipped with a RAG mechanism, generate fact-based, persuasive content when triggered. The core components of the framework are detailed below.

3.1 Overview of SITGNet

SITGNet aims to simulate the complete lifecycle of public opinion surrounding sudden social events on online platforms. In this framework, each event is modeled as a burst of public attention that spreads through user interactions such as posting, commenting, and sharing. The lifecycle of an event typically evolves through phases of emergence, outbreak, peak, and decline, during which users continuously exchange views and emotions. These interactions collectively shape the trajectory of public sentiment, amplifying both positive discussions and potential negative cascades. Meanwhile, they also provide opportunities for guided interventions to redirect discourse toward more rational and constructive directions. As illustrated in Fig. 1, the system is comprised of three synergistic modules:

To capture these dynamics, SITGNet integrates three synergistic modules: the Multi-agent Simulation Environment,

the LLM-based Agent Design, and the RAG-based Guiding Agent. The first module establishes a digital arena that defines how information propagates, how events evolve, and how agents interact. The second module introduces LLM-powered agents endowed with cognition, memory, and emotion, enabling human-like engagement in the simulated social space. The third module equips guiding agents with retrieval-augmented generation (RAG) capabilities, allowing them to produce evidence-based and persuasive content to stabilize public sentiment when negativity escalates.

During a typical simulation run, an event is initialized and user agents begin to interact, forming an evolving opinion landscape. Overall, SITGNet provides a controllable yet realistic social simulation framework that connects cognitive modeling with large language model reasoning. It offers an experimental platform for analyzing how information spreads, how sentiment evolves, and how intelligent interventions can guide online discourse toward a balanced and resilient state.

3.2 Multi-agent simulation environment

The SITGNet simulation environment is designed to reproduce the complex lifecycle of public opinion on social media

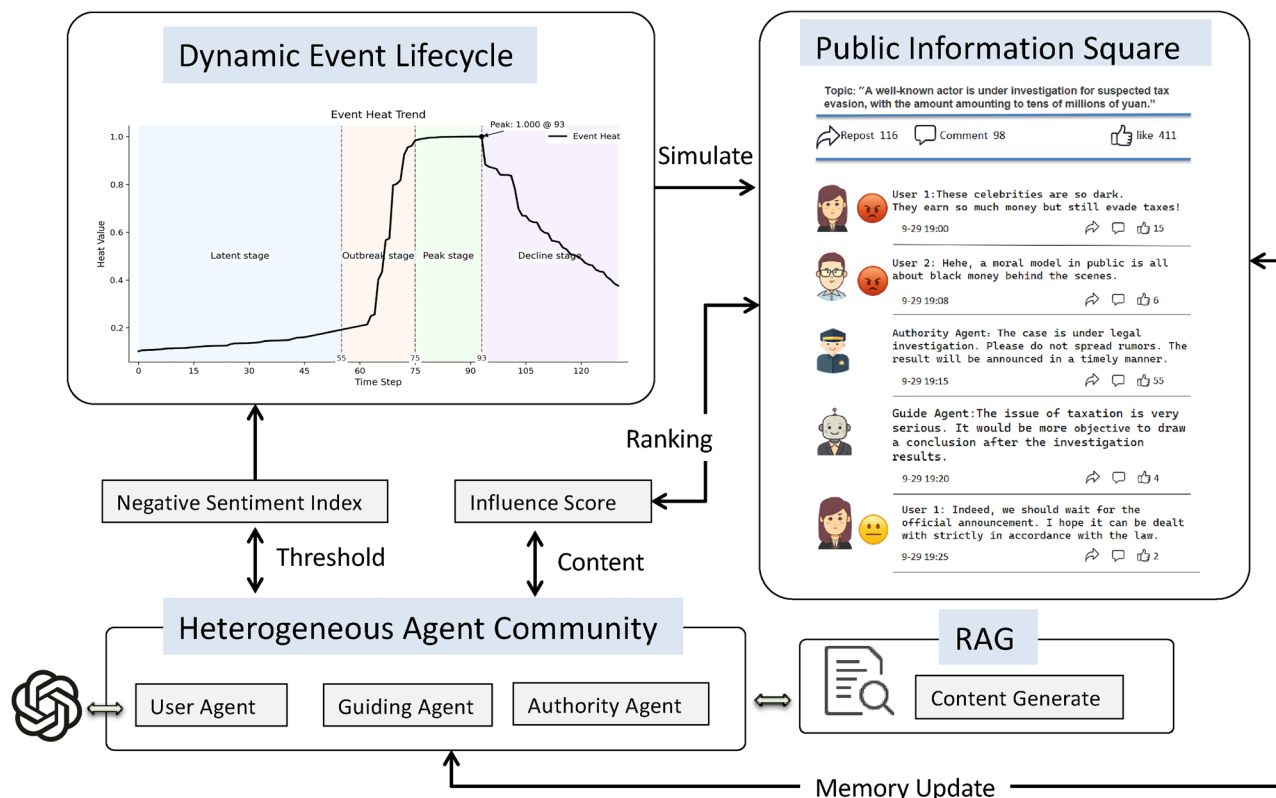


Fig. 1 Overview of the SITGNet system, which integrates the interaction between the environment and agents. The environment maintains the event state and public information. Agents perceive this informa-

tion, process it through their cognitive loops, and execute actions that feed back into the environment, creating a dynamic simulation cycle

platforms. Unlike previous works that mainly describe opinion diffusion as a uniform cascade, SITGNet emphasizes three interacting dimensions: the evolving salience of events, the mechanisms of information exposure, and the adaptive regulation of sentiment. By explicitly linking these components, the framework is able to capture the surge–peak–decline dynamics of attention while also providing a natural entry point for authority-driven guidance.

3.2.1 Information system

All content generated by agents is aggregated into a shared Public Information Square, which serves as the primary arena for opinion exchange and sentiment evolution. Within this environment, content visibility and propagation strength are determined by an Influence Ranking System that integrates both user authority and community engagement. The influence score for each content item c at time t is calculated as:

$$\text{Score}(c, t) = (w_f \cdot F + w_l \cdot L + w_r \cdot R) \cdot D(t), \quad (1)$$

where F is the author's follower count, L the number of likes, R the number of reposts, and $D(t)$ a time decay factor. The weights w_f , w_l , and w_r are empirically set to 0.4, 0.2, and 0.4 respectively. These weights were set based on the common understanding that a repost w_r signifies a stronger endorsement and wider reach than a like w_l , while the fans's influence w_f provides a foundational score. To prevent the excessive dominance of large accounts, a fan-based weighting mechanism scales influence according to follower count:

$$F = \begin{cases} \frac{fans}{100}, & fans \leq 1000, \\ 10 + \frac{fans - 1000}{10000}, & 1000 < fans \leq 100000, \\ 20 + \min\left(\frac{fans - 100000}{100000}, 30\right), & fans > 100000. \end{cases} \quad (2)$$

This scaling ensures that smaller accounts have proportionally limited impact, mid-sized accounts gain gradual influence, and large accounts remain capped, resulting in a balanced and realistic distribution of social power.

To monitor and regulate the emotional health of the discussion space, SITGNet integrates an LLM-powered Negative Sentiment Index (NSI). Unlike traditional keyword-based sentiment tracking, the NSI employs large language models to assess the discourse atmosphere holistically, considering emotional polarity, intensity, and propagation trends. The NSI is updated dynamically at each simulation step according to:

$$\text{NSI}(t) = \text{NSI}(t-1) + \frac{\Delta_{\text{sentiment}}}{100}, \quad (3)$$

where $\Delta_{\text{sentiment}}$ represents the LLM-assessed impact of new comments, ranging from -50 (strongly positive) to $+50$ (strongly negative). Once the NSI surpasses a pre-defined threshold Θ (typically 0.4), authoritative agents are automatically triggered to inject corrective or factual information. This mechanism creates a closed feedback loop between content visibility, user behavior, and emotional monitoring, enabling SITGNet to maintain a stable and rational public discourse environment.

3.2.2 Dynamic event lifecycle

At the core of the simulation is the Event Heat $H(t)$, a time-varying function that quantifies the prominence of a topic in public discourse. Crucially, $H(t)$ not only governs the visibility and spread of an event but also modulates the frequency of agent interactions with the environment. Higher event heat leads to intensified posting, commenting, and sharing, thereby mimicking the collective surges of activity observed during real-world crises.

The lifecycle of $H(t)$ is divided into four characteristic phases:

Latent stage. During the early phase, the event remains largely unnoticed, with visibility increasing only gradually. We model this period as exponential growth:

$$H(t) = H_0 e^{\alpha t}, \quad t < T_{\text{outbreak}}, \quad (4)$$

where H_0 is the baseline visibility and α controls the rate of awareness accumulation.

Outbreak stage. Once the event reaches the outbreak threshold, attention rises sharply, following a corrected S-shaped trajectory. To ensure continuity at T_{outbreak} , the curve is shifted so that the exponential and logistic functions align smoothly:

$$H(t) = \frac{H_{\text{peak}}}{1 + e^{-k(t-t_{\text{mid}})}} + \Delta, \quad T_{\text{outbreak}} \leq t < T_{\text{decline}}, \quad (5)$$

where H_{peak} denotes the saturation level of attention, k determines the steepness of escalation, t_{mid} is the midpoint of the outbreak-to-decline interval, and Δ is a corrective offset ensuring continuity at T_{outbreak} . This captures the rapid yet bounded surge of attention as the event becomes widely recognized.

Peak stage. As the logistic curve approaches saturation, attention levels off at the maximum intensity H_{peak} . In this stage, nearly all agents are exposed to the event, and new interactions primarily reinforce existing narratives rather than introducing novel perspectives. Formally, the heat stabilizes as

$$H(t) \approx H_{\text{peak}}, \quad t \approx T_{\text{decline}}, \quad (6)$$

highlighting the transient equilibrium before decline.

Decline stage. After reaching its peak, collective attention wanes, following an exponential decay:

$$H(t) = H_{\text{peak}} e^{-\lambda(t-T_{\text{decline}})}, \quad t \geq T_{\text{decline}}, \quad (7)$$

where λ governs the forgetting rate, reflecting the natural fading of salience as other events emerge and dominate discourse.

By dynamically coupling $H(t)$ with agent activity scheduling, SITGNet ensures that interaction intensity rises and falls in tandem with event salience, thereby reproducing the authentic rhythm of online discourse evolution.

3.2.3 Agent–environment feedback

The integration of these modules forms a dynamic feedback loop that closely replicates real-world social media interactions. At each simulation step, agents (i) perceive top-ranked content from the Public Information Square, (ii) interpret the information through LLM-based cognition that reflects their personality traits and emotional states, (iii) generate appropriate behavioral responses such as commenting, reposting, or liking, and (iv) collectively reshape the overall discourse environment by influencing event heat $H(t)$ and the Negative Sentiment Index (NSI). When the NSI exceeds a defined threshold, guiding agents are automatically activated to provide corrective interventions.

To capture realistic temporal patterns, SITGNet employs an adaptive scheduling mechanism where each agent's activity frequency is influenced by the current event heat. As event salience rises, agents become more active, leading to spontaneous bursts of engagement similar to those observed during real-world crises. Conversely, when sentiment stabilizes, interaction rates gradually decline, restoring equilibrium in the system.

This closed-loop framework allows SITGNet to simulate the co-evolution between attention dynamics, emotional climate, and strategic interventions. By maintaining temporal continuity and realistic feedback dependencies, the system achieves high ecological validity (the degree to which the simulation reflects real-world behaviors and outcomes in social environments), offering a controlled yet lifelike environment for studying public opinion evolution and testing various guidance strategies.

3.3 LLM-based agent design

The foundation of our heterogeneous agent community is the LLM-based Agent, which adopts a sophisticated cognitive

architecture to emulate human-like behavior in social contexts. Each agent operates through a closed cognitive loop built on five interconnected modules—Profile, Perception, Memory, Action, and Reflection—with extended Goal and Planning components for higher-level reasoning, as illustrated in Fig. 2.

3.3.1 Profile

Each agent is initialized with a detailed profile to ensure realism and behavioral diversity. The profile includes demographic information such as ID, nickname, age, gender, occupation, and region, together with personality traits based on the *Big Five* model. The use of personality traits is supported by empirical evidence rather than arbitrary choice. Analysis of large-scale social media data (a 10% sample of a 20-million-user dataset) showed clear links between personality and user activity. In particular, users with higher *Openness* were more active, while those with higher *Neuroticism* tended to participate less. Moreover, *Conscientiousness* was more common among key opinion leaders (KOLs), highlighting its role in sustaining influence (Fig. 3).

Based on these findings, SITGNet assigns personality traits according to user roles: influential agents are initialized with higher *Conscientiousness* and *Openness*, while ordinary agents show more variation across all traits. Each agent is also given a short biographical sketch consistent with its personality, and social activity metrics such as post and follower counts are initialized following observed data distributions. These steps help ensure that SITGNet's agent population reflects the diversity and behavioral realism of actual social media users.

3.3.2 Perception

The Perception module serves as the agent's sensory system that bridges external information and internal cognition. At each timestep, an agent browses the top- K ranked posts from the Public Information Square and forms an *impression* of the current discourse. The LLM extracts entities, topics, and sentiment cues, producing a filtered impression conditioned on the agent's prior experiences and personality traits. Consequently, agents exhibit individualized attention: for instance, optimistic agents tend to emphasize positive information, while pessimistic ones may focus on negative aspects of the same message.

3.3.3 Memory

The Memory module endows agents with identity, continuity, and adaptive cognition, ensuring behavioral diversity

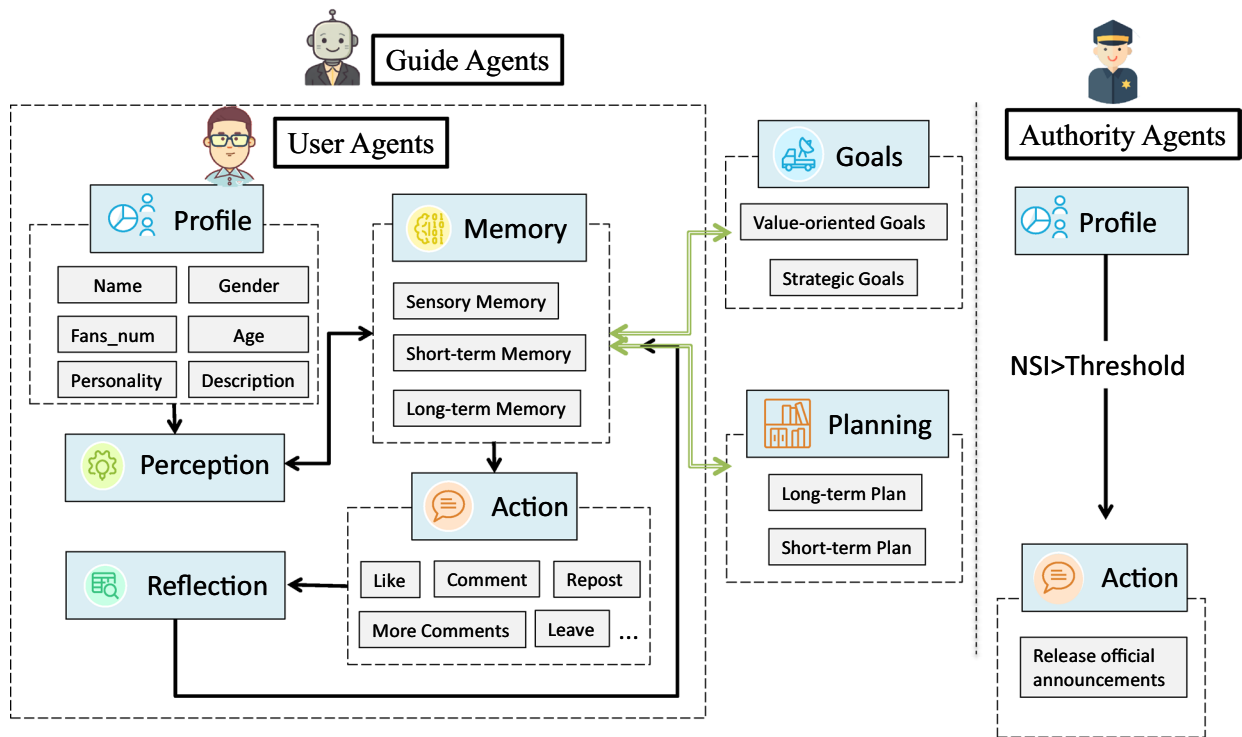


Fig. 2 Overview of the LLM-based Agent architecture. The agent follows a cognitive loop that begins with Perception, proceeds through Memory and Goal-driven Planning, and culminates in Action and Reflection. The framework integrates five core modules (Profile, Per-

ception, Memory, Action, Reflection) together with extended Goal and Planning components, enabling both ordinary and guiding agents to simulate opinion dynamics

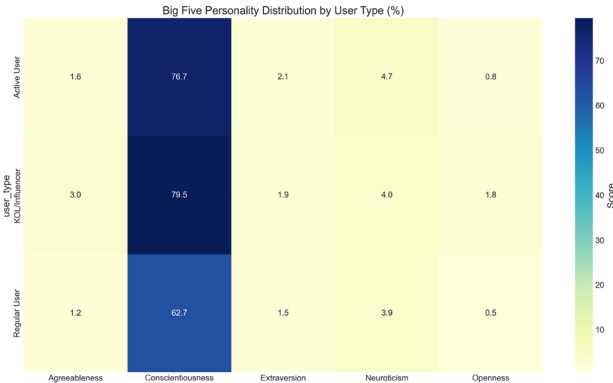


Fig. 3 Heatmap of Big Five personality distributions across user types (%). Conscientiousness is notably higher among KOLs, informing the initialization of influential agent profiles

across the simulated population. Drawing inspiration from cognitive psychology, the memory structure is organized into three tiers: (1) *long-term memory*, encoding stable beliefs, values, and identity features, with temporal decay modeled by an Ebbinghaus-inspired forgetting coefficient λ ; (2) *short-term memory*, maintaining transient emotions, opinions, and contextual states that evolve dynamically during the simulation; and (3) *sensory memory*, functioning as a temporary buffer that holds immediate perceptual stimuli

for up to 20 steps. The forgetting of long-term memories follows an exponential decay:

$$w(t) = \exp(-\lambda \Delta t) \cdot \alpha, \quad \lambda > 0, \alpha \in (0, 1]. \quad (8)$$

Through the interaction of these components, each agent continuously integrates new impressions with stored knowledge, allowing personal biases and emotional tendencies to emerge naturally in the simulation.

3.3.4 Action

The Action module translates cognition into observable behavior. Given its perceptions and memories, the agent reacts by choosing among commenting, liking, reposting, or ignoring a post. The LLM synthesizes the agent's emotional response with its personality traits and long-term beliefs to select the most plausible action. To reflect realistic social media dynamics, action frequency is governed by a Pareto distribution, ensuring that highly active users coexist with a larger population of less active ones:

$$P(\text{act}) \sim \text{Pareto}(x_m, k), \quad x \geq x_m, k > 1. \quad (9)$$

3.3.5 Reflection

Finally, the Update and Reflection module enables adaptation. After an action, the agent records the outcome—such as likes or replies—into memory, updating its short-term emotional state and, in some cases, refining its long-term beliefs. The reflection mechanism allows the LLM to evaluate whether outcomes align with the agent's goals, adjusting or discarding plans as necessary. To avoid unrealistic drift, however, long-term beliefs are only modified when persistent signals are observed. Collectively, these five modules provide agents with autonomy and adaptability within the simulated environment.

3.4 RAG-based guiding agent design

Building upon the base agent architecture, the Guiding Agent is designed to serve as an intelligent coordinator that bridges individual cognition and collective influence. It extends the core agent framework by incorporating explicit goal-setting, hierarchical planning, and a retrieval-augmented generation RAG mechanism to support informed and adaptive interventions. These components together enable guiding agents to pursue sustained objectives, formulate context-aware strategies, and generate evidence-based content that steers public discourse toward rational and constructive directions. The following subsections introduce the design of the *Goals*, *Planning*, and *RAG* modules in detail.

3.4.1 Goals

The Goals module defines the central purpose of the Guiding Agent—to guide public discussions toward more rational, balanced, and positive directions. Unlike ordinary agents that react passively to information, guiding agents are driven by clear objectives aimed at reducing misinformation, calming emotional polarization, and promoting constructive engagement. These goals evolve dynamically as the simulation progresses, allowing the agent to adjust its behavior in response to topic changes and sentiment

fluctuations. Table 1 contrasts SITGNet's goal-driven and hierarchical planning architecture with prior LLM-based frameworks, which often lack explicit goal-setting or restrict planning to narrow, task-oriented scenarios.

3.4.2 Planning

The Planning module enables guiding agents to transform their goals into coherent and executable actions. It combines long-term and short-term planning to maintain both consistency and flexibility. Long-term plans define sustained objectives such as stabilizing sentiment or improving information quality, while short-term plans translate them into concrete actions including posting frequency, tone, and timing. When sudden events occur, such as emotional surges or the spread of misinformation, plans are automatically adjusted to generate appropriate responses. This dynamic structure ensures that guiding agents remain aligned with their overall objectives while adapting effectively to real-time social dynamics. The resulting design allows guiding agents to act strategically, responsively, and coherently, achieving a stable yet adaptive influence on public discourse.

3.4.3 RAG framework

To enhance the reliability of Guiding Agents when utilizing external documents (e.g., .txt, .docx, .pdf, .xlsx), we designed and implemented a multi-stage RAG framework. The central objective of this framework is to mitigate the risk of LLM hallucinations and to ensure that generated content strikes an appropriate balance between factual accuracy, domain-specificity, and communicability. The overall architecture, as illustrated in Fig. 4, consists of five tightly integrated stages: Query Restructuring, Parallel Multi-Channel Retrieval, Fine-grained Reranking, Guidance-oriented Content Generation, and Incremental Knowledge Base Updates.

Restructuring At the entry point of the RAG pipeline, user queries are often short, abstract, or underspecified, which creates a semantic gap between the expressed intent and relevant document passages. To bridge this gap, our framework incorporates the Hypothetical Document Embeddings (HyDE) technique. Specifically, the LLM is prompted to generate a hypothetical document that directly addresses the query. This synthetic text is then encoded into a dense semantic vector and used as the retrieval query. By expanding the original query into a semantically richer representation, HyDE reduces ambiguity, improves alignment between user intent and corpus content, and leverages the LLM's world knowledge to create a high-quality probe document. This process substantially improves retrieval robustness, particularly in zero-shot and out-of-domain settings.

Table 1 Comparison of cognitive modules across representative LLM-based agent frameworks

Module	Generative Agents	MetaGPT	AutoAgents	§ SITGNet (Ours)
Profile	✓	✓	✓	✓ ✓
Perception	✓	✓	✓	✓ ✓
Memory	✓	✓	✓	✓ ✓
Action	✓	✓	✓	✓ ✓
Reflection	✓	✓	✓	× ✓
Planning	×	✓	✓	✓ ✓
Goals	×	×	×	× ✓

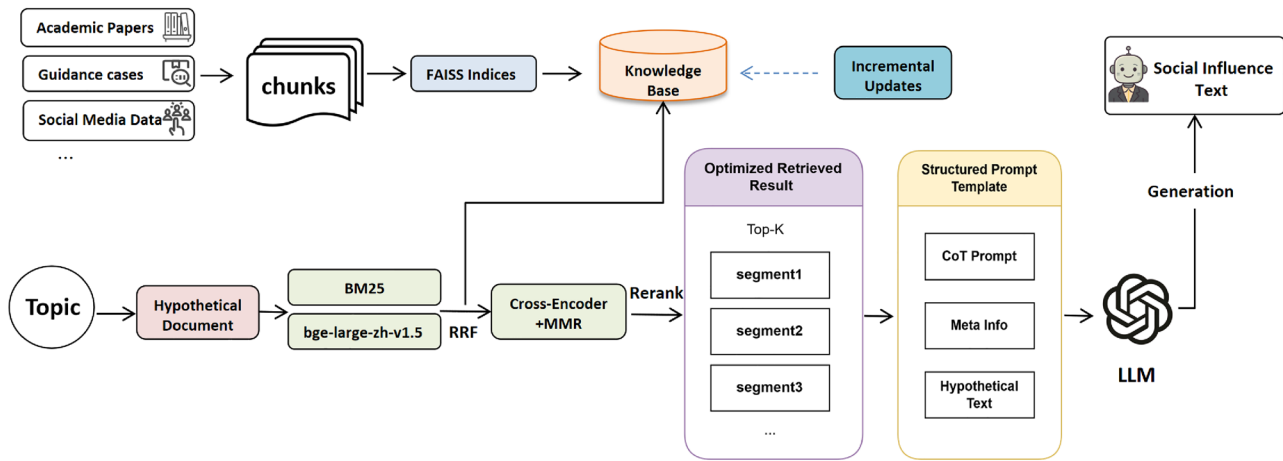


Fig. 4 Overview of the RAG framework. It consists of five stages: query restructuring, multi-channel retrieval, fine-grained reranking, guidance-oriented generation, and incremental knowledge updates

Parallel Multi-Channel Retrieval In the candidate document retrieval stage, we adopt a parallel multi-channel retrieval strategy to maximize recall. Concretely, the system performs sparse keyword-based retrieval using the BM25 algorithm and dense semantic retrieval using the *bge-large-zh-v1.5* embedding model. In addition, a hybrid retrieval mode is supported, in which results from both dense and sparse channels are merged via Reciprocal Rank Fusion (RRF) to combine their complementary strengths. To broaden coverage, we employ an over-fetch strategy (e.g., retrieving the top 20 candidates per channel before fusion), after which the merged set is deduplicated and narrowed to form a high-quality candidate pool.

Efficient Indexing and Storage At the indexing layer, our framework relies on the FAISS library for efficient similarity search. In the current implementation, we employ an IndexFlatL2 index, which is well-suited for medium-scale corpora and ensures exact nearest neighbor search. This setup provides fast and reliable retrieval, while keeping the system lightweight and easy to maintain. The framework is designed to support future extensions to more advanced FAISS indices (e.g., HNSW, OPQ, PQ) should the corpus scale increase significantly, enabling further optimization of the speed–accuracy–memory trade-off.

Fine-grained Reranking To ensure that the final retrieved set maintains both relevance and diversity, we employ a two-stage reranking process. First, a Cross-Encoder model (*bge-reranker-base*) is used to compute pairwise relevance scores between the query and candidate documents, yielding a refined ranking from which the top 10 passages are selected. Second, we apply the Maximal Marginal Relevance (MMR) algorithm to re-rank these candidates. By adjusting the trade-off parameter λ (set to 0.8 in our implementation), the reranker balances query relevance against content diversity. This process produces a final subset of

3–5 documents that serve as the knowledge foundation for subsequent generation.

Chain-of-Thought Based Guidance Content Generation During the generation phase, we integrate a Chain-of-Thought (CoT) based prompting strategy to prevent the model from mechanically reproducing retrieved content. Instead, the model is guided through a structured reasoning process. The CoT template prompts the model to (i) analyze the retrieved evidence to extract key themes and sentiment cues, (ii) derive potential guidance strategies that are contextually appropriate, and (iii) synthesize these into a coherent narrative. The final output is constrained to follow an “evidence–analysis–guidance” structure, ensuring that the generated content remains factually grounded, logically consistent, and communicatively effective. Moreover, the generated guidance is enforced to adopt a concise, positive, and persuasive tone, suitable for real-world public opinion guidance scenarios, as shown in Fig. 5.

Incremental Updates and Quality Assessment To maintain the timeliness and relevance of the knowledge base, we implement an incremental update mechanism. Each document is first segmented into semantically coherent chunks with controlled length and overlap to preserve contextual integrity. These chunks are enriched with metadata (e.g., source filename, position index, timestamp) before being embedded and stored in the FAISS index. A lightweight file monitoring module tracks modification timestamps to detect changes and trigger partial index rebuilding when necessary. Retrieval quality is periodically evaluated using standard IR metrics such as NDCG@k, Recall@k, and MRR, while user feedback is continuously incorporated to identify missing or outdated content. This feedback loop enables iterative refinement of chunking, metadata enrichment, and parameter tuning, thereby ensuring sustained retrieval quality and adaptability.

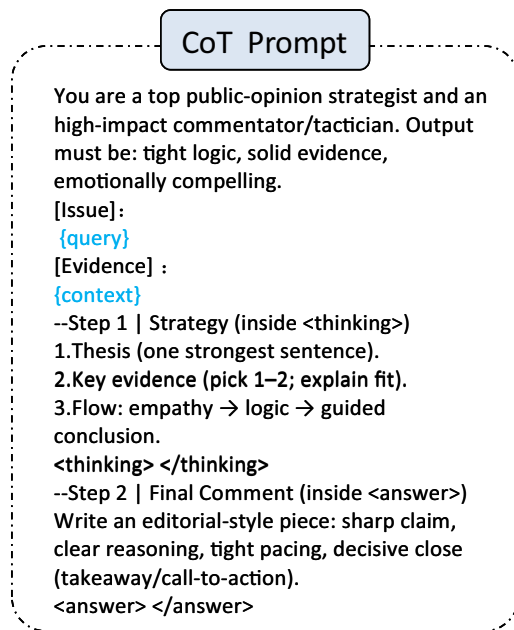


Fig. 5 CoT-based prompting template

To illustrate the RAG pipeline, consider a negative rumor about “milk quality issues”. The agent would: (1) Generate a hypothetical document; (2) Retrieve actual official statements and news reports using this query; (3) Rerank the results for relevance and finally generate a fact-based comment like, “I checked the official sources; the brand has released a valid quality report showing all batches are safe.”

4 Experimental setups

To systematically evaluate the effectiveness of SITGNet, we adopt a question-driven experimental design. Our experiments are structured around four central research questions:

- *Q1*: Can our LLM-based agents authentically emulate human-like behavior in online social interactions?
- *Q2*: Does our RAG framework significantly enhance the factuality and persuasiveness of guidance content?
- *Q3*: At the macro level, can SITGNet effectively suppress negative sentiment propagation and guide public opinion toward rational consensus?
- *Q4*: Is SITGNet computationally efficient and scalable for large-scale social simulation?

In the following subsections, we present detailed experimental setups and results corresponding to these questions.

Table 2 Sources of the RAG Knowledge Base

Source Type	Description	Num
Academic Papers	Research articles on opinion guidance and communication	122
Think Tank Cases	Guidance cases and policy analyses from research institutes	185
Media Discourses	Guiding news reports, editorials, and commentaries	340
Policy Documents	Government and international policy papers and guidelines	10
Industry Reports	Market and communication reports from associations and firms	15
Social Media Data	Guiding discourses from platforms (e.g., Weibo, Twitter)	1870

4.1 Data source

The construction of a realistic simulation environment relies on high-quality, representative data. We prepared two primary datasets: one for generating the user agent profiles and another to serve as the knowledge base for the RAG framework.

4.1.1 User profile dataset

The user profiles were derived from the `weibo.active_user_pool`, a large-scale dataset on Sina Weibo. The preparation pipeline included Data Cleaning and Filtering to ensure validity of key fields (e.g., ID, gender, bio, follower counts), Categorization and Sampling into three user tiers (KOLs/Influencers, Active Users, Normal Users) with a 5:20:75 stratified ratio, and Profile Augmentation and Anonymization using LLM-based biographical expansion, Big Five personality mapping, and hashing of identifiers. The final dataset contains 2,000 users (1,500 Normal, 400 Active, 100 KOLs), with a gender distribution of 1,068 male and 932 female, providing both realism and diversity while ensuring privacy.

4.1.2 RAG knowledge base

To enable retrieval-augmented generation, we constructed a heterogeneous knowledge base covering six source types, including academic papers, think tank cases, media discourses, policy documents, industry reports, and social media data. The preparation involved Collection and Filtering, Standardization, and Deduplication, ensuring consistency and quality across formats. As shown in Table 2, the final corpus contains 2,761 entries, providing both authoritative references and dynamic real-world discourses.

4.2 Implementation details

The SITGNet framework was implemented as a multi-agent social simulation system using Python 3.9 framework. All experiments were executed on a high-performance computing platform equipped with NVIDIA A800-SXM4-80GB GPUs. The simulation environment was designed as a custom framework to model dynamic public opinion evolution. A centralized controller manages the simulation clock, orchestrates event transitions, and records global states at each discrete timestep. Each event lifecycle is formalized as a finite-state machine with four transition phases (*Latent*, *Outbreak*, *Peak*, and *Decline*), governed by the global Event Heat $H(t)$, which is recalculated per iteration. Public discourse is represented as an Information Square that maintains and ranks posts according to an adaptive influence score updated after each round of agent interactions.

All agents are instantiated from a unified Agent base class, with cognitive functions supported by the DeepSeek-V3.1 model through asynchronous API calls. Each agent possesses a procedurally generated profile including demographics and Big Five personality traits. The cognitive cycle follows a Perception–Memory–Action–Update process. The memory module adopts a three-layer hierarchy: (1) long-term memory with an Ebbinghaus-inspired forgetting coefficient $\lambda = 0.8$ to model belief decay over time, (2) short-term memory for retaining recent experiences within limited timesteps, and (3) sensory memory serving as a buffer for immediate perceptual inputs. Each agent receives the top- $K = 10$ posts per step, integrates them with prior memory traces, and generates structured JSON-formatted actions (e.g., comments, likes, reposts). These actions are deterministically executed by the simulator to update the discourse state.

Guiding agents extend the base class by embedding explicit goal representations and hierarchical planning mechanisms, enabling proactive rather than reactive behavior. When triggered, guiding agents activate the RAGservice to construct evidence-grounded interventions. The RAG module is deployed as a standalone microservice using LangChain and LlamaIndex. Both dense and sparse retrieval are supported: documents are chunked into 512 tokens, embedded using `bge-large-zh-v1.5`, and indexed with FAISS; sparse retrieval employs BM25 via Elasticsearch. Retrieved results are re-ranked with `bge-reranker-base`, and the top-5 candidates are passed to the generator. The LLM produces structured outputs under a two-stage template (analysis and guidance). The RAG knowledge base contains 2,761 entries drawn from academic, media, and policy sources.

Each simulation scenario involves 100 agents. The system updates the Event Heat $H(t)$ and NSI after each step.

For all simulations, we set the maximum context length to 4,096 tokens, and the generation temperature to 0.7.

A centralized logging module continuously records both micro- and macro-level indicators, including Event Heat $H(t)$, NSI, agent activity distributions, post-level sentiment statistics, and engagement intensity by agent type. The system outputs JSON-formatted logs at each run for post-analysis using pandas, visualization in matplotlib, and statistical inspection via seaborn. Each experiment is repeated five times, and the mean and standard deviation are reported to account for stochastic effects.

5 Experiments results

5.1 Q1: Human-likeness of LLM agents

The first research question examines whether LLM-based agents can convincingly emulate human-like behavior in online social interactions, which is critical for the credibility of the simulation. If the agents' behavior diverges from human norms, the emergent opinion dynamics would lose validity, making this evaluation a foundational step. To assess this, we focused on two complementary dimensions: the consistency of ordinary user agents and the credibility of guiding agents.

For ordinary user agents, we evaluated whether their actions remained aligned with the demographic and psychological attributes encoded in their profiles. This required the agents to demonstrate coherence across three aspects: maintaining behaviors consistent with their social roles, expressing emotions congruent with their personality traits, and articulating stances consistent with their ideological leanings. Results summarized in Table 3 show that these agents sustain strong alignment across all dimensions, producing behaviors that mirror realistic heterogeneity in human societies. Importantly, the integration of SITGNet further enhances this alignment, reducing fluctuations and strengthening stability across diverse agent configurations. In practice, this means that a simulated doctor or teacher not only retains role-appropriate behaviors but also shows emotional and attitudinal responses consistent with real-world expectations, making the simulation's micro-level dynamics believable.

Guiding agents were evaluated with a focus on two complementary qualities. The first is content consistency, meaning their interventions must remain aligned with their assigned professional identity—for example, a health advisor consistently providing factually sound wellness information. The second is evasion capability, which refers to their ability to blend into the online environment without being recognized as artificial “water army” accounts. This quality

Table 3 User agent consistency evaluation across different LLMs.

Method	GPT-5			Gemini2.5-pro			Claude3.7-sonnet			Avg.
	Beh	Psy	Stn	Beh	Psy	Stn	Beh	Psy	Stn	Overall
DeepSeek-V3.1	0.83	0.82	0.66	0.91	0.87	0.92	0.81	0.80	0.82	0.83
Qwen-Plus	0.86	0.85	0.77	0.92	0.88	0.84	0.83	0.84	0.84	0.85
GPT-4o-mini	0.82	0.71	0.77	0.86	0.84	0.91	0.78	0.76	0.78	0.80
GLM-4.5-Flash	0.80	0.82	0.57	0.88	0.88	0.91	0.53	0.52	0.52	0.72
SITGNet (+DS-V3.1)	0.88	0.87	0.82	0.92	0.93	0.94	0.84	0.83	0.86	0.87
SITGNet (+Qwen-Plus)	0.86	0.89	0.79	0.93	0.92	0.86	0.82	0.86	0.85	0.86
SITGNet (+GPT-4o-mini)	0.84	0.73	0.81	0.89	0.90	0.96	0.80	0.78	0.80	0.83
SITGNet (+GLM-4.5-Flash)	0.82	0.84	0.75	0.89	0.89	0.88	0.65	0.61	0.72	0.78

Beh. (Behavior), *Psy.* (Psychology), *Stn.* (Stance). The best result in each column is highlighted in bold. Higher values are better

Table 4 Guiding agent consistency evaluation across different LLMs.

Method	GPT-5		Gemini2.5-pro		Claude3.7		Average	
	Con	Eva	Con	Eva	Con	Eva	Con	Eva
DeepSeek-V3.1	0.86	0.59	0.93	0.62	0.58	0.35	0.79	0.52
Qwen-Plus	0.66	0.54	0.94	0.56	0.56	0.51	0.72	0.54
GPT-4o-mini	0.64	0.56	0.86	0.55	0.13	0.65	0.54	0.59
GLM-4.5-Flash	0.44	0.52	0.94	0.61	0.08	0.29	0.49	0.47
SITGNet(+DS-V3.1)	0.89	0.68	0.91	0.72	0.67	0.58	0.82	0.66
SITGNet(+Qwen-Plus)	0.71	0.71	0.88	0.68	0.69	0.61	0.76	0.67
SITGNet(+GPT-4o-mini)	0.68	0.66	0.89	0.62	0.35	0.42	0.64	0.57
SITGNet(+GLM-4.5-Flash)	0.63	0.58	0.85	0.69	0.41	0.33	0.63	0.53

Con. (Content), *Eva.* (Evasiveness). The best result in each column is highlighted in bold. Higher values are better

is critical because once detected as inauthentic, the credibility of interventions would collapse and the overall guidance strategy would lose effectiveness. Results in Table 4 indicate that SITGNet guiding agents not only preserve content reliability but also achieve strong concealment, allowing them to operate credibly while exerting subtle influence within the network.

Therefore, in response to Q1, our results confirm that the agents' behaviors are consistent with their predefined profiles, demonstrating a high degree of human-likeness in their interactions. Taken together, these findings demonstrate that SITGNet agents achieve a balance between authenticity and guidance. User agents are sufficiently realistic to emulate the noisy diversity of online communities, while guiding agents act as credible yet unobtrusive influencers. This dual capacity ensures that emergent patterns in the simulation are grounded in plausible human behavior, lending confidence to downstream analyses of intervention strategies. Furthermore, the consistency of results across evaluators underscores the robustness of the framework, with DeepSeek emerging as the most reliable base model due to its strong overall stability and adaptability within SITGNet.

5.2 Q2: Effectiveness of the RAG framework

The second research question examines whether the RAG framework can enhance the factuality, coherence, and

persuasiveness of guidance content. Large language models excel at fluent generation, yet they remain vulnerable to hallucinations and factual drift.

To address this, we performed three sets of experiments. First, we compared retrieval strategies under standard metrics such as NDCG@k and Recall@k. As shown in Table 5, traditional sparse or dense methods alone achieve only partial coverage, while hybrid retrieval improves recall but still suffers from ranking errors. Our approach, which integrates query restructuring, multi-channel retrieval, and fine-grained reranking, consistently achieves the highest scores, particularly at low values of k . This demonstrates that reranking is essential to ensure that the most relevant documents occupy the top positions, thereby grounding generation on reliable evidence.

Second, we evaluated the impact of SITGNet augmentation on guidance quality. Results in Table 6 show clear improvements in overall scores when SITGNet is enabled. Gains are observed across all tested models, confirming that the framework strengthens both factual grounding and rhetorical coherence while remaining model-agnostic.

Finally, we conducted an ablation study to isolate the role of each module. Table 7 reveals that removing the reranking stage leads to the sharpest drop in factual accuracy, underscoring its importance for retrieval precision. Excluding HyDE reduces coverage and factual reliability, while disabling Chain-of-Thought (CoT) decreases naturalness and

Table 5 Performance comparison of retrieval baselines on NDCG@k and Recall@k. Bold indicates the best performance

Method	NDCG@k				Recall@k			
	@1	@3	@5	@10	@1	@3	@5	@10
BM25 (Sparse)	0.050	0.163	0.163	0.163	0.005	0.197	0.207	0.233
BM25+ (Sparse, variant)	0.050	0.153	0.152	0.152	0.009	0.188	0.200	0.217
SPLADE (Sparse, learned)	0.050	0.153	0.152	0.152	0.005	0.180	0.191	0.217
SBERT (Dense)	0.933	0.928	0.932	0.929	0.130	0.323	0.505	0.950
ANCE (Dense)	0.233	0.371	0.410	0.441	0.047	0.217	0.330	0.667
ColBERT (Dense, late interaction)	0.917	0.924	0.928	0.925	0.135	0.339	0.516	0.950
Hybrid (RRF: BM25 + Dense)	0.367	0.756	0.764	0.768	0.061	0.257	0.446	1.000
Vanilla RAG (Retriever + Generator)	0.933	0.945	0.944	0.933	0.147	0.339	0.510	0.983
Ours	0.970	0.966	0.948	0.940	0.152	0.348	0.524	0.986

Higher is better

Table 6 Performance comparison across domains: baseline LLMs versus models augmented with SITGNet (percent scale). Bold indicates the best performance

Model	Condition	Tech.	Edu.	Pub.	Bus.	Env.	Overall
DeepSeekV3.1	Base	83.60	84.30	84.90	85.70	72.30	82.40
	SITGNet	85.50	84.50	85.50	85.00	87.00	87.00
Qwen-Plus	Base	82.00	80.00	80.90	80.00	80.60	80.80
	SITGNet	82.80	83.60	80.50	82.20	80.80	82.00
GPT-4	Base	82.20	83.40	81.60	81.70	81.40	82.10
	SITGNet	84.80	83.20	83.00	85.00	82.80	83.80
GLM-4.5	Base	84.60	85.40	84.40	85.40	84.30	84.80
	SITGNet	86.20	84.20	86.50	86.20	85.80	86.00

Table 7 System-level performance metrics of SITGNet framework. Bold indicates the best performance

Config	Gui.	Fac.	Nat.	Per.	Com.	Overall
w/o HyDE	8.85	7.95	8.50	8.25	8.60	8.43
w/o CoT	8.55	8.40	7.55	8.55	9.10	8.43
w/o Rerank	8.95	7.98	8.65	8.15	8.40	8.43
Full RAG	9.00	8.50	8.70	8.60	9.15	8.79

Table 8 Core performance metrics comparison across different intervention strategies

Strategy	NSI Control				Emotional		Social Confidence	
	Peak NSI	Avg NSI	Final NSI	Reduction (%)	Final Emo.	Variance	Final Con.	Variance
No Intervention	0.850	0.670	0.720	—	0.320	0.0452	0.410	0.0385
Guiding Agents	0.710	0.520	0.480	33.3 ↓	0.450	0.0321	0.520	0.0283
Authority Intervention	0.480	0.380	0.310	56.9 ↓	0.580	0.0187	0.680	0.0152
Hybrid Strategy	0.420	0.310	0.250	65.3 ↓	0.640	0.0123	0.740	0.0098

persuasiveness, as the model tends to produce stitched rather than reasoned content. The full pipeline consistently outperforms all ablated variants, demonstrating that each component contributes a distinct and indispensable improvement.

Thus, for Q2, the ablation study shows that the complete RAG framework is crucial for generating high-quality, persuasive guidance, significantly outperforming its variants. Taken together, these findings confirm that SITGNet's RAG framework substantially enhances both reliability and persuasive power. By combining query restructuring, fine-grained reranking, and CoT prompting, SITGNet provides guiding agents with high-quality context that enables accurate, coherent, and compelling interventions.

5.3 Q3: Opinion guidance in simulations

While Q1 and Q2 confirmed the micro-level credibility of agents and the factual reliability of generated content, Q3 evaluates whether these local mechanisms translate into desirable emergent outcomes when scaled to complex simulations.

We compared four representative intervention configurations under the same environment, scheduling, and trigger rules: No Intervention, Guiding Agents, Authority Intervention (authoritative only), and the combined Hybrid Strategy. Macro performance results are summarized in Table 8. The evidence shows a

clear hierarchy: the baseline produces the highest peak and average NSI values, with unstable emotional and confidence patterns; guiding-only interventions improve outcomes but struggle to contain peak escalation; authority-only interventions achieve stronger suppression and stabilization; and the hybrid approach consistently outperforms all others, yielding the lowest final NSI, the greatest reduction rate, and the most stable emotional and confidence trajectories. Importantly, variance declines without a loss of engagement, indicating that improvements stem from stabilizing the discourse rather than suppressing participation.

Further evidence is provided by the diagnostic visualizations in Fig. 6 and Fig. 7. The multi-dimensional radar profile (Fig. 6) shows that the hybrid configuration achieves a balanced outcome: higher trust, engagement, rationality, and compliance, alongside reduced emotional intensity and skepticism. In parallel, the sentiment intensity distribution (Fig. 7) reveals a clear depolarization pattern: extreme negatives shrink, the neutral-to-moderate region expands, and positive stances gradually increase. This synergistic effect can be explained by the complementary roles these agents play. The authority agent likely influences the public's focus by providing a reliable starting point. Subsequently,



Fig. 6 Multi-dimensional radar across six dimensions (Emotional Intensity, Trust, Engagement, Rationality, Compliance, Skepticism), highlighting the hybrid strategy's superior profile

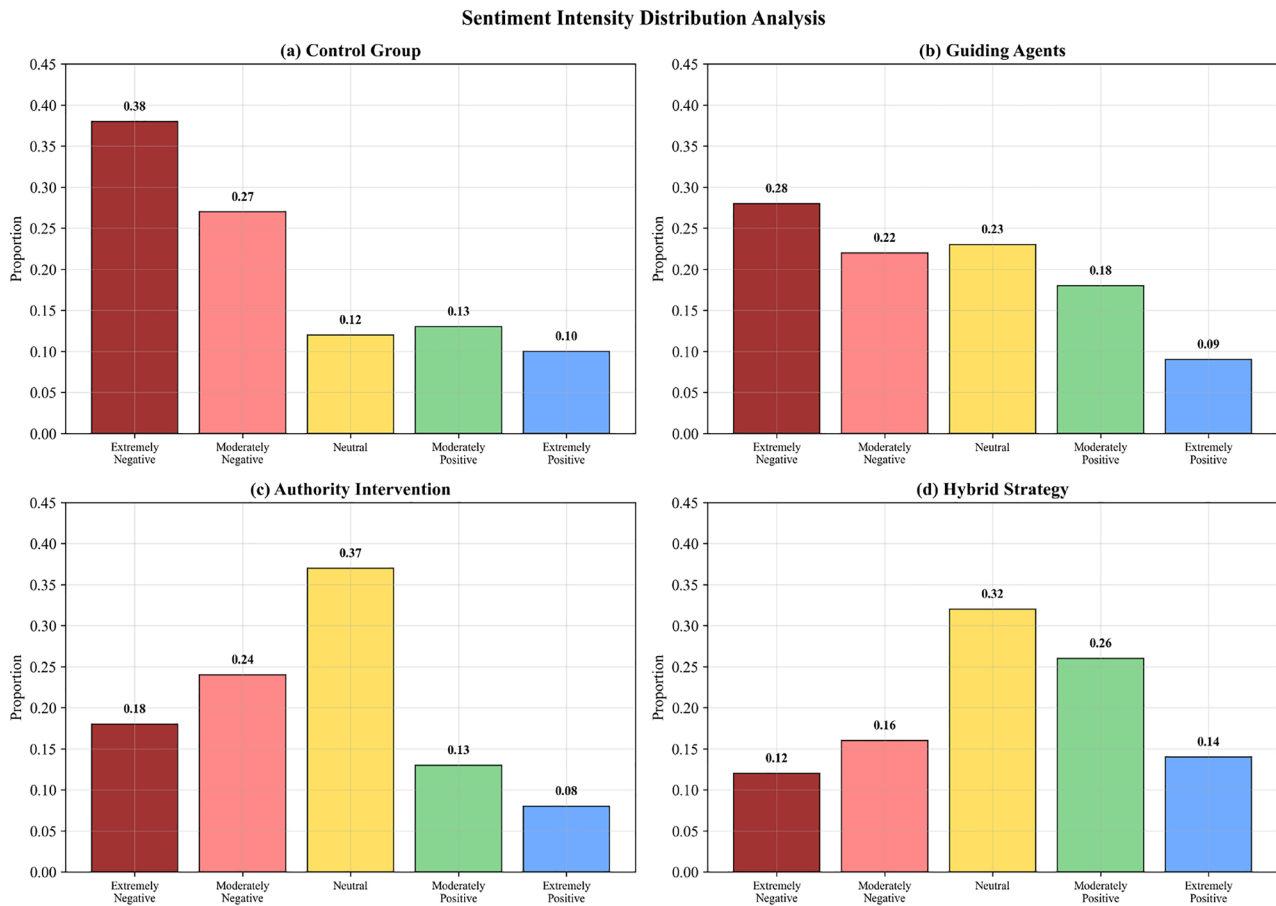


Fig. 7 Sentiment intensity distributions under four strategies. Extremes contract, the middle expands, and positives increase, indicating depolarization and guided conversion

the guiding agents enhance message credibility and reach through peer-to-peer diffusion, a mechanism consistent with social influence theories. Taken together, these observations suggest a two-phase mechanism. Authority-driven clarifications act as early containment, preventing unchecked escalation, while guiding agents diffuse constructive signals across the network, converting neutrality into support. This complementary dynamic aligns with the event-heat scheduler and demonstrates robustness across diverse contexts.

In conclusion, addressing Q3, the 'Hybrid Strategy' proves to be the most effective approach for mitigating negative sentiment, highlighting the importance of combining authoritative and grassroots interventions. Overall, the macro-level findings confirm that SITGNet can steer collective opinion toward positive and rational outcomes. By combining authoritative credibility with organic peer influence, the hybrid strategy produces not only quantitative improvements but also a structural shift in discourse dynamics, establishing a resilient equilibrium that generalizes across different crisis scenarios.

5.4 Q4: Efficiency and scalability

The fourth research question explores the scalability and computational feasibility of deploying SITGNet in large-scale simulation environments. As the number of agents and interaction density grow, both communication overhead and model inference latency become critical factors. This evaluation assesses the system's runtime efficiency, token throughput, and memory footprint under increasing simulation complexity, thereby determining the framework's suitability for practical deployment in research and policy analysis.

We scaled the number of agents from 10 to 500 under identical environment and model settings, and evaluated system-level performance across simulation metrics, token efficiency, and step performance. Table 9 summarizes the results. The findings show that token throughput remains stable at approximately 4k–4.5k tokens per minute across all scales, indicating that prompt and memory management prevent context inflation. Average time per step is consistently bounded within 22–27 s, and wall-clock time

Table 9 System-level performance metrics of SITGNet framework

Agents	Time (min)	Tokens/min	API Calls	Total Steps
10	31.1	4092.9	1038	84
20	75.7	4220.0	1850	170
50	165.1	4077.3	5120	401
100	322.5	4468.0	9210	834
200	737.2	4049.5	21463	1721
500	1829.4	4325.7	50980	4210

increases roughly linearly with agent count, demonstrating that no exponential or quadratic bottlenecks appear. Similarly, API calls, comments, and total steps grow proportionally, confirming that workload expands predictably with the population size.

Answering Q4, SITGNet demonstrates good scalability, with computation time increasing linearly with the number of agents, making larger-scale simulations feasible. These results indicate that SITGNet achieves near-linear scalability with stable per-step latency and predictable token budgets. In practice, efficiency can be further optimized through strategies like caching LLM responses and implementing a selective agent activation schedule. By allocating proportionate compute resources, the framework can support thousands of agents, enabling SITGNet to serve not only as a research platform but also as a practical system for real-time public opinion monitoring and strategy testing.

The experimental results have demonstrated the capabilities of our framework. In the following section, we discuss the broader implications and limitations of these findings.

6 Related work

6.1 Large language models for agent simulation

Recent advances in large language models (LLMs) have stimulated research on LLM-powered autonomous agents. Unlike traditional rule-based agents, LLM-based agents integrate modules such as memory, reflection, and planning, enabling them to exhibit adaptive, human-like behaviors. Work such as Generative Agents [23] demonstrated that dozens of LLM agents can simulate daily life in a virtual town with believable social interactions. Building on this paradigm, LLM agents have been applied to contexts including software development and collaborative problem-solving [24, 25], as well as behavioral simulations of social interaction, trust, and competition [26–28].

Beyond small-scale prototypes, platforms like AgentSociety [29], GenSim [30], and SocialVerse [31] explore larger environments, where agents are configured with heterogeneous profiles, long-term memory, and emotion-aware reasoning to model phenomena such as polarization and

misinformation spread. These works highlight the potential of LLM-based agents to reproduce complex human-like behaviors, while also revealing challenges of scalability, long-term consistency, and alignment with real-world user data. Our work extends this line of research by constructing LLM-based agents grounded in empirical attributes to more faithfully simulate online opinion dynamics.

6.2 Social simulation

Social simulation provides a foundation for understanding how individual interactions generate collective phenomena. Early approaches relied on system dynamics and cellular automata, but agent-based modeling (ABM) soon became the dominant paradigm. Classical models such as the Bounded Confidence Model [32] and the Axelrod Cultural Dissemination Model [33] offered abstractions for studying polarization and consensus, but assumed overly simple decision mechanisms and often failed to capture the richness of real-world social interactions. Inductive approaches based on human experiments provided empirical grounding but were costly and limited in scale [34].

To overcome these limitations, recent work has combined ABM with empirical data and adaptive architectures. Hybrid models have incorporated psychological mechanisms such as emotional contagion and selective exposure [35, 36], while data-driven approaches integrate social media traces to replicate opinion dynamics under real-world conditions [37]. New platforms support more flexible simulations: Sotopia [38] offers a benchmark for evaluating social intelligence, while S³ [39] leverages LLM agents to approximate public opinion evolution. Other systems such as OASIS [40] and SocioVerse [31] focus on large-scale simulations, incorporating dynamic contexts, recommendation mechanisms, and reconstructed real-world information to model polarization and misinformation cascades.

6.3 Other work

Another related line of research concerns misinformation detection and content monitoring in social networks. Work has addressed rumor detection, fake news classification, and bot identification using linguistic features, propagation patterns, and metadata [7, 41]. These efforts provide defensive mechanisms for identifying harmful information at scale, but remain largely reactive and do not alter opinion trajectories once misinformation spreads. More recently, retrieval-augmented generation (RAG) has been explored to enhance factual grounding in generative models [42, 43]. By integrating external knowledge into text generation, RAG offers a pathway for producing reliable counter-narratives, though applications to opinion guidance remain limited.

Research on user behavior guidance in social networks has focused mainly on monitoring and analysis. Typical tasks include event and topic detection, clustering, and sentiment analysis such as polarity classification and stance tracking [8, 10]. Intervention studies have emphasized identifying opinion leaders or network nodes to maximize message reach [11], and monitoring programs on Twitter have been used to capture public mood for policy communication. Beyond monitoring, some attempts at intervention rely on recommendation or exposure strategies, such as personalized content pushes or suppression of harmful content [9]. However, these approaches often lack principled computational models of influence and do not capture how individuals integrate new information into opinions. As noted in recent surveys [44, 45], platforms still provide limited means for steering user behavior, underscoring the need for proactive and computationally grounded opinion guidance.

7 Conclusions

In this paper, we propose SITGNet, an LLM-based multi-agent framework for simulating opinion dynamics and proactive guidance on social media. By incorporating empirically grounded agent profiles, cognitively rich interaction loops, and a retrieval-augmented generation pipeline, we create a digital social sandbox capable of reproducing realistic discourse evolution under complex scenarios. We also design specialized guiding agents, including authority and ordinary types, to deliver timely interventions that blend credibility with subtle influence.

Our evaluations confirm the effectiveness of SITGNet. The agents exhibit human-like consistency in behavior, the RAG framework significantly improves factuality and persuasiveness of guidance content, and hybrid intervention strategies effectively suppress negative sentiment and mitigate polarization. Furthermore, SITGNet demonstrates near-linear scalability, making it practical for large-scale social simulations.

While this study provides valuable insights, we acknowledge its limitations. First, our agent profiles are based on data from a single social media platform (Weibo), and their generalizability to other platforms needs further investigation. Second, the event lifecycle model is a simplification of complex real-world events. Future work could explore more dynamic and multi-faceted event models. Finally, addressing the inherent biases of LLMs remains a critical challenge for future research in this domain.

In future work, we aim to reinforce our agents with strategies that balance short-term responsiveness and long-term stability, ensuring more sustainable influence. We will also explore multi-perspective guidance mechanisms

that combine factual clarification, emotional support, and community engagement, while leveraging user profiling to deliver personalized and precise interventions. These directions will further enhance SITGNet's ability to support nuanced and effective opinion guidance in complex online environments.

8 Ethical impacts

We acknowledge the profound ethical implications and dual-use nature of a system capable of simulating and influencing public opinion. The primary and intended purpose of SITGNet is to serve as a defensive research sandbox, designed to help institutions understand and counter the spread of harmful narratives, such as disinformation and panic-inducing rumors, thereby enhancing societal resilience. However, we also recognize the inherent risks of misuse for malicious manipulation or the reinforcement of societal biases.

To mitigate these risks, we advocate for transparency and responsible use. The RAG mechanism, which grounds responses in factual knowledge bases, is an inherent design choice to reduce the generation of misinformation. Furthermore, we recommend that any real-world application of such systems be subject to strict ethical oversight and public scrutiny. We strongly emphasize that SITGNet is strictly a tool for modeling and understanding, not for unchecked real-world deployment. Its development and application should be governed by principles of transparency, accountability, and rigorous ethical standards, with the ultimate goal of fostering a healthier and more resilient online ecosystem through responsible innovation.

Acknowledgements This work is supported by the NSFC through grants U25B2029, 62322202, 62441612, and 62432006, Beijing Natural Science Foundation through grant L253021, the Pioneer and Leading Goose R&D Program of Zhejiang through grant 2025C02044, Local Science and Technology Development Fund of Hebei Province Guided by the Central Government of China through grant 254Z9902G, Hebei Natural Science Foundation through grant F2024210008, Major Science Technology Special Projects of Yunnan Province through grants 202502AD080012 and 202502AD080006, and the Fundamental Research Funds for the Central Universities.

Author contributions Kun Zhang conceived and designed the overall study, developed the methodological framework, and drafted the manuscript. Xiaoyan Yu revised and refined the manuscript for clarity, coherence, and consistency. Hao Peng, as the corresponding author, provided conceptual guidance, supervised the research process, and offered critical insights throughout the study design and manuscript development. Zhe Yang implemented and optimized the RAG-based experimental modules. Ye Tian was responsible for data preprocessing and integration. Hao Jin and Hui Lin collected and organized the original datasets used in the study. Tuoyu Feng contributed to the visualization and graphical presentation of experimental results. All authors read and approved the final version of the manuscript.

Data availability No datasets were generated or analysed during the current study.

Declarations

Conflict of interest The authors declare no Conflict of interest.

References

- Lazer D, Baum M, Benkler Y et al (2018) The science of fake news. *Science* 359(6380):1094–1096
- Vosoughi S, Roy D, Aral S (2018) The spread of true and false news online. *Science* 359(6380):1146–1151
- Starbird K (2019) Disinformation's spread: bots, trolls and all of us. *Nature* 571(7766):449–450
- Cinelli M, Quattrocioni W, Galeazzi A et al (2020) The COVID-19 social media infodemic. *Sci Rep* 10(1):16598
- Allcott H, Gentzkow M (2017) Social media and fake news in the 2016 election. *J Econ Perspect* 31(2):211–236
- Del Vicario M, Bessi A, Zollo F et al (2016) The spreading of misinformation online. *Proc Natl Acad Sci USA* 113(3):554–559
- Shu K, Sliva A, Wang S et al (2017) Fake news detection on social media: a data mining perspective. *ACM SIGKDD Explor Newsl* 19(1):22–36
- Mohammad SM (2016) Sentiment analysis: Detecting valence, emotions, and other affectual states from text. *Emotion Measurement*. Woodhead Publishing, pp 201–237
- Ferrara E, Varol O, Davis C et al (2016) The rise of social bots. *Commun ACM* 59(7):96–104
- Cambria E, Poria S, Hazarika D, et al (2018) Senticnet 5: Discovering conceptual primitives for sentiment analysis by means of context embeddings. In: *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18)*, pp 1795–1802
- Bakshy E, Hofman JM, Mason WA, et al (2011) Everyone's an influencer: Quantifying influence on twitter. In: *Proceedings of the fourth ACM international conference on Web search and data mining*, pp 65–74
- Weng L, Menczer F, Ahn YY (2013) Virality prediction and community structure in social networks. In: *Proceedings of the 7th International AAAI Conference on Weblogs and Social Media (ICWSM)*, pp 536–545
- Ziems C, Zhang Y, Chen X, et al (2023) Can large language models transform computational social science? [arXiv:https://arxiv.org/abs/2305.03514](https://arxiv.org/abs/2305.03514)
- OpenAI (2023) Gpt-4 technical report. [arXiv:https://arxiv.org/abs/2303.08774](https://arxiv.org/abs/2303.08774)
- Touvron H, Lavril T, Izacard G, et al (2023) Llama: Open and efficient foundation language models. [arXiv:https://arxiv.org/abs/2302.13971](https://arxiv.org/abs/2302.13971)
- Conte R, Gilbert N, Bonelli G et al (2012) Manifesto of computational social science. *EPJ Data Science* 1(1):1–24
- Epstein JM (1999) Agent-based computational models and generative social science. *Complexity* 4(5):41–60
- Macy MW, Willer R (2002) From factors to actors: computational sociology and agent-based modeling. *Annu Rev Sociol* 28(1):143–166
- Jin B et al (2025) How to find opinion leader on the online social network? *Appl Intell* 55(7):1–15
- Kang M et al (2023) Detection of opinion leaders: Static vs. dynamic evaluation. *PLOS ONE* 18(5):e028xxx
- Chu M, et al (2024) Emotional contagion on social media and the simulation of online emotional transmission. *Human Behavior and Social Media (Nature's journal series)*
- Yu Y, et al (2025) Emotions in online content diffusion. *INFORMS journal (ISRE)*
- Park JS, O'Brien JC, Cai CJ, et al (2023) Generative agents: Interactive simulacra of human behavior. In: *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST)*, pp 1–22
- Hong Z, Zhang Z, et al (2023) Metagpt: Meta programming for multi-agent collaborative framework. In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*
- Qian C, Wang X, Shi W, et al (2023) Communicative agents for software development. [arXiv preprint arXiv:2307.07924](https://arxiv.org/abs/2307.07924)
- Li C, Zhang H, et al (2023) Exploring social behaviors of llm-based agents in simulated environments. In: *Proceedings of the 2023 International Conference on Computational Social Science (IC2S2)*
- Xie W, Liu S, et al (2024) Modeling trust dynamics with llm-based agents. [arXiv preprint arXiv:2401.01234](https://arxiv.org/abs/2401.01234)
- Zhao Y, Chen L, et al (2023) Competitive behaviors of llm agents in multi-agent economic games. [arXiv preprint arXiv:2311.02145](https://arxiv.org/abs/2311.02145)
- Gao T, Huang Y, et al (2024) Agentsociety: Large-scale social simulation with llm-based agents. [arXiv preprint arXiv:2403.10258](https://arxiv.org/abs/2403.10258)
- Wang Z, Zhou R, et al (2024) Gensim: A general framework for llm-based multi-agent simulation. [arXiv preprint arXiv:2402.09999](https://arxiv.org/abs/2402.09999)
- Liu Y, Zhang H, et al (2024) Socialverse: Towards realistic multi-agent social simulation with llms. [arXiv preprint arXiv:2404.01567](https://arxiv.org/abs/2404.01567)
- Deffuant G, Neau D, Amblard F et al (2000) Mixing beliefs among interacting agents. *Advances in Complex Systems* 3(01n04):87–98
- Axelrod R (1997) The dissemination of culture: a model with local convergence and global polarization. *J Conflict Resolut* 41(2):203–226
- Chuang E, Rogers EM (2023) Opinion dynamics with heterogeneous thresholds: empirical and simulation-based insights. *Sci Rep* 13(1):1143
- Flache A, Mäs M, Feliciani T et al (2017) Models of social influence: towards the next frontiers. *J Artif Soc Soc Simul* 20(4):2
- Lorenz J, Rauhut H (2021) Computational approaches to social dynamics. *Annu Rev Sociol* 47:485–508
- Perra N, Rocha LE (2019) Modelling human behaviour in social networks: a survey. *EPJ Data Sci* 8(1):1–23
- Zhou S, Wang Z, et al (2023) Sotopia: A benchmark for social intelligence in llm agents. In: *Proceedings of the 37th Conference on Neural Information Processing Systems (NeurIPS)*
- Gao T, Chen Y, et al (2023) S3: Social-network simulation system with llm agents. In: *Proceedings of the 2023 International Conference on Computational Social Science (IC2S2)*
- Yang Z, Zhang Z, Zheng Z, et al (2024) Oasis: Open agent social interaction simulations with one million agents. In: *NeurIPS Workshop on Open-World Agents*
- Zhou X, Zafarani R (2020) Fake news detection via nlp is vulnerable to adversarial attacks. In: *Proceedings of the 36th Annual Computer Security Applications Conference (ACSAC)*, pp 1–12
- Lewis P, Perez E, Piktus A, et al (2020) Retrieval-augmented generation for knowledge-intensive nlp tasks. In: *NeurIPS 2020*
- Shuster K, Piktus A, Komeili M, et al (2022) Language models that seek for knowledge: Modular search & generation for dialogue and beyond. In: *ACL 2022*
- Friederike H (2021) How public opinion evolves on social media: mechanisms and intervention strategies. *Curr Opin Psychol* 44:36–42
- Galam S (2022) *Opinion Dynamics: Theoretical and Practical Perspectives*. Springer

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.