Rethink Rumor Detection in the Era of LLMs: A Review

Chang Yang, Peng Zhang*, Jing Zhang, Hui Gao, Changhao Song
College of Intelligence and Computing, Tianjin University, Tianjin, China
{yangchang, pzhang}@tju.edu.cn

Abstract

The rise of large language models (LLMs) has fundamentally reshaped the technological paradigm of rumor detection, offering transformative opportunities to construct adaptive detection systems while simultaneously ushering in new threats, such as "logically perfect rumors". This paper aims to unify existing methods in the field of rumor detection and reveal the logical mechanisms behind them. From the perspective of complex systems, we innovatively propose a Cognition-Interaction-Behavior (CIB) tri-level framework for rumor detection based on collective intelligence and explore the synergistic relationship between LLMs and collective intelligence in rumor governance. We identify promising future research directions, including advancing agent-based modeling to capture complex rumor dynamics, addressing emerging challenges unique to the LLM era, and interdisciplinary perspectives. We hope this work lays a theoretical foundation for next-generation rumor detection paradigms and offers valuable insights for advancing the field.

1 Introduction

In the digital era, the widespread adoption of social media and the explosion of user-generated content have enabled rumors to threaten public safety and social trust at unprecedented speeds, scales, and levels of complexity (Kim and Dennis, 2019). Meanwhile, the rapid advancements in large language models (LLMs) have demonstrated remarkable performance across various fields (Tan et al., 2023; Poldrack et al., 2023), but it has also brought challenges that cannot be ignored. Models like GPT-4 (Achiam et al., 2023) and DeepSeek (Guo et al., 2025), known for their deep semantic understanding and reasoning capabilities, can generate highly credible and logically coherent professional content. However, this ability can also be used to generate "logically perfect rumors" (such as false arguments based on chain reasoning), which are far more concealed and misleading than traditional generation methods.(Bommasani et al., 2021; Kreps et al., 2022). For example, studies have shown that ChatGPT, when provided with malicious prompts, can not only optimize deceptive text but also proactively enhance their disguise by incorporating additional misleading details (Augenstein et al., 2024). Thus, leveraging the powerful capabilities of LLMs while addressing their inherent limitations has emerged as an urgent challenge in the field of rumor detection.

Existing rumor detection surveys primarily focus on the dissemination mechanisms of rumors on social media (Shu et al., 2017; Del Vicario et al., 2016; Johnson et al., 2020), the psychological mechanisms underlying belief in rumors (Roozenbeek et al., 2020), and effective intervention strategies (Zubiaga et al., 2015; Guess et al., 2020). However, most existing frameworks primarily rely on feature-based or technical classifications, which result in two primary issues: (1) the failure to thoroughly explore the theoretical and logical connections between detection methods and rumor propagation mechanisms, and (2) the inability to effectively reveal the intrinsic relationships between features, especially in the context of research on LLMs in this field (Chen and Shu, 2024).

To bridge this gap, we introduce a Cognition-Interaction-Behavior (CIB) tri-level framework to systematically elucidate the underlying logic of rumor propagation and detection on social networks. The specific contributions of this work include: (1) A new theoretical paradigm for rumor detection. The construction of the CIB framework unifies existing rumor detection methods and uncovers the multi-scale coupling mechanisms underlying rumor propagation, including collective knowledge emergence, interactive network evolution, and iterative behavioral patterns. (2) A systematic ex-

* Corresponding author

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 8730–8749
November 4-9, 2025

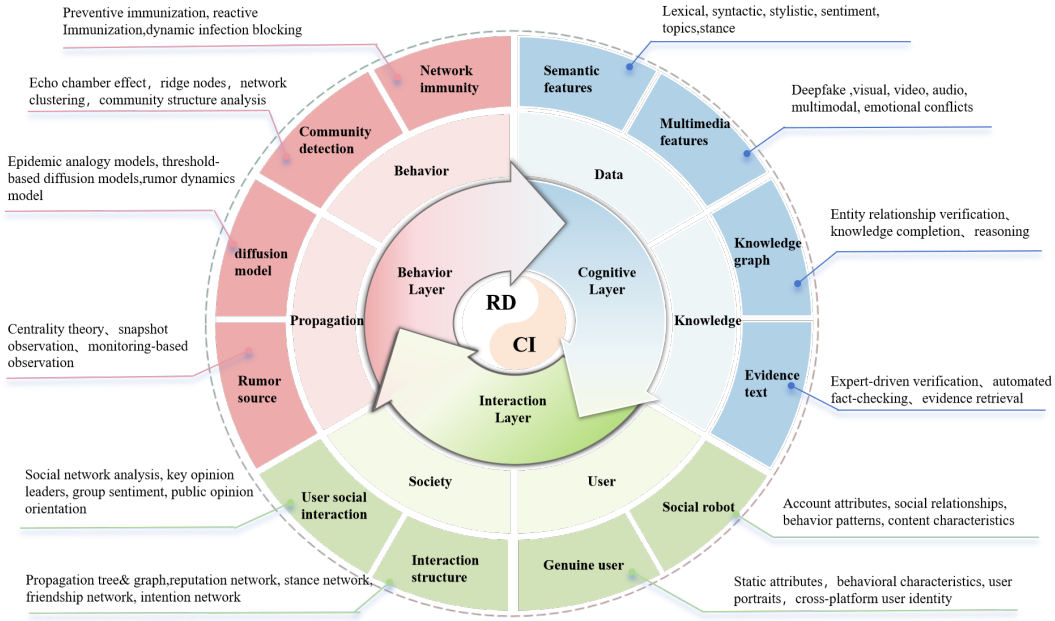

Figure 1: The three-layer architecture operates collaboratively. The cognition layer integrates multi-source evidence to provide informational support for the interaction layer. Through user interactions, the interaction layer facilitates the formation of the behavior layer. The behavior layer, in turn, continuously refines the cognition layer through accumulated experiences and collective cognitive feedback. (RD is rumor detection; CI is collective intelligence, driving information's dynamic reconstruction and optimization).

ploration of LLMs' multifaceted roles in rumor detection and the synergies with collective intelligence, forming a more comprehensive adaptive governance system. (3) A summary of the core opportunities and challenges of rumor detection in the LLM era and an outline of future development pathways for rumor detection.

2 Collective Intelligence-Based Rumor Detection Framework

In the social media ecosystem, user communities serve dual roles as both disseminators and evaluators, forming the self-organizing foundation of the networked information ecology. Studies have shown that through cross-validation among users and the interplay of opinions, social networks can facilitate collective cognitive correction (Ma et al., 2018). Compared to individual cognition, collective intelligence leverages the integration of diverse knowledge and dynamic interactions, demonstrating superior cognitive capabilities in addressing complex information (Castillo et al., 2011), thereby offering a novel approach to advancing rumor detection (Phan et al., 2023).

From the perspective of complex systems, the emergence of collective intelligence is essentially a self-organizing process driven by the reduction of information entropy. During this process, social media users' diverse cognition, social connections, and dynamic behaviors interact, facilitating information flow and collaborative evolution. Rumor diffusion, as a specific form of information dissemination, is often constrained by individuals' cognitive thresholds (e.g., cognitive abilities, emotional biases) and the topological structure of the social network. At its core, rumor diffusion can be viewed as a staged state of cognitive imbalance: it arises when users, driven by information uncertainty and emotional impetus, engage in social interactions to reduce uncertainty, which in turn drives the continuous evolution of network structures (Allcott and Gentzkow, 2017). This process generates macro-level dissemination behaviors (potentially unintentionally promoting rumor propagation). However, collective intelligence can dynamically correct such imbalanced states in social networks through multi-level knowledge sharing and interaction.

Based on the above theoretical construction, this study proposes a tri-level framework for rumor detection based on collective intelligence, as shown in Figure 1. The cognition layer facilitates the construction of crowd knowledge for rumor identification through knowledge sharing and evidence integration among users. It serves as the foundational

support and aggregates multidimensional evidence. The interaction layer analyzes users' social relationships and interaction behaviors within social networks to capture rumor signals. The behavior layer models the evolution of information dissemination and collective behavior. Finally, the feedback mechanism based on collective intelligence optimizes the dissemination path and reduces the spread of rumors.

2.1 Cognition Layer

The cognition layer constructs the crowd knowledge for rumor identification through two complementary analytical pathways: data-driven and knowledge-driven analysis.

Data-driven analysis extracts features from diverse social network data (Aich et al., 2022; Horne and Adali, 2017). Semantically, rumor texts exhibit distinctive linguistic patterns: lexically avoiding deep information expression, syntactically trending toward simplification, and stylistically employing exaggerated headlines and emotionally stimulating content to enhance propagation by triggering negative public emotions(Vosoughi et al., 2018). As media formats diversify, detection technologies have expanded to visual content analysis(Vaccari and Chadwick, 2020), evolving from early pixel-level analysis to deep learning methods that effectively address challenges posed by deepfake technologies (Hao et al., 2021; Khan et al., 2022). Furthermore, cross-modal feature fusion and consistency verification enhance complex rumor content identification by analyzing conflicts between text-image emotions and audio-visual inconsistencies (Agarwal et al., 2020; Chugh et al., 2020).

However, feature analysis alone struggles with semantically complex or factually questionable rumors, prompting the development of **knowledge-driven analysis**. Knowledge graphs provide background verification and logical reasoning capabilities through structured entities and relationships, mapping textual entity relationships, verifying content accuracy, and identifying potential contradictions (Hu et al., 2021). For semantically ambiguous or information-deficient cases, knowledge graphs can perform semantic completion to fill critical elements in vague statements (Sun et al., 2022; Zhang et al., 2019). Complementarily, evidence-based verification directly connects to authoritative information sources to fact-check rumor content (Wouters and Opgenhaffen, 2024). The technical approach has evolved from traditional expert manual verification to modern automated fact-checking that extracts evidence from authoritative data sources through multi-source data retrieval, semantic alignment, and logical reasoning to evaluate support for or refutation of rumor claims (Das et al., 2023; Guo et al., 2022).

2.2 Interaction Layer

The interaction layer reveals rumor's social dynamics by analyzing user features and social contexts in social networks. From the perspective of user feature analysis, user interactions in social media can help identify abnormal users who spread rumors; at the same time, social context analysis can identify abnormal rumor propagation patterns by exploring the process of information dissemination in different network structures.

User groups in social networks, as core drivers of rumor propagation, exhibit diverse characteristics revealed through comprehensive **user features analysis**. Automated social bots manipulate public opinion through high-frequency content delivery and synchronized interactions, requiring detection that considers both non-human attribute features (standardized avatars, high-frequency posting) and network structural anomalies (high-density interconnections) (Guo et al., 2021; Haider et al., 2023). Real users display more complex behavioral patterns: malicious users deliberately spread rumors driven by interests, while ordinary users may unconsciously participate in dissemination due to cognitive limitations or emotional factors. As rumor propagation crosses platform boundaries, cross-platform user identity correlation analysis has become a research focus (Nie et al., 2016), identifying disguised behaviors through multi-dimensional feature matching. Combining deep learning and network analysis techniques (Hamdi et al., 2020; Zhang et al., 2015; Zhou et al., 2015), researchers have increasingly integrated analysis of static and dynamic identity features into broader network environments, enhancing user modeling adaptability in cross-platform scenarios.

The **social context analysis** enables a comprehensive understanding of how rumors form, spread, and evolve in social networks. The user identities are embedded within social contexts, where network structures and interaction patterns jointly shape rumor propagation dynamics. Network structures directly influence propagation efficiency (Vosoughi et al., 2018): sparse networks, lacking effective supervision mechanisms, easily form

flat diffusion structures, accelerating rumor spread, while dense networks build information filtering barriers through strong connection characteristics. Network user role heterogeneity further increases propagation complexity (Raponi et al., 2022), and the multi-level propagation patterns are captured by modeling recursive tree structures and propagation graphs (Bian et al., 2020; Min et al., 2022). Based on network structure, user interactions form social response mechanisms: different user types form homogeneous clusters, emotional factors catalyze group polarization, and key opinion leaders amplify effects in bridging communities (Wei and Meng, 2021).

2.3 Behavior Layer

The behavior layer focuses on dynamic analysis from rumor propagation modeling to group behaviors monitoring, to predict and intervene in rumor propagation (Xuan et al., 2019; Alkhodair et al., 2020).

Propagation pattern analysis views rumor propagation as a complex social dynamic system, constructing multi-level diffusion models. Early research borrowed from epidemiological models (Kermack and McKendrick, 1927; Dong and Huang, 2018; Zhao et al., 2013; Wan et al., 2017) to describe node state transitions, and information diffusion threshold models characterize propagation mechanisms from audience decision perspectives (Yan et al., 2019). As complex network theory deepens, research has expanded to multidimensional influencing factors: the temporal dimension of propagation evolution, user characteristic moderation effects, network structure diffusion constraints, and content attribute acceptance impact (Xiao et al., 2019; Hosni et al., 2020). Building on this foundation, rumor source detection techniques based on propagation models have evolved from global traversal based on centrality theory to snapshot observation and real-time monitoring methods, to network decoupling strategies addressing multi-source concurrent propagation, further enhancing adaptability to dynamic propagation environments(Zhu et al., 2022a; Qiu et al., 2022). End-to-end frameworks leveraging graph neural networks integrate propagation paths, temporal dynamics, and node features, significantly enhancing rumor source detection's robustness and accuracy (Wang et al., 2022; Cheng et al., 2024), providing effective support for addressing complex propagation environments and data noise challenges. These technologies collectively form a complete analysis chain from mechanism understanding to source tracing.

Closely associated with propagation patterns is **behavioral pattern analysis**, focusing on how network structures and group behaviors influence rumor diffusion. Nodes in social networks cluster into tight communities where rumors flow efficiently, while cross-community diffusion depends on bridging nodes or weak ties. When bridging points are scarce, propagation is limited to local communities, but rapidly spreads once reaching critical density (Zhang et al., 2018; Yang et al., 2016). Based on understanding community structures, network immunization strategies form two complementary approaches: preventive immunization monitors high-risk nodes and pre-transmits truth through advance analysis of network topology and community characteristics (Petrescu et al., 2021); adversarial immunization implements real-time intervention for known propagation sources, selecting key nodes for isolation or filtering to block transmission chains at minimal cost (Tariq et al., 2017).

3 Collective Intelligence-based Rumor Detection in the LLM Era

In this chapter, we explore the synergistic relationship between LLM and collective intelligence in rumor detection, that is, how LLM enhances the CIB framework to play various important roles in rumor detection, while the collective intelligence mechanism drives LLM agent-based modeling to enhance the simulation capabilities of rumor detection. We present an agent-based modeling approach within the CIB framework and establish a macro-micro feedback loop that can link cognitive processes and behavioral outcomes in rumor detection.

3.1 LLM-enhanced CIB Framework

LLMs play multifaceted roles in rumor detection, transforming rumor detection from static pattern recognition to dynamic reasoning, as shown in Figure 2. Next, we systematically analyze the multifaceted roles of LLM in CIB to reveal its enhancement effect on rumor detection.

At the cognition layer, LLMs enhance rumor detection capabilities through two primary technical forms: deep knowledge representation and dynamic fact verification, enabling models to more effectively address challenges in complex rumor

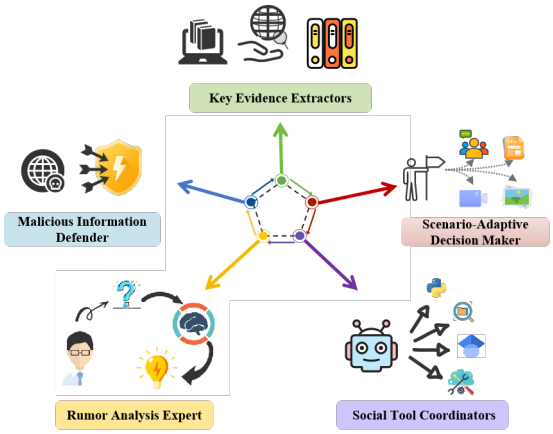

Figure 2: Multiple roles of LLM in rumor detection

scenarios. First, through large-scale pretraining, LLMs function as efficient **Key Evidence Extractors**. While traditional rumor detection systems rely on static structured methods like knowledge graphs to expand knowledge, LLMs' implicit encoding capabilities can capture deep semantic associations in unstructured information, providing an evidential foundation for rumor content verification and helping improve the generalization capabilities of smaller models (Nan et al., 2024; Yang et al., 2023). Second, LLMs serve as **Scenario-Adaptive Decision Makers**, with zero-shot reasoning capabilities allowing them to efficiently handle diverse rumor scenarios without fine-tuning (Li et al., 2023c; Wu et al., 2023). Combined with Retrieval-Augmented Generation(RAG) and external knowledge bases (Peng et al., 2023; Niu et al., 2024), LLMs can dynamically integrate the latest knowledge, addressing limitations of traditional methods in knowledge breadth and real-time capability, effectively reducing the probability of "hallucination phenomena" and enhancing detection credibility (Ji et al., 2023; Rawte et al., 2023).

At the interaction layer, LLMs primarily improve the simulation capability of information flow in social networks and the identification efficiency of interaction signals, providing social context information for behavioral prediction. LLMs function as **Social Tool Coordinators** by coordinating external tools (such as search engines, deepfake detectors) through Agent frameworks, further extending rumor detection capabilities (Chern et al., 2023; Wan et al., 2024; Li et al., 2024b). Unlike traditional static social network analysis and modeling, LLM Agents can perceive social environments, combine short-term memory (context learning) and long-term memory (external knowledge retrieval), plan and invoke tools, improving analytical performance. Generative Agents (Park et al., 2023) drive rumor detection to achieve technical upgrades from static network topology analysis to dynamic behavioral simulation through the simulation of interactive behaviors between users.

At the behavior layer, LLMs significantly enhance rumor analysis capabilities and the precision of intervention strategies, providing solid support for increasingly complex information environments. First, as **Rumor Analysis Experts**, LLMs excel in advanced reasoning and cross-domain background knowledge tasks. While traditional rumor detection centers on classification, relying on carefully annotated large datasets, LLMs leverage their emergent capabilities, with chain-of-thought(COT) reasoning decomposing complex problems into a series of intermediate reasoning steps, substantially enhancing logical transparency and explainability (Zhang and Gao, 2023). LLMs possess cross-scenario transfer capabilities(Cao et al., 2023b,a) and can conduct unified reasoning by combining text, image, audio, and other multimodal data (Yao et al., 2023), overcoming the limitations of traditional methods in multimodal fusion processing, enabling detection mechanisms to leap from pattern classification to causal inference (Zhu et al., 2022b; Nan et al., 2021). Furthermore, LLMs can serve as **Malicious Information Defenders**, demonstrating robust performance in adversarial social network environments. By combining adversarial training and red-teaming methods (Bhardwaj and Poria, 2023; OpenAI, 2023), LLMs can rapidly adapt to continuously evolving new forgery techniques, addressing the lag in model iteration and processing capacity in traditional methods (Wu et al., 2024b; Sun et al., 2024). For instance, this dynamic adaptability further enhances the robustness of rumor detection when dealing with complex tasks such as rumor diffusion, stylized language attacks, and deepfake information.

3.2 Collective Intelligence-driven CIB Framework

In the CIB framework, complex systems theory provides a foundation for understanding rumor propagation in social networks. Its core characteristic is the emergence of "collective intelligence" from non-linear interactions between components—system behaviors that cannot be predicted through simple aggregation of constituent elements. Agent-based Modeling (ABM)

implements this approach by focusing on autonomous decision-making entities, connecting micro-individual behaviors with macro-system outcomes through dynamic agent-environment interactions, enabling analysis of complex systems across multiple scales within human societies.

Research demonstrates that multi-agent systems effectively reproduce collective behavior patterns from classical sociological and economic theories while spontaneously developing error-correction mechanisms in collaborative tasks (Li et al., 2023a). EconAgent (Li et al., 2024a) simulates macroeconomic mechanisms, reproducing inflation and labor market unemployment fluctuations. AgentSociety (Piao et al., 2025) constructs realistic social environment simulations for modeling opinion propagation, cognitive polarization, and public policy responses. RLLNC (Ma et al., 2024) applies agents to urban governance challenges, including traffic control, pandemic intervention, and power system scheduling with notable success. For social network information governance issues like rumor detection, agent-based modeling shows distinctive advantages—generating social interaction simulations highly consistent with actual community behavior patterns (Park et al., 2022) and modeling trust relationship formation (Xie et al., 2024) and information propagation dynamics (Törnberg et al., 2023). Some researchers (Zhang et al., 2024a; Hu et al., 2025) use LLM-based multi-agent collaborative frameworks to explore rumor propagation mechanisms, reflecting information diffusion trends and optimizing intervention strategies. At the same time, FactAgent (Li et al., 2024b) enables real-time information credibility assessment by analyzing shallow linguistic features such as expression style and consistency with common-sense rules.

3.3 ABM in the CIB Framework

From the perspective of future development, we propose a roadmap for agent-based modeling in the CIB framework, as illustrated in Figure 3. It utilizes cross-layer dynamic feedback to establish a bidirectional **Macro-Micro Feedback Loop**, fostering an evolution driven by collective intelligence.

At the cognition layer, agents can utilize LLMs and multimodal analysis tools to achieve a deep semantic understanding of texts, images, videos, and other content associated with rumors, also performing real-time monitoring and dynamic analysis of content flow on social media platforms. By incorporating psychological models, agents can dynamically assess the potential intent behind information and quantify user cognitive biases. By analyzing user historical behavior, a dynamic user cognitive profile can be constructed to predict the susceptibility of different user groups to specific rumors, including multi-dimensional features such as their knowledge level, thinking ability and professional background in specific fields. Beyond surface feature detection, the cognition layer provides a solid foundation for subsequent interactive simulation and behavioral intervention.

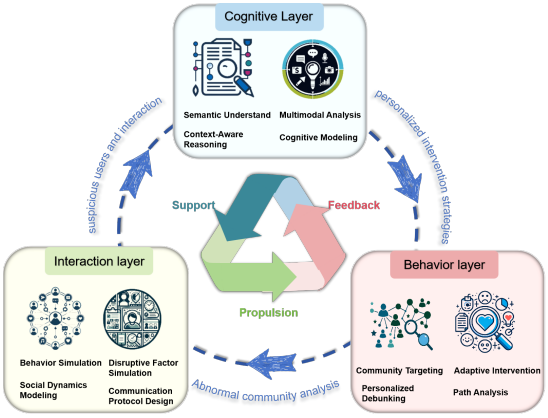

Figure 3: The CIB framework establishes a Macro-Micro Feedback Loop that integrates cross-layer dynamic feedback to bridge macro-level information dissemination with micro-level individual cognition.

At the interaction layer, agents can simulate the diversity of user behaviors in social networks (such as sharing, commenting, and reporting) and external factors such as social bots, constructing dynamic environments that reflect real-world propagation patterns. The interaction layer can also simulate complex information dissemination environmental factors, including the recommendation mechanism of the platform algorithm, the popularity ranking rules, and the impact of the review policy on information visibility. The dynamic network structure formed by the continuous interaction of multiple agents can not only reflect the information diffusion path in real social media, but also capture the network topology structure adjustment caused by various emergencies or changes in topic popularity, thus providing a reliable experimental environment for understanding the complexity of rumor propagation in the real world.

At the behavior layer, agents can capture the nonlinear propagation paths of collective behavior and identify the unique propagation characteristics

of different types of rumors, such as the explosive spread of panic-type rumors, the progressive spread of conspiracy theories, and the targeted and precise delivery of commercially induced rumors. Adaptively identify user groups with similar propagation behavior patterns in the network, and monitor the splitting, merging and evolution of the community structure in real time.

In addition to rumor detection, the framework implements reverse belief intervention. The behavior layer detects abnormal communities as the initial goal. Subsequently, the interaction layer analyzes suspicious users and their interaction structures in the community. On this basis, the cognition layer conducts collective knowledge analysis and social context reasoning on suspicious conversation clues to identify key evidence. Finally, the behavior layer intervenes on time to generate personalized rumor-refuting content that meets the cognitive characteristics of the target audience. Precision delivery is implemented through the optimal intervention nodes pre-calculated by the interaction layer, and finally, the user's belief state is updated at the cognition layer. This feedback mechanism enables the rumor detection and intervention model to continuously improve and optimize itself, enhancing its adaptability to the dynamic evolution characteristics of rumor propagation.

4 Future Research

We further explore two important frontiers in this chapter: advancing agent-based modeling to capture complex rumor dynamics, and addressing emerging challenges unique to the LLM era, combining the CIB framework to give the future direction of rumor detection, and we also provide more interdisciplinary perspectives in Appendix B.

4.1 ABM in Rumor Detection

Looking ahead, agent-based modeling for rumor detection presents several promising research directions that address current limitations and embrace the multi-dimensional complexity of this domain.

Deepening the Understanding of Rumor Propagation Mechanisms. Future research should move beyond shallow feature classification to integrate complex cognitive patterns and social dynamics into ABM. By incorporating concepts from cognitive consistency theory, we can model how individuals preferentially accept information that aligns with existing beliefs (Nickerson, 1998). Agents could simulate emotional drivers like fear and anger that amplify rumor spread, alongside social pressure dynamics, and capture hedonistic motivations and group identity factors in information sharing would better reflect real-world behavior patterns (Jiwa et al., 2023; Lewandowsky, 2022; Wanless and Berk, 2020). Additionally, embedding the decentralized nature of social networks into agent environments would enable more accurate simulation of information cascades and extreme attitude formation, similar to the promising work on echo chamber modeling (Wang et al., 2024a).

Dynamic Modeling of Rumor Diffusion and Intervention. ABM should evolve beyond static propagation roles to capture the fluid nature of individual behavior throughout the rumor lifecycle. Drawing inspiration from epidemiological modeling techniques, researchers could incorporate complex social variables such as education levels and forgetting mechanisms to precisely characterize dynamic spread processes. Intervention strategies require more sophisticated modeling that adapts to the progressive and complex changes in rumor propagation patterns. Agents with adaptive strategies for accuracy verification and influence blocking would significantly enhance intervention effectiveness across varied rumor scenarios.

Holistic Integration of Rumor Dynamics. A critical advancement would be developing frameworks that systematically integrate content, propagation, and interaction dimensions within logical contexts. Future ABM should comprehensively simulate the combined effects of multimodal information in rumor propagation and the pivotal role of social bots as propagation drivers. Incorporating user information-seeking behaviors and individual differences in response to controversial information would enrich model fidelity. Particularly promising is the modeling of relationships between active verification behaviors and subsequent actions like sharing, reporting, or ignoring information.

Simulating Human Cognitive and Decision-Making Processes. Although agents have memory and planning modules, the next step should authentically simulate human cognitive biases and decision-making mechanisms. Models that account for cognitive limitations—including memory capacity, knowledge levels, computational ability, and reasoning capacity—would more accurately reflect human information processing. Incorporating working memory constraints that influence information storage and processing efficiency represents

a significant advancement opportunity. Additionally, modeling instinctive psychological traits like loss aversion, which creates asymmetric perceptions of losses versus gains, would enhance behavioral realism in rumor response simulations. Personality trait modeling could further differentiate individual decision-making patterns in information evaluation and sharing contexts.

4.2 CIB in the Era of LLM

Enhancing Credibility in the Cognition Layer Future research should address the convergence of AI-generated content and the "dual-source risk" problem (Pan et al., 2023b; Chen and Shu, 2024; Shu et al., 2021), where techniques for detecting LLM-generated text become crucial for rumor identification. Current technical approaches are categorized into white-box and black-box detection methodologies. White-box detection, exemplified by watermarking technology (Liu and Bu, 2024; Liu et al., 2024; Zhao et al., 2023; Kuditipudi et al., 2024), ensures content traceability through distinctive identification markers embedded during generation. These can be implemented at various stages: incorporating trigger words and watermark labels in training samples, adjusting word distribution during logits generation, or employing predetermined random seeds for word-based sampling. Black-box detection encompasses zero-shot methods (Mitchell et al., 2023; Ippolito et al., 2020), classifier and neural network-based approaches (Mireshghallah et al., 2024; Mitrović et al., 2023), and online detection tools (AIT). For sensitive applications like rumor detection, maintaining low false alarm rates while ensuring generalization across base models remains paramount.

The convergence of unintentional hallucinations and deliberately fabricated rumors significantly complicates content reliability assessment. Hallucinations stem from inadequate model knowledge or insufficient reasoning capabilities, which are attributed to data quality issues, training methodology deficiencies, and reasoning process errors, ultimately leading to factually inconsistent content generation (Zhou et al., 2023a; Chuang et al., 2024; Ji et al., 2023). Mitigation strategies include corpus quality enhancement (Yu et al., 2022; Yang et al., 2024), targeted parameter modification through model editing (Wang et al., 2024b; Zhang et al., 2024b), real-time external knowledge integration via retrieval-augmented generation (RAG) (Feng et al., 2024; Jiang et al., 2023), and multi-step reasoning verification mechanisms (Dhuliawala et al., 2024; Pan et al., 2024). Furthermore, more challenging is detecting intentionally fabricated misinformation. Advanced LLM generation capabilities have enabled maliciously crafted rumors to achieve unprecedented sophistication in logical coherence, linguistic expression, and persuasive power (Pan et al., 2023a; Spitale et al., 2023). Traditional rumor detection methodologies often fail against such 'logically perfect rumors.' Consequently, developing innovative detection systems adapted to the LLM era becomes imperative, requiring continuous adaptation to maintain pace with evolving generative models (Ayoobi et al., 2023; Zhou et al., 2023b). Furthermore, detection must be attributable and explainable (Huang and Sun, 2024). Beyond the technical challenge of 'deepfakes,' the LLM era introduces the social risk of reverse stigmatization, where authentic information is deliberately mischaracterized as AI-synthesized and subsequently discredited (SCHIFF et al., 2025). This complex interplay necessitates both technical forensics and social verification, demonstrating that rumor detection transcends technical challenges to become a systemic issue affecting social cognition and information ecology.

Optimizing Cognitive Alignment in the Interaction Layer LLM alignment enables models to follow human instructions across real-world scenarios while generating high-quality text that adheres to societal values and safety constraints. Despite continuous technological advancement, the semantic gap between LLMs and human cognitive systems has widened in social media environments (Kidd and Birhane, 2023). LLMs demonstrate superior deductive reasoning compared to inductive reasoning, significantly affecting their reliability in fact-based reasoning tasks. This disparity manifests in biased interpretations of cultural contexts (Fedorenko et al., 2024), limited capacity for analyzing complex causal relationships (Guo et al., 2020), and logical reasoning inconsistencies (Binz and Schulz, 2023). These limitations create significant concerns in user interactions, as the model's flattering behavior reinforces existing cognitive biases and potentially induces false memory formation [(Chan et al., 2024; Acerbi and Stubbersfield, 2023). Such vulnerabilities provide exploitable opportunities for information manipulation, exacerbating misinformation's persistent impact.

Establishing protective barriers against cognitive infiltration requires utilizing personalized and

persuasive debunking content generated by LLMs to implement belief interventions, elevating rumor intervention from informational to cognitive levels. The human-like cognitive abilities (Hagendorff et al., 2023) and theory of mind characteristics (Kosinski, 2023) demonstrated by LLMs introduce novel possibilities for rumor intervention. Through personalized dialogue and semantic guidance that attenuate user identification with false beliefs, LLMs can generate tailored persuasive content addressing specific psychological predispositions (Costello et al., 2024). Compared with conventional rumor interventions (Chan et al., 2017; Johansson et al., 2022), LLM-facilitated belief interventions demonstrate sustained efficacy and broader applicability (Matz et al., 2024). Future research should prioritize integrating LLMs' semantic generation capabilities with contextual cognitive advantages, establishing cognitive enhancement loops in critical processes such as information traceability and fact verification. Optimal rumor intervention strategies must balance enhanced user engagement and trust while preventing models from succumbing to overconfidence or emphasizing stylistic elements over substantive content (Lee et al., 2022), thereby establishing a comprehensive intervention framework emphasizing both proactive defense mechanisms and cognitive recalibration.

Improving Technical Adaptability in the Behavioral Layer Current LLM security vulnerabilities manifest through prompt injection attacks (Liu et al., 2023b,a) that employ sophisticated context manipulation techniques to bypass security boundaries via command concatenation, role-playing, and context confusion, and model generation attacks (Chao et al., 2025; Shah et al., 2023; Yao et al., 2024; Deng et al., 2023) that leverage auxiliary LLMs to automatically generate deceptive prompts through iterative optimization, modular generation, fuzzy testing, and defense analysis. These approaches significantly enhance attack scale and adaptability. However, LLM robustness and dynamic adaptability in complex scenarios remain insufficient, with defense measures consistently lagging behind rapidly evolving attack technologies (Wolf et al., 2023). Future rumor detection research should integrate proactive defense strategies, combining red team testing (Perez et al., 2022; Ganguli et al., 2022) with fact verification mechanisms (Lee et al., 2022) to dynamically improve LLM attack-defense adaptation capabilities. By simulating adversarial scenarios to continuously optimize security boundaries and introducing external knowledge verification to enhance factual compliance, LLMs can achieve self-updating and dynamic correction capabilities in information confrontation environments.

Existing benchmarks primarily assess model performance on static data, inadequately addressing the sophisticated deception inherent in LLM-generated false information characterized by enhanced language complexity, logical coherence, and refinement (Bang et al., 2023). Current LLM hallucination benchmarks (Fu et al., 2023; Li et al., 2023b) fail to capture the unique psychological and behavioral characteristics essential to rumor detection and refutation processes (Lewandowsky et al., 2012). Therefore, rumor detection evaluation systems should incorporate dynamic indicators addressing complex contexts and user behavioral characteristics, while assessing models' early recognition capabilities for emerging rumors, cross-domain transfer learning effectiveness, and adversarial robustness. Additionally, specialized security evaluation systems must address adversarial challenges in rumor detection by developing assessment methods targeting rumor propagation characteristics (Liu et al., 2025), focusing on multimodal rumor generation technologies, domain-specific attack strategies, and social psychology-oriented attack mechanisms to provide reliable safeguards for model applications in complex social environments .

5 Conclusion

Based on the complex system characteristics of collective intelligence, we have reconstructed a rumor detection paradigm—the Cognition-Interaction-Behavior (CIB) framework—adapted to the era of LLMs. We thoroughly explored the multidimensional roles of LLMs in enhancing rumor detection capabilities and their synergistic relationship with collective intelligence. Innovatively, the CIB framework enables dynamic bidirectional rumor detection and intervention, providing a roadmap for applying agent-based modeling (ABM) in rumor detection. We analyzed the emerging challenges in the LLM era and proposed feasible future research directions, providing theoretical foundations and developmental pathways for rumor detection.

6 Limitations

In the future research, we propose a roadmap for rumor detection under the CIB framework, providing a comprehensive analysis of potential research challenges and corresponding directions. However, further exploration is needed to evaluate in large-scale social media environments. Additionally, polarized contexts or anomalous interactions may introduce more significant complexities. To refine and optimize the framework, we will consider enhancing robustness and dynamic adaptability in complex scenarios.

References

Alberto Acerbi and Joseph M Stubbersfield. 2023. Large language models show human-like content biases in transmission chain experiments. *Proceedings of the National Academy of Sciences*, 120(44):e2313790120.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

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

Shruti Agarwal, Hany Farid, Ohad Fried, and Maneesh Agrawala. 2020. Detecting deep-fake videos from phoneme-viseme mismatches. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 660–661.

Ankit Aich, Souvik Bhattacharya, and Natalie Parde. 2022. Demystifying neural fake news via linguistic feature-based interpretation. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6586–6599.

Sarah A Alkhodair, Steven HH Ding, Benjamin CM Fung, and Junqiang Liu. 2020. Detecting breaking news rumors of emerging topics in social media. *Information Processing & Management*, 57(2):102018.

Hunt Allcott and Matthew Gentzkow. 2017. Social media and fake news in the 2016 election. *Journal of economic perspectives*, 31(2):211–236.

Gordon W Allport. 1947. The psychology of rumor. *Henry Holt*.

Isabelle Augenstein, Timothy Baldwin, Meeyoung Cha, Tanmoy Chakraborty, Giovanni Luca Ciampaglia, David Corney, Renee DiResta, Emilio Ferrara, Scott Hale, Alon Halevy, et al. 2024. Factuality challenges in the era of large language models and opportunities for fact-checking. *Nature Machine Intelligence*, 6(8):852–863.

Navid Ayoobi, Sadat Shahriar, and Arjun Mukherjee. 2023. The looming threat of fake and llm-generated linkedin profiles: Challenges and opportunities for detection and prevention. In *Proceedings of the 34th ACM Conference on Hypertext and Social Media*, HT '23, New York, NY, USA. Association for Computing Machinery.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *Preprint*, arXiv:2302.04023.

Rishabh Bhardwaj and Soujanya Poria. 2023. Red-teaming large language models using chain of utterances for safety-alignment. *arXiv preprint arXiv:2308.09662*.

Tian Bian, Xi Xiao, Tingyang Xu, Peilin Zhao, Wenbing Huang, Yu Rong, and Junzhou Huang. 2020. Rumor detection on social media with bi-directional graph convolutional networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 549–556.

Marcel Binz and Eric Schulz. 2023. Using cognitive psychology to understand gpt-3. *Proceedings of the National Academy of Sciences*, 120(6):e2218523120.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Rui Cao, Ming Shan Hee, Adriel Kuek, Wen-Haw Chong, Roy Ka-Wei Lee, and Jing Jiang. 2023a. Procap: Leveraging a frozen vision-language model for hateful meme detection. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 5244–5252.

Rui Cao, Roy Ka-Wei Lee, Wen-Haw Chong, and Jing Jiang. 2023b. Prompting for multimodal hateful meme classification. *arXiv preprint arXiv:2302.04156*.

Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. 2011. Information credibility on twitter. In *Proceedings of the 20th international conference on World wide web*, pages 675–684.

Man-pui Sally Chan, Christopher R Jones, Kathleen Hall Jamieson, and Dolores Albarracín. 2017. Debunking: A meta-analysis of the psychological efficacy of messages countering misinformation. *Psychological science*, 28(11):1531–1546.

Samantha Chan, Pat Pataranutaporn, Aditya Suri, Wazeer Zulfikar, Pattie Maes, and Elizabeth F Loftus. 2024. Conversational ai powered by large language models amplifies false memories in witness interviews. *arXiv preprint arXiv:2408.04681*.

Patrick Chao, Alexander Robey, Edgar Dobriban, Hamed Hassani, George J Pappas, and Eric Wong. 2025. Jailbreaking black box large language models in twenty queries. In *2025 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pages 23–42. IEEE.

Canyu Chen and Kai Shu. 2024. Combating misinformation in the age of llms: Opportunities and challenges. *AI Magazine*, 45(3):354–368.

Le Cheng, Peican Zhu, Keke Tang, Chao Gao, and Zhen Wang. 2024. Gin-sd: source detection in graphs with incomplete nodes via positional encoding and attentive fusion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 55–63.

I Chern, Steffi Chern, Shiqi Chen, Weizhe Yuan, Kehua Feng, Chunting Zhou, Junxian He, Graham Neubig, Pengfei Liu, et al. 2023. Factool: Factuality detection in generative ai–a tool augmented framework for multi-task and multi-domain scenarios. *arXiv preprint arXiv:2307.13528*.

Lynn Chua, Badih Ghazi, Yangsibo Huang, Pritish Kamath, Ravi Kumar, Daogao Liu, Pasin Manurangsi, Amer Sinha, and Chiyuan Zhang. 2024. Mind the privacy unit! user-level differential privacy for language model fine-tuning. *arXiv preprint arXiv:2406.14322*.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James Glass, and Pengcheng He. 2024. Dola: Decoding by contrasting layers improves factuality in large language models. *Preprint*, arXiv:2309.03883.

Komal Chugh, Parul Gupta, Abhinav Dhall, and Ramanathan Subramanian. 2020. Not made for each other-audio-visual dissonance-based deepfake detection and localization. In *Proceedings of the 28th ACM international conference on multimedia*, pages 439–447.

Daniel M Cornforth, David JT Sumpter, Sam P Brown, and Åke Brännström. 2012. Synergy and group size in microbial cooperation. *The American Naturalist*, 180(3):296–305.

Thomas H. Costello, Gordon Pennycook, and David G. Rand. 2024. Durably reducing conspiracy beliefs through dialogues with ai. *Science*, 385(6714):eadq1814.

Anubrata Das, Houjiang Liu, Venelin Kovatchev, and Matthew Lease. 2023. The state of human-centered nlp technology for fact-checking. *Information processing & management*, 60(2):103219.

Michela Del Vicario, Alessandro Bessi, Fabiana Zollo, Fabio Petroni, Antonio Scala, Guido Caldarelli, H Eugene Stanley, and Walter Quattrociocchi. 2016. The spreading of misinformation online. *Proceedings of the national academy of Sciences*, 113(3):554–559.

Gelei Deng, Yi Liu, Yuekang Li, Kailong Wang, Ying Zhang, Zefeng Li, Haoyu Wang, Tianwei Zhang, and Yang Liu. 2023. Masterkey: Automated jailbreak across multiple large language model chatbots. *arXiv preprint arXiv:2307.08715*.

Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. 2024. Chain-of-verification reduces hallucination in large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 3563–3578, Bangkok, Thailand. Association for Computational Linguistics.

Suyalatu Dong and Yong-Chang Huang. 2018. Sis rumor spreading model with population dynamics in online social networks. In *2018 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET)*, pages 1–5. IEEE.

Yogesh K Dwivedi, Nir Kshetri, Laurie Hughes, Emma Louise Slade, Anand Jeyaraj, Arpan Kumar Kar, Abdullah M Baabdullah, Alex Koohang, Vishnupriya Raghavan, Manju Ahuja, et al. 2023. Opinion paper:"so what if chatgpt wrote it?" multidisciplinary perspectives on opportunities, challenges and implications of generative conversational ai for research, practice and policy. *International Journal of Information Management*, 71:102642.

Alexandros Efstratiou and Emiliano De Cristofaro. 2022. Adherence to misinformation on social media through socio-cognitive and group-based processes. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2):1–35.

Yahya H Ezzeldin, Shen Yan, Chaoyang He, Emilio Ferrara, and A Salman Avestimehr. 2023. Fairfed: Enabling group fairness in federated learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pages 7494–7502.

Evelina Fedorenko, Steven T Piantadosi, and Edward AF Gibson. 2024. Language is primarily a tool for communication rather than thought. *Nature*, 630(8017):575–586.

Zhangyin Feng, Xiaocheng Feng, Dezhi Zhao, Maojin Yang, and Bing Qin. 2024. Retrieval-generation synergy augmented large language models. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 11661–11665. IEEE.

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2023. Gptscore: Evaluate as you desire. *arXiv preprint arXiv:2302.04166*.

Deep Ganguli, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Ben Mann, Ethan Perez, Nicholas Schiefer, Kamal Ndousse, et al. 2022. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. *arXiv preprint arXiv:2209.07858*.

Andrew M Guess, Michael Lerner, Benjamin Lyons, Jacob M Montgomery, Brendan Nyhan, Jason Reifler, and Neelanjan Sircar. 2020. A digital media literacy intervention increases discernment between mainstream and false news in the united states and india. *Proceedings of the National Academy of Sciences*, 117(27):15536–15545.

Bin Guo, Yasan Ding, Lina Yao, Yunji Liang, and Zhiwen Yu. 2020. The future of false information detection on social media: New perspectives and trends. *ACM Computing Surveys (CSUR)*, 53(4):1–36.

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

Qinglang Guo, Haiyong Xie, Yangyang Li, Wen Ma, and Chao Zhang. 2021. Social bots detection via fusing bert and graph convolutional networks. *Symmetry*, 14(1):30.

Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. 2022. A survey on automated fact-checking. *Transactions of the Association for Computational Linguistics*, 10:178–206.

Thilo Hagendorff, Sarah Fabi, and Michal Kosinski. 2023. Human-like intuitive behavior and reasoning biases emerged in large language models but disappeared in chatgpt. *Nature Computational Science*, 3(10):833–838.

Samar Haider, Luca Luceri, Ashok Deb, Adam Badawy, Nanyun Peng, and Emilio Ferrara. 2023. Detecting social media manipulation in low-resource languages. In *Companion Proceedings of the ACM Web Conference 2023*, pages 1358–1364.

Tarek Hamdi, Hamda Slimi, Ibrahim Bounhas, and Yahya Slimani. 2020. A hybrid approach for fake news detection in twitter based on user features and graph embedding. In *Distributed Computing and Internet Technology: 16th International Conference, ICDCIT 2020, Bhubaneswar, India, January 9–12, 2020, Proceedings 16*, pages 266–280. Springer.

Jing Hao, Zhixin Zhang, Shicai Yang, Di Xie, and Shiliang Pu. 2021. Transforensics: image forgery localization with dense self-attention. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15055–15064.

Benjamin Horne and Sibel Adali. 2017. This just in: Fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news. In *Proceedings of the international AAAI conference on web and social media*, volume 11, pages 759–766.

Adil Imad Eddine Hosni, Kan Li, and Sadique Ahmad. 2020. Minimizing rumor influence in multiplex online social networks based on human individual and social behaviors. *Information Sciences*, 512:1458–1480.

Beizhe Hu, Qiang Sheng, Juan Cao, Yuhui Shi, Yang Li, Danding Wang, and Peng Qi. 2024. Bad actor, good advisor: Exploring the role of large language models in fake news detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 22105–22113.

Linmei Hu, Tianchi Yang, Luhao Zhang, Wanjun Zhong, Duyu Tang, Chuan Shi, Nan Duan, and Ming Zhou. 2021. Compare to the knowledge: Graph neural fake news detection with external knowledge. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 754–763.

Tianrui Hu, Dimitrios Liakopoulos, Xiwen Wei, Radu Marculescu, and Neeraja J Yadwadkar. 2025. Simulating rumor spreading in social networks using llm agents. *arXiv preprint arXiv:2502.01450*.

Linan Huang and Quanyan Zhu. 2023. An introduction of system-scientific approaches to cognitive security. *arXiv preprint arXiv:2301.05920*.

Yue Huang and Lichao Sun. 2024. Fakegpt: fake news generation, explanation and detection of large language models. *arxiv. org*.

Hongwen Hui, Chengcheng Zhou, Xing Lü, and Jiarong Li. 2020. Spread mechanism and control strategy of social network rumors under the influence of covid-19. *Nonlinear Dynamics*, 101:1933–1949.

Daphne Ippolito, Daniel Duckworth, Chris Callison-Burch, and Douglas Eck. 2020. Automatic detection of generated text is easiest when humans are fooled. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1808–1822, Online. Association for Computational Linguistics.

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

Zhengbao Jiang, Frank F Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. Active retrieval augmented generation. *arXiv preprint arXiv:2305.06983*.

Matthew Jiwa, Patrick S Cooper, Trevor TJ Chong, and Stefan Bode. 2023. Hedonism as a motive for information search: biased information-seeking leads to biased beliefs. *Scientific Reports*, 13(1):2086.

Pica Johansson, Florence Enock, Scott Hale, Bertie Vidgen, Cassidy Bereskin, Helen Margetts, and Jonathan Bright. 2022. How can we combat online misinformation? a systematic overview of current interventions and their efficacy. *Preprint*, arXiv:2212.11864.

Neil F Johnson, Nicolas Velásquez, Nicholas Johnson Restrepo, Rhys Leahy, Nicholas Gabriel, Sara El Oud, Minzhang Zheng, Pedro Manrique, Stefan Wuchty, and Yonatan Lupu. 2020. The online competition between pro-and anti-vaccination views. *Nature*, 582(7811):230–233.

S Mo Jones-Jang, Tara Mortensen, and Jingjing Liu. 2021. Does media literacy help identification of fake news? information literacy helps, but other literacies don't. *American behavioral scientist*, 65(2):371–388.

William Ogilvy Kermack and Anderson G McKendrick. 1927. A contribution to the mathematical theory of epidemics. *Proceedings of the royal society of london. Series A, Containing papers of a mathematical and physical character*, 115(772):700–721.

Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah. 2022. Transformers in vision: A survey. *ACM computing surveys (CSUR)*, 54(10s):1–41.

Celeste Kidd and Abeba Birhane. 2023. How ai can distort human beliefs. *Science*, 380(6651):1222–1223.

Antino Kim and Alan R Dennis. 2019. Says who? the effects of presentation format and source rating on fake news in social media. *Mis quarterly*, 43(3):1025–1039.

Michal Kosinski. 2023. Theory of mind may have spontaneously emerged in large language models. *arXiv preprint arXiv:2302.02083*, 4:169.

Sarah Kreps, R Miles McCain, and Miles Brundage. 2022. All the news that's fit to fabricate: Ai-generated text as a tool of media misinformation. *Journal of experimental political science*, 9(1):104–117.

Weirui Kuang, Bingchen Qian, Zitao Li, Daoyuan Chen, Dawei Gao, Xuchen Pan, Yuexiang Xie, Yaliang Li, Bolin Ding, and Jingren Zhou. 2024. Federatedscope-llm: A comprehensive package for fine-tuning large language models in federated learning. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 5260–5271.

Rohith Kuditipudi, John Thickstun, Tatsunori Hashimoto, and Percy Liang. 2024. Robust distortion-free watermarks for language models. *Preprint*, arXiv:2307.15593.

KP Krishna Kumar and G Geethakumari. 2014. Detecting misinformation in online social networks using cognitive psychology. *Human-centric Computing and Information Sciences*, 4(1):14.

Nayeon Lee, Wei Ping, Peng Xu, Mostofa Patwary, Pascale N Fung, Mohammad Shoeybi, and Bryan Catanzaro. 2022. Factuality enhanced language models for open-ended text generation. *Advances in Neural Information Processing Systems*, 35:34586–34599.

Stephan Lewandowsky. 2022. Fake news and participatory propaganda. In *Cognitive Illusions*, pages 324–340. Routledge.

Stephan Lewandowsky, Ullrich KH Ecker, Colleen M Seifert, Norbert Schwarz, and John Cook. 2012. Misinformation and its correction: Continued influence and successful debiasing. *Psychological science in the public interest*, 13(3):106–131.

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

Junyi Li, Xiaoxue Cheng, Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023b. HaluEval: A large-scale hallucination evaluation benchmark for large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6449–6464, Singapore. Association for Computational Linguistics.

Miaoran Li, Baolin Peng, Michel Galley, Jianfeng Gao, and Zhu Zhang. 2023c. Self-checker: Plug-and-play modules for fact-checking with large language models. *arXiv preprint arXiv:2305.14623*.

Nian Li, Chen Gao, Mingyu Li, Yong Li, and Qingmin Liao. 2024a. EconAgent: Large language model-empowered agents for simulating macroeconomic activities. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15523–15536, Bangkok, Thailand. Association for Computational Linguistics.

Xinyi Li, Yongfeng Zhang, and Edward C Malthouse. 2024b. Large language model agent for fake news detection. *arXiv preprint arXiv:2405.01593*.

Gang Liang, Wenbo He, Chun Xu, Liangyin Chen, and Jinquan Zeng. 2015. Rumor identification in microblogging systems based on users' behavior. *IEEE Transactions on Computational Social Systems*, 2(3):99–108.

Aiwei Liu, Leyi Pan, Yijian Lu, Jingjing Li, Xuming Hu, Xi Zhang, Lijie Wen, Irwin King, Hui Xiong, and Philip Yu. 2024. A survey of text watermarking in the era of large language models. *ACM Comput. Surv.*, 57(2).

Songyang Liu, Chaozhuo Li, Jiameng Qiu, Xi Zhang, Feiran Huang, Litian Zhang, Yiming Hei, and Philip S Yu. 2025. The scales of justitia: A comprehensive survey on safety evaluation of llms. *arXiv preprint arXiv:2506.11094*.

Yepeng Liu and Yuheng Bu. 2024. Adaptive text watermark for large language models. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Yi Liu, Gelei Deng, Yuekang Li, Kailong Wang, Zihao Wang, Xiaofeng Wang, Tianwei Zhang, Yepang Liu, Haoyu Wang, Yan Zheng, et al. 2023a. Prompt injection attack against llm-integrated applications. *arXiv preprint arXiv:2306.05499*.

Yi Liu, Gelei Deng, Zhengzi Xu, Yuekang Li, Yaowen Zheng, Ying Zhang, Lida Zhao, Tianwei Zhang, Kailong Wang, and Yang Liu. 2023b. Jailbreaking chatgpt via prompt engineering: An empirical study. *arXiv preprint arXiv:2305.13860*.

Chengdong Ma, Aming Li, Yali Du, Hao Dong, and Yaodong Yang. 2024. Efficient and scalable reinforcement learning for large-scale network control. *Nature Machine Intelligence*, 6(9):1006–1020.

Jing Ma, Wei Gao, and Kam-Fai Wong. 2018. Rumor detection on twitter with tree-structured recursive neural networks. Association for Computational Linguistics.

Peihua Mai, Ran Yan, Zhe Huang, Youjia Yang, and Yan Pang. 2023. Split-and-denoise: Protect large language model inference with local differential privacy. *arXiv preprint arXiv:2310.09130*.

SC Matz, JD Teeny, Sumer S Vaid, H Peters, GM Harari, and M Cerf. 2024. The potential of generative ai for personalized persuasion at scale. *Scientific Reports*, 14(1):4692.

William J McGuire and Demetrios Papageorgis. 1961. The relative efficacy of various types of prior belief-defense in producing immunity against persuasion. *The Journal of Abnormal and Social Psychology*, 62(2):327.

William Menegas, Korleki Akiti, Ryunosuke Amo, Naoshige Uchida, and Mitsuko Watabe-Uchida. 2018. Dopamine neurons projecting to the posterior striatum reinforce avoidance of threatening stimuli. *Nature neuroscience*, 21(10):1421–1430.

Erxue Min, Yu Rong, Yatao Bian, Tingyang Xu, Peilin Zhao, Junzhou Huang, and Sophia Ananiadou. 2022. Divide-and-conquer: Post-user interaction network for fake news detection on social media. In *Proceedings of the ACM web conference 2022*, pages 1148–1158.

Niloofar Mireshghallah, Justus Mattern, Sicun Gao, Reza Shokri, and Taylor Berg-Kirkpatrick. 2024. Smaller language models are better zero-shot machine-generated text detectors. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 278–293, St. Julian's, Malta. Association for Computational Linguistics.

Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D. Manning, and Chelsea Finn. 2023. Detectgpt: zero-shot machine-generated text detection using probability curvature. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Sandra Mitrović, Davide Andreoletti, and Omran Ayoub. 2023. Chatgpt or human? detect and explain. explaining decisions of machine learning model for detecting short chatgpt-generated text. *Preprint*, arXiv:2301.13852.

Qiong Nan, Juan Cao, Yongchun Zhu, Yanyan Wang, and Jintao Li. 2021. Mdfend: Multi-domain fake news detection. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 3343–3347.

Qiong Nan, Qiang Sheng, Juan Cao, Beizhe Hu, Danding Wang, and Jintao Li. 2024. Let silence speak: Enhancing fake news detection with generated comments from large language models. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 1732–1742.

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

Yuanping Nie, Yan Jia, Shudong Li, Xiang Zhu, Aiping Li, and Bin Zhou. 2016. Identifying users across social networks based on dynamic core interests. *Neurocomputing*, 210:107–115.

Richard E Nisbett and Timothy D Wilson. 1977. The halo effect: Evidence for unconscious alteration of judgments. *Journal of personality and social psychology*, 35(4):250.

Cheng Niu, Yang Guan, Yuanhao Wu, Juno Zhu, Juntong Song, Randy Zhong, Kaihua Zhu, Siliang Xu, Shizhe Diao, and Tong Zhang. 2024. Veract scan: Retrieval-augmented fake news detection with justifiable reasoning. *arXiv preprint arXiv:2406.10289*.

R OpenAI. 2023. Gpt-4 technical report. arxiv 2303.08774. *View in Article*, 2(5).

Liangming Pan, Michael Saxon, Wenda Xu, Deepak Nathani, Xinyi Wang, and William Yang Wang. 2024. Automatically correcting large language models: Surveying the landscape of diverse automated correction strategies. *Transactions of the Association for Computational Linguistics*, 12:484–506.

Yikang Pan, Liangming Pan, Wenhu Chen, Preslav Nakov, Min-Yen Kan, and William Wang. 2023a. On the risk of misinformation pollution with large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1389–1403, Singapore. Association for Computational Linguistics.

Yikang Pan, Liangming Pan, Wenhu Chen, Preslav Nakov, Min-Yen Kan, and William Yang Wang. 2023b. On the risk of misinformation pollution with large language models. *arXiv preprint arXiv:2305.13661*.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra

of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pages 1–22.

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

Peter S Park, Simon Goldstein, Aidan O'Gara, Michael Chen, and Dan Hendrycks. 2024. Ai deception: A survey of examples, risks, and potential solutions. *Patterns*, 5(5).

E Parliament. 2023. Artificial intelligence act: deal on comprehensive rules for trustworthy ai. *Pressemitteilung vom*, 9.

Baolin Peng, Michel Galley, Pengcheng He, Hao Cheng, Yujia Xie, Yu Hu, Qiuyuan Huang, Lars Liden, Zhou Yu, Weizhu Chen, et al. 2023. Check your facts and try again: Improving large language models with external knowledge and automated feedback. *arXiv preprint arXiv:2302.12813*.

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

Matjaž Perc, Mahmut Ozer, and Janja Hojnik. 2019. Social and juristic challenges of artificial intelligence. *Palgrave Communications*, 5(1).

Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving. 2022. Red teaming language models with language models. *arXiv preprint arXiv:2202.03286*.

Heinrich Peters and Sandra Matz. 2024. Large language models can infer psychological dispositions of social media users. *PNAS Nexus*, page pgae231.

Alexandru Petrescu, Ciprian-Octavian Truică, Elena-Simona Apostol, and Panagiotis Karras. 2021. Sparse shield: Social network immunization vs. harmful speech. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 1426–1436.

Huyen Trang Phan, Ngoc Thanh Nguyen, and Dosam Hwang. 2023. Fake news detection: A survey of graph neural network methods. *Applied Soft Computing*, 139:110235.

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

Russell A Poldrack, Thomas Lu, and Gašper Beguš. 2023. Ai-assisted coding: Experiments with gpt-4. *arXiv preprint arXiv:2304.13187*.

Liqing Qiu, Shiqi Sai, and Moji Wei. 2022. Bpsl: a new rumor source location algorithm based on the timestamp back propagation in social networks. *Applied Intelligence*, pages 1–13.

Antonio Rangel, Colin Camerer, and P Read Montague. 2008. Neuroeconomics: The neurobiology of value-based decision-making. *Nature Reviews. Neuroscience*, 9(7):545.

Simone Raponi, Zeinab Khalifa, Gabriele Oligeri, and Roberto Di Pietro. 2022. Fake news propagation: A review of epidemic models, datasets, and insights. *ACM Transactions on the Web (TWEB)*, 16(3):1–34.

Vipula Rawte, Amit Sheth, and Amitava Das. 2023. A survey of hallucination in large foundation models. *arXiv preprint arXiv:2309.05922*.

Jon Roozenbeek, Claudia R Schneider, Sarah Dryhurst, John Kerr, Alexandra LJ Freeman, Gabriel Recchia, Anne Marthe Van Der Bles, and Sander Van Der Linden. 2020. Susceptibility to misinformation about covid-19 around the world. *Royal Society open science*, 7(10):201199.

Dietram A Scheufele and Nicole M Krause. 2019. Science audiences, misinformation, and fake news. *Proceedings of the National Academy of Sciences*, 116(16):7662–7669.

KAYLYN JACKSON SCHIFF, DANIEL S. SCHIFF, and NATÁLIA S. BUENO. 2025. The liar's dividend: Can politicians claim misinformation to evade accountability? *American Political Science Review*, 119(1):71–90.

Rusheb Shah, Soroush Pour, Arush Tagade, Stephen Casper, Javier Rando, et al. 2023. Scalable and transferable black-box jailbreaks for language models via persona modulation. *arXiv preprint arXiv:2311.03348*.

Cuihua Shen, Mona Kasra, and James O'Brien. 2021. This photograph has been altered: Testing the effectiveness of image forensic labeling on news image credibility. *arXiv preprint arXiv:2101.07951*.

Kai Shu, Limeng Cui, Suhang Wang, Dongwon Lee, and Huan Liu. 2019. defend: Explainable fake news detection. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 395–405.

Kai Shu, Yichuan Li, Kaize Ding, and Huan Liu. 2021. Fact-enhanced synthetic news generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13825–13833.

Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake news detection on social media: A data mining perspective. *ACM SIGKDD explorations newsletter*, 19(1):22–36.

Giovanni Spitale, Nikola Biller-Andorno, and Federico Germani. 2023. Ai model gpt-3 (dis)informs us better than humans. *Science Advances*, 9(26):eadh1850.

Mengzhu Sun, Xi Zhang, Jiaqi Zheng, and Guixiang Ma. 2022. Ddgcn: Dual dynamic graph convolutional networks for rumor detection on social media. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 4611–4619.

Yanshen Sun, Jianfeng He, Limeng Cui, Shuo Lei, and Chang-Tien Lu. 2024. Exploring the deceptive power of llm-generated fake news: A study of real-world detection challenges. *arXiv preprint arXiv:2403.18249*.

Tetsuro Takahashi and Nobuyuki Igata. 2012. Rumor detection on twitter. In *The 6th International Conference on Soft Computing and Intelligent Systems, and The 13th International Symposium on Advanced Intelligence Systems*, pages 452–457. IEEE.

Yiming Tan, Dehai Min, Yu Li, Wenbo Li, Nan Hu, Yongrui Chen, and Guilin Qi. 2023. Can chatgpt replace traditional kbqa models? an in-depth analysis of the question answering performance of the gpt llm family. In *International Semantic Web Conference*, pages 348–367. Springer.

Juvaria Tariq, Muhammad Ahmad, Imdadullah Khan, and Mudassir Shabbir. 2017. Scalable approximation algorithm for network immunization. *arXiv preprint arXiv:1711.00784*.

Kassym-Jomart Tokayev. 2023. Ethical implications of large language models a multidimensional exploration of societal, economic, and technical concerns. *International Journal of Social Analytics*, 8(9):17–33.

Sabrina M Tom, Craig R Fox, Christopher Trepel, and Russell A Poldrack. 2007. The neural basis of loss aversion in decision-making under risk. *Science*, 315(5811):515–518.

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

Cecilie S Traberg, Jon Roozenbeek, and Sander van der Linden. 2022. Psychological inoculation against misinformation: Current evidence and future directions. *The ANNALS of the American Academy of Political and Social Science*, 700(1):136–151.

Cristian Vaccari and Andrew Chadwick. 2020. Deepfakes and disinformation: Exploring the impact of synthetic political video on deception, uncertainty, and trust in news. *Social media+ society*, 6(1):2056305120903408.

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

Chen Wan, Tao Li, and Zhicheng Sun. 2017. Global stability of a seir rumor spreading model with demographics on scale-free networks. *Advances in Difference Equations*, 2017(1):253.

Herun Wan, Shangbin Feng, Zhaoxuan Tan, Heng Wang, Yulia Tsvetkov, and Minnan Luo. 2024. Dell: Generating reactions and explanations for llm-based misinformation detection. *arXiv preprint arXiv:2402.10426*.

Chenxi Wang, Zongfang Liu, Dequan Yang, and Xiuying Chen. 2024a. Decoding echo chambers: Llm-powered simulations revealing polarization in social networks. *arXiv preprint arXiv:2409.19338*.

Junxiang Wang, Junji Jiang, and Liang Zhao. 2022. An invertible graph diffusion neural network for source localization. In *Proceedings of the ACM Web Conference 2022*, pages 1058–1069.

Song Wang, Yaochen Zhu, Haochen Liu, Zaiyi Zheng, Chen Chen, and Jundong Li. 2024b. Knowledge editing for large language models: A survey. *ACM Comput. Surv.*, 57(3).

Alicia Wanless and Michael Berk. 2020. The audience is the amplifier: Participatory propaganda. *The SAGE handbook of propaganda*, pages 85–104.

Jianliang Wei and Fei Meng. 2021. How opinion distortion appears in super-influencer dominated social network. *Future Generation Computer Systems*, 115:542–552.

Ming Wei, Xin Wang, Longzhao Liu, Hongwei Zheng, Yishen Jiang, Yajing Hao, Zhiming Zheng, Feng Fu, and Shaoting Tang. 2025. Indirect reciprocity in the public goods game with collective reputations. *Journal of the Royal Society Interface*, 22(225):20240827.

Yotam Wolf, Noam Wies, Oshri Avnery, Yoav Levine, and Amnon Shashua. 2023. Fundamental limitations of alignment in large language models. *arXiv preprint arXiv:2304.11082*.

Ferre Wouters and Michaël Opgenhaffen. 2024. Regional facts matter: A comparative perspective of sub-state fact-checking initiatives in europe. *Media and Communication*, 12.

Feijie Wu, Zitao Li, Yaliang Li, Bolin Ding, and Jing Gao. 2024a. Fedbiot: Llm local fine-tuning in federated learning without full model. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 3345–3355.

Guangyang Wu, Weijie Wu, Xiaohong Liu, Kele Xu, Tianjiao Wan, and Wenyi Wang. 2023. Cheap-fake detection with llm using prompt engineering. In *2023 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*, pages 105–109. IEEE.

Jiaying Wu, Jiafeng Guo, and Bryan Hooi. 2024b. Fake news in sheep's clothing: Robust fake news detection against llm-empowered style attacks. In *Proceedings*

of the 30th ACM SIGKDD conference on knowledge discovery and data mining, pages 3367–3378.

Ke Wu, Song Yang, and Kenny Q Zhu. 2015. False rumors detection on sina weibo by propagation structures. In *2015 IEEE 31st international conference on data engineering*, pages 651–662. IEEE.

Yunpeng Xiao, Diqiang Chen, Shihong Wei, Qian Li, Haohan Wang, and Ming Xu. 2019. Rumor propagation dynamic model based on evolutionary game and anti-rumor. *Nonlinear Dynamics*, 95:523–539.

Chengxing Xie, Canyu Chen, Feiran Jia, Ziyu Ye, Kai Shu, Adel Bibi, Ziniu Hu, Philip Torr, Bernard Ghanem, and Guohao Li. 2024. Can large language model agents simulate human trust behaviors? *arXiv preprint arXiv:2402.04559*.

Qi Xuan, Xincheng Shu, Zhongyuan Ruan, Jinbao Wang, Chenbo Fu, and Guanrong Chen. 2019. A self-learning information diffusion model for smart social networks. *IEEE Transactions on Network Science and Engineering*, 7(3):1466–1480.

Ruidong Yan, Yi Li, Weili Wu, Deying Li, and Yongcai Wang. 2019. Rumor blocking through online link deletion on social networks. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 13(2):1–26.

Chang Yang, Peng Zhang, Wenbo Qiao, Hui Gao, and Jiaming Zhao. 2023. Rumor detection on social media with crowd intelligence and chatgpt-assisted networks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5705–5717.

Jian Yang, Xinyu Hu, Gang Xiao, and Yulong Shen. 2024. A survey of knowledge enhanced pre-trained language models. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* Just Accepted.

Liang Yang, Xiaochun Cao, Dongxiao He, Chuan Wang, Xiao Wang, and Weixiong Zhang. 2016. Modularity based community detection with deep learning. In *IJCAI*, volume 16, pages 2252–2258.

Barry Menglong Yao, Aditya Shah, Lichao Sun, Jin-Hee Cho, and Lifu Huang. 2023. End-to-end multimodal fact-checking and explanation generation: A challenging dataset and models. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2733–2743.

Dongyu Yao, Jianshu Zhang, Ian G Harris, and Marcel Carlsson. 2024. Fuzzllm: A novel and universal fuzzing framework for proactively discovering jailbreak vulnerabilities in large language models. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4485–4489. IEEE.

Wenhao Yu, Chenguang Zhu, Zaitang Li, Zhiting Hu, Qingyun Wang, Heng Ji, and Meng Jiang. 2022. A survey of knowledge-enhanced text generation. *ACM Comput. Surv.*, 54(11s).

Huaiwen Zhang, Quan Fang, Shengsheng Qian, and Changsheng Xu. 2019. Multi-modal knowledge-aware event memory network for social media rumor detection. In *Proceedings of the 27th ACM international conference on multimedia*, pages 1942–1951.

Mingqing Zhang, Haisong Gong, Qiang Liu, Shu Wu, and Liang Wang. 2024a. Breaking event rumor detection via stance-separated multi-agent debate. *arXiv preprint arXiv:2412.04859*.

Ningyu Zhang, Yunzhi Yao, Bozhong Tian, Peng Wang, Shumin Deng, Mengru Wang, Zekun Xi, Shengyu Mao, Jintian Zhang, Yuansheng Ni, Siyuan Cheng, Ziwen Xu, Xin Xu, Jia-Chen Gu, Yong Jiang, Pengjun Xie, Fei Huang, Lei Liang, Zhiqiang Zhang, Xiaowei Zhu, Jun Zhou, and Huajun Chen. 2024b. A comprehensive study of knowledge editing for large language models. *Preprint*, arXiv:2401.01286.

Xuan Zhang and Wei Gao. 2023. Towards llm-based fact verification on news claims with a hierarchical step-by-step prompting method. *arXiv preprint arXiv:2310.00305*.

Yuan Zhang, Tianshu Lyu, and Yan Zhang. 2018. Cosine: Community-preserving social network embedding from information diffusion cascades. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Yutao Zhang, Jie Tang, Zhilin Yang, Jian Pei, and Philip S Yu. 2015. Cosnet: Connecting heterogeneous social networks with local and global consistency. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1485–1494.

Laijun Zhao, Hongxin Cui, Xiaoyan Qiu, Xiaoli Wang, and Jiajia Wang. 2013. Sir rumor spreading model in the new media age. *Physica A: Statistical Mechanics and its Applications*, 392(4):995–1003.

Xuandong Zhao, Prabhanjan Ananth, Lei Li, and Yu-Xiang Wang. 2023. Provable robust watermarking for ai-generated text. *arXiv preprint arXiv:2306.17439*.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. 2023a. Lima: less is more for alignment. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Jiawei Zhou, Yixuan Zhang, Qianni Luo, Andrea G Parker, and Munmun De Choudhury. 2023b. Synthetic lies: Understanding ai-generated misinformation and evaluating algorithmic and human solutions.

In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–20.

Xiaoping Zhou, Xun Liang, Haiyan Zhang, and Yuefeng Ma. 2015. Cross-platform identification of anonymous identical users in multiple social media networks. *IEEE transactions on knowledge and data engineering*, 28(2):411–424.

Peican Zhu, Le Cheng, Chao Gao, Zhen Wang, and Xuelong Li. 2022a. Locating multi-sources in social networks with a low infection rate. *IEEE Transactions on Network Science and Engineering*, 9(3):1853–1865.

Yongchun Zhu, Qiang Sheng, Juan Cao, Shuokai Li, Danding Wang, and Fuzhen Zhuang. 2022b. Generalizing to the future: Mitigating entity bias in fake news detection. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2120–2125.

Arkaitz Zubiaga, Maria Liakata, Rob Procter, Kalina Bontcheva, and Peter Tolmie. 2015. Towards detecting rumours in social media. In *Workshops at the Twenty-Ninth AAAI conference on artificial intelligence*.

A Rumor Detection and Related Tasks

The core characteristic of rumors lies in their "unverified ambiguity and uncertainty," which makes them highly prone to misinterpretation or misuse during the dissemination process. Unlike debunked false information (misinformation) (Scheufele and Krause, 2019; Kumar and Geethakumari, 2014), deliberately fabricated falsehoods (disinformation) (Guo et al., 2020), or fake news that adopts the form of journalistic reporting to deliberately mislead the public (Shu et al., 2017, 2019) (Detecting fake news with NLP)], the uniqueness of rumors lies in the dynamic evolution of their verification status. Currently, rumor detection in a broad sense largely focuses on the verification of rumor veracity, emphasizing the description of the potential risks posed by false rumors to societal trust (Takahashi and Igata, 2012; Wu et al., 2015; Liang et al., 2015). From a narrower perspective, studies on rumor also consider their dissemination characteristics and societal impacts (Allport, 1947; Zubiaga et al., 2015). This provides theoretical support for uncovering the deeper logic underpinning rumor propagation while laying the foundational framework for research in rumor detection.

B Interdisciplinary research

Advancing rumor detection technology necessitates the development of a comprehensive interdisciplinary theoretical framework. This cross-disciplinary collaboration can illuminate the fundamental tension in information ecological governance—the dynamic interplay between bounded rationality in technical systems and the inherent complexity of social cognition processes.

B.1 Individual Cognition and Neuroscience

From the neuroscientific perspective, individuals evaluate the value of information when confronted with rumors, a process regulated by dopamine system activity (Rangel et al., 2008). The dopamine system comprises three subsystems: the Pavlovian system dominates instinctive responses, potentially prompting individuals to directly react to emotionally charged rumors; the habit system guides the acceptance of specific types of information based on information consumption patterns; and the goal-directed system supports more rational information analysis and verification. However, research (Tom et al., 2007) indicates that decision-making processes are often more influenced by the Pavlovian and habit systems, leading to simplified or biased rumor identification and verification. This explains why, when facing complex or ambiguous rumors, people tend to rely on intuition or existing cognitive frameworks rather than engaging in resource-intensive deep verification. Future research could utilize LLMs to simulate user cognition and decision-making, analyzing the potential impact of specific information on populations with different cognitive characteristics. For instance, studies (Menegas et al., 2018) have found that the brain's posterior striatum influences dopamine system activity, causing individuals to avoid threatening stimuli, which may explain why rumors related to public safety and health spread more rapidly. Based on this understanding, targeted defense tools can be developed, such as assistive tools that identify information overload states in real-time or detect emotional manipulation components, thereby enhancing individual and group capabilities to recognize attacks.

B.2 Group Dynamics and Game Theory

Within the framework of game theory, group gaming behaviors can reshape their environment, which in turn influences group strategy selection. Rumor propagation on social media exhibits complex group interaction characteristics. Incorporating game theory into rumor detection research enables more precise characterization of strategic compe-

tition, cooperation, and conflict between different groups, as well as analysis of the emergent conditions for information verification behavior, providing a theoretical foundation for achieving regulatory control throughout the entire process from initial diffusion to effective suppression. Research (Cornforth et al., 2012) demonstrates that the structural characteristics of groups significantly impact the emergence and maintenance of cooperative behavior, supporting the design of differentiated intervention measures and analysis of their effectiveness. Furthermore, beyond explicit behaviors such as cooperation, coordination, and betrayal, it is necessary to explore implicit causal relationships and long-term behavioral patterns. For example, indirect reciprocity reflects the common phenomenon of social groups using reputation and other indirect information to assist decision-making, constituting one of the fundamental mechanisms promoting the emergence of cooperation. In fast-paced, label-oriented online environments, users often rely on heuristic cues (such as information sources or group reputation) rather than independently evaluating each piece of information when making trust and dissemination decisions. Appropriate group evaluation mechanisms can break the "information verification dilemma" (Wei et al., 2025) by introducing social benefits (such as reputation enhancement) to incentivize more users to participate in verification. This has important theoretical and practical implications for utilizing LLMs to construct social simulation systems that better align with human social behavior logic and for guiding multi-agent systems in effective rumor identification and intervention decisions.

B.3 Intervention Strategies and Decision Science

The complexity of rumor propagation stems from the interactive effects of social psychology and group behavior dynamics. Communication theory (Nisbett and Wilson, 1977) reveals that rumor texts activate audience cognitive schemas through emotionally charged narrative structures (such as crisis-rendering rhetoric and moral binary opposition), inducing cognitive shortcuts based on existing beliefs and lowering information verification thresholds. In the LLMs era, these cognitive biases are amplified across multiple dimensions, making it difficult to effectively contain information diffusion chains through passive detection models alone. Future research needs to shift toward more proactive, personalized intervention strategies to suppress rumor propagation at its source. Taking individual cognition and group behavior as entry points, theories such as "cognitive immunity" and psychological positioning have demonstrated their ability to help users build "immunity" against future rumors (McGuire and Papageorgis, 1961; Efstratiou and De Cristofaro, 2022). These methods weaken false information by pre-exposure to induce resistant cognition, combining educational games, warning labels, and accuracy prompts to strengthen information critical capabilities (Traberg et al., 2022; Shen et al., 2021; Pennycook et al., 2021; Jones-Jang et al., 2021). However, single-frontend interventions are limited by cognitive inertia resulting from the "continued influence effect" and the identification gap caused by educational differences (higher education levels correlate with stronger identification abilities)(Lewandowsky et al., 2012; Afassinou, 2014; Hui et al., 2020). Therefore, integrated dynamic intervention strategies are needed, such as "friction-based interventions" that suppress conformity and impulsive behavior by increasing the "processing costs" of user decisions (e.g., information verification prompts, segmented content presentation, user interaction design). Simultaneously, utilizing LLM technology to provide real-time knowledge enhancement services for less educated groups creates a synergistic effect between educational gradient compensation and behavioral intervention, representing an important direction for future research.

C Information Ecosystem Governance Under Multi-Multi-dimensionalraints

In the context of rumor governance, the synergistic governance of legal, ethical, and technological constraints emerges as a necessary approach.

Legal measures should focus on regulating data usage while ensuring privacy protection. Privacy-preserving technologies(such as differential privacy (Chua et al., 2024; Mai et al., 2023) and federated learning(Kuang et al., 2024; Wu et al., 2024a; Ezzeldin et al., 2023)), combined with compliance frameworks(General Data Protection Regulation (GDPR) and the Artificial Intelligence Act (AI Act) (Parliament, 2023)), enhance model performance while safeguarding data security. These measures serve as a foundation for responsible rumor detection and governance in the digital age.

LLMs have been shown to possess the capa-

bility of inferring psychological tendencies from user-generated texts (Peters and Matz, 2024; Perc et al., 2019), potentially influencing users' false memories (Chan et al., 2024; Acerbi and Stubbersfield, 2023). Furthermore, LLM-generated content could be weaponized for privacy infringements, cognitive attacks, and social media manipulation (Huang and Zhu, 2023; Park et al., 2024). To address these challenges, platforms, and developers should proactively disclose algorithm designs, ensure data sources and security measures, and establish transparent accountability chains to enhance transparency and responsibility allocation (Dwivedi et al., 2023).

At the social governance level, advancing multi-stakeholder collaborative mechanisms is essential. This involves building a governance ecosystem that includes developers, policymakers, and sociologists, aimed at enhancing the transparency and societal adaptability of LLM technologies and achieving a comprehensive balance between technological efficiency and societal impact (Hu et al., 2024; Tokayev, 2023), which ensures that the governance of false information can be effectively expanded in different social and technical environments.