Rethink Rumor Detection in the Era of LLM: A Review

Anonymous ACL submission

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 flawless" rumors. This paper focuses on modeling rumor detection in the era of LLMs by unifying existing methods in the field of rumor detection and uncovering their underlying logical mechanisms. From the perspective of complex systems, we innovatively propose a "Cognition-Interaction-Behavior" (CIB) trilevel framework for rumor detection based on collective intelligence and explore the synergistic relationship between LLMs and collective intelligence in rumor governance. Furthermore, we analyze the core challenges in the LLM era and outline future development pathways for social simulation agents. 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 has enabled rumors to threaten public safety and social trust at unprecedented speeds, scales, and levels of complexity (Shao et al., 2016; 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 three-tiered "Cognition-Interaction-Behavior" (CIB) 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 (as shown in Figure 4) and uncovers the multi-scale coupling mechanisms underlying rumor propagation, including collective knowledge emergence, interactive network evolution, and iterative behavioral patterns. (2) A systematic exploration of LLMs' multi-faceted roles in rumor detection and their synergies

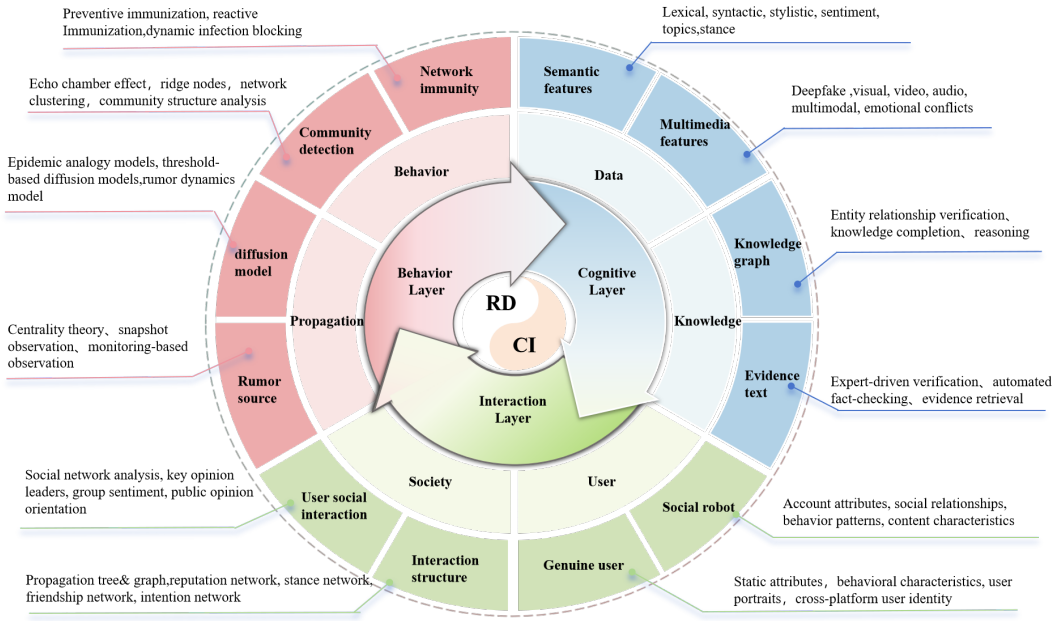

Figure 1: The three-layer architecture operates collaboratively. The cognition layer integrates multi-source evidence to provide informational support for the interaction layer. Through user interactions, the interaction layer facilitates the formation of the behavior layer. The behavior layer, in turn, 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).

with collective intelligence, forming a more comprehensive adaptive governance system. (3) A summary of the core challenges of rumor detection in the LLM era and an outline of future development pathways for socially simulated agents.

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 procesocial media users' the diverse cognition, social connections, and dynamic behaviers 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 three-order model framework for rumor detection based on collective intelligence, as shown in Figure 1. The cognitive layer facilitates the construction of a collective knowledge system for rumor identification through knowledge sharing and evidence integration among users. It serves as the foundational support for detecting rumors and aggregating multidmulti-dimensionalnce. The in-

teraction 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 Cognitive Layer

The Cognitive Layer leverages **data-driven analysis (§2.1.1)** to explore the features of multi-source information on social networks deeply, while **knowledge-driven analysis (§2.1.2)** focuses on verifying information sources and effectively integrating multi-perspective evidence. Together, these processes construct collective knowledge as the cognitive foundation for rumor detection.

2.1.1 Data-Driven Analysis

Data-driven analysis focuses on extracting features from multi-source data within social networks, aiming to develop multidmulti-dimensionalaches for assessing information veracity. Research on linguistic semantics, visual content, and multimodal characteristics provides critical support for more in-depth rumor detection efforts.

In terms of semantic analysis, rumor texts often exhibit multi-multi-dimensionallies. Studies have revealed distinct patterns at various levels by integrating linguistic and psychological theories (Undeutsch, 1967; Zuckerman, 1981). Lexical Level: rumolevelts tend to avoid expressing in-depth information (Aich et al., 2022; Horne and Adali, 2017), characterized by lower usage rates of first-person pronouns and higher proportions of adverbs and emotional words. Syntactic Level: rumolevelts exhibit a trend toward simplification (Pérez-Rosas et al., 2017), such as reduced lexical diversity and shorter average sentence lengths, abnormal frequency distributions of function words and punctuation, further highlight the "readability" characteristic of rumor texts. Stylistic Level: rumolevelts often feature exaggerated headlines, informal language, and frequent insertion of URLs or hashtags to enhance virality and attractiveness (Blass, 1984). These linguistic abnormalities also reveal rumors' strategic use of language to evoke negative emotions (e.g., anger, fear) in public, amplifying their dissemination, particularly in sensitive topics such as politics and health (Vosoughi et al., 2018). Studies have developed various detection methods through linguistic pattern analysis to leverage these characteristics to capture deeper semantic features.

In addition, deepfake technologies have significantly increased the deceptive capabilities and dissemination risks associated with image-based rumors (Vaccari and Chadwick, 2020). Early methods relied on spatial domain (Popescu and Farid, 2004) and frequency domain (Fridrich et al., 2003) analysis to extract pixel-level features for identifying common local manipulations in forged images but suffered from insufficient sensitivity to localized manipulations. Studies (Marra et al., 2019) have shown that statistical properties and spectral responses of GAN-generated content systematically differ from authentic images, providing a technological breakthrough for detection. The introduction of deep learning has further enhanced detection performance. Techniques such as local feature extraction (Bayar and Stamm, 2016; Wang et al., 2020a), temporal modeling (Güera and Delp, 2018), and the use of pre-trained models (Hao et al., 2021; Khan et al., 2022) have been effective in capturing complex visual forgery patterns. For video forgeries, beyond traditional frame-level classification methods (Montserrat et al., 2020), advancements have been made through inter-frame consistency analysis (Amerini et al., 2019), metadata validation (Huh et al., 2018), detection of visual artifacts (Matern et al., 2019), and leveraging biometric signal characteristics (Li et al., 2018). These approaches improve forgery detection performance by revealing inherent flaws in dynamic modeling. To more comprehensively capture the multimodal characteristics of rumor information, multimodal analysis focuses on feature fusion and consistency verification (Wang et al., 2018a; Jin et al., 2017). These methods balance modality-specific features and improve cross-modal detection performance by employing different fusion strategies (Singhal et al., 2019; Qian et al., 2021). Consistency verification is utilized to identify cross-modal information conflicts, including text-image inconsistencies (e.g., emotional conflicts) and audio-visual mismatches (e.g., forged videos) (Agarwal et al., 2020; Chugh et al., 2020). These approaches effectively enhance the performance of multimodal rumor detection.

2.1.2 Knowledge-Driven Analysis

The knowledge-driven analysis leverages external information resources to enrich and validate rumor content, offering critical support for social net-

works' noisy, concise content. Verification methods based on **knowledge graphs** and **evidence texts** are the core research content.

As structured data systems, knowledge graphs provide a networked organization of large-scale entities and relationships, supporting rumor detection through contextual verification and logical reasoning. By matching entities and relationships within the text, knowledge graphs can quickly validate content accuracy and identify potential contradictions (Cui et al., 2020). Several studies (Hu et al., 2021) further integrate semantic analysis and graph reasoning techniques to uncover implicit associations or supporting evidence, thereby improving verification reliability. For rumor content characterized by semantic ambiguity or missing information, knowledge graphs leverage their capabilities in semantic completion and reasoning to explore hidden, deeper-level entity associations. By incorporating contextual information (Dun et al., 2021) and multimodal data (Wang et al., 2020b), knowledge graphs can fill in critical gaps within implicit or ambiguous statements. Furthermore, they have demonstrated robust adaptability in cross-domain rumor detection tasks (Sun et al., 2022; Zhang et al., 2019).

On the other hand, evidence-based text verification focuses on fact-checking, aiming to validate the factuality of rumor content through authoritative information sources such as news articles, scientific literature, and fact-checking platforms. Traditionally, this task has relied on expert-driven manual verification. Internationally accredited organizations such as the International Fact-Checking Network (IFCN) (Porter and Wood, 2021), the European Fact-Checking Standards Network (EFCSN) (Wouters and Opgenhaffen, 2024), and government platforms (e.g., China's Internet Joint Rumor Debunking Platform, Tencent Fact-Check Platform) provide standardized evaluation procedures to assess the authenticity and timeliness of evidence, delivering high-quality validation services to the public (Vlachos and Riedel, 2014).

However, the expert-driven model faces efficiency bottlenecks and struggles to respond rapidly to large-scale, real-time information dissemination demands (Das et al., 2023; Guo et al., 2022). Automated fact-checking has increasingly become a focus of research to address these limitations. Key processes in these systems include multi-source data retrieval, semantic alignment, and logical reasoning. By leveraging deep learning models combined with information retrieval methods (Hanselowski et al., 2019), semantic relevance between rumors and evidence is extracted from authoritative data sources such as Wikipedia and scientific literature (Schuster et al., 2021; Wadden et al., 2021). Additionally, semantic alignment techniques are employed to assess the extent to which the retrieved evidence supports or refutes the claims embedded in the rumor. For complex, multi-layered claims, deep learning methodologies (Zhong et al., 2019) can generate reliable verification results. Explainability-enhancing techniques are also applied to extract key logic and evidence chains, improving users' understanding of and trust in the verification outcomes (Lu and Li, 2020).

2.2 Interaction layer

The Interaction Layer emphasizes the analysis of **user feature (§2.2.1))** and **social context (§2.2.2)** to provide comprehensive contextual information, including individual behavior patterns, interactions between groups, and the dynamic formation of group consensus. This not only reveals the processes through which information spreads among users but also captures the collaboration and conflicts involved in users' efforts to discern and verify rumors.

2.2.1 User Feature Analysis

User groups within social networks serve as the core driving force behind rumor propagation. They are comprised of genuine users and social bots tasked with content generation and dissemination. Analyzing user characteristics and behavioral patterns can effectively uncover rumor dissemination's underlying mechanisms and risks.

As specialized accounts, social bots often manipulate public opinion through high-frequency content posting and synchronized interactions, significantly accelerating rumor dissemination. Bot detection methods can be broadly categorized into feature-based and network-based approaches. The feature-based analysis identifies non-human attributes, such as anomalous metadata (e.g., default profile pictures, short-lived accounts), high-frequency posting behaviors, and polarized content (e.g., repetitive sentence structures, simple semantics). Network-based detection detects bots by identifying abnormalities in dissemination structures. Social bots often amplify their influence by forming densely interconnected groups or fabricating community structures. In recent years, deep learning techniques, such as Graph Neural

Networks (GNNs), have been employed to model the dependencies within network topologies (Guo et al., 2021). When combined with content features, pre-trained language models have further enhanced the generalization performance of social bot detection in heterogeneous environments (Haider et al., 2023).

Genuine users are the central driving force behind information dissemination (Aral and Walker, 2012). Among them, malicious users deliberately spread rumors for personal or organizational gain (Kahneman, 1979), while ordinary users may inadvertently propagate rumors due to cognitive limitations or emotional triggers. Research has delved into the dissemination patterns of these users by examining their static attributes (e.g., registration time, geographical location, number of friends) and dynamic behavioral characteristics (e.g., posting frequency, emotional traits such as anger and fear, and account credibility scores) (Chu et al., 2012). For example, anomalies such as short-term, high-frequency interactions and highly repeated content releases are often considered potential signals of rumor spreading (Zhao et al., 2014). Additionally, as rumor dissemination increasingly transcends platform boundaries, cross-platform user identity association analysis has emerged as a research focus. For example, by matching user profiles (Iofciu et al., 2011), content geolocation (Riederer et al., 2016), and personalized traits such as posting style (Goga et al., 2013) and interest preferences (Nie et al., 2016), researchers can identify attempts by malicious users to disguise their identities across platforms. Studies have increasingly applied unified analyses of static and dynamic behaviors across broader network ecosystems by integrating deep learning and network analysis techniques (Hamdi et al., 2020; Zhang et al., 2015; Zhou et al., 2015).

2.2.2 Social Context Analysis

Rumor propagation is influenced by individual behaviors and the deeper constraints imposed by social network structures and user interaction patterns. By uncovering the synergies between user interactions and network structures, social context analysis provides critical support for understanding the mechanisms underlying the diffusion of rumors.

The interaction characteristics within social network structures exhibit distinct patterns in rumor propagation. Due to a lack of supervision and information verification mechanisms, Sparse networks tend to form low-cohesion, flat diffusion structures, which accelerate the spread of false information (Vosoughi et al., 2018). In contrast, dense networks, with their strong connectivity, can partially filter or curb the propagation of rumors. Moreover, the heterogeneity of user roles within a network (e.g., "messengers" bridging multiple communities or "skeptics" constructing local verification networks) dynamically impacts interaction structures (Raponi et al., 2022). For example, user comment chains and resharing behaviors can be modeled as propagation tree structures (Kwon et al., 2013), which are utilized to capture both top-down and bottom-up propagation patterns (Alrubaian et al., 2016; Ma et al., 2015). To more comprehensively represent such complex social contexts, multiple entities such as users, posts, and hashtags can be modeled as propagation graphs (Nguyen et al., 2020; Shu et al., 2020). Studies also introduce graph neural networks (Bian et al., 2020; Min et al., 2022), which aggregate node features to capture complex interaction relationships and diffusion dynamics among users. These significantly enhance the efficacy of modeling rumor propagation patterns and their underlying mechanisms.

User interaction patterns serve as critical clues for uncovering rumor social context. Studies have shown that genuine users, fake news producers, and hybrid users tend to form homogeneous clusters within collaborative networks. The polarization between different clusters reflects regional differences and political and ideological divisions. This phenomenon is particularly pronounced in rumor propagation, where emotion acts as a catalyst. Emotionally charged content, by evoking negative emotions such as anger and fear, triggers group polarization and amplifies the speed and scope of rumor dissemination (Zeng and Zhu, 2019; Pröllochs et al., 2021). According to the "two-step flow" theory in communication studies (Katz, 1957), information flows first from mass media to key opinion leaders (KOLs), who then influence broader audiences. KOLs often act as community bridges (Yang et al., 2018), playing a significant amplifying role in rumor dissemination while also having the potential to contribute effectively to rumor debunking. Modeling KOLs is crucial for understanding their influence mechanisms in rumor propagation (Wei and Meng, 2021). The most common approaches for KOL modeling involve social network analysis techniques, utilizing centrality metrics (Opsahl et al., 2010), statistical models (Amor et al., 2016),

network topology analysis (Zhao et al., 2016), and deep learning methods (Shafiq et al., 2013). These methods aid in designing effective intervention strategies.

2.3 Behavior Layer

Rampant deepfakes and false rumors are often blamed as key culprits in influencing voter behavior. Studies suggest that while changing people's political opinions is challenging, influencing their actions is comparatively easier (Adam). The spread of rumors on social networks has been extensively studied from both macro and micro perspectives (Xuan et al., 2019). At the macro level, reselevelleverages **propagation pattern modeling (§2.3.1)** to analyze the dynamic processes of rumor propagation in complex networks (Zhu and Huang, 2019). At the micro level, studlevelocus on **behavioral pattern analysis (§2.3.2)** to analyze community characteristics and predict rumor dissemination (Alkhodair et al., 2020).

2.3.1 Propagation Pattern Modeling

Propagation modeling aims to uncover the diffusion patterns of rumors within the framework of a complex sociodynamic system, focusing on the coupled process of information dissemination and collective behaviors. Epidemic analogy models serve as the foundational approach in early studies, where state transition mechanisms (e.g., Susceptible-Infectious) are used to describe the spread of rumors across network topologies (e.g., SI (Kermack and McKendrick, 1927), SIS (Dong and Huang, 2018), SIR (Zhao et al., 2013), and their variants (Wan et al., 2017). Threshold-based diffusion models, such as the Linear Threshold Model (LT) (Chen et al., 2012) and the independent cascade model (IC), shift toward modeling the propagation process from the perspective of audience decision-making. These models are also used to design intervention strategies, such as node blocking or link disruption, to suppress diffusion (Yan et al., 2019). With the development of complex network theory, more studies focus on exploring multidmulti-dimensionalrs that influence rumor propagation in online social networks: temporal dimension (Tripathy et al., 2010), user dimension (Hosni et al., 2018, 2020), network dimension (Wang et al., 2018b), and information dimension (Xiao et al., 2019), providing a more systematic theoretical foundation for studying the diffusion of rumors in complex networks.

Rumor source detection based on propagation models plays a pivotal role in tracing the origins of information dissemination, providing crucial insights for intervention strategies. Early methods, grounded in centrality theory, estimate the importance of the nodes by traversing the global topology of social networks to identify the source nodes (Ali et al., 2020). However, these approaches often suffer from high computational complexity. Snapshot observation methods (Louni and Subbalakshmi, 2018) improve detection efficiency by extracting limited node infection states or propagation paths. These methods effectively perform on homogeneous and heterogeneous propagation models (Cai et al., 2018). Additionally, monitoring-based observation techniques (Qiu et al., 2022) leverage sensor nodes to capture real-time dissemination data, enhancing adaptability to dynamic propagation environments. In scenarios involving multi-source concurrent propagation (Zhu et al., 2022a), research focuses on decoupling infection networks into several independent regions. A divide-and-conquer strategy is then employed to locate the sources iteratively, reducing the computational complexity of detection. With the maturation of deep learning techniques (Wang et al., 2022; Ling et al., 2022) and Graph Neural Networks (GNNs) (Cheng et al., 2024), end-to-end frameworks have emerged as highly effective tools. By integrating propagation paths, temporal dynamics, and node characteristics, these approaches significantly enhance the robustness and accuracy of rumor source detection.

2.3.2 Behavioral Pattern Analysis

In social networks, nodes often cluster into tightly-knit communities, where rumors spread efficiently within communities but rely on bridge nodes or weak ties between communities for cross-community dissemination. Research shows that when bridge nodes are scarce, rumor dissemination remains localized, but when sufficiently abundant, it penetrates communities and reaches a broader audience (Zanette, 2001). To identify these critical nodes, community detection methods (Newman, 2004; Blondel et al., 2008) usually adopt heuristics to find closely related subgroups in the network (Zhang et al., 2018; Yang et al., 2016). Targeted interventions and immunization strategies can then be applied to these key nodes to minimize the spread of rumors.

Network immunization strategies are typically categorized into preventive and counteractive im-

munization: Preventive Immunization focuses on proactively optimizing network structures and key node distributions before harmful information emerges. By analyzing topological properties (e.g., node centrality, spectral attributes) and community structure characteristics, high-risk nodes, and cross-community bridge points can be identified to disseminate accurate information (Petrescu et al., 2021). Counteractive immunization emphasizes real-time intervention during the propagation process. When initial sources or infected nodes are known, dynamic detection and analysis of infected nodes and their neighboring communities are performed. Key high-dissemination nodes are then selected for removal or blocking to efficiently disrupt the propagation chain with minimal cost (Fan et al., 2013). It prioritizes local optimization under resource constraints, such as minimizing the dissemination of malicious information within the network (Tariq et al., 2017), for instance, reducing the probability of some users sharing false content with their network connections.

3 Collective Intelligence-based Rumor Detection in the LLM Era

With the advancement of Natural Language Processing (NLP), rumor detection techniques have evolved from traditional statistical methods (Lim et al., 2017; Rayana and Akoglu, 2015) to deep learning models (Ma et al., 2016; Chen et al., 2019), and more recently to pre-trained language models (PLMs) (Kaliyar et al., 2021; Pelrine et al., 2021), progressively transitioning from static feature extraction to automated dynamic semantic analysis. However, traditional methods often suffer from limited adaptability in scenarios such as early-stage rumor detection and complex environments due to the scarcity of annotated data (He et al., 2021). LLMs further transform the rumor detection landscape with their extensive pre-training and contextual reasoning capabilities, enabling efficient reasoning in resource-poor environments and significantly improving the transparency and interpretability of detection results.

3.1 LLM-enhanced CIB Framework

LLMs play a multifaceted role in existing rumor detection approaches, deeply integrating into various roles ranging from knowledge analysis to adversarial defense, as shown in Figure 2. LLMs have brought revolutionary advancements to traditional methods, demonstrating significant technical potential and promising application prospects.

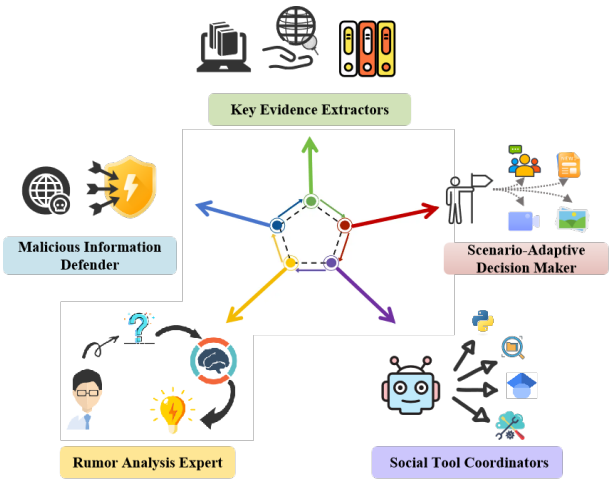

Figure 2: Multiple roles of LLM in rumor detection

At the cognitive layer, LLMs serve as highly efficient **Key Evidence Extractors** through large-scale pretraining. Unlike traditional rumor detection systems that rely on static, structured approaches(such as knowledge graphs) to expand knowledge, LLMs leverage their implicit encoding capabilities to capture deep semantic associations in unstructured information, which is further used as evidence to enhance the generalization ability of traditional language models (Nan et al., 2024; Yang et al., 2023a). Additionally, LLMs operate as **Scenario-Adaptive Decision Makers**, leveraging their zero-shot reasoning abilities to address diversified rumor scenarios efficiently without requiring fine-tuning (Li et al., 2023c; Wu et al., 2023a). Combining Retrieval-Augmented Generation (RAG) with external knowledge bases (Peng et al., 2023; Niu et al., 2024), LLMs can dynamically integrate up-to-date knowledge, resolving limitations in knowledge coverage and timeliness inherent to traditional methods. This integration also effectively reduces the likelihood of hallucination phenomena, thereby enhancing the reliability of detection outputs (Ji et al., 2023; Rawte et al., 2023).

At the interaction layer, LLMs function as **Social Tool Coordinators**, coordinating external tools (e.g., search engines, deepfake detectors) through agents to expand rumor detection capabilities further (Chern et al., 2023; Wan et al., 2024; Li et al., 2024a). Unlike traditional, static social network analysis and modeling methods, LLM-based agents can perceive social environments by combining

short-term (contextual learning) and long-term (external knowledge retrieval) memory. These agents are capable of planning and calling external tools dynamically, improving analytical performance. Furthermore, generative agents (Park et al., 2023) can simulate user interaction behaviors, driving a paradigm shift in rumor detection from static feature modeling to dynamic interaction simulation.

At the behavior layer, LLMs act as **Rumor Analysis Experts** with superior performance in tasks requiring advanced reasoning and cross-domain contextual knowledge. Traditional rumor detection approaches relied heavily on classification tasks, often requiring manually labeled large datasets. In contrast, the emergent abilities of LLMs, such as Chain of Thought (CoT) reasoning, can decompose complex problems into intermediate reasoning steps, thereby significantly improving logical transparency and explainability (Zhang and Gao, 2023a). LLMs also exhibit robust cross-domain transferability (Cao et al., 2023c,b), enabling unified reasoning across multimodal inputs, including text, images, and audio (Yao et al., 2023a). This addresses the limitations of traditional methods in multimodal fusion and shifts detection mechanisms from pattern classification to causal inference (Zhu et al., 2022b; Nan et al., 2021). Additionally, LLMs can act as **Malicious Information Defenders**, showcasing strong robustness in adversarial social network environments. By integrating adversarial training and red-teaming methods (Bhardwaj and Poria, 2023; OpenAI, 2023), LLMs can quickly adapt to evolving forgery techniques, overcoming the lag in model iteration and processing capabilities of traditional approaches (Wu et al., 2024b; Sun et al., 2024). For example, when tackling tasks such as detecting rumor dissemination, stylized language attacks, and deepfake content, this dynamic adaptability further enhances the robustness of rumor detection systems.

3.2 Collective Intelligence-driven LLM Detection

Network users are the primary agents of information dissemination and important participants in rumor detection. The propagation of rumors relies on user attention, trust, and further sharing behaviors. In contrast, user reports and feedback can effectively constrain this diffusion effect, which is critical in rumor detection. This user feedback-based supervisory mechanism helps address the limitations of LLMs in adapting to dynamic scenarios by introducing ethical constraints and social consensus at the human cognitive level, thelevelinfusing greater flexibility and human-centricity into the rumor detection process (Kou et al., 2022).

Collective intelligence injects new technical implications into the development of LLM agents. Research demonstrates (Li et al., 2023b) that agents, by simulationing group behaviors modeled in sociology and economics theories, can exhibit emergent corrective mechanisms during collaborative tasks. These emergent behaviors provide theoretical support for rumor detection agents (Zhang et al., 2024b), demonstrating their significant potential in improving the simulation of social media ecosystems and other complex societal environments. For example, agents can produce socially simulated content that is indistinguishable from real-world community behavior (Park et al., 2022), simulate trust-building interactions in social dynamics (Xie et al., 2024), and facilitate harmless discussions that bridge biases and political divides, offering valuable insights into real-world phenomena (Törnberg et al., 2023).

In the field of rumor detection, existing research (Hu et al., 2025) further uses LLM-based multi-agent simulation to explore the trend of rumor propagation and optimize intervention strategies and also uses tools to implement (Li et al., 2024a) real-time evaluation of information credibility based on shallow features such as language style and common sense rules. However, the complexity of the rumor detection task requires a more comprehensive approach that considers multi-level features. We discussed this in the future research route (Appendix B), paving the way for future exploration.

4 Conclusion

Based on the complex system characteristics of collective intelligence, we reconstructed the rumor detection paradigm adapted to the era of LLMs, the "Cognition-Interaction-Behavior" (CIB) framework. We emphasized the important role of LLM in rumor detection and its complementary relationship with collective intelligence. In addition, CIB can use cross-layer dynamic feedback to establish a "Macro-Micro Feedback Loop" to dynamically realize two-way rumor detection and intervention, providing a roadmap for applying LLM-based multi-agent social simulation in rumor detection.

5 Limitations

In the future research section of this paper, we propose a roadmap for collective intelligence-based rumor detection agents under the CIB framework, providing a comprehensive analysis of potential research challenges and corresponding directions. However, further exploration is needed to evaluate the practical application of this framework 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*.

David Adam. Misinformation might sway elections-but not in the way that you think. *Nature*.

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.

Muhammad Ahmad, Sarwan Ali, Juvaria Tariq, Imdadullah Khan, Mudassir Shabbir, and Arif Zaman. 2020. Combinatorial trace method for network immunization. *Information Sciences*, 519:215–228.

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.

Abdulrahman I Al-Ghadir, Aqil M Azmi, and Amir Hussain. 2021. A novel approach to stance detection in social media tweets by fusing ranked lists and sentiments. *Information Fusion*, 67:29–40.

Syed Shafat Ali, Tarique Anwar, and Syed Afzal Murtaza Rizvi. 2020. A revisit to the infection source identification problem under classical graph centrality measures. *Online Social Networks and Media*, 17:100061.

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*.

Majed Alrubaian, Muhammad Al-Qurishi, Mohammad Mehedi Hassan, and Atif Alamri. 2016. A credibility analysis system for assessing information on twitter. *IEEE Transactions on Dependable and Secure Computing*, 15(4):661–674.

Irene Amerini, Leonardo Galteri, Roberto Caldelli, and Alberto Del Bimbo. 2019. Deepfake video detection through optical flow based cnn. In *Proceedings of the IEEE/CVF international conference on computer vision workshops*, pages 0–0.

Benjamin RC Amor, Sabine I Vuik, Ryan Callahan, Ara Darzi, Sophia N Yaliraki, and Mauricio Barahona. 2016. Community detection and role identification in directed networks: understanding the twitter network of the care. data debate. In *Dynamic networks and cyber-security*, pages 111–136. World Scientific.

Markus Anderljung, Joslyn Barnhart, Anton Korinek, Jade Leung, Cullen O'Keefe, Jess Whittlestone, Shahar Avin, Miles Brundage, Justin Bullock, Duncan Cass-Beggs, et al. 2023. Frontier ai regulation: Managing emerging risks to public safety. *arXiv preprint arXiv:2307.03718*.

Christopher MJ André, Helene FL Eriksen, Emil J Jakobsen, Luca CB Mingolla, and Nicolai B Thomsen. 2023. Detecting ai authorship: Analyzing descriptive features for ai detection.

Fares Antaki, Samir Touma, Daniel Milad, Jonathan El-Khoury, and Renaud Duval. 2023. Evaluating the performance of chatgpt in ophthalmology: an analysis of its successes and shortcomings. *Ophthalmology science*, 3(4):100324.

Dimosthenis Antypas, Jose Camacho-Collados, Alun Preece, and David Rogers. 2021. Covid-19 and misinformation: A large-scale lexical analysis on twitter. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Student Research Workshop*, pages 119–126.

Elena-Simona Apostol, Özgur Coban, and Ciprian-Octavian Truică. 2023. Contain: A community-based algorithm for network immunization. *arXiv preprint arXiv:2303.01934*.

Sinan Aral and Dylan Walker. 2012. Identifying influential and susceptible members of social networks. *Science*, 337(6092):337–341.

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.

Anton Bakhtin, Sam Gross, Myle Ott, Yuntian Deng, Marc'Aurelio Ranzato, and Arthur Szlam. 2019. Real or fake? learning to discriminate machine from human generated text. *arXiv preprint arXiv:1906.03351*.

Eytan Bakshy, Itamar Rosenn, Cameron Marlow, and Lada Adamic. 2012. The role of social networks in information diffusion. In *Proceedings of the 21st international conference on World Wide Web*, pages 519–528.

Ritwik Banerjee, Song Feng, Jun Seok Kang, and Yejin Choi. 2014. Keystroke patterns as prosody in digital writings: A case study with deceptive reviews and essays. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1469–1473.

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. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 675–718, Nusa Dua, Bali. Association for Computational Linguistics.

Jawadul H Bappy, Amit K Roy-Chowdhury, Jason Bunk, Lakshmanan Nataraj, and BS Manjunath. 2017. Exploiting spatial structure for localizing manipulated image regions. In *Proceedings of the IEEE international conference on computer vision*, pages 4970–4979.

Anthony M Barrett, Dan Hendrycks, Jessica Newman, and Brandie Nonnecke. 2022. Actionable guidance for high-consequence ai risk management: Towards standards addressing ai catastrophic risks. *arXiv preprint arXiv:2206.08966*.

Belhassen Bayar and Matthew C Stamm. 2016. A deep learning approach to universal image manipulation detection using a new convolutional layer. In *Proceedings of the 4th ACM workshop on information hiding and multimedia security*, pages 5–10.

Yoshua Bengio, Geoffrey Hinton, Andrew Yao, Dawn Song, Pieter Abbeel, Yuval Noah Harari, Ya-Qin Zhang, Lan Xue, Shai Shalev-Shwartz, Gillian Hadfield, et al. 2023. Managing ai risks in an era of rapid progress. *arXiv preprint arXiv:2310.17688*, page 18.

Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michal Podstawski, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Hubert Niewiadomski, Piotr Nyczyk, et al. 2024. Graph of thoughts: Solving elaborate problems with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17682–17690.

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

Xiuli Bi, Yang Wei, Bin Xiao, and Weisheng Li. 2019. Rru-net: The ringed residual u-net for image splicing forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0.

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.

Thomas Blass. 1984. Social psychology and personality: Toward a convergence. *Journal of Personality and Social Psychology*, 47(5):1013.

Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. 2008. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008.

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*.

Asgarali Bouyer, Hamid Ahmadi Beni, Bahman Arasteh, Zahra Aghaee, and Reza Ghanbarzadeh. 2023. Fip: A fast overlapping community-based influence maximization algorithm using probability coefficient of global diffusion in social networks. *Expert systems with applications*, 213:118869.

Samuel R Bowman, Jeeyoon Hyun, Ethan Perez, Edwin Chen, Craig Pettit, Scott Heiner, Kamilė Lukošiūtė, Amanda Askell, Andy Jones, Anna Chen, et al. 2022. Measuring progress on scalable oversight for large language models. *arXiv preprint arXiv:2211.03540*.

Arastoo Bozorgi, Saeed Samet, Johan Kwisthout, and Todd Wareham. 2017. Community-based influence maximization in social networks under a competitive linear threshold model. *Knowledge-Based Systems*, 134:149–158.

Jean-Flavien Bussotti, Luca Ragazzi, Giacomo Frisoni, Gianluca Moro, and Paolo Papotti. 2024. Unknown claims: Generation of fact-checking training examples from unstructured and structured data. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12105–12122.

Kechao Cai, Hong Xie, and John CS Lui. 2018. Information spreading forensics via sequential dependent snapshots. *IEEE/ACM Transactions on Networking*, 26(1):478–491.

Han Cao, Lingwei Wei, Mengyang Chen, Wei Zhou, and Songlin Hu. 2023a. Are large language models good fact checkers: A preliminary study. *arXiv preprint arXiv:2311.17355*.

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

Nicholas Carlini, Milad Nasr, Christopher A Choquette-Choo, Matthew Jagielski, Irena Gao, Pang Wei W Koh, Daphne Ippolito, Florian Tramer, and Ludwig Schmidt. 2023. Are aligned neural networks adversarially aligned? *Advances in Neural Information Processing Systems*, 36:61478–61500.

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.

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. 2023. Chateval: Towards better llm-based evaluators through multi-agent debate. *arXiv preprint arXiv:2308.07201*.

Michael Chan, Francis LF Lee, and Hsuan-Ting Chen. 2021. Examining the roles of multi-platform social media news use, engagement, and connections with news organizations and journalists on news literacy: A comparison of seven democracies. *Digital Journalism*, 9(5):571–588.

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*.

Canyu Chen and Kai Shu. 2023. Can llm-generated misinformation be detected? *arXiv preprint arXiv:2309.13788*.

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

Hung-Hsuan Chen, Yan-Bin Ciou, and Shou-De Lin. 2012. Information propagation game: A tool to acquire humanplaying data for multiplayer influence maximization on social networks. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1524–1527.

Junyi Chen, Leyuan Liu, and Fan Zhou. 2025. Do not wait: Preemptive rumor detection with cooperative llms and accessible social context. *Information Processing & Management*, 62(3):103995.

Lei Chen, Guanying Li, Zhongyu Wei, Yang Yang, Baohua Zhou, Qi Zhang, and Xuanjing Huang. 2022. A progressive framework for role-aware rumor resolution. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2748–2758, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Sanxing Chen, Yukun Huang, and Bhuwan Dhingra. 2024. Real-time fake news from adversarial feedback. *arXiv preprint arXiv:2410.14651*.

Weixuan Chen and Daniel McDuff. 2018. Deepphys: Video-based physiological measurement using convolutional attention networks. In *Proceedings of the european conference on computer vision (ECCV)*, pages 349–365.

Yimin Chen, Niall J Conroy, and Victoria L Rubin. 2015. Misleading online content: recognizing clickbait as" false news". In *Proceedings of the 2015 ACM on workshop on multimodal deception detection*, pages 15–19.

Yixuan Chen, Jie Sui, Liang Hu, and Wei Gong. 2019. Attention-residual network with cnn for rumor detection. In *Proceedings of the 28th ACM international conference on information and knowledge management*, pages 1121–1130.

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*.

Zi Chu, Steven Gianvecchio, Haining Wang, and Sushil Jajodia. 2012. Detecting automation of twitter accounts: Are you a human, bot, or cyborg? *IEEE Transactions on dependable and secure computing*, 9(6):811–824.

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*.

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.

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

Chaoqun Cui and Caiyan Jia. 2024. Propagation tree is not deep: Adaptive graph contrastive learning approach for rumor detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 73–81.

Jian Cui, Kwanwoo Kim, Seung Ho Na, and Seungwon Shin. 2022. Meta-path-based fake news detection leveraging multi-level social context information. In *Proceedings of the 31st ACM international conference on information & knowledge management*, pages 325–334.

Limeng Cui, Haeseung Seo, Maryam Tabar, Fenglong Ma, Suhang Wang, and Dongwon Lee. 2020. Deterrent: Knowledge guided graph attention network for detecting healthcare misinformation. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 492–502.

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.

Soumita Das, Ravi Kishore Devarapalli, and Anupam Biswas. 2024. Leveraging cascading information for community detection in social networks. *Information Sciences*, 674:120696.

Luigi De Angelis, Francesco Baglivo, Guglielmo Arzilli, Gaetano Pierpaolo Privitera, Paolo Ferragina, Alberto Eugenio Tozzi, and Caterina Rizzo. 2023. Chatgpt and the rise of large language models: the new ai-driven infodemic threat in public health. *Frontiers in public health*, 11:1166120.

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.

Natalia Díaz-Rodríguez, Javier Del Ser, Mark Coeckelbergh, Marcos López de Prado, Enrique Herrera-Viedma, and Francisco Herrera. 2023. Connecting the dots in trustworthy artificial intelligence: From ai principles, ethics, and key requirements to responsible ai systems and regulation. *Information Fusion*, 99:101896.

Peter Sheridan Dodds, Eric M Clark, Suma Desu, Morgan R Frank, Andrew J Reagan, Jake Ryland Williams, Lewis Mitchell, Kameron Decker Harris, Isabel M Kloumann, James P Bagrow, et al. 2015. Human language reveals a universal positivity bias. *Proceedings of the national academy of sciences*, 112(8):2389–2394.

Ming Dong, Bolong Zheng, Nguyen Quoc Viet Hung, Han Su, and Guohui Li. 2019. Multiple rumor source detection with graph convolutional networks. In *Proceedings of the 28th ACM international conference on information and knowledge management*, pages 569–578.

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.

John Dougrez-Lewis, Elena Kochkina, Maria Liakata, and Yulan He. 2024. Knowledge graphs for real-world rumour verification. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 9843–9853.

Yaqian Dun, Kefei Tu, Chen Chen, Chunyan Hou, and Xiaojie Yuan. 2021. Kan: Knowledge-aware attention network for fake news detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 81–89.

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.

Owain Evans, Owen Cotton-Barratt, Lukas Finnveden, Adam Bales, Avital Balwit, Peter Wills, Luca Righetti, and William Saunders. 2021. Truthful ai: Developing and governing ai that does not lie. *arXiv preprint arXiv:2110.06674*.

Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Song. 2018. Robust physical-world attacks on deep learning visual classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1625–1634.

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.

Lidan Fan, Zaixin Lu, Weili Wu, Bhavani Thuraisingham, Huan Ma, and Yuanjun Bi. 2013. Least cost rumor blocking in social networks. In *2013 IEEE 33rd International Conference on Distributed Computing Systems*, pages 540–549. IEEE.

Mahmoud Fawzi and Walid Magdy. 2024. " pinocchio had a nose, you have a network!": On characterizing fake news spreaders on arabic social media. *Proceedings of the ACM on Human-Computer Interaction*, 8(CSCW1):1–20.

Shangbin Feng, Zhaoxuan Tan, Herun Wan, Ningnan Wang, Zilong Chen, Binchi Zhang, Qinghua Zheng, Wenqian Zhang, Zhenyu Lei, Shujie Yang, et al. 2022. Twibot-22: Towards graph-based twitter bot detection. *Advances in Neural Information Processing Systems*, 35:35254–35269.

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.

Santo Fortunato. 2010. Community detection in graphs. *Physics reports*, 486(3-5):75–174.

Santo Fortunato and Marc Barthelemy. 2007. Resolution limit in community detection. *Proceedings of the national academy of sciences*, 104(1):36–41.

Santo Fortunato and Darko Hric. 2016. Community detection in networks: A user guide. *Physics reports*, 659:1–44.

Gabriel Freedman, Adam Dejl, Deniz Gorur, Xiang Yin, Antonio Rago, and Francesca Toni. 2024. Argumentative large language models for explainable and contestable decision-making. *arXiv preprint arXiv:2405.02079*.

Jessica Fridrich, David Soukal, Jan Lukas, et al. 2003. Detection of copy-move forgery in digital images. In *Proceedings of digital forensic research workshop*, volume 3, pages 652–63. Cleveland, OH.

Rinaldo Gagiano, Maria Myung-Hee Kim, Xiuzhen Jenny Zhang, and Jennifer Biggs. 2021. Robustness analysis of grover for machine-generated news detection. In *Proceedings of the 19th Annual Workshop of the Australasian Language Technology Association*, pages 119–127.

ŁG Gajewski, Krzysztof Suchecki, and JA Hołyst. 2019. Multiple propagation paths enhance locating the source of diffusion in complex networks. *Physica A: Statistical Mechanics and its Applications*, 519:34–41.

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*.

Maryanne Garry, Way Ming Chan, Jeffrey Foster, and Linda A Henkel. 2024. Large language models (llms) and the institutionalization of misinformation. *Trends in cognitive sciences*.

Jiahui Geng, Fengyu Cai, Yuxia Wang, Heinz Koeppl, Preslav Nakov, and Iryna Gurevych. 2024. A survey of confidence estimation and calibration in large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6577–6595.

Moumita Ghosh, Samhita Das, and Pritha Das. 2022. Dynamics and control of delayed rumor propagation through social networks. *Journal of Applied Mathematics and Computing*, pages 1–30.

David Glukhov, Ilia Shumailov, Yarin Gal, Nicolas Papernot, and Vardan Papyan. 2023. Llm censorship: A machine learning challenge or a computer security problem? *arXiv preprint arXiv:2307.10719*.

Oana Goga, Howard Lei, Sree Hari Krishnan Parthasarathi, Gerald Friedland, Robin Sommer, and Renata Teixeira. 2013. Exploiting innocuous activity for correlating users across sites. In *Proceedings of the 22nd international conference on World Wide Web*, pages 447–458.

Jian Guan, Jesse Dodge, David Wadden, Minlie Huang, and Hao Peng. 2023. Language models hallucinate, but may excel at fact verification. *arXiv preprint arXiv:2310.14564*.

David Güera and Edward J Delp. 2018. Deepfake video detection using recurrent neural networks. In *2018 15th IEEE international conference on advanced video and signal based surveillance (AVSS)*, pages 1–6. IEEE.

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.

Philipp Hacker. 2024. Sustainable ai regulation. *Common Market Law Review*, 61(2).

Philipp Hacker, Andreas Engel, and Marco Mauer. 2023. Regulating chatgpt and other large generative ai models. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 1112–1123.

Thilo Hagendorff. 2024. Deception abilities emerged in large language models. *Proceedings of the National Academy of Sciences*, 121(24):e2317967121.

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.

Patrick Haller, Ansar Aynetdinov, and Alan Akbik. 2023. Opiniongpt: Modelling explicit biases in instruction-tuned llms. *arXiv preprint arXiv:2309.03876*.

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.

Ahmed Abdeen Hamed. 2023. Improving detection of chatgpt-generated fake science using real publication text: Introducing xfakebibs a supervised learning network algorithm.

Kateřina Haniková, David Chudán, Vojtěch Svátek, Peter Vajdečka, Raphaël Troncy, Filip Vencovskỳ, and Jana Syrovátková. 2024. Towards fact-check summarization leveraging on argumentation elements tied to entity graphs. In *Companion Proceedings of the ACM on Web Conference 2024*, pages 1473–1481.

Hans WA Hanley and Zakir Durumeric. 2024. Machine-made media: Monitoring the mobilization of machine-generated articles on misinformation and mainstream news websites. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 18, pages 542–556.

Andreas Hanselowski, Christian Stab, Claudia Schulz, Zile Li, and Iryna Gurevych. 2019. A richly annotated corpus for different tasks in automated fact-checking. *arXiv preprint arXiv:1911.01214*.

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.

Bing He, Mustaque Ahamad, and Srijan Kumar. 2023. Reinforcement learning-based counter-misinformation response generation: a case study of covid-19 vaccine misinformation. In *Proceedings of the ACM Web Conference 2023*, pages 2698–2709.

Zhenyu He, Ce Li, Fan Zhou, and Yi Yang. 2021. Rumor detection on social media with event augmentations. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, pages 2020–2024.

Da Silva Gameiro Henrique, Andrei Kucharavy, and Rachid Guerraoui. 2023. Stochastic parrots looking for stochastic parrots: Llms are easy to fine-tune and hard to detect with other llms. *arXiv preprint arXiv:2304.08968*.

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.

Adil Imad Eddine Hosni, Kan Li, Cangfeng Ding, and Sadique Ahmed. 2018. Least cost rumor influence minimization in multiplex social networks. In *International Conference on Neural Information Processing*, pages 93–105. Springer.

Saghar Hosseini, Hamid Palangi, and Ahmed Hassan Awadallah. 2023. An empirical study of metrics to measure representational harms in pre-trained language models. *arXiv preprint arXiv:2301.09211*.

Dongpeng Hou, Shu Yin, Chao Gao, Xianghua Li, and Zhen Wang. 2024. Propagation dynamics of rumor vs. non-rumor across multiple social media platforms driven by user characteristics. *arXiv preprint arXiv:2401.17840*.

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.

Nan Hu, Zirui Wu, Yuxuan Lai, Xiao Liu, and Yansong Feng. 2022. Dual-channel evidence fusion for fact verification over texts and tables. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5232–5242.

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*.

Di Huang, Jinbao Song, and Xingyu Zhang. 2025a. Semi-supervised social bot detection with relational graph attention transformers and characteristics of the social environment. *Information Fusion*, page 102956.

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. 2023. Harnessing the power of chatgpt in fake news: An in-depth exploration in generation, detection and explanation. *arXiv preprint arXiv:2310.05046*.

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

Yuheng Huang, Jiayang Song, Zhijie Wang, Shengming Zhao, Huaming Chen, Felix Juefei-Xu, and Lei Ma. 2025b. Look before you leap: An exploratory study of uncertainty analysis for large language models. *IEEE Transactions on Software Engineering*.

Minyoung Huh, Andrew Liu, Andrew Owens, and Alexei A Efros. 2018. Fighting fake news: Image splice detection via learned self-consistency. In *Proceedings of the European conference on computer vision (ECCV)*, pages 101–117.

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.

Tereza Iofciu, Peter Fankhauser, Fabian Abel, and Kerstin Bischoff. 2011. Identifying users across social tagging systems. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 5, pages 522–525.

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.

Kathleen H Jamieson. 2008. *Echo chamber: Rush Limbaugh and the conservative media establishment*. Oxford University Press.

Danish Javed, NZ Jhanjhi, Navid Ali Khan, Sayan Kumar Ray, Alanoud Al Mazroa, Farzeen Ashfaq, and Shampa Rani Das. 2024. Towards the future of bot detection: A comprehensive taxonomical review and challenges on twitter/x. *Computer Networks*, 254:110808.

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*.

Zhiwei Jin, Juan Cao, Han Guo, Yongdong Zhang, and Jiebo Luo. 2017. Multimodal fusion with recurrent neural networks for rumor detection on microblogs. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 795–816.

Zhiwei Jin, Juan Cao, Yongdong Zhang, and Jiebo Luo. 2016. News verification by exploiting conflicting social viewpoints in microblogs. In *Proceedings of the AAAI conference on artificial intelligence*, volume 30.

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.

Anna Jobin, Marcello Ienca, and Effy Vayena. 2019. The global landscape of ai ethics guidelines. *Nature machine intelligence*, 1(9):389–399.

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.

Daniel Kahneman. 1979. Prospect theory: An analysis of decisions under risk. *Econometrica*, 47:278.

Rohit Kumar Kaliyar, Anurag Goswami, and Pratik Narang. 2021. Fakebert: Fake news detection in social media with a bert-based deep learning approach. *Multimedia tools and applications*, 80(8):11765–11788.

Mohammadamin Kanaani. 2024. Triple-r: Automatic reasoning for fact verification using language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 16831–16840.

Hamid Karimi and Jiliang Tang. 2019. Learning hierarchical discourse-level structure for fake news detection. *arXiv preprint arXiv:1903.07389*.

Elihu Katz. 1957. The two-step flow of communication: An up-to-date report on an hypothesis. *Public opinion quarterly*, 21(1):61–78.

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.

Nida Saddaf Khan, Maira Ata, and Quratulain Rajput. 2015. Identification of opinion leaders in social network. In *2015 International Conference on Information and Communication Technologies (ICICT)*, pages 1–6. IEEE.

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.

Dhruv Khattar, Jaipal Singh Goud, Manish Gupta, and Vasudeva Varma. 2019. Mvae: Multimodal variational autoencoder for fake news detection. In *The world wide web conference*, pages 2915–2921.

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.

Camille Koenders, Johannes Filla, Nicolai Schneider, and Vinicius Woloszyn. 2021. How vulnerable are automatic fake news detection methods to adversarial attacks? *arXiv preprint arXiv:2107.07970*.

Ziyi Kou, Lanyu Shang, Yang Zhang, Zhenrui Yue, Huimin Zeng, and Dong Wang. 2022. Crowd, expert & ai: A human-ai interactive approach towards natural language explanation based covid-19 misinformation detection. In *IJCAI*, pages 5087–5093.

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.

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.

Raghvendra Kumar, Bhargav Goddu, Sriparna Saha, and Adam Jatowt. 2025. Silver lining in the fake news cloud: Can large language models help detect misinformation? *IEEE Transactions on Artificial Intelligence*, 6(1):14–24.

Tharindu Kumarage and Huan Liu. 2023. Neural authorship attribution: Stylometric analysis on large language models. In *2023 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC)*, pages 51–54. IEEE.

Sejeong Kwon, Meeyoung Cha, Kyomin Jung, Wei Chen, and Yajun Wang. 2013. Prominent features of rumor propagation in online social media. In *2013 IEEE 13th international conference on data mining*, pages 1103–1108. IEEE.

An Lao, Qi Zhang, Chongyang Shi, Longbing Cao, Kun Yi, Liang Hu, and Duoqian Miao. 2024. Frequency spectrum is more effective for multimodal representation and fusion: A multimodal spectrum rumor detector. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18426–18434.

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.

Bingxin Li and Linhe Zhu. 2024. Turing instability analysis of a reaction–diffusion system for rumor propagation in continuous space and complex networks. *Information Processing & Management*, 61(3):103621.

Bo Li, Peng Qi, Bo Liu, Shuai Di, Jingen Liu, Jiquan Pei, Jinfeng Yi, and Bowen Zhou. 2023a. Trustworthy ai: From principles to practices. *ACM Computing Surveys*, 55(9):1–46.

Chen Li, Hao Peng, Jianxin Li, Lichao Sun, Lingjuan Lyu, Lihong Wang, S Yu Philip, and Lifang He. 2021. Joint stance and rumor detection in hierarchical heterogeneous graph. *IEEE transactions on neural networks and learning systems*, 33(6):2530–2542.

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

Jiarong Li, Haijun Jiang, Xuehui Mei, Cheng Hu, and Guoliang Zhang. 2020. Dynamical analysis of rumor spreading model in multi-lingual environment and heterogeneous complex networks. *Information Sciences*, 536:391–408.

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*.

Sha Li, Ruining Zhao, Manling Li, Heng Ji, Chris Callison-Burch, and Jiawei Han. 2023d. Open-domain hierarchical event schema induction by incremental prompting and verification. *arXiv preprint arXiv:2307.01972*.

Weimin Li, Xiaokang Zhou, Chao Yang, Yuting Fan, Zhao Wang, and Yanxia Liu. 2022. Multi-objective optimization algorithm based on characteristics fusion of dynamic social networks for community discovery. *Information Fusion*, 79:110–123.

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

Yuezun Li, Ming-Ching Chang, and Siwei Lyu. 2018. In ictu oculi: Exposing ai created fake videos by detecting eye blinking. In *2018 IEEE International workshop on information forensics and security (WIFS)*, pages 1–7. Ieee.

Yupeng Li, Haorui He, Jin Bai, and Dacheng Wen. 2024b. Mcfend: a multi-source benchmark dataset for chinese fake news detection. In *Proceedings of the ACM on Web Conference 2024*, pages 4018–4027.

Zhuohang Li, Jiaxin Zhang, Chao Yan, Kamalika Das, Sricharan Kumar, Murat Kantarcioglu, and Bradley A Malin. 2024c. Do you know what you are talking about? characterizing query-knowledge relevance for reliable retrieval augmented generation. *arXiv preprint arXiv:2410.08320*.

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.

Wee Yong Lim, Mong Li Lee, and Wynne Hsu. 2017. Ifact: An interactive framework to assess claims from tweets. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 787–796.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Teaching models to express their uncertainty in words. *arXiv preprint arXiv:2205.14334*.

Chen Ling, Junji Jiang, Junxiang Wang, and Zhao Liang. 2022. Source localization of graph diffusion via variational autoencoders for graph inverse problems. In *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining*, pages 1010–1020.

Aiwei Liu, Qiang Sheng, and Xuming Hu. 2024a. Preventing and detecting misinformation generated by large language models. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3001–3004.

Hui Liu, Wenya Wang, Haoru Li, and Haoliang Li. 2024b. Teller: A trustworthy framework for explainable, generalizable and controllable fake news detection. *arXiv preprint arXiv:2402.07776*.

Xuannan Liu, Peipei Li, Huaibo Huang, Zekun Li, Xing Cui, Jiahao Liang, Lixiong Qin, Weihong Deng, and Zhaofeng He. 2024c. Fka-owl: Advancing multimodal fake news detection through knowledge-augmented lvlms. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 10154–10163.

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

Alireza Louni and KP Subbalakshmi. 2018. Who spread that rumor: Finding the source of information in large online social networks with probabilistically varying internode relationship strengths. *IEEE transactions on computational social systems*, 5(2):335–343.

Yi-Ju Lu and Cheng-Te Li. 2020. Gcan: Graph-aware co-attention networks for explainable fake news detection on social media. *arXiv preprint arXiv:2004.11648*.

Qing Lyu, Shreya Havaldar, Adam Stein, Li Zhang, Delip Rao, Eric Wong, Marianna Apidianaki, and Chris Callison-Burch. 2023. Faithful chain-of-thought reasoning. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 305–329, Nusa Dua, Bali. Association for Computational Linguistics.

Jing Ma, Wei Gao, Shafiq Joty, and Kam-Fai Wong. 2019. Sentence-level evidence embedding for claim verification with hierarchical attention networks. Association for Computational Linguistics.

Jing Ma, Wei Gao, Prasenjit Mitra, Sejeong Kwon, Bernard J Jansen, Kam-Fai Wong, and Meeyoung Cha. 2016. Detecting rumors from microblogs with recurrent neural networks.

Jing Ma, Wei Gao, Zhongyu Wei, Yueming Lu, and Kam-Fai Wong. 2015. Detect rumors using time series of social context information on microblogging websites. In *Proceedings of the 24th ACM international on conference on information and knowledge management*, pages 1751–1754.

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

Zihan Ma, Minnan Luo, Hao Guo, Zhi Zeng, Yiran Hao, and Xiang Zhao. 2024. Event-radar: Event-driven multi-view learning for multimodal fake news detection. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5809–5821.

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*.

Shrikant Malviya and Stamos Katsigiannis. 2024. Evidence retrieval for fact verification using multi-stage reranking. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7295–7308, Miami, Florida, USA. Association for Computational Linguistics.

Francesco Marra, Diego Gragnaniello, Luisa Verdoliva, and Giovanni Poggi. 2019. Do gans leave artificial fingerprints? In *2019 IEEE conference on multimedia information processing and retrieval (MIPR)*, pages 506–511. IEEE.

David Martín-Gutiérrez, Gustavo Hernández-Peñaloza, Alberto Belmonte Hernández, Alicia Lozano-Diez, and Federico Álvarez. 2021. A deep learning approach for robust detection of bots in twitter using transformers. *IEEE Access*, 9:54591–54601.

Falko Matern, Christian Riess, and Marc Stamminger. 2019. Exploiting visual artifacts to expose deepfakes and face manipulations. In *2019 IEEE Winter Applications of Computer Vision Workshops (WACVW)*, pages 83–92. IEEE.

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.

Mohit Mayank, Shakshi Sharma, and Rajesh Sharma. 2022. Deap-faked: Knowledge graph based approach for fake news detection. In *2022 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 47–51. IEEE.

Nikhil Mehta, María Leonor Pacheco, and Dan Goldwasser. 2022. Tackling fake news detection by continually improving social context representations using graph neural networks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1363–1380.

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.

Daniel Mas Montserrat, Hanxiang Hao, Sri K Yarlagadda, Sriram Baireddy, Ruiting Shao, János Horváth, Emily Bartusiak, Justin Yang, David Guera, Fengqing Zhu, et al. 2020. Deepfakes detection with automatic face weighting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 668–669.

Xuemei Mou, Wei Xu, Yangfu Zhu, Qian Li, and Yunpeng Xiao. 2022. A social topic diffusion model based on rumor, anti-rumor, and motivation-rumor. *IEEE Transactions on Computational Social Systems*, 10(5):2644–2659.

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.

Mark EJ Newman. 2004. Fast algorithm for detecting community structure in networks. *Physical Review E—Statistical, Nonlinear, and Soft Matter Physics*, 69(6):066133.

Lynnette Hui Xian Ng and Kathleen M. Carley. 2022. Online coordination: Methods and comparative case studies of coordinated groups across four events in the united states. In *Proceedings of the 14th ACM Web Science Conference 2022*, WebSci '22, page 12–21, New York, NY, USA. Association for Computing Machinery.

Van-Hoang Nguyen, Kazunari Sugiyama, Preslav Nakov, and Min-Yen Kan. 2020. Fang: Leveraging social context for fake news detection using graph representation. In *Proceedings of the 29th ACM international conference on information & knowledge management*, pages 1165–1174.

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.

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*.

Cheng Niu, Yuanhao Wu, Juno Zhu, Siliang Xu, Kashun Shum, Randy Zhong, Juntong Song, and Tong Zhang. 2023. Ragtruth: A hallucination corpus for developing trustworthy retrieval-augmented language models. *arXiv preprint arXiv:2401.00396*.

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

Tore Opsahl, Filip Agneessens, and John Skvoretz. 2010. Node centrality in weighted networks: Generalizing degree and shortest paths. *Social networks*, 32(3):245–251.

Liangming Pan, Xiaobao Wu, Xinyuan Lu, Anh Tuan Luu, William Yang Wang, Min-Yen Kan, and Preslav Nakov. 2023a. Fact-checking complex claims with program-guided reasoning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6981–7004, Toronto, Canada. Association for Computational Linguistics.

Yikang Pan, Liangming Pan, Wenhu Chen, Preslav Nakov, Min-Yen Kan, and William Wang. 2023b. 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. 2023c. 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.

Kellin Pelrine, Jacob Danovitch, and Reihaneh Rabbany. 2021. The surprising performance of simple baselines for misinformation detection. In *Proceedings of the Web Conference 2021*, pages 3432–3441.

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*.

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*.

Verónica Pérez-Rosas, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea. 2017. Automatic detection of fake news. *arXiv preprint arXiv:1708.07104*.

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.

Moritz Pilarski, Kirill Olegovich Solovev, and Nicolas Pröllochs. 2024. Community notes vs. snoping: How the crowd selects fact-checking targets on social media. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 18, pages 1262–1275.

Sanjaikanth E Vadakkethil Somanathan Pillai. 2024. Enhancing misinformation detection through semantic analysis and knowledge graphs. In *2024 4th International Conference on Data Engineering and Communication Systems (ICDECS)*, pages 1–5. IEEE.

Clara Pizzuti. 2017. Evolutionary computation for community detection in networks: A review. *IEEE Transactions on Evolutionary Computation*, 22(3):464–483.

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

Pascal Pons and Matthieu Latapy. 2005. Computing communities in large networks using random walks. In *Computer and Information Sciences-ISCIS 2005: 20th International Symposium, Istanbul, Turkey, October 26-28, 2005. Proceedings 20*, pages 284–293. Springer.

Alin C Popescu and Hany Farid. 2004. Exposing digital forgeries by detecting duplicated image regions.

Ethan Porter and Thomas J Wood. 2021. The global effectiveness of fact-checking: Evidence from simultaneous experiments in argentina, nigeria, south africa, and the united kingdom. *Proceedings of the National Academy of Sciences*, 118(37):e2104235118.

Martin Potthast, Johannes Kiesel, Kevin Reinartz, Janek Bevendorff, and Benno Stein. 2017. A stylometric inquiry into hyperpartisan and fake news. *arXiv preprint arXiv:1702.05638*.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A Smith, and Mike Lewis. 2022. Measuring and narrowing the compositionality gap in language models. *arXiv preprint arXiv:2210.03350*.

Nicolas Pröllochs, Dominik Bär, and Stefan Feuerriegel. 2021. Emotions in online rumor diffusion. *EPJ Data Science*, 10(1):51.

Peng Qi, Zehong Yan, Wynne Hsu, and Mong Li Lee. 2024. Sniffer: Multimodal large language model for explainable out-of-context misinformation detection. *Preprint*, arXiv:2403.03170.

Xiangyu Qi, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal, and Peter Henderson. 2023. Fine-tuning aligned language models compromises safety, even when users do not intend to! *Preprint*, arXiv:2310.03693.

Shengsheng Qian, Jinguang Wang, Jun Hu, Quan Fang, and Changsheng Xu. 2021. Hierarchical multi-modal contextual attention network for fake news detection. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, pages 153–162.

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.

Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. 2023. In-context retrieval-augmented language models. *Transactions of the Association for Computational Linguistics*, 11:1316–1331.

Yuan Rao and Jiangqun Ni. 2016. A deep learning approach to detection of splicing and copy-move forgeries in images. In *2016 IEEE international workshop on information forensics and security (WIFS)*, pages 1–6. IEEE.

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.

Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. 2017. Truth of varying shades: Analyzing language in fake news and political fact-checking. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pages 2931–2937.

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

Shebuti Rayana and Leman Akoglu. 2015. Collective opinion spam detection: Bridging review networks and metadata. In *Proceedings of the 21th acm sigkdd international conference on knowledge discovery and data mining*, pages 985–994.

Veronica Red, Eric D Kelsic, Peter J Mucha, and Mason A Porter. 2011. Comparing community structure to characteristics in online collegiate social networks. *SIAM review*, 53(3):526–543.

Zhicheng Ren, Zhiping Xiao, and Yizhou Sun. 2024. Do we trust what they say or what they do? a multimodal user embedding provides personalized explanations. *arXiv preprint arXiv:2409.02965*.

Christopher Riederer, Yunsung Kim, Augustin Chaintreau, Nitish Korula, and Silvio Lattanzi. 2016. Linking users across domains with location data: Theory and validation. In *Proceedings of the 25th international conference on world wide web*, pages 707–719.

Hamid Roghani, Asgarali Bouyer, and Esmaeil Nourani. 2021. Pldls: A novel parallel label diffusion and label selection-based community detection algorithm based on spark in social networks. *Expert Systems with Applications*, 183:115377.

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.

Martin Rosvall and Carl T Bergstrom. 2008. Maps of random walks on complex networks reveal community structure. *Proceedings of the national academy of sciences*, 105(4):1118–1123.

Shailendra Sahu and T Sobha Rani. 2022. A neighboursimilarity based community discovery algorithm. *Expert Systems with Applications*, 206:117822.

Chiman Salavati, Alireza Abdollahpouri, and Zhaleh Manbari. 2019. Ranking nodes in complex networks based on local structure and improving closeness centrality. *Neurocomputing*, 336:36–45.

Amine Sallah, Said Agoujil, Mudasir Ahmad Wani, Mohamed Hammad, Ahmed A Abd El-Latif, Yassine Maleh, et al. 2024. Fine-tuned understanding: Enhancing social bot detection with transformer-based classification. *IEEE Access*.

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. 2023. The liar's dividend: Can politicians claim misinformation to evade accountability? *American Political Science Review*, pages 1–20.

Tal Schuster, Adam Fisch, and Regina Barzilay. 2021. Get your vitamin c! robust fact verification with contrastive evidence. *arXiv preprint arXiv:2103.08541*.

M Zubair Shafiq, Muhammad U Ilyas, Alex X Liu, and Hayder Radha. 2013. Identifying leaders and followers in online social networks. *IEEE Journal on Selected Areas in Communications*, 31(9):618–628.

Chintan Shah, Nima Dehmamy, Nicola Perra, Matteo Chinazzi, Albert-László Barabási, Alessandro Vespignani, and Rose Yu. 2020. Finding patient zero: Learning contagion source with graph neural networks. *arXiv preprint arXiv:2006.11913*.

Chengcheng Shao, Giovanni Luca Ciampaglia, Alessandro Flammini, and Filippo Menczer. 2016. Hoaxy: A platform for tracking online misinformation. In *Proceedings of the 25th international conference companion on world wide web*, pages 745–750.

Chengcheng Shao, Giovanni Luca Ciampaglia, Onur Varol, Kai-Cheng Yang, Alessandro Flammini, and Filippo Menczer. 2018. The spread of low-credibility content by social bots. *Nature communications*, 9(1):1–9.

Junming Shao, Zhichao Han, Qinli Yang, and Tao Zhou. 2015. Community detection based on distance dynamics. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, page 1075–1084, New York, NY, USA. Association for Computing Machinery.

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, Deepak Mahudeswaran, Suhang Wang, and Huan Liu. 2020. Hierarchical propagation networks for fake news detection: Investigation and exploitation. In *Proceedings of the international AAAI conference on web and social media*, volume 14, pages 626–637.

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.

Ronit Singhal, Pransh Patwa, Parth Patwa, Aman Chadha, and Amitava Das. 2024. Evidence-backed fact checking using rag and few-shot in-context learning with llms. *arXiv preprint arXiv:2408.12060*.

Shivangi Singhal, Rajiv Ratn Shah, Tanmoy Chakraborty, Ponnurangam Kumaraguru, and Shin'ichi Satoh. 2019. Spotfake: A multi-modal framework for fake news detection. In *2019 IEEE fifth international conference on multimedia big data (BigMM)*, pages 39–47. IEEE.

Xing Su, Shan Xue, Fanzhen Liu, Jia Wu, Jian Yang, Chuan Zhou, Wenbin Hu, Cecile Paris, Surya Nepal, Di Jin, et al. 2022. A comprehensive survey on community detection with deep learning. *IEEE Transactions on Neural Networks and Learning Systems*.

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.

Shulong Tan, Ziyu Guan, Deng Cai, Xuzhen Qin, Jiajun Bu, and Chun Chen. 2014. Mapping users across networks by manifold alignment on hypergraph. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 28.

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.

Zhen Tan, Chengshuai Zhao, Raha Moraffah, Yifan Li, Song Wang, Jundong Li, Tianlong Chen, and Huan Liu. 2024. Glue pizza and eat rocks - exploiting vulnerabilities in retrieval-augmented generative models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1610–1626, Miami, Florida, USA. Association for Computational Linguistics.

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.

Guangmo Tong, Weili Wu, and Ding-Zhu Du. 2018. Distributed rumor blocking with multiple positive cascades. *IEEE Transactions on Computational Social Systems*, 5(2):468–480.

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*.

Rudra M Tripathy, Amitabha Bagchi, and Sameep Mehta. 2010. A study of rumor control strategies on social networks. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 1817–1820.

U Undeutsch. 1967. Beurteilung der glaubhaftigkeit von aussagen [veracity assessment of statements]. undeutsch. *Handbuch der Psychologie*, 11:26–181.

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.

Karthik Valmeekam, Matthew Marquez, Alberto Olmo, Sarath Sreedharan, and Subbarao Kambhampati. 2023. Planbench: An extensible benchmark for evaluating large language models on planning and reasoning about change. *Advances in Neural Information Processing Systems*, 36:38975–38987.

Neeraj Varshney, Wenlin Yao, Hongming Zhang, Jianshu Chen, and Dong Yu. 2023. A stitch in time saves nine: Detecting and mitigating hallucinations of llms by validating low-confidence generation. *arXiv preprint arXiv:2307.03987*.

Nikhita Vedula and Srinivasan Parthasarathy. 2021. Face-keg: Fact checking explained using knowledge graphs. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, pages 526–534.

Andreas Vlachos and Sebastian Riedel. 2014. Fact checking: Task definition and dataset construction. In *Proceedings of the ACL 2014 workshop on language technologies and computational social science*, pages 18–22.

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

Ivan Vykopal, Matúš Pikuliak, Ivan Srba, Robert Moro, Dominik Macko, and Maria Bielikova. 2023. Disinformation capabilities of large language models. *arXiv preprint arXiv:2311.08838*.

David Wadden, Kyle Lo, Lucy Lu Wang, Arman Cohan, Iz Beltagy, and Hannaneh Hajishirzi. 2021. Multivers: Improving scientific claim verification with weak supervision and full-document context. *arXiv preprint arXiv:2112.01640*.

Nathan Walter and Riva Tukachinsky. 2020. A meta-analytic examination of the continued influence of misinformation in the face of correction: How powerful is it, why does it happen, and how to stop it? *Communication research*, 47(2):155–177.

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*.

Biao Wang, Ge Chen, Luoyi Fu, Li Song, and Xinbing Wang. 2017. Drimux: Dynamic rumor influence minimization with user experience in social networks. *IEEE Transactions on Knowledge and Data Engineering*, 29(10):2168–2181.

Bo Wang, Jing Ma, Hongzhan Lin, Zhiwei Yang, Ruichao Yang, Yuan Tian, and Yi Chang. 2024a. Explainable fake news detection with large language model via defense among competing wisdom. In *Proceedings of the ACM on Web Conference 2024*, pages 2452–2463.

Bo Wang, Jing Ma, Hongzhan Lin, Zhiwei Yang, Ruichao Yang, Yuan Tian, and Yi Chang. 2024b. Explainable fake news detection with large language model via defense among competing wisdom. In *Proceedings of the ACM Web Conference 2024*, WWW '24, page 2452–2463, New York, NY, USA. Association for Computing Machinery.

Chenxi Wang, Zongfang Liu, Dequan Yang, and Xiuying Chen. 2024c. 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.

Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A Efros. 2020a. Cnn-generated images are surprisingly easy to spot... for now. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8695–8704.

Yaqing Wang, Fenglong Ma, Zhiwei Jin, Ye Yuan, Guangxu Xun, Kishlay Jha, Lu Su, and Jing Gao. 2018a. Eann: Event adversarial neural networks for multi-modal fake news detection. In *Proceedings of the 24th acm sigkdd international conference on knowledge discovery & data mining*, pages 849–857.

Yi Wang, Jinde Cao, Xiaodi Li, and Ahmed Alsaedi. 2018b. Edge-based epidemic dynamics with multiple routes of transmission on random networks. *Nonlinear Dynamics*, 91:403–420.

Youze Wang, Shengsheng Qian, Jun Hu, Quan Fang, and Changsheng Xu. 2020b. Fake news detection via knowledge-driven multimodal graph convolutional networks. In *Proceedings of the 2020 international conference on multimedia retrieval*, pages 540–547.

Zecong Wang, Jiaxi Cheng, Chen Cui, and Chenhao Yu. 2023. Implementing bert and fine-tuned roberta to detect ai generated news by chatgpt. *arXiv preprint arXiv:2306.07401*.

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

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

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

Teresa Weikmann and Sophie Lecheler. 2023. Visual disinformation in a digital age: A literature synthesis and research agenda. *New Media & Society*, 25(12):3696–3713.

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. 2023a. 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.

Jun Wu, Xuesong Ye, and Chengjie Mou. 2023b. Botshape: A novel social bots detection approach via behavioral patterns. *arXiv preprint arXiv:2303.10214*.

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.

Siyuan Wu, Yue Huang, Chujie Gao, Dongping Chen, Qihui Zhang, Yao Wan, Tianyi Zhou, Xiangliang Zhang, Jianfeng Gao, Chaowei Xiao, et al. 2024c. Unigen: A unified framework for textual dataset generation using large language models. *arXiv preprint arXiv:2406.18966*.

Yang Xia, Haijun Jiang, Shuzhen Yu, and Zhiyong Yu. 2024. The dynamic analysis of the rumor spreading and behavior diffusion model with higher-order interactions. *Communications in Nonlinear Science and Numerical Simulation*, 138:108186.

Bin Xiao, Yang Wei, Xiuli Bi, Weisheng Li, and Jianfeng Ma. 2020. Image splicing forgery detection combining coarse to refined convolutional neural network and adaptive clustering. *Information Sciences*, 511:172–191.

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.

Yunpeng Xiao, Jinsong Yang, Wanjing Zhao, Qian Li, and Yucai Pang. 2024. Cross-domain social rumor-propagation model based on transfer learning. *IEEE Transactions on Neural Networks and Learning Systems*.

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*.

Jierui Xie, Stephen Kelley, and Boleslaw K Szymanski. 2013. Overlapping community detection in networks: The state-of-the-art and comparative study. *Acm computing surveys (csur)*, 45(4):1–35.

Yueqi Xie, Jingwei Yi, Jiawei Shao, Justin Curl, Lingjuan Lyu, Qifeng Chen, Xing Xie, and Fangzhao Wu. 2023. Defending chatgpt against jailbreak attack via self-reminders. *Nature Machine Intelligence*, 5(12):1486–1496.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. 2023. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. *arXiv preprint arXiv:2306.13063*.

Junhao Xu, Longdi Xian, Zening Liu, Mingliang Chen, Qiuyang Yin, and Fenghua Song. 2024a. The future of combating rumors? retrieval, discrimination, and generation. *arXiv preprint arXiv:2403.20204*.

Rongwu Xu, Brian Lin, Shujian Yang, Tianqi Zhang, Weiyan Shi, Tianwei Zhang, Zhixuan Fang, Wei Xu, and Han Qiu. 2024b. The earth is flat because...: Investigating LLMs' belief towards misinformation via persuasive conversation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16259–16303, Bangkok, Thailand. Association for Computational Linguistics.

Rongwu Xu, Brian S Lin, Shujian Yang, Tianqi Zhang, Weiyan Shi, Tianwei Zhang, Zhixuan Fang, Wei Xu, and Han Qiu. 2023. The earth is flat because...: Investigating llms' belief towards misinformation via persuasive conversation. *arXiv preprint arXiv:2312.09085*.

Ziwei Xu, Sanjay Jain, and Mohan Kankanhalli. 2024c. Hallucination is inevitable: An innate limitation of large language models. *arXiv preprint arXiv:2401.11817*.

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, Xia Yu, JiaYi Wu, BoZhen Zhang, and HaiBo Yang. 2024a. Graph-aware multi-feature interacting network for explainable rumor detection on social network. *Expert Systems with Applications*, 249:123687.

Chang Yang, Peng Zhang, Hui Gao, and Jing Zhang. 2024b. Deciphering rumors: A multi-task learning approach with intent-aware hierarchical contrastive learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4471–4483.

Chang Yang, Peng Zhang, Wenbo Qiao, Hui Gao, and Jiaming Zhao. 2023a. 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.

Fan Yang, Yang Liu, Xiaohui Yu, and Min Yang. 2012. Automatic detection of rumor on sina weibo. In *Proceedings of the ACM SIGKDD Workshop on Mining Data Semantics*, MDS '12, New York, NY, USA. Association for Computing Machinery.

Fan Yang, Shiva K Pentyala, Sina Mohseni, Mengnan Du, Hao Yuan, Rhema Linder, Eric D Ragan, Shuiwang Ji, and Xia Hu. 2019. Xfake: Explainable fake news detector with visualizations. In *The world wide web conference*, pages 3600–3604.

Jaewon Yang and Jure Leskovec. 2013. Overlapping community detection at scale: a nonnegative matrix factorization approach. In *Proceedings of the sixth ACM international conference on Web search and data mining*, pages 587–596.

Kaijun Yang, Qing Bao, and Hongjun Qiu. 2023b. Identifying multiple propagation sources with motif-based graph convolutional networks for social networks. *IEEE Access*, 11:61630–61645.

Li Yang, Yafeng Qiao, Zhihong Liu, Jianfeng Ma, and Xinghua Li. 2018. Identifying opinion leader nodes in online social networks with a new closeness evaluation algorithm. *Soft Computing*, 22:453–464.

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.

Ruichao Yang, Jing Ma, Wei Gao, and Hongzhan Lin. 2025. Llm-enhanced multiple instance learning for joint rumor and stance detection with social context information. *ACM Transactions on Intelligent Systems and Technology*.

Tianbao Yang, Rong Jin, Yun Chi, and Shenghuo Zhu. 2009. Combining link and content for community detection: a discriminative approach. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 927–936.

Xianjun Yang, Xiao Wang, Qi Zhang, Linda Petzold, William Yang Wang, Xun Zhao, and Dahua Lin. 2023c. Shadow alignment: The ease of subverting safely-aligned language models. *arXiv preprint arXiv:2310.02949*.

Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. 2023d. The dawn of lmms: Preliminary explorations with gpt-4v (ision). *arXiv preprint arXiv:2309.17421*, 9(1):1.

Barry Menglong Yao, Aditya Shah, Lichao Sun, Jin-Hee Cho, and Lifu Huang. 2023a. 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.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2023b. Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems*, 36:11809–11822.

Xiaopeng Yao, Yue Gu, Chonglin Gu, and Hejiao Huang. 2022. Fast controlling of rumors with limited cost in social networks. *Computer Communications*, 182:41–51.

Yining Ye, Xin Cong, Shizuo Tian, Jiannan Cao, Hao Wang, Yujia Qin, Yaxi Lu, Heyang Yu, Huadong Wang, Yankai Lin, et al. 2023. Proagent: From robotic process automation to agentic process automation. *arXiv preprint arXiv:2311.10751*.

Shuzhen Yu, Zhiyong Yu, Haijun Jiang, and Shuai Yang. 2021. The dynamics and control of 2i2sr rumor spreading models in multilingual online social networks. *Information Sciences*, 581:18–41.

Damián H Zanette. 2001. Critical behavior of propagation on small-world networks. *Physical Review E*, 64(5):050901.

Runxi Zeng and Di Zhu. 2019. A model and simulation of the emotional contagion of netizens in the process of rumor refutation. *Scientific reports*, 9(1):14164.

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.

Litian Zhang, Xiaoming Zhang, Ziyi Zhou, Feiran Huang, and Chaozhuo Li. 2024a. Reinforced adaptive knowledge learning for multimodal fake news detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 16777–16785.

Litian Zhang, Xiaoming Zhang, Ziyi Zhou, Xi Zhang, S Yu Philip, and Chaozhuo Li. 2025. Knowledge-aware multimodal pre-training for fake news detection. *Information Fusion*, 114:102715.

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

Qi Zhang, Yuan Li, Jialing Zou, Jianming Zhu, Dingning Liu, and Jianbin Jiao. 2024c. Unifying multi-modal interactions for rumor diffusion prediction with global hypergraph modeling. *Knowledge-Based Systems*, 301:112246.

Si Zhang and Hanghang Tong. 2016. Final: Fast attributed network alignment. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1345–1354.

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

Xuan Zhang and Wei Gao. 2023b. Towards LLM-based fact verification on news claims with a hierarchical step-by-step prompting method. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 996–1011, Nusa Dua, Bali. Association for Computational Linguistics.

Yao Zhang and B Aditya Prakash. 2015. Data-aware vaccine allocation over large networks. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 10(2):1–32.

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.

Jian Zhao, Nan Cao, Zhen Wen, Yale Song, Yu-Ru Lin, and Christopher Collins. 2014. # fluxflow: Visual analysis of anomalous information spreading on social media. *IEEE transactions on visualization and computer graphics*, 20(12):1773–1782.

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.

Ruochen Zhao, Hailin Chen, Weishi Wang, Fangkai Jiao, Xuan Long Do, Chengwei Qin, Bosheng Ding, Xiaobao Guo, Minzhi Li, Xingxuan Li, et al. 2023a. Retrieving multimodal information for augmented generation: A survey. *arXiv preprint arXiv:2303.10868*.

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

Yiran Zhao, Jinghan Zhang, I Chern, Siyang Gao, Pengfei Liu, Junxian He, et al. 2023c. Felm: Benchmarking factuality evaluation of large language models. *Advances in Neural Information Processing Systems*, 36:44502–44523.

Yuxin Zhao, Shenghong Li, and Feng Jin. 2016. Identification of influential nodes in social networks with community structure based on label propagation. *Neurocomputing*, 210:34–44.

Wanjun Zhong, Jingjing Xu, Duyu Tang, Zenan Xu, Nan Duan, Ming Zhou, Jiahai Wang, and Jian Yin. 2019. Reasoning over semantic-level graph for fact checking. *arXiv preprint arXiv:1909.03745*.

Jiawei Zhou, Yixuan Zhang, Qianni Luo, Andrea G Parker, and Munmun De Choudhury. 2023. 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.

Jie Zhou, Xu Han, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, and Maosong Sun. 2019a. Gear: Graph-based evidence aggregating and reasoning for fact verification. *arXiv preprint arXiv:1908.01843*.

Peng Zhou, Xintong Han, Vlad I Morariu, and Larry S Davis. 2018. Learning rich features for image manipulation detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1053–1061.

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.

Xinyi Zhou and Reza Zafarani. 2018. Fake news: A survey of research, detection methods, and opportunities. *arXiv preprint arXiv:1812.00315*, 2.

Zhixuan Zhou, Huankang Guan, Meghana Moorthy Bhat, and Justin Hsu. 2019b. Fake news detection via nlp is vulnerable to adversarial attacks. *arXiv preprint arXiv:1901.09657*.

Biru Zhu, Xingyao Zhang, Ming Gu, and Yangdong Deng. 2021. Knowledge enhanced fact checking and verification. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3132–3143.

Lingxuan Zhu, Weiming Mou, and Peng Luo. 2024. Potential of large language models as tools against medical disinformation. *JAMA Internal Medicine*, 184(4):450–450.

Linhe Zhu and Xiaoyuan Huang. 2019. Sis model of rumor spreading in social network with time delay and nonlinear functions. *Communications in Theoretical Physics*, 72(1):015002.

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.

Linlin Zong, Jiahui Zhou, Wenmin Lin, Xinyue Liu, Xianchao Zhang, and Bo Xu. 2024. Unveiling opinion evolution via prompting and diffusion for short video fake news detection. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 10817–10826.

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*.

M Zuckerman. 1981. Verbal and nonverbal communication of deception. *Advances in experimental social psychology/Academic Press*.

A Rumor Definition 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 rumors 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 Future Research Directions

B.1 LLM-based Multi-agent Social Simulation in Rumor Detection

While existing studies have shed light on the potential of rumor detection agents, there remains a lack of in-depth research into the complexities of the task. Critical gaps persist in addressing the key aspects of handling the complexity of rumor propagation, which can be summarized into the following three areas:

Lack of a deep understanding of the mechanisms of rumor propagation. Current research often relies on shallow features to classify information but fails to account for rumor dissemination's intricate cognitive behavioral patterns and sociodynamic characteristics. For instance, individuals tend to process information in ways that align with their pre-existing beliefs(cognitive consistency theory) while exhibiting significant non-objective and selective cognitive biases (Nickerson, 1998). Emotional drivers (such as fear and anger)(Chan et al., 2021), and social pressures (Blass, 1984) amplify the effects of rumor propagation. Moreover, some users spread false information not purely based on judgments of its truthfulness but rather due to hedonistic motives (Jiwa et al., 2023) or a sense of group identity (Lewandowsky, 2022; Wanless and Berk, 2020). Social networks' distributed and decentralized nature further reduces individuals' capacity to discern false information. It increases the likelihood of social cascades, forming extreme attitudes (Jamieson, 2008). This aspect has also received attention, e.g., simulating social phenomena

like echo chambers (Wang et al., 2024c).

Over-simplified modeling of rumor propagation and intervention. Current models are often restricted to static propagation roles (e.g., spreaders, bystanders, fact-checkers) and fail to comprehensively capture the dynamic changes in individual behaviors during the dissemination process and the dynamic role shifts in group interactions. In contrast, the information propagation dynamics extensively studied in information epidemiology offer valuable inspiration, such as integrating complex social variables (educational levels, forgetting mechanisms) to model dynamic processes more accurately. At the same time, intervention measures also lack flexibility and dynamic optimization regarding accurate verification and influencer blocking. This limits their ability to adapt to the gradual and complex evolution of rumor propagation, which often requires intervention strategies that are adaptable and sensitive to contextual changes.

Insufficient global modeling of the dynamic attributes of rumors. Existing studies are often confined to a specific dimension of rumors, such as content, propagation, or interaction, without systematically integrating these interdependent elements within a logical framework. Furthermore, the combined effect of multimodal information on rumor dissemination and the critical role of social bots as key propagation drivers have not been sufficiently considered. Additionally, active user behaviors, such as retrieval and learning of controversial information, remain underexplored. Although individual differences exist, such behaviors (such as actively verifying the authenticity of information and subsequently choosing to share, report, or ignore it) significantly impact information perception and dissemination.

B.1.1 Multi-Agent Systems Under the CIB Framework

To tackle existing challenges, we propose a research roadmap for multi-agent systems within the CIB framework, as illustrated in Figure 3. This approach utilizes cross-layer dynamic feedback to establish a **Macro-Micro Feedback Loop**, enabling the modeling of macro-level information dissemination based on micro-level individual cognition. It facilitates dynamic, bidirectional interventions between macro and micro levels, fostering evolution driven by deeper cognitive insights.

At the cognitive layer, agents can utilize LLMs and multimodal analysis tools to achieve a deep semantic understanding of text, 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 authenticity of information and precisely quantify user cognitive biases (e.g., selective processing or emotion-driven behaviors). For instance, agents can infer malicious intents behind false information through context-aware reasoning and identify the potential motivations of target users for spreading such information. This enables agents to provide information support for subsequent collaborative actions.

At the interaction layer, multi-agent systems leverage sophisticated communication protocols to collaborate efficiently, simulating the diversity of user behaviors in social networks, such as information sharing, commenting, and reporting. Moreover, they can also simulate external factors such as social bots, constructing dynamic environments that better reflect real-world propagation patterns. By comprehensively modeling collective knowledge and social interactions, these dynamic simulations help address critical questions, such as: How can we generate more targeted intervention strategies with greater accuracy? How can we improve collaborative efficiency when users participate in rumor reporting or debunking efforts?

At the behavior layer, multi-agent systems can capture the nonlinear propagation paths of collective behavior and dynamically generate personalized debunking content and intervention strategies through cognitive modeling. For example, By analyzing a target audience's cognitive and behavioral characteristics, agents can produce debunking content that is more persuasive and tailored to the audience. In response to the dynamic evolution of rumor propagation, agents can adaptively adjust intervention models, enabling more efficient and precise countermeasures.

Furthermore, in addition to modeling information dissemination of cognition to behavior, the framework enables a dynamic rumor detection and intervention mechanism in reverse. At the behavior layer, agents detect anomalous communities as initial targets. Subsequently, agents at the interaction layer analyze the suspicious users and interaction structures within these communities. Building on this, agents in the cognition layer perform collective knowledge analysis and social context reasoning on suspicious conversational threads to identify

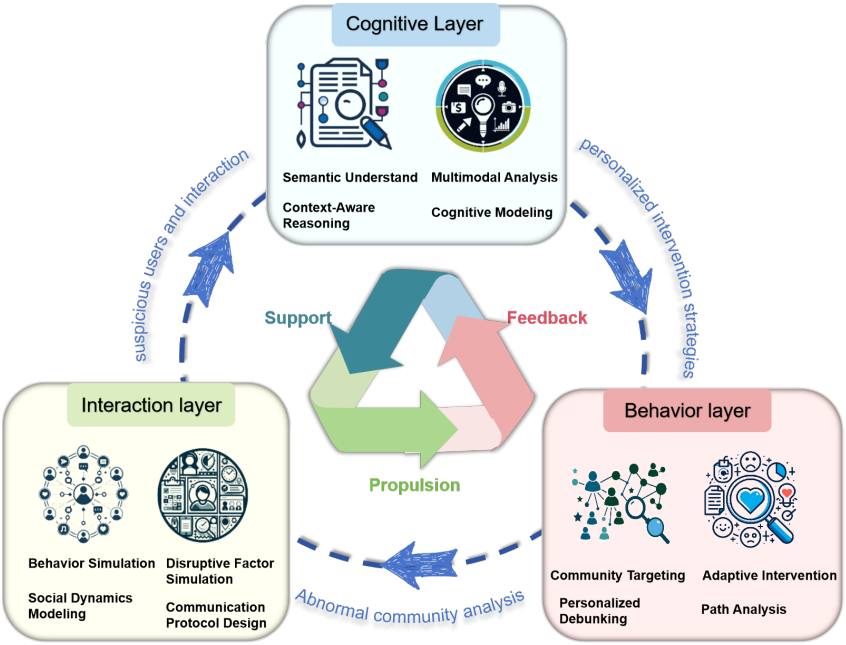

Figure 3: The CIB framework establishes a Macro-Micro Feedback Loop that integrates cross-layer dynamic feedback to bridge macro-level information dissemination with micro-level individual cognition. This enables bidirectional interventions, wherein macro-level propagation dynamics inform micro-level behavior modeling, while individual insights refine system-wide strategies. Beyond modeling the flow from cognition to behavior, the framework supports a reverse feedback mechanism for dynamic rumor detection and intervention, driving continuous adaptation and iterative improvement in complex, evolving misinformation environments.

critical evidence. Finally, the behavior layer intervenes promptly, generating personalized debunking content tailored to the cognitive characteristics of the target audience as part of belief-based interventions. This feedback mechanism allows rumor detection and intervention models to continuously improve and optimize themselves, enhancing their adaptability to the dynamically evolving nature of rumor propagation.

B.2 Cross-disciplinary Collaborative Optimization

LLMs also face significant challenges across three dimensions: content credibility, cognitive alignment, and technology adaptability (Liu et al., 2024a; De Angelis et al., 2023). First, the intertwining of hallucinated content generated by LLMs with malicious misinformation substantially complicates assessing content reliability (Pan et al., 2023c; Chen and Shu, 2024; Shu et al., 2021). Technologies such as deepfakes mislead the public and stigmatize genuine content, further eroding transparency in public discourse (e.g., real but negative content dismissed as synthetic and subsequently discredited (Schiff et al., 2023)). Second, limitations in LLMs' semantic alignment and ability to perform complex reasoning may reinforce users' preexisting cognitive biases (Xu et al., 2023; Garry et al., 2024; Hosseini et al., 2023). In complex scenes, the robustness and dynamic adaptability of the model (Carlini et al., 2023; Haller et al., 2023), as well as early benchmarks (Zhou et al., 2023; Chen and Shu, 2023), show weak performance. (He et al., 2023; Li et al., 2024b). To address these challenges, future research must promote cross-disciplinary collaborative optimization:

Cognition layer: Enhancing content credibility and knowledge attribution. Future research should focus on building dual mechanisms that combine technological forensics and social validation to improve content credibility, particularly for attributing and explaining the trustworthiness of LLM-generated content (André et al., 2023; Kumarage and Liu, 2023). Another area of importance is addressing "technology-amplified cognitive biases" and their impact on information credibility perception and cognitive schema activation. For instance, The high fluency of LLM-generated content amplifies the halo effect (Augenstein et al., 2024); Multimodal synthetic content, such as deepfakes, enhances emotional infiltration, making misinformation harder to detect.

Interaction layer: strengthening cognitive alignment and belief intervention in social interactions. Integrating psychology and behavioral science insights can enable belief interventions using LLM-generated personalized and persuasive debunking content (Costello et al., 2024; Matz et al., 2024). For example, Employing dynamic intervention strategies, such as the "Friction Strategy," can suppress blind adherence and impulsive information sharing by increasing the cognitive processing cost of user decisions; LLMs could also provide real-time knowledge enhancement services for low-education user groups, helping mitigate the continued influence effect (Lewandowsky et al., 2012; Walter and Tukachinsky, 2020) (where misinformation persists even after being debunked) and the recognition gap driven by educational disparities (Afassinou, 2014; Hui et al., 2020). This approach fosters a synergistic effect between educational compensation and behavioral interventions.

Behavior layer: Enhancing robustness and adaptability in dynamic environments. To improve the adaptability of LLMs in dynamic environments, future research can leverage LLM-agent architectures that perform multi-phase reasoning chains (e.g., planning-execution-reflection), enabling zero-shot automatic feature annotation and latent rumor identification in real-time. Additionally, a more comprehensive rumor baseline assessment should be established to address the evolving characteristics of misinformation in complex environments.

By integrating beneficial insights from across disciplines into a rumor detection framework based on collective intelligence, this research line can promote the development of efficient and highly adaptable systems, pushing the boundaries of rumor detection methods.

B.3 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 capability 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.

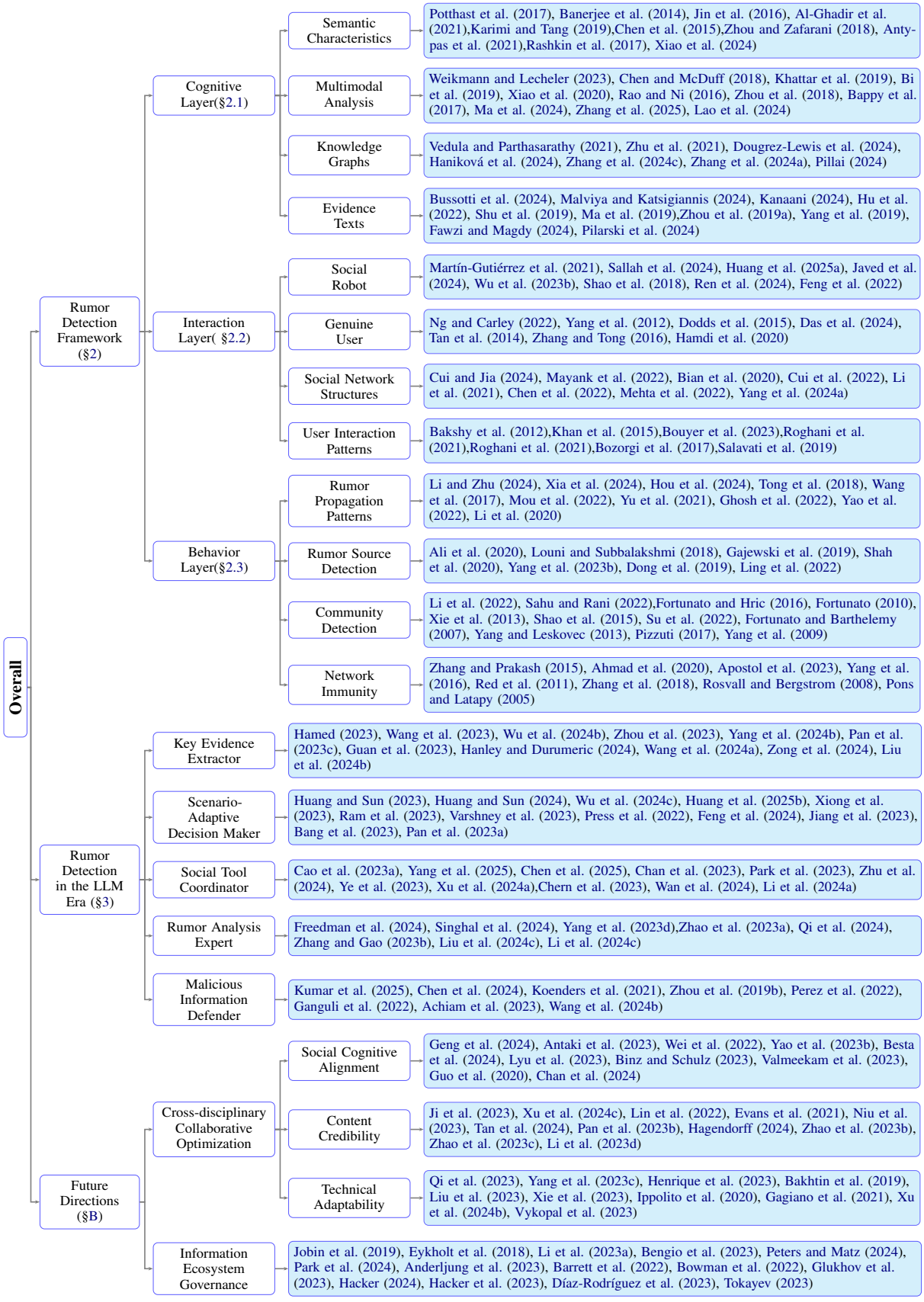

Figure 4: Classification of Rumor Detection Methods, Applications in the LLM Era, Challenges, and Future Directions under the CIB Framework