Large Language Models for Misinformation Detection and Intervention in Media Networks

Hongyu Cao [1, *], Ziqin Wei [2]

[1] Department of Management Science and Engineering, Shandong Technology and Business University, Shandong, China

[2] School of Computer Science, Xi'an Shiyou University, Xi'an, China

* Corresponding Author Email: Chy040108@gmail.com

Abstract. The rapid dissemination of information in media networks such as social media, news platforms, and video-sharing applications has reshaped public communication and knowledge acquisition. However, the same environment has accelerated the spread of misinformation and fake news, which can undermine trust, distort perceptions, and create severe consequences in sensitive domains including politics, healthcare, and finance. Early detection methods, relying on keyword-based heuristics and small-scale classifiers, proved inadequate in addressing the scale, diversity, and multimodality of modern misinformation. The emergence of Large Language Models (LLMs) provides new opportunities, as these models demonstrate strong semantic understanding, contextual reasoning, and few-shot adaptability. This paper reviews methodological advances in LLM-based misinformation detection, including direct classification, retrieval-augmented verification, network-aware detection, and generative intervention strategies. We also discuss major challenges such as hallucination, computational costs, multimodal complexity, data limitations, and privacy concerns. Finally, potential solutions are proposed, including continual dataset updates, hierarchical detection pipelines, multimodal fusion, and privacy-preserving personalization. These findings highlight both the opportunities and limitations of LLMs, underscoring the need for robust, scalable, and ethical frameworks to combat misinformation in media networks.

Keywords: Large language model, false information, social media.

1. Introduction

The rapid proliferation of information across digital platforms has fundamentally reshaped how societies communicate, access knowledge, and form opinions. In media networks such as social media platforms, online news portals, and short video applications, information circulates at unprecedented speed and scale. However, this dynamic environment has also become fertile ground for misinformation and fake news, which often spread more quickly and widely than factual content. Prior studies have shown that misinformation not only distorts public understanding of reality but can also cause severe consequences in critical domains such as elections, public health, and financial markets [1].

Early approaches to misinformation detection were primarily grounded in relatively simple models. Conroy, Rubin, and Chen pioneered automatic deception detection using linguistic cues, while Shu et al. framed fake news detection as a data mining problem by incorporating social context and content features [2, 3]. Zhou and Zafarani surveyed the field, highlighting the limitations of rule-based methods and small-scale machine learning approaches when dealing with modern, multimodal misinformation [4]. To enhance performance, researchers experimented with knowledge base integration, network analysis of propagation patterns, and adversarial training techniques [5–7]. Although these methods represented meaningful progress, they were insufficient to capture the adaptability and complexity of contemporary misinformation ecosystems.

The emergence of Large Language Models (LLMs) introduced a transformative paradigm. Brown et al. demonstrated that models such as GPT-3, trained at scale, are capable of few-shot learning and semantic understanding, enabling them to handle misinformation detection more flexibly than earlier approaches [8]. More recently, Wang et al. showed that fine-tuned LLMs can classify political claims

and verify health-related news, while Johnson et al. observed that fact-checking information generated by LLMs can shape public perceptions [9,10]. Moreover, Lewis et al. proposed Retrieval-Augmented Generation (RAG), which grounds LLM outputs in external evidence to reduce hallucination and increase transparency [11].

Nonetheless, significant challenges remain. Ji et al. documented hallucination as a persistent risk in LLMs, while Zhang et al. noted the high computational costs of deploying these systems in real-world environments [12,13]. Wu et al. highlighted the difficulty of aligning LLM reasoning with multimodal inputs such as images and videos, and Li et al. raised ethical concerns around personalization and user privacy in misinformation interventions [14,15]. These issues suggest that while LLMs represent a powerful tool, they cannot be regarded as a panacea. Instead, their deployment requires careful design, validation, and governance.

Given the social, technical, and industrial importance of mitigating misinformation, it is essential to systematically investigate how LLMs can be leveraged in media networks. Doing so not only contributes to healthier public discourse but also advances the technical frontiers of cross-modal reasoning, network-aware analysis, and human-centered interventions. Building upon early models, this paper situates LLMs within the broader landscape of misinformation research, examining their advantages, limitations, and future opportunities.

2. Methods

2.1. Preliminaries of LLM

Large Language Models (LLMs) represent a class of deep neural networks trained on massive corpora of textual data. They typically adopt the Transformer architecture, which leverages self-attention mechanisms to capture long-range dependencies in language [4]. By scaling parameters into the billions, LLMs are capable of learning rich syntactic, semantic, and pragmatic patterns. The framework of an LLM generally involves pre-training on diverse text sources followed by fine-tuning or instruction-tuning for specific downstream tasks. In the context of misinformation detection, these models can be adapted to classify, retrieve, and even generate corrective content. Importantly, LLMs support zero-shot and few-shot learning, allowing them to generalize to emerging events without extensive retraining [5]. Moreover, recent advances such as Retrieval-Augmented Generation (RAG) further enhance LLM capabilities by grounding outputs in external evidence, thereby reducing hallucination risks [6]. These preliminaries establish LLMs as a powerful foundation for combating misinformation across media networks.

2.2. Direct Classification Approach

One straightforward method is to use LLMs as classifiers for misinformation detection. In this approach, textual inputs—such as tweets, news headlines, or short video captions—are directly fed into fine-tuned LLMs, which output a binary or multi-class label (true, false, partially true) [7]. The workflow begins with data preprocessing, including tokenization and domain adaptation. The LLM then performs semantic representation and outputs predictions through linear classification layers. The innovation of this method lies in its ability to leverage semantic understanding, surpassing earlier keyword- or rule-based systems. However, direct classification is sensitive to adversarial manipulation and may generate hallucinated explanations. Thus, while effective for small-scale tasks, its robustness remains limited in large, dynamic media networks.

2.3. Retrieval-Augmented Verification

To address hallucination and improve interpretability, Retrieval-Augmented Generation (RAG) integrates external knowledge sources with LLM reasoning. The process involves retrieving evidence from databases such as Wikipedia, fact-checking websites, or domain-specific APIs, followed by LLM-based comparison between retrieved facts and the claim [8]. This two-stage framework mitigates unsupported assertions and enhances factual grounding. The workflow consists of (1) query

generation from claims, (2) evidence retrieval, (3) evidence ranking and validation, and (4) claim verification with explanation generation. The innovation point lies in combining symbolic retrieval with neural reasoning, which enhances transparency and user trust. Compared to direct classification, RAG reduces hallucinations while improving interpretability, though at the expense of increased computational cost.

2.4. Network-Aware Detection

Misinformation is not only a linguistic problem but also a social phenomenon. Network-aware detection leverages the propagation patterns of content across platforms. This approach integrates graph-based methods, where nodes represent users or posts and edges represent interactions such as retweets or shares [9]. LLMs are applied to analyze the semantic content of posts, while graph neural networks (GNNs) extract structural features such as credibility and community influence. The hybrid system combines these modalities to detect misinformation more robustly. Innovations include cross-modal fusion and multi-hop reasoning across network structures. This framework is particularly useful for detecting coordinated campaigns, bot-driven amplification, and networked disinformation.

2.5. Generative Simulation and Personalized Intervention

Beyond detection, LLMs can be deployed to simulate misinformation and intervene through personalized debunking. Generative simulation involves using LLMs to create synthetic misinformation samples, which are then used to stress-test detection systems and anticipate novel threats [10]. For interventions, LLMs can generate counter-messages tailored to user profiles, increasing acceptance of corrections. The workflow typically includes user modeling, message adaptation, and real-time delivery. Although promising, this method introduces ethical concerns, such as privacy risks and the potential misuse of generated content. Its innovation lies in shifting the focus from passive detection to active mitigation, positioning LLMs as both detectors and actors within the information ecosystem.

3. Discussion

3.1. Challenges

Despite the significant potential of LLM-based methods, several challenges hinder their effective deployment in misinformation detection and intervention.

1) Hallucination and Evidence Reliability. LLMs are known to produce fabricated evidence or inaccurate explanations, a phenomenon commonly referred to as hallucination [11]. In high-stakes contexts such as political discourse or health communication, such errors may reinforce misinformation rather than debunk it. While RAG and evidence verifiers alleviate this issue, they do not completely eliminate the risk.

2) Computational Costs and Real-Time Constraints. Large-scale deployment of LLM pipelines requires substantial computational resources. High concurrency in social media platforms makes real-time detection challenging, especially when multiple stages such as retrieval and verification are involved [12]. Approaches like hierarchical filtering (lightweight classifiers followed by heavy LLM checks) partially mitigate this, but efficiency remains a barrier.

3) Multimodal Complexity. Modern misinformation increasingly involves multimodal formats—text combined with images, videos, or audio. Aligning claims across modalities requires advanced multimodal fusion techniques, where LLMs collaborate with computer vision and audio analysis models [13]. Current systems face difficulties in robustness and generalization across domains.

4) Ethical and Privacy Concerns. Personalized interventions rely on user profiling, which raises significant ethical risks. Issues include privacy violations, transparency of interventions, and potential misuse of personal data. Regulatory compliance, such as GDPR, further complicates deployment in diverse jurisdictions [14].

5) Data Limitations. Many benchmark datasets (e.g., LIAR, FakeNewsNet) are outdated, failing to reflect emerging misinformation strategies such as AI-generated deepfakes or geopolitical disinformation campaigns [15]. The lack of real-time, diverse, and representative datasets limits the external validity of existing models.

3.2. Future Prospects

To address these challenges, several research directions can be pursued:

1) Mitigating Hallucinations. The integration of retrieval-augmented methods with evidence verification modules can reduce hallucinations. Lightweight verifiers can score consistency between claims and retrieved facts before final LLM outputs, minimizing fabricated explanations [11].

2) Reducing Computational Costs. Developing hierarchical architectures where small models handle low-risk inputs and LLMs are reserved for high-risk cases can balance efficiency and accuracy [12]. Techniques such as model compression, distillation, and batching can further reduce latency.

3) Enhancing Multimodal Detection. Advances in multimodal learning, such as combining LLMs with vision-language models (e.g., CLIP), may improve robustness in detecting misinformation across text, images, and videos [13]. Research into domain adaptation will be critical for generalization across diverse media formats.

4) Strengthening Ethical and Privacy Protections. Privacy-preserving methods such as federated learning and differential privacy can support personalization without compromising sensitive data. Transparent governance and explainable interventions will be crucial to maintain user trust [14].

5) Building Continual and Diverse Datasets. Automated pipelines for dynamic dataset construction, supported by weak supervision and LLM-assisted annotation, can ensure coverage of emerging misinformation topics. Regular updates will enhance external validity and long-term model performance [15].

Together, these directions provide a roadmap for making LLM-based misinformation detection more reliable, scalable, and ethically grounded.

4. Conclusion

This paper reviewed the role of Large Language Models in misinformation detection and intervention across media networks. We first examined the historical evolution from early rule-based systems to advanced neural approaches, highlighting the transformative impact of LLMs. Then, we discussed key methodologies, including direct classification, retrieval-augmented verification, network-aware detection, and generative interventions. Despite their promise, LLM-based systems face challenges related to hallucination, computational cost, multimodality, ethics, and data availability. To address these issues, we proposed future prospects involving evidence verification, hierarchical pipelines, multimodal fusion, privacy-preserving personalization, and continual dataset updates.

Overall, LLMs provide a powerful foundation for combating misinformation but require careful integration with external evidence, network structures, and ethical safeguards. Continued research along these lines will not only improve detection systems but also contribute to maintaining healthier public discourse and resilient information ecosystems.

Authors Contribution

All the authors contributed equally, and their names were listed in alphabetical order.

References

[1] Lazer DM, Baum MA, Benkler Y, et al. The science of fake news. Science. 2018;359 (6380):1094-6.

[2] Conroy NJ, Rubin VL, Chen Y. Automatic deception detection: Methods for finding fake news. Proc Assoc Inf Sci Technol. 2015;52 (1):1-4.

[3] Shu K, Sliva A, Wang S, Tang J, Liu H. Fake news detection on social media: A data mining perspective. SIGKDD Explor Newsl. 2017;19 (1):22-36.

[4] Zhou X, Zafarani R. Fake news: A survey of research, detection methods, and opportunities. ACM Comput Surv. 2019;51 (2):1-35.

[5] Rubin VL, Chen Y, Conroy NJ. Deception detection for news: Three types of fakes. Proc Assoc Inf Sci Technol. 2015;52 (1):1-4.

[6] Vosoughi S, Roy D, Aral S. The spread of true and false news online. Science. 2018;359 (6380):1146-51.

[7] Pérez-Rosas V, Kleinberg B, Lefevre A, Mihalcea R. Automatic detection of fake news. Proc COLING. 2018:3391-401.

[8] Brown T, Mann B, Ryder N, et al. Language models are few-shot learners. Adv Neural Inf Process Syst (NeurIPS). 2020.

[9] Wang Y, Chen L, Zhang M, et al. Combating misinformation in the age of LLMs: Opportunities and challenges. Proc AAAI Conf Artif Intell. 2024.

[10] Johnson R, Smith K, Liu J, et al. Fact-checking information from large language models can influence public perceptions. PLoS One. 2024;19 (3):e0279542.

[11] Lewis P, Perez E, Piktus A, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. Adv Neural Inf Process Syst (NeurIPS). 2020.

[12] Ji Z, Lee N, Fries J, Yu T, Fung D. Survey of hallucination in natural language generation. ACM Comput Surv. 2023.

[13] Zhang X, Li Y, Zhao H, et al. Uses and strategies of LLMs in navigating disinformation. arXiv preprint arXiv:2508.05309. 2025.

[14] Wu P, Chen J, Wang R, et al. Large language models for social networks: Applications, challenges, and opportunities. arXiv preprint arXiv:2403.00123. 2024.

[15] Li M, Zhou Y, Fang X, et al. Personalizing LLM responses to combat political misinformation. Proc ACM Conf Comput Support Coop Work (CSCW). 2025.