Proceedings of the 4th International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/2026.39295

Large Language Model-Driven Misinformation Detection: Methods, Trustworthiness Evaluation, and Challenges

Shuyao Li

School of Computer and Control Engineering, Qiqihar University, Qiqihar, China
lishuyao99@163.com

Abstract. Large language models (LLMs) have offered a new path for misinformation detection, but the reliability of model-generated auxiliary information and the applicability of detection results in new scenarios have yet to be verified. Some representative studies and related methods have been classified into three types in this paper: content-driven, context- and relation-driven, and knowledge- and evidence-driven methods. Factual reliability, generalisation reliability and the reliability of explanations and evidence are introduced. According to the comparison of previous experiments, semantic augmentation, simulation of user behaviour and logical reasoning can all improve the detection rate in specific circumstances; however, the level of improvement differs depending on the dataset and model parameters. This suggests that method selection should depend on the available information and the specific detection scenario, rather than classification scores alone. Based on the above results, this article analyzes and evaluates issues such as data contamination, synthetic data bias, and external evidence reliability, providing reference for the selection and evaluation of detection methods.

Keywords: Large language models, misinformation detection, trustworthiness, fact-checking, retrieval augmentation

1. Introduction

The spread of false information has increased the difficulty for the public to obtain reliable information and has also drawn attention to automatic detection methods. Zhou and Zafarani reviewed fake news detection methods from the perspectives of knowledge, writing style, propagation patterns, and source credibility [1]. With the involvement of Large Language Models (LLMs) in content generation, information written manually, generated by machines, and rewritten by machines coexist. Su et al. found that the composition of news from different sources in the training data can affect the detector's recognition performance of machine-generated news [2]. LLMs therefore change both the content to be detected and the tools available for semantic extraction and auxiliary judgment.

Previous studies have attempted to use LLMs to enhance semantic representation, generate auxiliary information, or organize external evidence. Ma et al. supplemented news semantics through topics and entity relationships. The comparison and ablation experiments show that modeling topic and entity information helps improve detection performance on the tested dataset

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.39295

[3]. Wan et al. used LLM-generated user responses and proxy task explanations to assist in detection. The ablation analysis shows that this auxiliary information can help improve the detection performance in the evaluated setting, but the generated response cannot be used as evidence of real user behavior [4]. Zeng et al. reviewed automated fact-checking through claim detection and claim validation, including evidence retrieval and claim verification, providing a framework for examining claims and their supporting evidence [5]. Chen and Shu reviewed the research on false information detection before the emergence of LLMs, and discussed opportunities and challenges from two aspects: using LLMs to assist in detection and responding to false information generated by LLMs, providing reference for analyzing changes in detection methods and their risks [6]. These studies use different information and evaluation conditions, and it is difficult to judge the reliability of the method solely based on classification scores. It is also necessary to examine the authenticity of the generated information, the quality of evidence, and cross-scenario performance.

This article summarizes representative studies based on the main sources of information and discusses the applicable conditions of different methods. The main contributions of this article include the following points. Firstly, this article clarifies the connections and differences between content-driven, context- and relationship-driven, and knowledge- and evidence-driven methods. Secondly, it compares the reliability of the detection results from three aspects: facts, generalization, and interpretation and evidence. Finally, it analyzes the advantages and limitations of representative methods based on publicly available experimental results.

2. Background and preliminaries

This article mainly discusses the veracity classification of news articles or posts, and also includes factual verification studies directly related to the utilization of external evidence. These tasks overlap, but their evaluation targets are not identical: news detection can use content and propagation cues, whereas fact-checking typically retrieves evidence for specific claims and determines whether it supports or refutes them [5].

Among them are semantic expansion, context construction, evidence collection and reasoning, etc., of LLMs. Accuracy measures the proportion of overall predictions that are correct. Using false information as the positive class, Precision indicates how many predicted false samples are actually false, and Recall shows how many actual false samples have been found. F1 is the harmonic mean of the two. Macro-F1 takes the arithmetic mean of each category F1 to give each category the same weight. Classification metrics do not establish trustworthiness. Cross-scenario stability, factual grounding, and the reliability of explanations and evidence also need to be examined. The ablation experiment analyzes the impact of removing or replacing modules on performance.

3. Methods and technical classification of LLM-driven misinformation detection

Figure 1 shows the method classification and trustworthiness evaluation framework of this article. The three types of methods focus on content characteristics, reaction and propagation information, and external factual basis, which can complement each other and do not constitute a fixed processing flow. Using a graph structure does not necessarily make a method propagation-based. For example, the graph in LESS4FD mainly represents the relationship between news, topics, and entities.

DOI: 10.54254/2755-2721/2026.39295

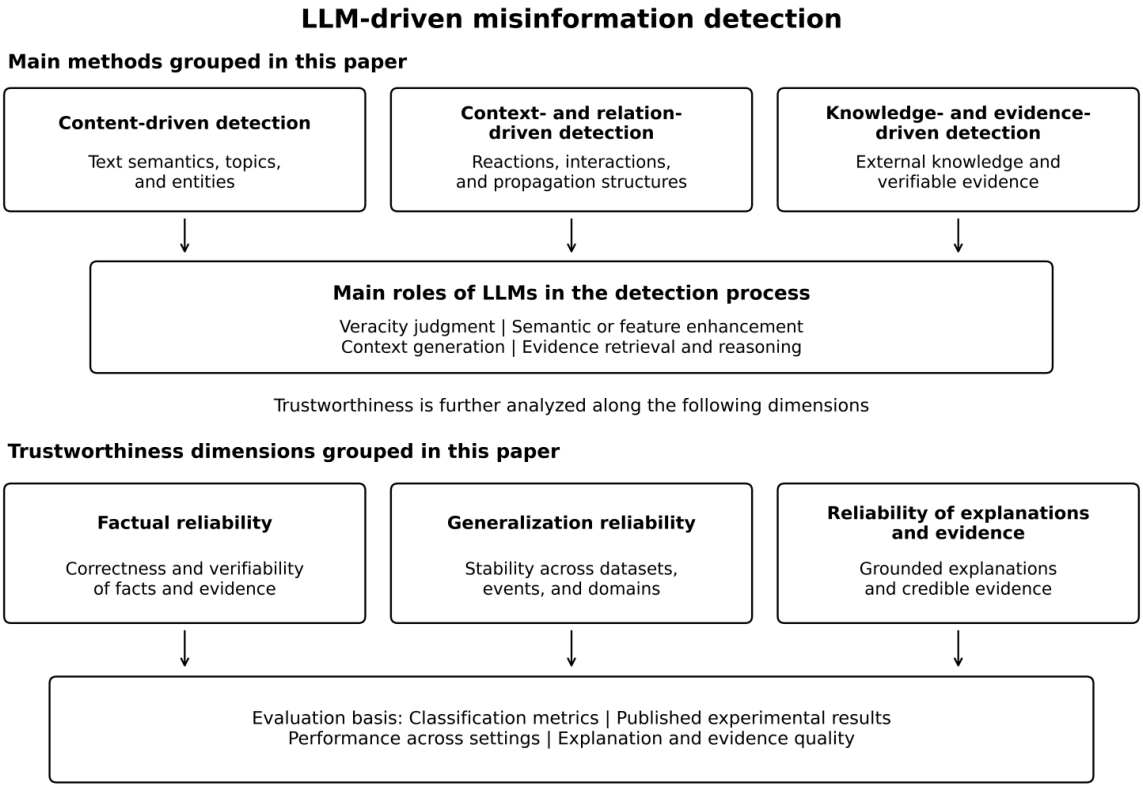

Figure 1. Framework of LLM-driven misinformation detection methods and trustworthiness evaluation (picture credit: original)

3.1. Content-driven detection

The main types of content-driven methods extract discriminative information from news text or image-text data. LESS4FD uses a heterogeneous graph that contains news articles, topics and entities. LLMs can extract entities and semantics, while BERTopic is used to discover topics [3]. The various nodes in the graph are connected and exchange meaning. Ma and others compared and ablated the results to find that topic and entity information improved the detection accuracy on the evaluation set.

Distribution gap for image-text misinformation detection exists between synthetic and real data. Zeng and others combined three existing synthetic datasets, then employed data selection to address the resulting deficiencies and enhance generalisation performance on real-world data [7].

3.2. Context- and relation-driven detection

Context- and relation-driven methods use user reactions, propagation structures and interaction relationships outside of the news content to help detect. DELL uses LLM-generated user reactions and explanations for proxy tasks as auxiliary information in news detection [4]. Proxy tasks include sentiment analysis and stance detection. Detectors trained on this data are task experts and combine their predictions. The above user behaviour is simulated and should not be considered real data.

When the dissemination record is missing, adding only the comment text will not show the structure of interaction. Chen and others' StruSP considers comment reply relationships and the structure of propagation trees in generating propagation information, and combines semantic and

DOI: 10.54254/2755-2721/2026.39295

structural information to aid detection [8]. The construction of synthetic propagation should be consistent with actual cases.

3.3. Knowledge- and evidence-driven detection

Knowledge- and evidence-driven methods use external materials to verify facts. First, RAFTS retrieves and organises the evidence to select materials related to the claim, then synthesizes supporting and refuting arguments, and finally makes a decision [9]. Guan et al. studied how evidence conditions affect the fact-verification performance of language models [10]. Ev2R is a reference-based evaluation method for the retrieved evidence that sets verdict-level proxy scores of evidence quality [11].

External evidence can supplement the model's knowledge, but retrieving related materials does not guarantee the accuracy of the evidence. Sources of evidence, their relevance to the claim being verified, and conflicts among the pieces of evidence can all affect the final judgment.

Table 1 compares the information sources, roles of LLMs, and limitations of representative methods.

Table 1. Comparison of representative LLM-based misinformation detection and related fact verification methods

Method	Main information	Role of the LLM	Main strengths	Points to consider
LESS4FD [3]	Text, topics, entities	Semantic enhancement	Adds high-level semantic and relational information	Performance differs considerably across datasets
DELL [4]	News, reactions, explanations	Auxiliary information generation and expert ensembling	Uses reactions and explanations to enhance detection	Generated reactions are not equivalent to real propagation
RAFTS [9]	Claims and external evidence	Forms arguments based on retrieved evidence	Provides external grounding for fact verification	A fact verification task; depends on evidence quality
TELLER [12]	News content and logical information	Assists in generating explainable information for judgments	Focuses on explainability, generalizability, and controllability	Cross-domain performance depends on the transfer setting
StruSP [8]	Propagation semantics and structure	Generates structure-aware propagation	Supplements incomplete propagation information	Synthetic propagation needs alignment with real structures

The methods in Table 1 are various applications of LLMs: LESS4FD improves semantic representation; DELL and StruSP generate supplementary information; RAFTS organises external evidence. TELLER instantiates logical predicates into logic atoms of the news articles, uses an LLM to assign truth values to these atoms, and then learns logical rules for classification. Therefore, content analysis and logical reasoning are combined [12].

4. Trustworthiness evaluation of LLM-based misinformation detection

The three types of trustworthiness are: the actual facts that are being referenced; the stability of their operation after a change in the distribution of data; and the capacity to trace back the supporting reasons.

Agreement with labels should be distinguished from factual grounding. Classification indicators can be used to measure the degree of agreement with labels, and it must also be verified that the

DOI: 10.54254/2755-2721/2026.39295

claims to be verified are consistent with the supporting evidence. Guan and others' research have shown that the effect of verification is related to the quality of the evidence provided [10].

Generalisation ability refers to the ability of a model to perform well in new situations, other areas, and various texts obtained from different sources. TELLER discusses trustworthy detection from the perspectives of explainability, generalizability, and controllability, and sets up cross-domain experiments [12]; Zeng et al. investigated the distribution differences between synthesized data and real multimodal data [7].

The Reliability of Explanations and Evidence refers to the verifiability of the reasoning process. Huang and others classify LLM hallucinations into factuality hallucinations and faithfulness hallucinations, the latter of which includes inconsistencies with instructions, context or internal logic [13]. Ev2R assesses the evidence itself and can be used to evaluate the quality of retrieved evidence [11]. Provide corresponding support for the reasons. Differentiate between model-generated reasons and independently verifiable data, and the readability of an explanation should not be assumed to be its correctness.

5. Comparison of experimental results for representative methods

This section presents the results published in the original studies; no new model training or experiment replication has been carried out. Table 2 shows the results on different datasets, Table 3 compares detection methods in the same study, and Table 4 presents cross-domain test results. Since the experimental conditions in the various studies are not the same, these results have not been compared to determine a general performance order.

5.1. LESS4FD results across datasets

Table 2. Accuracy and Macro-F1 of LESS4FD on five datasets

Dataset	Llama2Accuracy	Llama2Macro-F1	GPT-3.5Accuracy	GPT-3.5Macro-F1
MM COVID	0.973±0.011	0.972±0.011	0.974±0.010	0.973±0.010
ReCOVery	0.917±0.017	0.897±0.020	0.938±0.020	0.929±0.017
MC Fake	0.883±0.006	0.787±0.008	0.894±0.012	0.833±0.013
LIAR	0.689±0.034	0.658±0.035	0.678±0.021	0.672±0.019
PAN2020	0.731±0.037	0.727±0.037	0.771±0.017	0.769±0.017

Source: Table 3 in Ma et al., p. 514 [3]. Values are means ± standard deviations of ten-fold cross-validation, and F1 is Macro-F1. According to Section 3.2 and Appendix A.4 of the original paper, topics are obtained by BERTopic. The two varieties have the same entities and topics, and their respective embeddings are obtained by GPT-3.5. News embeddings are obtained from Llama2 and GPT-3.5, respectively; therefore, the comparison mainly focuses on the source of the news representations.

Table 2 shows that both indicators on MM COVID exceed 0.97, while the performance on LIAR is lower. The original paper used binary classification, so the performance differences cannot be directly attributed to LIAR's initial six-class labelling scheme. The Domains and Classes of these datasets are also not the same. MM COVID and ReCOVery concern health information; LIAR refers to political statements; MC Fake covers all of the above. As shown in Appendix A.1 of the original text, the number of real and fake news samples in MC Fake are 12,435 and 2,591, respectively, and a significant difference between Accuracy and Macro-F1 is also observed [3]. These differences suggest that when comparing results, both domain and class distributions should be considered simultaneously, as the scores in the table cannot distinguish the impact of the two.

DOI: 10.54254/2755-2721/2026.39295

5.2. Comparison of DELL with different baselines

Table 3. Macro-F1 of DELL and baselines on Pheme and LLM-mis

Method	Pheme Macro-F1	LLM-mis Macro-F1
Zero-shot	0.459	0.597
Few-shot	0.490	0.565
Retrieval	0.464	0.624
DeBERTa	0.779	0.887
DELL Single	0.810	0.928
DELL Selective	0.820	0.897

Source: Table 1 in Wan et al., p. 2641 [4]. This table reports binary classification results. Single denotes the best-performing individual task expert. In Selective, an LLM selects relevant experts based on the news content and combines their predictions and confidence scores to make a judgment. In the main experiments, Mistral-7B was primarily used to generate reactions and perform proxy tasks, while ChatGPT was used for expert ensembling.

In Table 3, Zero-shot provides no labeled examples, whereas Few-shot provides a small number. Retrieval incorporates knowledge retrieved from Wikipedia, while DeBERTa uses a pretrained text encoder and a classifier [4]. Both DELL settings outperform the listed prompting and retrieval baselines as well as DeBERTa, although the gains over the trained DeBERTa detector are smaller. Selective exceeds Single by 0.010 on Pheme but falls below it by 0.031 on LLM-mis, showing that selective expert ensembling does not necessarily outperform the best individual expert.

5.3. Cross-domain results of TELLER

Table 4. Cross-domain Macro-F1 (%) of TELLER using Llama2-13B

Training sources → test dataset	Direct judgment	TELLER
Constraint + PolitiFact → GossipCop	52.27	60.90
GossipCop + PolitiFact → Constraint	51.75	84.87
Constraint + GossipCop → PolitiFact	77.10	79.41

Source: Table 3 in Liu et al., p. 15563 [12]. Train TELLER's Decision System with the two source datasets and test on the third. Direct is the basic way to ask an LLM to predict news labels.

As shown in Table 4, when Llama2-13B is used, TELLER's Macro-F1 in all three transfer directions is higher than that of direct judgment, but the improvement amounts are 8.63, 33.12 and 2.31 percentage points, respectively, and they vary considerably. In the original study, Macro-F1 with Llama2-7B decreased from 81.73% to 81.04% after transferring from GossipCop and PolitiFact to Constraint, showing that the method does not improve performance under all circumstances [12].

6. Existing problems and development trends

Contamination of evaluation data will reduce the reliability of detection results. Some of the test data have been used for pre-training of the model; therefore, the scores on static benchmarks may not fully reflect its evaluation performance in new environments. TripleFact reduces the risk through

DOI: 10.54254/2755-2721/2026.39295

dynamic acquisition of claims and evaluations of temporal generalisation [14]. In the future, evaluations should record the time of data collection, the base model version and train-test split, and check for duplicate or near-duplicate texts. Report the results of new events separately to evaluate performance on unseen content.

The Quality of the external evidence and the accuracy of the generated explanations should be verified independently. The retrieved corpus may contain false data, and the generated explanations may also exceed the scope of the evidence. Zou and others have shown that by adding malicious texts to a RAG knowledge base, it is possible to make the language model output attacker-selected answers in the evaluation conditions [15]. Du and others have studied attacks that add fake documents to the evidence repository or insert false information into the existing documents to reduce the effectiveness of fact verification in their experiments [16]. Evidence-based detection should record the origin and time of the information, find inconsistencies among the evidence, and distinguish between retrieved original texts and model-generated arguments.

Cross-scenario generalisation and bias in synthetic data are also problems with application. As shown in Table 4, the gains under the different transfer settings are not the same, and there are also distribution gaps between synthetic and real-world multimodal data [7]. In the future, researchers will investigate the impact of each factor independently, determine what information is observed and what is generated by the model, and use classification accuracy and error analysis to evaluate the system.

Therefore, based on the availability of data, a simple model can be selected. The advantages of selective expert ensembling vary among the datasets in Table 3, and the cross-domain gains in Table 4 are not uniform. Future research will investigate when the addition of extra information is necessary and when it is not.

7. Conclusion

Review of LLM-based Misinformation Detection and Grouping of Methods by Primary Information Source: Content-driven, Context- and Relation-driven, and Knowledge- and Evidence-driven Approaches. Trustworthiness is the correctness of facts, generalisation performance and reliability of explanations and supporting evidence. Comparison of published LESS4FD, DELL, and TELLER results shows that performance varies based on the dataset, model application mode, and transfer settings. Therefore, only one experiment cannot prove that the method works in other circumstances.

In addition to the division result, subsequent work should also present the reasons for this division, confirm that the reasons are based on verifiable data, and list cases in which the method failed. Dynamic Evaluation, evidence tracing and synthetic data bias analysis can be used to support all-around assessments of detection results.

References

[1] Zhou, X., & Zafarani, R. (2020). A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Computing Surveys, 53*(5), Article 109, 1–40. https://doi.org/10.1145/3395046

[2] Su, J., Cardie, C., & Nakov, P. (2024). Adapting fake news detection to the era of large language models. In *Findings of the Association for Computational Linguistics: NAACL 2024* (pp. 1473–1490). Mexico City, Mexico: Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.findings-naacl.95

[3] Ma, X., Zhang, Y., Ding, K., et al. (2024). On fake news detection with LLM enhanced semantics mining. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing* (pp. 508–521). Miami, Florida, USA: Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.emnlp-main.31

DOI: 10.54254/2755-2721/2026.39295

[4] Wan, H., Feng, S., Tan, Z., et al. (2024). DELL: Generating reactions and explanations for LLM-based misinformation detection. In *Findings of the Association for Computational Linguistics: ACL 2024* (pp. 2637–2667). Bangkok, Thailand: Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.findings-acl.155

[5] Zeng, X., Abumansour, A. S., & Zubiaga, A. (2021). Automated fact-checking: A survey. *Language and Linguistics Compass, 15*(10), e12438. https://doi.org/10.1111/lnc3.12438

[6] Chen, C., & Shu, K. (2024). Combating misinformation in the age of LLMs: Opportunities and challenges. *AI Magazine, 45*(3), 354–368. https://doi.org/10.1002/aaai.12188

[7] Zeng, F., Li, W., Gao, W., et al. (2024). Multimodal misinformation detection by learning from synthetic data with multimodal LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2024* (pp. 10467–10484). Miami, Florida, USA: Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.findings-emnlp.613

[8] Chen, M., Wei, L., Zhou, W., et al. (2025). Structure-aware propagation generation with large language models for fake news detection. In *Findings of the Association for Computational Linguistics: EMNLP 2025* (pp. 13258–13272). Suzhou, China: Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.findings-emnlp.714

[9] Yue, Z., Zeng, H., Shang, L., et al. (2024). Retrieval augmented fact verification by synthesizing contrastive arguments. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics, Volume 1: Long Papers* (pp. 10331–10343). Bangkok, Thailand: Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.acl-long.556

[10] Guan, J., Dodge, J., Wadden, D., et al. (2024). Language models hallucinate, but may excel at fact verification. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1: Long Papers* (pp. 1090–1111). Mexico City, Mexico: Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.naacl-long.62

[11] Akhtar, M., Schlichtkrull, M., & Vlachos, A. (2026). Ev2R: Evaluating evidence retrieval in automated fact-checking. *Transactions of the Association for Computational Linguistics, 14*, 530–561. https://doi.org/10.1162/tacl.a.647

[12] Liu, H., Wang, W., Li, H., et al. (2024). TELLER: A trustworthy framework for explainable, generalizable and controllable fake news detection. In *Findings of the Association for Computational Linguistics: ACL 2024* (pp. 15556–15583). Bangkok, Thailand: Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.findings-acl.919

[13] Huang, L., Yu, W., Ma, W., et al. (2025). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems, 43*(2), Article 42, 1–55. https://doi.org/10.1145/3703155

[14] Xu, C., & Yan, N. (2025). TripleFact: Defending data contamination in the evaluation of LLM-driven fake news detection. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics, Volume 1: Long Papers* (pp. 8808–8823). Vienna, Austria: Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.acl-long.431

[15] Zou, W., Geng, R., Wang, B., & Jia, J. (2025). PoisonedRAG: Knowledge corruption attacks to retrieval-augmented generation of large language models. In *34th USENIX Security Symposium (USENIX Security 25)* (pp. 3827–3844). USENIX Association.Seattle, WA, USA: https://www.usenix.org/conference/usenixsecurity25/presentation/zou-poisonedrag

[16] Du, Y., Bosselut, A., & Manning, C. D. (2022). Synthetic disinformation attacks on automated fact verification systems. *Proceedings of the AAAI Conference on Artificial Intelligence, 36*(10), 10581–10589. https://doi.org/10.1609/aaai.v36i10.21302