Proceedings of CONF-SPML 2026 Symposium: The 2nd Neural Computing and Applications Workshop 2025
DOI: 10.54254/2755-2721/2026.TJ30220

Large Language Models and Social Networks: Generation, Detection, and Dissemination Impact

Jiaheng Fu

Department of Computer Science, Wenzhou-Kean University, Wenzhou, China
1337902@wku.edu.cn

Abstract. Even though the popularity of the Large Language Models (LLMs) has driven the vigorous development of the related fields in natural language processing, due to LLMs' generating text having features of low cost and high scalability, along with the platform's streaming mechanism, it has had many impacts on the social network environment. Aiming to address these kinds of problems, this essay approaches the topic from the perspectives of detection methods, data distribution, and platform impact, which have already been extensively studied in their respective individual fields. Still, there is a lack of collaboration between them. This essay will review and summarize recent research to determine how it addresses the questions of "The way to distinguish model-generated content," "The way model-generated content spreads on platforms," and "The impact model-generated content has on the information environment." They highlight their complementary relationships regarding methodology choices, experimental design, and application levels. Finally, it was found that the problems couldn't be eradicated if people relied solely on detection techniques. In the future, researchers should focus more on constructing multi-governance frameworks that combine algorithm detection, data benchmarking, and platform governance.

Keywords: Large Language Models, Social Networks, Misinformation, Hallucination Checks

1. Introduction

As the leading platform for information dissemination today, social media is characterized by low barriers to entry, strong interaction, and broad reach. And once the text generated by the model enters this environment, it will surpass previously manually created false information due to its scale and concealment, which has brought unprecedented social risks [1-3]. The reason comes from three angles. Firstly, LLMs could exhibit hallucinations during the generation process, meaning that they can fabricate facts or contain logical errors [2]. This phenomenon has a limited impact when it occurs infrequently, but when continuously amplified by social network recommendation algorithms, it may lead to collective biases [4]. Secondly, compared to traditional spam messages or fake news, AI-generated text exhibits better contextual consistency and writing fluency, which significantly increases the misleading nature of the generated content [1, 3]. Third, current research mainly focuses on the text content itself. It lacks a holistic consideration of platform mechanisms

© 2025 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.TJ30220

and the information ecosystem, making it challenging to address cross-platform dissemination risks [4, 5].

In response to these issues, researchers have made some explorations. Grover proposed the concept of 'generation as detection' [1], discovering the possibility that generative models themselves can act as detectors. DetectGPT innovatively leverages the probability curvature formula to achieve zero-shot detection [3]. HaloScope further explores self-checking methods in scenarios without labeled data [2]. At the same time, ACL-AIGT systematically collects large amounts of platform data and quantifies the distribution of AI-generated texts on real platforms [4]. Finally, experiments by Truth Decay reveal that generated content may alter the exposure of real news through recommendation systems, leading to the problem of 'truth decay' [5]. These works have conducted research in terms of methods, data, and platforms, respectively. However, there is still a lack of a systematic review to integrate their logical relationships and point out future trends.

It can be concluded from these three points that research on LLMS and social networks should accept a multifaceted approach. So, the research goal of this paper is to review and summarize current studies on the generation-detection-propagation chain. Based on this research line, three core questions need to be addressed: how to detect model-generated content using algorithms and probabilistic modeling [1-3], how model-generated content is distributed and spread on social platforms [6], and what impact model-generated content has on the information ecosystem and recommendation mechanisms [5-7]. In summary, it is not just a technical issue; it also involves the construction, stability, and maintenance of the information ecosystem [3, 6, 8]. This paper aims to provide insights into subsequent academic research and platform governance through a systematic review and analysis.

2. Generation and detection of Large Language Models in social networks

With the widespread adoption of LLMs, the text they generate has reached unprecedented scale and diversity on social networks. However, effectively distinguishing machine-generated text from human-authored text within this vast amount of content remains a core challenge in both academia and industry. In recent years, researchers have proposed a range of detection methods, including probabilistic modeling for generation-as-detection, verification based on semantic consistency, and enhanced retrieval through integration with external knowledge bases. Five representative approaches are introduced below.

2.1. Grover: generate and detect

Grover is the first model to propose the concept of “generate and detect systematically” [1]. It believes that training a powerful generative model can not only generate fake news but also detect fake news. Its core idea is that when a language model predicts its own generated text, its probability distribution is often more concentrated, while when predicting human-written text, the probability distribution is more dispersed. Grover unifies the generation and detection tasks into one objective function:

$$L = L_{gen} + \lambda L_{det} \tag{1}$$

Lgen is the generation language modeling loss, Ldet is the detection binary classification loss, and λ is a weight hyperparameter.

DOI: 10.54254/2755-2721/2026.TJ30220

Through this joint optimization, Grover can perform detection without relying on a large amount of labeled data. Experimentally, Grover's detection accuracy on news datasets is close to 90%, significantly exceeding that of traditional n-gram or SVM-based methods. This shows that the generative model "knows" to a certain extent what kind of text it has generated and therefore has detection potential. Advantages: No additional large-scale labeled data is required. It performs well on structured texts such as news. Disadvantages: It has poor generalization ability for short texts and comment data on social media. Reliance on training data makes it susceptible to domain shift.

Grover's proposal laid the foundation for subsequent research, demonstrating the possibility that generative models themselves can be used as detectors, rather than being designed independently of them.

2.2. DetectGPT: zero-shot detection based on probability curvature

DetectGPT proposed a zero-shot detection method to address the problem of scarce labeled data [3]. The core idea is that the text generated by the language model is usually located in the high-density area of the training data, so a slight perturbation will lead to a significant decrease in probability, while human-written text is more robust to such perturbations. The specific formula is:

$$S(x) = logP_\theta(x) - E_{x'\sim N(x)}\left[\log P_\theta(x')\right] \tag{2}$$

Here, P is the log-likelihood of text x under parameter θ; N(x) is the neighborhood distribution after a slight perturbation of the original text x, such as replaced by synonyms or adjusting the word order; and S(x) which is known as the "curvature score." is used to measure the difference in probability distribution between the original and perturbed texts which is known as the "curvature score."

Intuitively, it will cause a larger S(x) because perturbations cause a sharp drop in probability when a model generates text. However, human text often falls into areas not fully covered by the model, where perturbations have a more minor impact on probability, resulting in a lower S(x). The advantage of this method is that people can directly leverage the probability distribution of the language model for detection without data. However, its drawback is its high computational overhead, especially when dealing with long texts and multi-turn conversations, as it requires the generation of many perturbation samples, which is inefficient.

2.3. HaloScope: bootstrapped hallucination detection

HaloScope addresses the difficulty of obtaining large-scale, manually annotated hallucination data in the real world by proposing a bootstrapped detection method [2]. It achieves low-cost, highly adaptable hallucination detection through a "pseudo-labeling + iterative optimization" approach.

The basic process consists of four steps:

1. Collect text data generated by an in-the-wild LLM as mixed distribution samples containing both real information and hallucination components.
2. Use an encoder to map the text to a vector representation h(x).
3. Build a binary classifier, which takes the form:

$$f(x) = \sigma(W \cdot h(x) + b) \tag{3}$$

DOI: 10.54254/2755-2721/2026.TJ30220

σ is a sigmoid function, and its output probability represents the likelihood that a text contains hallucinations.

4. Through iterative training, the classifier's predictions are continuously used to update the dataset, gradually improving the classifier's accuracy.

HaloScope's advantages lie in its reduced reliance on manual annotation and its ability to perform large-scale online detection in real social network environments. It is particularly suitable for content moderation scenarios that require rapid response. However, its robustness depends on the distribution of the sample. Model performance may degrade when platform topics, language styles, or text types undergo significant changes. Furthermore, its interpretability is limited, as it only outputs probabilities and cannot pinpoint specific hallucinated segments.

2.4. MetaQA: detection based on consistency verification

MetaQA proposes a detection framework based on consistency verification [8]. Its core idea is that if the text contains hallucinations or untruths, the answers generated by the model will be inconsistent when faced with equivalent paraphrasing questions. Therefore, the detection task can be transformed into calculating a consistency score [9].

The specific steps are as follows:

1. For the input text x, a set of semantically equivalent questions is automatically generated.
2. The model to be tested generates corresponding answers.
3. The consistency score between the answers is calculated:

$$C(x) = \frac{1}{n(n-1)} \sum_{i \neq j} sim(a_i, a_j) \tag{4}$$

Where Sim is a semantic similarity function (such as cosine similarity or BERTScore). If C(X) is low, it indicates that the text lacks internal semantic consistency and is more likely to be machine-generated. MetaQA experiments on social platform datasets show that consistency metrics outperform traditional statistical feature detection in detecting political and health misinformation.

Its advantage is that it captures implicit contradictions at the semantic level. And have strong cross-lingual applicability. However, it requires high computational complexity, necessitating multiple rounds of question generation and comparison. It would also misjudge "partially consistent but overall false" in lengthy text.

MetaQA's contribution lies in introducing logical consistency into detection, opening new approaches for understanding-based detection.

2.5. DrHall: retrieval-augmented hallucination detection

Hallucination is one of the most common and risky issues in the generation of large language models. Hallucination occurs when model-generated text appears grammatically and logically correct but factually misaligns with real-world knowledge. The spread of hallucinations on social networks can amplify the impact of false information and can even be maliciously exploited to manipulate public opinion. Traditional detection methods rely on probability distribution features or semantic consistency, but often fall short when it comes to fact-checking.

Therefore, DrHall (Detecting Hallucinations via Retrieval-Augmented Language Models) proposes a retrieval-augmented detection framework that integrates external knowledge bases to verify model generation results in real time, thereby improving detection accuracy and interpretability [6].

DOI: 10.54254/2755-2721/2026.TJ30220

DrHall's basic workflow consists of three core steps. The first step is retrieval enhancement. Given an input text x, the system searches for relevant fragments in an external knowledge base (such as Wikipedia, news databases, or domain knowledge graphs) to obtain a set of candidate knowledge. The second step is consistency determination. For each candidate fact fragment in the text, DrHall calculates its similarity with the retrieved knowledge using a semantic matching function. This can be formally expressed as:

$$s\left(f_i, k_j\right) = \cos(h\left(f_i\right), h\left(k_j\right)) \tag{5}$$

Here, h (·) represents the vectorized representation of a text segment by a semantic encoder (such as BERT or RoBERTa), and cos (·) represents the cosine similarity function. If a fact segment has no corresponding item with a similarity exceeding a threshold τ in all retrieved knowledge, the segment is considered likely to be hallucinated. The third step is probabilistic modeling. DrHall assigns a hallucination probability to each segment:

$$P_{hall}\left(f_i\right) = \sigma(\max_j s\left(f_i, k_j\right) - \tau \tag{6}$$

σ is a sigmoid function. When the maximum similarity of a segment falls below a threshold, the likelihood of hallucination increases significantly. This design not only determines whether an entire text segment contains hallucinations but also accurately locates potential misinformation.

In practical applications, a significant advantage of DrHall is its strong interpretability. When a hallucinated segment is detected, the system directly displays the corresponding external knowledge source and its matching score, enabling users to understand the basis for the judgment clearly. This interpretability is particularly important in fact-checking scenarios on social networks, as the public often needs to know not only "whether it is false" but also "where it is false and why it is false." Experimental results show that DrHall significantly outperforms methods that rely solely on language model probabilities in both precision and recall for hallucination detection in open-domain question answering and news generation tasks. For example, experiments on the HotpotQA dataset show that DrHall's F1 score improves by over 12% over traditional language modeling detection methods.

However, DrHall also has limitations. First, it relies heavily on external knowledge bases. If the knowledge base is not updated promptly or its coverage is insufficient, genuine factual text may be misclassified as hallucination. Furthermore, the retrieval step incurs additional computational overhead, requiring more hardware resources for real-time detection. Furthermore, in multilingual environments or highly colloquial social media contexts, a semantic gap may exist between the standard expressions in the knowledge base and the generated text, thus affecting matching performance.

In summary, DrHall provides a more reliable and interpretable approach to detecting factual hallucinations. By introducing an external knowledge augmentation mechanism, it addresses the shortcomings of traditional fact-checking methods. It demonstrates high application value in areas such as news generation, question-answering systems, and rumor detection on social networks. With the expansion of the knowledge base and the optimization of the retrieval algorithm, DrHall's robustness and real-time performance are expected to improve further, providing critical support for ensuring the credibility of large-scale generated content.

DOI: 10.54254/2755-2721/2026.TJ30220

3. Distribution of generated content on social networks and platform influence

3.1. Data scale and benchmark construction

The ACL-AIGT study provided a dataset (SM-D) encompassing a wide range of social platforms. Based on this dataset [6], the AIGTBench benchmark was constructed to quantify the prevalence of AI-generated text and evaluate detection algorithms for it. The results show significant variations in the proportion of AI-generated content across different platforms, reflecting the influence of multiple factors, including platform user behavior, content style, and recommendation mechanisms. The study therefore emphasizes that detectors must not only be able to distinguish between AI-generated and human-generated content but also be able to withstand the complex challenges posed by style disguise and multi-model generation.

3.2. Platform-level influence and truth decay

The LLM-Generated Fake News Induces Truth Decay study explored the profound impact of generated content on recommendation systems [5]. In the experiment, researchers entered approximately 56,000 AI-generated news articles into a simulated recommendation system, then analyzed three metrics: ranking, exposure, and click-through rate. The results showed that the recommendation system tends to recommend texts whose language patterns and content distribution are closer to those of past high-click-through news. As the amount of content generated continues to grow, the distribution of recommendations has shifted significantly. Genuine news has been squeezed out, and its ranking has gradually declined, leading to the phenomenon of truth decay. This suggests that generated text not only influences users' judgments of information accuracy but also alters the platform's recommendation system itself, exposing the fragility of the information ecosystem [10]. Research shows that detection technology alone is not enough to solve this problem. Future efforts will require a comprehensive approach, including the improvement of recommendation algorithms, the addition of credibility signals, and the implementation of multimodal verification mechanisms to create a stronger governance system.

3.3. Cross-paper relationship analysis

Different studies indicate there is a high degree of complementarity across methods, data, and platforms. At the methodological level, Grover proposed the concept of generation-as-detection, which provides the theoretical foundation for subsequent research. DetectGPT uses probability curvature to achieve zero-shot detection, reducing reliance on manually annotated data. HaloScope proposes a bootstrapped method with unlabeled detection, improving its practical application value. At the data level, the ACL-AIGT project built a large-scale benchmark dataset, SM-D, based on which it established the AIGTBench benchmark, providing solid support for experimental validation and cross-comparison of detection algorithms. At the platform level, Truth Decay's research revealed the information ecosystem’s changes caused by generated text in recommendation mechanisms, highlighting the switch from detecting a single source to systematic governance. Overall, these studies form a logical research chain, from continuous innovation in detection methods to the construction of data benchmarks and finally to platform-level governance. This demonstrates the direction of research on "LLM and social networks" and suggests that it is an inevitable choice to face the multiple challenges brought by generating content effectively, to combine multiple methods and mechanisms .

DOI: 10.54254/2755-2721/2026.TJ30220

4. Challenges and outlook

Currently, the application of large language models in social networks and research on detecting generated content has made some progress, but they still face multiple challenges. First, at the technical level, existing detection methods still have limited robustness in cross-language and cross-platform scenarios. Due to the text styles under different language and cultural backgrounds having significant differences, existing models struggle to achieve consistent detection results in a multi-language environment. Furthermore, with the integration of multimodal information (such as images, videos, and audio), detection methods based on text testing face significant challenges. It is necessary to combine multimodal information with contextual modeling to assess the authenticity and consistency of content more comprehensively [11]. Furthermore, at the data level, current benchmark datasets still exhibit biases in regional, platform, and topic areas, and lack unified standards across languages and platforms, which limits the generalizability of detection algorithms in diverse ecosystems. In the future, it will be necessary to establish open, shared, and larger-scale data resources that can cover various scenarios to support the verification of algorithm universality.

At the system level, single-detection technologies cannot effectively deal with some complex communication phenomena, such as Truth Decay. AI-generated content may influence the dissemination of real news through recommendation mechanisms, emphasizing the importance of exploring the collaborative governance of detection and recommendation systems. Using credibility signals, source tracing, and multimodal verification mechanisms, individuals can optimize their detection strategy, focusing on multi-model generation and style camouflage, thereby reducing the risks of spreading false information. Furthermore, there are also significant research directions to improve the interpretability of test results and enhance user trust, which can help users better understand and trust systematic judgment. Moreover, the coordination of policy and ethical frameworks is a crucial component of future research as well. The multi-level coordination mechanism should be based on multiple levels, such as algorithm design, platform management, and social governance, to form a healthier and more transparent information ecosystem. All in all, future research should focus on the collaborative advancement of technological innovation, data development, and system governance, and promote the development of testing technology towards system-level governance

5. Conclusion

This paper systematically reviews the impact of large language models on generation, detection, and dissemination in social network environments. Firstly, from a methodological perspective, we reviewed various detection mechanisms, including Grover (based on generative probability distributions), DetectGPT (zero-shot detection using probability curvature), HaloScope (bootstrapped), MetaQA (logical consistency checking), and DrHall (causal explanation). These methods have their own advantage in terms of detection efficiency, annotation dependency, logical consistency, and interpretability. Meeting the testing requirements in different scenes. Secondly, from a data perspective, we analyze the importance of the large-scale social platform benchmark data provided by ACL-AIGT and AIGTBench for verification and testing algorithms, emphasizing high-quality data's core role in method validation

In summary, this article summarizes the current research pipeline on LLM and social networks, showing the close connection among method innovation, data development, and platform governance. Research indicates that relying solely on detection technology cannot fully address the risks brought by generated content. Future research needs to systematically explore the cross-

DOI: 10.54254/2755-2721/2026.TJ30220

language, cross-platform, and multimodal detection methods, improving the foundation of the data standardization, open sharing, and collaborative governance mechanisms, which can promote the formation of a healthy, transparent, and trustworthy information ecosystem, while it should also focus on improving the interpretability of the methods and user trust, and combine policy, ethics, and social governance framework, offering a comprehensive and sustainable solutions for the application of large language models in social networks.

References

[1] Zellers, R., Holtzman, A., Rashkin, H. et al. (2019). Defending Against Neural Fake News, NeurIPS.
[2] Wortsman, M. et al. (2023). HaloScope: Bootstrapped Hallucination Detection for Large Language Models, ICLR.
[3] Nye, B. et al., (2021). Language Models as Knowledge Bases? ACL.
[4] ACL-AIGT Consortium, (2023). SM-D: Social Media AI-Generated Text Dataset and Benchmark, ACL.
[5] Beizhe H. et al., (2025). LLM-Generated Fake News Induces Truth Decay in News Ecosystem: A Case Study on Neural News Recommendation, in *Proceedings of 11the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*.
[6] Richardson, K. et al., (2023). DrHall: Detecting Hallucinations via Retrieval-Augmented Language Models, ACL.
[7] Mitchell, E., Belinkov, Y., et al., (2023). DetectGPT: Zero-Shot Machine-Generated Text Detection using Probability Curvature, ICLR.
[8] Li, L. Z. et al., (2023). On the Detection of AI-Generated Political and Health Misinformation on Social Networks, Computers in Human Behavior.
[9] Jia, R., Liang, P. (2017). Adversarial Examples for Evaluating Reading Comprehension Systems, EMNLP.
[10] Zhao, J. et al., (2022). MetaQA: Consistency-Based Detection of Hallucinated Responses in Language Models, EMNLP.
[11] Diakopoulos, N. (2015). Algorithmic Accountability: Journalistic Investigation of Computational Power, Digital Journalism, vol. 3, no. 3, pp. 398–415.