Conference track «11.4 Strategic Communication»

Detecting AI-generated text in Chinese media contexts

Academic supervisor – Kaspiarovich-Rynkevich Olga Nikolaevna

Wang Zhihao
Student (master)
Belarusian State University, Институт журналистики, Minsk, Belarus
E-mail: wangzhihao1214@gmail.com

The rapid development of generative large language models (LLMs) has significantly transformed contemporary digital communication and content production. As the quality of machine-generated texts continues to improve, the task of distinguishing between human-written and AI-generated content has become an important research challenge. This issue is particularly relevant for natural language processing, academic integrity assessment, and digital media governance. In Chinese-language media environments, reliable detection of AI-generated texts is also essential for maintaining information credibility and regulating online communication.

Existing research identifies several major methodological approaches to AI text detection, including statistical analysis based on language-model probability features, supervised machine-learning classification, zero-shot detection methods, and watermarking mechanisms embedded during text generation. In addition, linguistic and stylistic analysis provides complementary insights into structural differences between human and machine-produced texts.

One of the earliest detection approaches relies on statistical properties of language-model outputs. These methods employ metrics such as perplexity, which reflects how predictable a given text sequence is for a language model. Machine-generated texts tend to demonstrate a higher proportion of high-probability tokens compared with human writing. A widely known tool based on this principle is GLTR, which visualizes token probability distributions in order to reveal patterns characteristic of AI-generated text [1]. However, perplexity-based methods may be less reliable when applied to short or highly standardized texts.

Another widely used approach involves supervised machine-learning classification. In this framework, labeled datasets containing both human-written and AI-generated texts are used to train detection models. For instance, the GPT-2 Output Detector employs a RoBERTa-based classifier to distinguish machine-generated content from human-authored texts [2]. Although such models can demonstrate high accuracy under controlled conditions, they often suffer from distribution dependence, meaning that their performance decreases when applied to texts generated by different models, domains, or languages [3].

Recent studies increasingly focus on zero-shot detection methods, which do not require labeled training datasets. One representative example is DetectGPT, which identifies machine-generated texts by analyzing curvature in the probability landscape of language models [4].

Another approach involves watermarking, where identifiable statistical signals are embedded directly during the generation process. Kirchenbauer et al. proposed a watermarking technique based on preferential sampling of specific token subsets, enabling subsequent statistical detection of generated texts [5]. However, this approach requires implementation by model developers and may be weakened through paraphrasing or translation [6].

The analysis demonstrates that existing detection approaches—such as statistical methods based on perplexity, supervised machine-learning classifiers, zero-shot detection techniques, and watermarking mechanisms—each provide valuable contributions but also exhibit important limitations. Statistical methods are sensitive to text length and stylistic standardization, supervised models often face generalization challenges across domains and languages, while zero-shot approaches and watermarking depend on specific technical conditions or implementation by

model developers. No single method currently ensures universally reliable detection. These limitations highlight the necessity of combining multiple analytical approaches and conducting further empirical research, particularly in Chinese-language media environments, in order to improve the robustness and practical applicability of AI-generated text detection systems.

Источники и литература

1) Gehrmann, S. GLTR: Statistical Detection and Visualization of Generated Text / S. Gehrmann [et al.] // Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations. – 2019. – P. 111–116.
2) OpenAI. GPT 2: 1.5B release. – Mode of access: https://openai.com/index/gpt-2-1-5b-release/. – Date of access: 15.02.2026.
3) Wu, J. A Survey on LLM-Generated Text Detection: Necessity, Methods, and Future Directions / J. Wu [et al.] // Computational Linguistics. – 2025. – Vol. 51. – No. 1. – P. 275–338.
4) Mitchell, E. DetectGPT: Zero-Shot Machine-Generated Text Detection using Probability Curvature / E. Mitchell [et al.] // Proceedings of the 40th International Conference on Machine Learning. – 2023. – Vol. 202. – P. 24950–24962.
5) Kirchenbauer, J. A Watermark for Large Language Models / J. Kirchenbauer [et al.] // Proceedings of the 40th International Conference on Machine Learning. – 2023. – Vol. 202. – P. 17061–17084.
6) Chen, J. Imitate Before Detect: Aligning Machine Stylistic Preference for Machine-Revised Text Detection / J. Chen [et al.] // Proceedings of the AAAI Conference on Artificial Intelligence. – 2025. – P. 23559–23567.