Review

A Review of AIGC Text Detection in Academic Papers

Jie Wang[1,*]

[1] Department of Information Management, Nanjing University of Science and Technology, Nanjing, China
* Correspondence: Jie Wang, Department of Information Management, Nanjing University of Science and Technology, Nanjing, China

Abstract: With the widespread adoption of large language models in academic writing, the detection of artificial intelligence-generated content has become an important research topic. This paper reviews the detection of AI-generated text in academic papers. It defines the relevant concepts, surveys detection methods and commercial detection tools, and introduces experimental datasets and evaluation metrics. The review shows that current detection approaches face several challenges, including a shortage of suitable datasets, ambiguous labels for human–AI collaborative text, insufficient coverage of generative models, limited generalization, and inadequate interpretability. Future research should develop datasets covering multiple languages, disciplines, models, and generation methods; strengthen evaluation on unseen models and adversarial samples; and improve explainable detection and human-review mechanisms, thereby supporting the governance of academic integrity.

Keywords: artificial intelligence-generated content; academic papers; AI-generated text detection; large language models

1. Introduction

Rapid advances in computing have led to an explosive accumulation of data and substantial improvements in computational capacity. Artificial intelligence (AI) can now not only interact with humans but also perform tasks such as writing and video production. In light of the broader history of AI, the development of artificial intelligence-generated content (AIGC) can be divided into three stages: an early embryonic stage, a period of gradual accumulation, and a stage of rapid development [1].

During the early embryonic stage (1950s–1990s), AIGC was confined to small-scale experiments; algorithms were limited, investment was modest, and no major breakthroughs were achieved. During the accumulation stage (1990s–2010s), the rapid growth of the Internet, the expansion of data, and substantial gains in computing power drove marked progress in AI algorithms, although AIGC applications remained limited. In the rapid-development stage (2010s–present), advances in deep-learning architectures such as the Transformer accelerated the development of AIGC. Companies including Meta, Microsoft, and ByteDance began exploring AIGC use cases, while platforms such as ChatGPT emerged. Demand from academic research, advertising and marketing, e-commerce, film and entertainment, software development, and office productivity has since driven the deployment of AIGC applications.

AIGC has expanded from early unimodal text and image generation to audio, video, code, digital humans, and other forms of content. At the same time, problems such as hallucinations in large language models (LLMs) [2], deepfakes [3], and information pollution [4] have created a crisis of trust, generating an urgent need for effective mechanisms to identify AI-generated content.

This paper focuses on the detection of AI-generated text in academic papers, particularly content produced by generative AI systems such as ChatGPT and DeepSeek. It covers fundamental concepts, AIGC detection techniques, datasets and evaluation frameworks, current challenges, and future research directions.

Received: 03 July 2026
Revised: 19 August 2026
Accepted: 02 September 2026
Published: 09 September 2026

2. Core Concepts

2.1. Artificial Intelligence-Generated Content

The Measures for the Management of Generative Artificial Intelligence Services (Draft for Comment), issued by the Cyberspace Administration of China [5], defines generative AI technology as technology that generates text, images, audio, video, code, and other content using algorithms, models, and rules. Wu et al [6]. regard AIGC as a new mode of content creation in which AI assists or replaces humans in producing content in response to user-provided keywords or instructions. From the perspective of content modality, Foo et al [7]. define AIGC methods as AI algorithms used to generate text, images, video, three-dimensional assets, and other media content.

The concept of AIGC has three key dimensions. First, in terms of generation, it involves the recreation of content based on generative AI's understanding of data. Second, it encompasses multiple content modalities, including text, images, audio, and video. Third, it changes the human–AI relationship by moving AI from the role of an assistive tool toward that of a content producer.

2.2. AI-Generated Text Detection

Wu et al [8]. define AI-generated text detection as a text-classification task that determines whether a given text was written by a human or generated by an AI model. Fraser et al [9]. further note that this task is typically formulated as binary AI-versus-human classification, although it may be extended to multiclass classification to distinguish degrees of AI involvement or identify the specific source model. Hong [10] considers AIGC detection a technical application that uses specialized algorithms to analyze text, images, video, audio, and other content, identify features associated with AI generation, and output a generation probability or distribution.

Accordingly, AIGC detection has two central elements: specialized algorithms as the means of detection and AI-generation features as the detection target.

The remainder of this paper focuses on AI-generated text detection, including its methodological framework, evaluation metrics, existing problems, and future challenges.

3. AIGC Detection Techniques for Academic Papers

3.1. Classification of Detection Techniques

3.1.1. Pretrained-Language-Model-Based Methods

Pretrained-language-model-based methods detect AI-generated text by fine-tuning models such as BERT and RoBERTa to capture distributional differences between human-written and AI-generated text in semantic, syntactic, stylistic, and other nonstatistical features.

In 2019, Solaiman et al [11]. fine-tuned RoBERTa to detect text generated by GPT-2. Guo et al [12]. constructed the Human ChatGPT Comparison Corpus and fine-tuned RoBERTa to detect ChatGPT-generated text. Tian et al [13]. proposed a multiscale positive–unlabeled learning framework. Wang et al [14]. combined supervised machine learning, deep pretrained models, and quantitative analysis of textual features, supplemented by plagiarism-detection software. Zhang et al [15]. integrated machine learning, deep learning, and multidimensional text-evaluation metrics, used prompts to generate comparison samples, and applied ROUGE to analyze feature differences.

However, detectors based on pretrained language models often overfit in-domain data and therefore exhibit limited generalization.

3.1.2. Statistical-Feature-Based Methods

Statistical-feature-based methods identify AI-generated text by applying decision thresholds to statistical measures or by analyzing differences in feature distributions.

Solaiman et al [11]. developed an early detector that distinguished LLM-generated from human-written text based on differences in token-probability distributions. Mitchell et al [16]. developed DetectGPT, which uses differences in average log probability between human-written and machine-generated text. Wu et al [17]. extracted text from

selected LLMs, recorded the next-token probabilities of salient n-grams as features, and calculated proxy perplexity for each model; the source model was then identified through joint analysis of these proxy-perplexity values. Focusing on academic journal texts, Shen and Wang [18] constructed a detection framework based on differences between human-written and AI-generated text in academic conventions, content logic, and writing style.

Statistical-feature-based methods offer a degree of interpretability. However, their reliance on statistical signals can limit both detection performance and generalization.

3.1.3. Watermark-Based Methods

Watermark-based methods embed a watermark by influencing the LLM decoding process or directly modifying generated text, thereby simplifying the identification of LLM-generated content.

Topkara et al [19]. increased textual ambiguity through synonym substitution so that watermarks could survive subsequent text modification. Yang et al [20]. proposed a BERT-based contextual lexical-substitution scheme. Yang, Chen, et al [21]. introduced a digital-watermarking framework for text generated by black-box language models, using binary encoding and synonym substitution to embed watermarks. Abdelnabi and Fritz [22] proposed an adversarially trained watermarking transformer (AWT) that can conceal watermarks and trace text provenance. Hu et al [23]. developed an unbiased watermarking technique that does not degrade LLM output quality and is imperceptible to users.

Watermark-based detection determines whether a text contains a watermark regardless of its genre and can therefore provide strong general applicability. Nevertheless, it is a model-specific governance mechanism. Deployment generally requires access to the source model or its generation pipeline, making independent implementation by third parties difficult.

3.2. Commercial Detection Tools

3.2.1. CNKI AIGC Detection

The CNKI AIGC Detection Service is designed to identify suspected AI-generated content in academic texts and is used by universities and journals as a reference tool for AI-content screening.

Built on CNKI's large-scale scholarly literature resources, the service uses deep-learning models to analyze features such as the frequency of logical shifts between paragraphs and changes in argumentative depth. A warning may be triggered when the depth of argumentation is unusually uniform throughout a document. CNKI also maintains a dynamically updated lexicon of AI-associated features; when the density of such expressions exceeds a threshold, the estimated likelihood of AIGC increases. The service currently supports mixed Chinese–English text detection.

3.2.2. VIP AI Detection

VIP's AIGC detector relies on three technical components: dynamic fingerprint matching, linguistic-pattern analysis, and paragraph-level classification. VIP maintains an AI-text fingerprint database containing characteristic patterns associated with systems such as DeepSeek and Kimi. During detection, linguistic features are extracted and compared with the database; highly similar text is marked as potentially AI-generated. In addition to fingerprint matching, the system analyzes linguistic patterns to identify paragraphs with limited lexical diversity, highly regular sentence structures, or unusually tight logical connections.

Unlike CNKI's document-level scoring approach, VIP assigns an independent AI-probability score to each paragraph and then aggregates these scores into an overall AI-content rate.

3.2.3. Turnitin

Turnitin is an academic-integrity tool developed for educational settings. It introduced AI-writing detection in 2023 and has since become one of the tools used by overseas universities in academic-integrity review.

Turnitin determines whether content is AI-generated by analyzing sentence-level characteristics, such as word frequency and collocation patterns, together with semantic features such as coherence and contextual logic. It reports AI-writing indicators at the sentence level. The underlying premise is that AI-generated text may display a uniform style, overly regular sentence construction, and limited individual perspective. Its detection coverage includes text generated by mainstream models such as GPT-3.5, GPT-4, GPT-4o, and Claude. Turnitin also provides multilingual similarity-detection functions that can support cross-language manuscript screening.

3.2.4. Copyleaks

Founded in 2015, Copyleaks provides a cloud-based text-analysis engine built on machine-learning techniques. The system can automatically identify and compare textual content to detect potential plagiarism. Copyleaks supports plagiarism detection in more than 100 languages and AI-content detection in more than 30 languages.

Copyleaks identifies AI-generated text by analyzing grammatical patterns, lexical repetition, and logical fluency, with particular attention to the opening and concluding sections of a document. It currently supports detection of content generated by models including GPT-4.5, Gemini 2.5, DeepSeek V3, and Claude 4 Opus. Copyleaks reports may include source citations, AI-detection analysis, and suggestions for content improvement.

4. Datasets and Evaluation Framework

4.1. Public Datasets

Few public datasets have been developed specifically for detecting AI-generated text in academic papers, and most existing resources focus on English-language abstracts (As shown in Table 1).

Table 1. Public Datasets for Detecting AI-Generated Text in Academic Papers

Dataset	Academic-text coverage	Generation model/method	Size	Language
CHEAT	IEEE paper abstracts	Generation, Polish, and Mix	15,395 human abstracts; 35,304 synthetic abstracts	English
IDMGSP	Scientific papers	SCIgen, GPT-2, GPT-3, ChatGPT, and Galactica	29,000 samples	English
AI-GA	COVID-19-related paper abstracts	GPT-3 generation from paper titles	28,662 samples	English
AIGTxt	Introduction, background, and literature-review passages	Original human text, ChatGPT rewriting, and human–AI mixtures	3,000 samples in the original version	English
CAS-CS	Computer-science paper abstracts	Human–AI collaborative generation at varying levels	Approximately 55,000 samples plus a cross-model test set	English

4.1.1. Cheat

The human-written texts in the CHEAT dataset were collected from IEEE Xplore and mainly concern computer science and engineering. The dataset contains 15,395 human-written English abstracts and 35,304 ChatGPT-generated abstracts, with an average length of approximately 164 words per sample [24].

CHEAT uses ChatGPT to generate, polish, and mix human-written texts. In the Generation setting, complete abstracts are produced directly from paper titles and

keywords. In the Polish setting, ChatGPT edits or rewrites human-written abstracts. In the Mix setting, human-written and ChatGPT-generated passages are combined.

However, CHEAT is concentrated in computer science and engineering and does not cover other rhetorical sections, such as introductions and conclusions.

4.1.2. Idmgsp

The IDMGSP dataset comprises two broad categories: authentic human-authored papers and machine-generated fake papers. Each sample contains an abstract, an introduction, and a conclusion, and the dataset contains 29,000 samples in total [25] (As shown in Table 2).

Table 2. Sources and Sizes of the Idmgsp Dataset

Source	Number of samples	Description
arXiv parsing set 1 (authentic)	12k	Original arXiv papers; sections extracted from PDFs using PyMuPDF
arXiv parsing set 2 (authentic)	4k	A separate PDF-parsing pipeline used to exclude parsing artifacts
SCIgen (fake)	3k	Generated with a context-free grammar (CFG)
GPT-2 (fake)	3k	Three fine-tuned GPT-2 models separately generate abstracts, introductions, and conclusions
Galactica (fake)	3k	A scientific LLM using chained generation, with each later section conditioned on preceding sections
ChatGPT (fake)	3k	A paper title is provided once to generate all three sections
GPT-3 (fake)	1k	Fine-tuned GPT-3 with chained iterative generation
Total authentic papers	16k	
Total machine-generated papers	13k	
Total	29k	

4.1.3. AI-GA

AI-GA was developed specifically for the detection of scientific-paper abstracts and is entirely focused on COVID-19 [26]. Paper titles and abstracts were collected, and GPT-3 Davinci was prompted with each title alone to generate a corresponding abstract. The final dataset contains equal numbers of authentic and generated abstracts (As shown in Table 3).

Table 3. Composition of the AI-GA Dataset

Type	Number
Authentic human-written abstracts	14,331
Machine-generated abstracts	14,331
Total	28,662

AI-GA nevertheless has several limitations. First, it is restricted to the medical domain and specifically to COVID-19, with no coverage of other disciplines. Second, it contains only abstracts rather than introductions, conclusions, or full-text sections. Finally, all generated samples were produced by a single model, GPT-3 Davinci.

4.1.4. AIGTxt

AIGTxt covers ten fields, including computer science and AI, medicine and health, and the social sciences and humanities [27]. It draws on abstracts, introductions, and literature-review passages from papers published before 2022 and uses ChatGPT to generate corresponding samples. The resulting dataset contains three classes: human-written text, text rewritten by ChatGPT, and mixed human–AI text (As shown in Table 4).

Table 4. Composition of the AIGTxt Dataset

Type	Description	Number
Human samples	Original passages from human-authored papers	1,000
Machine samples	Corresponding passages rewritten by ChatGPT	1,000
Mixed samples	Mixtures of human-written and ChatGPT-generated text	1,000
Total		3,000

4.1.5. Cas-cs

CAS-CS is a computer-science academic-writing dataset introduced in 2025 [28]. Rather than providing binary human-versus-AI labels, it assigns continuous human-involvement scores ranging from 0 to 1 for the evaluation of human–AI collaborative writing. To produce each prompt, Z sentences are randomly selected from an abstract and appended to a fixed template; varying the amount of human-provided information produces a continuum of involvement levels (As shown in Table 5).

Table 5. Composition of the Cas-cs Dataset

Type	Description	Number
CAS-CS main dataset	Generated using ChatGPT, with continuous human-involvement scores from 0 to 1	55,000
Cross-model generalization test subset	Generated using Claude, Gemini, GPT-4, and Falcon	1,000
Total		56,000

Several datasets that were not created specifically for academic papers can nevertheless support pretraining, external evaluation, or cross-domain generalization experiments, including M4, HC3, HC3 Plus, OpenLLMText, the TweepFake Dataset, and the GPT-2 Output Dataset.

Existing public datasets generally use two approaches to generate LLM-written text. The first prompts a model to produce a complete text directly. The second provides an opening sentence and instructs the LLM to continue the passage. Among the public academic datasets identified in this review, none is designed specifically for Chinese academic papers. There is also a lack of datasets that simultaneously cover all major sections of a paper and multiple generation settings, including direct generation, rewriting and polishing, and adversarial generation.

4.2. Evaluation Framework

An evaluation framework for AI-generated text detection is a standardized set of methods used to measure a detector's ability to distinguish human-written from machine-generated text. Common metrics include accuracy, precision, recall, F1 score, and the area under the receiver operating characteristic curve (AUROC).

Accuracy provides an intuitive measure of overall effectiveness because it captures the proportion of both human-written and machine-generated texts that are classified correctly. In LLM-generated text detection, precision measures the proportion of texts predicted as AI-generated that are truly AI-generated, whereas recall measures the proportion of all AI-generated texts that are successfully identified. The F1 score is the harmonic mean of precision and recall and therefore balances false positives and false negatives [29]. AUROC summarizes performance across different decision thresholds and is especially useful for assessing robustness and behavior at different levels of detection sensitivity.

5. Current Challenges and Future Directions

5.1. Challenges in Dataset Construction

5.1.1. Scarcity of High-Quality Academic Detection Data

Most public datasets for AI-generated text detection consist of news, question-answering, and social-media content. Academic papers are underrepresented, and the

available datasets are predominantly in English, with insufficient coverage of other languages. Different sections of a paper serve distinct rhetorical functions, yet most existing datasets contain only abstracts. Consequently, detectors may fail to recognize section-specific patterns of expression.

Future datasets should cover multiple languages, disciplines, and rhetorical sections to provide a reliable foundation for cross-language and cross-domain detection.

5.1.2. Ambiguous Label Boundaries in Human–AI Collaboration

Traditional AIGC detection datasets generally apply binary labels—human-written or machine-generated—even though real-world academic writing is considerably more complex. Authors may use AI to condense content or rewrite paragraphs, or they may substantially revise an AI-generated draft through multiple rounds of human editing. Rigidly labeling all text involving AI assistance as machine-generated weakens the interpretability of the labels.

Datasets should therefore record not only the final label but also the model and version used, prompts, sampling parameters, degree of human revision, and location within the paper. They should also annotate the specific sentences and passages generated or modified by a machine.

5.1.3. Insufficient Coverage of Generative Models

Early datasets mainly used GPT models or early versions of ChatGPT to generate samples. In recent years, however, multiple model families—including Claude, Gemini, LLaMA, DeepSeek, and Qwen—have emerged, and existing datasets do not cover all of them. Linguistic patterns common to earlier models have become less pronounced in newer generations. A detector trained primarily on older-model samples may therefore lose performance as generative models evolve.

Datasets should consequently adopt open, versioned, and continuously updated designs. Metadata should include the generation date and model version, and new models and generation strategies should be incorporated regularly.

5.2. Model Bias and Adaptability

5.2.1. Limited Cross-Language Transferability

Current detectors rely heavily on English-language datasets, and the learned detection features may not transfer directly to other languages. Languages differ substantially in tokenization, morphology, word order, function-word use, and syntactic structure. Moreover, machine-translated text may display standardized phrasing and low-perplexity patterns resembling those of machine-generated text.

Cross-language evaluation should therefore go beyond testing another language after training. It should separately compare direct transfer, multilingual pretrained models, continued pretraining in the target language, and target-language fine-tuning.

5.2.2. Limited Generalization Across Models, Disciplines, and Time

Supervised detectors generally perform well on models, tasks, and topics similar to those represented in their training data. Their performance may decline substantially when the source model, text length, or discipline changes. A practical evaluation of multiple detectors by Tufts et al. showed that detectors struggle to maintain their original performance on unseen models, unseen tasks, and adversarial prompts [30].

Future studies should therefore adopt more rigorous experimental designs, including cross-model, cross-disciplinary, and cross-language evaluation, replication on independent datasets, and temporal holdout testing, with particular attention to performance under distribution shift.

5.3. Misclassification by Detection Models

5.3.1. False Classification of Text

Human-authored papers may exhibit statistical characteristics similar to those of AI-generated text because of standardized academic expression and conventional vocabulary. Abstracts and conclusions, in particular, are highly formulaic. If a detector treats low

perplexity or regular sentence patterns as direct evidence of machine generation, it may incorrectly classify some human-authored papers as machine-generated. In academic settings, such false positives may result in manuscript rejection or an investigation of academic misconduct.

Accuracy measured on a single dataset should therefore not be interpreted as evidence of a detector's universal reliability.

5.3.2. Limited Interpretability

Many detection tools report only an AIGC probability or suspected-AI proportion and do not explain which words, sentences, or discourse features support the decision. A probability score cannot by itself establish which model an author used, when it was used, or what specific operation was performed. When multiple detectors produce conflicting results for the same paper, scores without supporting explanations are difficult to evaluate or trust.

Future detection systems should identify influential features and suspicious passages and should communicate model uncertainty and scope of applicability. Samples with low confidence or inconsistent predictions across models should be referred for human review.

6. Conclusions and Outlook

6.1. Conclusions

This paper systematically reviewed the concepts, detection techniques, commercial tools, datasets, evaluation metrics, and current challenges associated with detecting AI-generated text in academic papers. The principal conclusions are as follows.

First, AI-generated text detection is moving beyond binary human-versus-machine classification toward finer-grained tasks, including generated-span localization, generation-method identification, model attribution, and estimation of the degree of AI involvement. Second, each existing approach has strengths and limitations. Pretrained models can capture deep semantic features but are constrained by their training data and application domains. Statistical methods offer some interpretability but are sensitive to text length, model version, and generation parameters. Watermarking provides strong provenance authentication but requires cooperation from platforms or model providers. No universal approach currently performs reliably across all models, disciplines, languages, and generation methods. Third, datasets for detecting AI-generated academic text remain inadequate; common limitations include monolingual coverage, insufficient representation of generative models, incomplete coverage of paper sections, and the absence of complex generation settings. Fourth, existing evaluation frameworks do not fully reflect practical detector performance, particularly generalization across models, disciplines, and time.

Overall, AI-generated text detection now encompasses several technical approaches, but it remains part of an evolving contest between generation and detection technologies. Limited generalization, difficulty identifying complex forms of AI-assisted text, insufficiently interpretable results, and the shortage of Chinese academic data are the principal challenges that must be addressed.

6.2. Outlook

This review has two principal limitations. First, it does not empirically evaluate specific detection algorithms and therefore cannot directly compare the performance of different technical approaches. Second, it does not examine in depth the ethical and legal issues surrounding AI-use disclosure, author responsibility, intellectual property, and appeals against detection results.

Future research can address these limitations in the following areas.

First, researchers should develop high-quality datasets covering multiple disciplines, rhetorical sections, generative models, and generation methods, while recording model versions, prompts, generation parameters, and the degree of human revision. Second, detector generalization and robustness should be improved through evaluation across

models, disciplines, sections, and time, with emphasis on identifying text from unseen models, machine-rewritten content, and adversarial samples. Third, interpretability should be strengthened by locating suspicious passages, reporting AI-generation probabilities, and explaining the basis of each decision. Fourth, a standardized evaluation and governance framework should be established. In addition to basic classification metrics, evaluations should report the false-positive rate for human-written text, the false-negative rate for AI-generated text, and performance degradation across scenarios.

As datasets, detection algorithms, and watermark-based authentication technologies continue to improve, AIGC detection may become an important supporting tool for academic-integrity governance. It can enhance the credibility and standardization of academic-content governance and contribute to a healthy and orderly scholarly ecosystem.

References

1. China Academy of Information and Communications Technology and JD Discovery Research Institute, *White Paper on Artificial Intelligence Generated Content (AIGC)*, Beijing, China, 2022.
2. Y. Wu, X. Qiu, Y. Wang, et al., "Deconstruction of the causes and governance of information pollution triggered by large language model hallucinations," *Journal of Intelligence*, 2025. [Online]. Available: https://link.cnki.net/urlid/61.1167.G3.20251208.0922.012
3. J. Qiu, T. Zhang, and Z. Xu, "Collaborative governance strategies for AIGC-generated false information based on a tripartite evolutionary game," *Library and Information Service*, vol. 70, no. 2, pp. 71–85, 2026, doi: 10.13266/j.issn.0252-3116.2026.02.006.
4. J. Chu and X. Fan, "Multidimensional causes and governance strategies of information pollution in the AIGC environment," *Information Studies: Theory & Application*, vol. 49, no. 2, pp. 38–46, 2026, doi: 10.16353/j.cnki.1000-7490.2026.02.005.
5. Cyberspace Administration of China, *Measures for the Management of Generative Artificial Intelligence Services (Draft for Comment)*, Apr. 11, 2023. [Online]. Available: https://www.cac.gov.cn/2023-04/11/c_1682854275475410.htm
6. J. Wu, W. Gan, Z. Chen, et al., "AI-generated content (AIGC): A survey," arXiv preprint arXiv:2304.06632, 2023. [Online]. Available: https://arxiv.org/abs/2304.06632
7. L. G. Foo, H. Rahmani, and J. Liu, "AI-generated content (AIGC) for various data modalities: A survey," *ACM Computing Surveys*, vol. 57, no. 9, Art. no. 243, 2025, doi: 10.1145/3728633.
8. J. Wu, S. Yang, R. Zhan, et al., "A survey on LLM-generated text detection: Necessity, methods, and future directions," *Computational Linguistics*, vol. 51, no. 1, pp. 275–338, 2025, doi: 10.1162/coli_a_00549.
9. K. C. Fraser, H. Dawkins, and S. Kiritchenko, "Detecting AI-generated text: Factors influencing detectability with current methods," *Journal of Artificial Intelligence Research*, vol. 82, pp. 2233–2278, 2025, doi: 10.1613/jair.1.16665.
10. T. Hong, "Functional reflection and normative application of AIGC detection," *Studies in Science of Science*, 2026. [Online]. Available: https://doi.org/10.16192/j.cnki.1003-2053.20260224.002
11. I. Solaiman, M. Brundage, J. Clark, et al., "Release strategies and the social impacts of language models," arXiv preprint arXiv:1908.09203, 2019. [Online]. Available: https://arxiv.org/abs/1908.09203
12. B. Guo, X. Zhang, Z. Wang, et al., "How close is ChatGPT to human experts? Comparison corpus, evaluation, and detection," arXiv preprint arXiv:2301.07597, 2023. [Online]. Available: https://arxiv.org/abs/2301.07597
13. Y. Tian, H. Chen, and X. Wang, "Multiscale positive-unlabeled detection of AI-generated texts," in *Proc. 12th Int. Conf. Learn. Represent.*, 2024.
14. Y. Wang, X. Guo, Z. Liu, et al., "Detection and comparative analysis of Chinese paper abstracts generated by AI and written by scholars," *Journal of Intelligence*, vol. 42, no. 9, pp. 127–134, 2023.
15. Q. Zhang, X. Wang, and Y. Gao, "A comparative study of literature abstracts generated by ChatGPT and written by scholars: A case study of the information resource management field," *Library and Information Service*, vol. 68, no. 8, pp. 35–47, 2024, doi: 10.13266/j.issn.0252-3116.2024.08.004.
16. E. Mitchell, Y. Lee, A. Khazatsky, et al., "DetectGPT: Zero-shot machine-generated text detection using probability curvature," in *Proc. 40th Int. Conf. Mach. Learn.*, Honolulu, HI, USA: PMLR, 2023, pp. 24950–24962.
17. K. Wu, L. Pang, H. Shen, et al., "LLMDet: A third party large language models generated text detection tool," in *Findings of the Association for Computational Linguistics: EMNLP 2023*, Singapore: Assoc. Comput. Linguistics, 2023, pp. 2113–2133, doi: 10.18653/v1/2023.findings-emnlp.139.
18. X. Shen and L. Wang, "Detection of artificial-intelligence-generated academic journal texts," *Science-Technology & Publication*, no. 8, pp. 56–62, 2023, doi: 10.16510/j.cnki.kjycb.2023.08.002.
19. S. Abdelnabi and M. Fritz, "Adversarial watermarking transformer: Towards tracing text provenance with data hiding," in *2021 IEEE Symp. Secur. Privacy*, Los Alamitos, CA, USA: IEEE Comput. Soc., 2021, pp. 121–140, doi: 10.1109/SP40001.2021.00083.
20. U. Topkara, M. Topkara, and M. J. Atallah, "The hiding virtues of ambiguity: Quantifiably resilient watermarking of natural language text through synonym substitutions," in *Proc. 8th Workshop Multimedia Secur.*, New York, NY, USA: ACM, 2006, pp. 164–174, doi: 10.1145/1161366.1161397.

21. X. Yang, J. Zhang, K. Chen, et al., "Tracing text provenance via context-aware lexical substitution," in *Proc. AAAI Conf. Artif. Intell.*, vol. 36, no. 10, 2022, pp. 11613–11621, doi: 10.1609/aaai.v36i10.21415.
22. X. Yang, K. Chen, W. Zhang, et al., "Watermarking text generated by black-box language models," arXiv preprint arXiv:2305.08883, 2023. [Online]. Available: https://arxiv.org/abs/2305.08883
23. Z. Hu, L. Chen, X. Wu, et al., "Unbiased watermark for large language models," arXiv preprint arXiv:2310.10669, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2310.10669
24. P. Yu, J. Chen, X. Feng, et al., "CHEAT: A large-scale dataset for detecting ChatGPT-written abstracts," *IEEE Trans. Big Data*, vol. 11, no. 3, pp. 898–906, 2025, doi: 10.1109/TBDATA.2025.3536929.
25. E. Mosca, M. H. I. Abdalla, P. Basso, et al., "Distinguishing fact from fiction: A benchmark dataset for identifying machine-generated scientific papers in the LLM era," in *Proc. 3rd Workshop Trustworthy Nat. Lang. Process.*, Toronto, ON, Canada: Assoc. Comput. Linguistics, 2023, pp. 190–207, doi: 10.18653/v1/2023.trustnlp-1.17.
26. P. C. Theocharopoulos, P. Anagnostou, P. Tsakanikas, et al., "Detection of fake generated scientific abstracts," arXiv preprint arXiv:2304.06148, 2023. [Online]. Available: https://arxiv.org/abs/2304.06148
27. B. Alhijawi, R. Jarrar, A. AbuAlRub, et al., "Deep learning detection method for large language models-generated scientific content," *Neural Comput. Appl.*, vol. 37, pp. 91–104, 2025, doi: 10.1007/s00521-024-10538-y.
28. Y. Guo, Z. Dou, H. H. Nguyen, et al., "Measuring human involvement in AI-generated text: A case study on academic writing," arXiv preprint arXiv:2506.03501, 2025. [Online]. Available: https://arxiv.org/abs/2506.03501
29. M. Sokolova, N. Japkowicz, and S. Szpakowicz, "Beyond accuracy, F-score and ROC: A family of discriminant measures for performance evaluation," in *AI 2006: Advances in Artificial Intelligence*, Berlin, Germany: Springer, 2006, pp. 1015–1021, doi: 10.1007/11941439_114.
30. B. Tufts, X. Zhao, and L. Li, "A practical examination of AI-generated text detectors for large language models," in *Findings of the Association for Computational Linguistics: NAACL 2025*, Albuquerque, NM, USA: Assoc. Comput. Linguistics, 2025, pp. 4839–4856, doi: 10.18653/v1/2025.findings-naacl.271.