Detection of AI-generated Academic Papers

Rachel So
rachel.so@4open.science

2025

1 Introduction/Background

The detection of AI-generated academic papers has become a critical concern in scholarly publishing as artificial intelligence technologies continue to advance. The issue of algorithmically generated scientific papers is not new—it dates back to 2005 when SCIgen was created to produce nonsensical papers that could pass peer review [11]. Despite the relative simplicity of these early generation methods, such papers have continued to appear in respected publications years later, depleting valuable reviewing resources and undermining trust in the scientific review process [11] [8].

Recent advancements in natural language generation (NLG) have dramatically increased the sophistication of AI-generated content. Modern AI tools can now produce coherent and seemingly authentic texts, including scholarly papers that are difficult to distinguish from human-authored content [30]. This evolution presents more complex challenges for academic publishers, who are struggling to adapt to the rapid and pervasive adoption of advanced AI text generators like ChatGPT [13].

The concerns extend beyond just fraudulent academic papers. In educational contexts, NLG models may be employed by students to cheat on language learning assignments through machine translation or to effortlessly produce essays on assigned topics [11] [27] [12] [15]. These issues have prompted many academic journals to update their publication guidelines, typically banning AI systems as authors and requiring disclosure when AI tools are used in content creation [13].

As AI-generated research becomes increasingly prevalent, distinguishing between human-authored and AI-assisted content is critical for maintaining transparency, academic integrity, and reproducibility in scholarly work [7]. The lack of reliable detection methods has created an urgent need for effective tools and standards to identify AI-generated content, particularly in professional academic writing [13]. Addressing these challenges requires robust detection techniques and potentially new attribution standards to ensure the integrity of scholarly literature in the age of advanced AI text generation.

The emergence of AI-generated academic papers represents a significant challenge to academic integrity, with concerns dating back to 2005 when SCIgen was created to produce nonsensical papers. As advanced language models like ChatGPT have become more sophisticated, the detection of AI-generated content has become increasingly important for maintaining research standards and transparency.

2 Methods and Approaches for Detection

The detection of AI-generated academic papers relies on a variety of methods that analyze different aspects of text to identify machine-generated content. One prominent approach examines linguistic patterns through metrics such as "perplexity" and "burstiness." Perplexity measures how similar the text is to what the AI language model has previously encountered, while burstiness analyzes the variability of sentences [6]. These features form the basis of detection tools like GPTZero, developed by a Princeton University student, which has received considerable attention for its ability to distinguish AI-generated content [6] [31].

Technical analysis methods extend beyond these basic metrics to include stylometric analysis, examining vocabulary, grammar, and writing style patterns characteristic of AI-generated text [18]. Some researchers are developing more sophisticated approaches that analyze logic flows within full paragraphs or entire papers, moving beyond sentence-level analysis [31]. Another novel approach called "self-detection" explores whether AI systems can identify their own output when presented with a mix of human and AI-generated texts, though this method has shown limitations [9].

AI detection models vary significantly in their performance. DetectGPT, for example, has demonstrated an ability to distinguish between human-written and large language model (LLM)-generated text with over 95% accuracy in controlled tests, though it requires specific tuning to the source model being evaluated [6]. In contrast, OpenAI's AI Text Classifier has shown much lower accuracy at around 26%, suggesting it should be used alongside other detection methods [31]. This variability in performance presents a challenge for reliable detection.

When tested against human reviewers, AI detection tools have shown promising results. In a study where both AI detection software and human reviewers evaluated abstracts created by ChatGPT, the AI detection tools outperformed human reviewers who only identified 68% of AI-generated abstracts correctly [25] [17]. Traditional plagiarism detection tools, however, proved almost completely ineffective at identifying AI-generated content [25].

More advanced transformer-based models are increasingly being employed for detection tasks. Researchers have developed BERT-based detectors specifically for identifying AI-generated academic texts, with one study reporting 84% accuracy in detecting Chinese academic papers generated by AI [37]. These

neural-network-based binary classifiers represent a promising direction for automated detection, especially when trained on domain-specific content.

For journal editors and reviewers, a multi-layered approach is recommended that combines specialized AI detection tools with updated review guidelines and author disclosure requirements [20]. This comprehensive strategy is necessary because detection is increasingly challenging due to AI's ability to generate content that lacks obvious errors, can adapt to various writing styles, covers a broad range of subjects, and produces text at high speed [18].

Despite these advances, detection remains an evolving challenge characterized as a "cat-and-mouse game" between detection systems and increasingly sophisticated text generators [31]. As detection techniques improve, so do the text-generating models, necessitating continuous development and refinement of detection approaches to maintain the integrity of academic publishing [30] [36] [26].

Current AI-generated text detection employs multiple approaches including technical analysis of linguistic patterns, stylometric features, and specialized AI models trained to identify machine-generated content. Detection techniques range from examining perplexity and burstiness to more advanced methods like BERT-based classifiers, though these systems continue to evolve in a cat-and-mouse game with increasingly sophisticated text generators.

3 Existing AI Detection Tools and Their Performance

Several AI detection tools have emerged to address the challenge of identifying AI-generated academic content, though their performance varies considerably. Among the most effective solutions, Originality.AI has demonstrated high accuracy rates between 94-98% across multiple studies when identifying content from various AI models including GPT-3, GPT-J, and GPT-NEO [2] [1]. This tool employs a specialized version of the BERT model specifically designed to detect AI-written content with minimal bias [2] [24].

Other high-performing tools include Copyleaks and TurnItIn, which along with Originality.AI have demonstrated strong accuracy in distinguishing between both GPT-3.5 and GPT-4 generated content from human writing [28]. In contrast, most other detection tools perform reasonably well with GPT-3.5 content but struggle significantly with identifying more sophisticated GPT-4 generated papers [28].

The GPT-2 Output Detector has shown promising results in some studies, with one evaluation reporting 94% accuracy in distinguishing between human-written and AI-generated medical research abstracts [19] [16]. In this study, AI-generated abstracts were assigned high "false" scores (median 99.98%) compared to human-written abstracts (median 0.02%), demonstrating strong discriminatory ability [1] [17].

However, a comprehensive evaluation of AI detection tools presents a less optimistic picture. When examining 12 publicly available tools and two commercial systems (Turnitin and PlagiarismCheck), researchers concluded that most available detection tools were "neither accurate nor reliable" and showed bias toward classifying output as human-written rather than detecting AI-generated text [29]. This aligns with findings that OpenAI's own AI text classifier demonstrated low accuracy rates [13].

A concerning issue with many detection tools is their potential bias against non-native English speakers, with seven common web-based AI detection tools more frequently identifying non-native English writers' works as AI-generated compared to native English speakers' writing [13]. This bias appears related to "perplexity" measures used in many detection systems, highlighting the need for alternative approaches.

Some researchers have explored using alternative machine learning approaches, with one study comparing Logistic Regression, Support Vector Machine, and Naive Bayes classifiers for AI-content detection. Logistic regression performed best with accuracy exceeding 90%, while SVM achieved approximately 70% accuracy, and Naive Bayes performed poorly at around 50% [34].

More innovative approaches are also being developed, such as a method that compares the semantic similarity of peer reviews to reference AI-generated reviews for the same paper, which reportedly outperforms existing approaches in detecting GPT-4o and Claude-written reviews [33].

When tested against human reviewers, AI detection tools often outperform humans in identifying AI-generated content. In one notable study, human reviewers correctly identified only 68% of ChatGPT-generated abstracts while incorrectly flagging 14% of human-written abstracts as AI-generated [1] [17]. This finding underscores the difficulty humans face in differentiating between AI and human-written academic content.

Current detection tools also face challenges with content obfuscation techniques, which can significantly worsen their performance [29]. Additionally, traditional plagiarism detection tools have proven ineffective for identifying AI-generated content, highlighting the need for specialized AI detection systems [1].

While academic publishers and editors increasingly turn to AI content detection tools to maintain academic integrity [5], their effectiveness remains inconsistent, with ongoing research needed to enhance their accuracy and reliability [5] [14]. The integration of these tools into academic writing workflows continues to evolve, with most platforms showing similar sensitivity levels currently but expected to improve with future upgrades [21].

Various AI detection tools have been developed with significantly different accuracy levels, ranging from highly effective platforms like Originality.AI (94-98% accuracy) to less reliable tools like OpenAI's classifier. Testing shows most tools struggle with detecting output from more advanced models like GPT-4 and often produce concerning levels of false positives, par-

ticularly with non-native English writers.

4 Challenges and Limitations in AI Text Detection

Despite the development of various AI detection tools, significant challenges and limitations hinder their effectiveness in academic settings. A fundamental limitation is the accuracy of these tools, with studies concluding that most available detection systems are "neither accurate nor reliable" and tend to demonstrate bias toward classifying output as human-written rather than identifying AI-generated content [29]. Even OpenAI's own AI Text Classifier shows limited effectiveness with a success rate of only about 26%, though it may provide some value when used alongside other detection methods [31].

The dynamic nature of AI development creates an ongoing challenge for detection systems. As new language models continue to emerge, detection tools must constantly adapt, since classifiers are often specific to particular LLM models [6]. While some tools like DetectGPT have shown strong performance with over 95% accuracy in controlled tests with specific models, they require tuning to the source model being evaluated, making comprehensive coverage increasingly difficult as the number of potential LLMs grows [6].

Content obfuscation techniques present another significant challenge, with research showing these methods can substantially diminish the performance of detection tools [29]. Paraphrasing approaches, in particular, have demonstrated effectiveness in evading detection systems, highlighting vulnerabilities in current methodologies [9]. Furthermore, traditional plagiarism detection software has proven largely ineffective at identifying AI-generated text, leaving a significant gap in existing academic integrity systems [6].

The sophistication of newer AI models creates additional challenges. Current detection tools primarily analyze vocabulary, grammar, and sentence-level style patterns, but researchers note the need to develop more advanced approaches that examine logic flows across full paragraphs or entire papers [31]. This situation creates what researchers describe as a "cat-and-mouse game" between detection systems and increasingly sophisticated text generators, similar to the ongoing battle between cybersecurity professionals and attackers [31].

Practical implementation challenges also exist for academic publishers and institutions. Journal editors and reviewers often lack awareness about the latest developments in AI text generation and detection technologies, and the time constraints of the review process may prevent the in-depth analysis needed to reliably identify AI-generated content [20]. Many editors and reviewers also lack access to specialized AI detection tools, particularly when dealing with high volumes of submissions [20].

Novel approaches to detection continue to emerge, including self-detection methods that attempt to leverage AI systems' ability to identify their own output when presented with mixed human and AI-generated texts. However, research has demonstrated significant limitations with this approach, reinforcing

the need for more robust detection techniques [9]. For detection systems to be effective, they must analyze language patterns across multiple levels, including syntax, semantics, and pragmatics, while continuously adapting to advances in generation technology [6].

> Current AI detection tools face significant accuracy limitations, with most demonstrating bias toward classifying content as human-written and struggling particularly with newer language models like GPT-4. Detection systems must continuously evolve in response to increasingly sophisticated AI text generators, creating an ongoing technological arms race.

5 Datasets and Benchmarks for Evaluation

The development of effective AI detection methods requires specialized datasets and benchmarks designed specifically for evaluating detection performance in academic contexts. To address this need, Liyanage et al. created two datasets consisting of artificially generated research content: a completely synthetic dataset generated by GPT-2 and a hybrid dataset with partial text substitution where sentences in original abstracts were replaced with machine-generated content. When evaluating detection methods on these datasets, existing state-of-the-art classification models achieved a maximum accuracy of only 70.2%, highlighting the significant challenges in this domain and the substantial room for improvement in detection accuracy. [22]

For the Chinese academic community, which includes millions of researchers producing extensive literature, Zhu et al. constructed a large-scale Chinese academic paper dataset called TK2A. Using this dataset, they trained a BERT-based detector to distinguish between AI-generated and authentic Chinese academic texts. Their experiments demonstrated promising results with detection accuracy reaching 84%, verifying the feasibility of automated detection in non-English academic contexts. [37]

Other researchers have focused on creating specialized datasets for modern AI models. Alhijawi et al. developed AIGTxt, described as the first dataset specifically designed to enhance AI-generated scientific text detection tools, with a particular focus on ChatGPT-generated content. Along with this dataset, they introduced AI-Catcher, a detection method utilizing deep learning and natural language processing (NLP) techniques specifically aimed at identifying academic scientific text generated by ChatGPT. [4]

The academic community has also established dedicated competitions to accelerate progress in detection capabilities. The "DAGPap24: Detecting Automatically Generated Scientific Papers" competition, held in conjunction with the 4th Workshop on Scholarly Document Processing at ACL 2024, challenged participants to build detection models that could accurately distinguish between human-written fragments, synonym replacement fragments, ChatGPT rewrites, and generated summaries within scientific papers. [35]

Similarly, the Academic Essay Authenticity Challenge was organized as part

of the GenAI Content Detection shared tasks associated with COLING 2025. This challenge specifically focused on distinguishing between machine-generated and human-authored academic essays, with a straightforward task definition: "Given an essay, identify whether it is generated by a machine or authored by a human." [10] The challenge incorporated both English and Arabic essays, reflecting the multilingual nature of academic writing and the need for detection methods that work across different languages. [3]

These datasets and benchmarking initiatives represent crucial resources for advancing the field of AI-generated content detection in academic contexts. By providing standardized evaluation frameworks, they enable researchers to compare different detection approaches and track progress over time, ultimately supporting the development of more robust and reliable detection methods.

Several specialized datasets have been developed to advance research in AI-generated academic content detection, including synthetic paper collections, hybrid datasets with partially machine-generated content, and language-specific benchmarks. These datasets serve as critical evaluation tools for detection methods, with current benchmark competitions showing detection accuracies ranging from 70-84%, indicating significant room for improvement.

6 Future Directions and Recommendations

As AI-generated academic content becomes increasingly sophisticated, future detection methods will need to evolve beyond current approaches to maintain academic integrity. One promising direction involves the development of advanced AI-driven plagiarism detection technologies specifically designed to identify the subtle characteristics and structures unique to AI-generated text. These next-generation detection tools could significantly alter the current landscape by focusing on features that conventional plagiarism checkers typically miss during review phases [23].

An innovative approach that shows potential is based on semantic similarity comparison. Rather than focusing solely on textual patterns, this method compares a given document to reference AI-generated content for the same subject matter. Research has demonstrated that this approach outperforms existing detection methods when identifying peer reviews written by advanced models like GPT-4o and Claude, suggesting a viable path forward for detection technology [33]. This approach is particularly important as most current detection methods struggle to robustly identify AI-generated content while maintaining low false positive rates [33].

Beyond improved detection technologies, structural solutions may also be necessary. The Research Attribution Markup Language (RAML) has been proposed as a standardized method for tagging AI-generated content within academic manuscripts. Unlike existing solutions for image watermarking or document metadata, RAML specifically addresses the need for systematic at-

tribution in text-based research outputs, potentially providing a framework for transparency, academic integrity, and reproducibility [7].

The development of more specialized evaluation frameworks represents another important future direction. Current research has established methods for creating AI-generated peer reviews for academic papers and evaluating various detection models on these datasets [32]. This approach of developing specialized datasets and evaluation methods for specific academic contexts (like peer review) could be expanded to other areas of scholarly communication.

To effectively address the challenges of AI-generated academic content, a multi-faceted approach will likely be necessary, combining improved detection technologies, standardized attribution frameworks, and potentially new publication processes that accommodate the reality of AI assistance while maintaining transparency and academic integrity . As the capabilities of language models continue to advance, ongoing research in detection methodologies will remain essential to preserve trust in scholarly communication.

Future advancements in AI detection will require both technical innovations like specialized AI-driven detection systems and potential structural changes to academic publishing processes. Proposed solutions include semantic similarity-based detection methods, standardized attribution frameworks like Research Attribution Markup Language (RAML), and improved AI-driven plagiarism detection technologies.

Acknowledgements

Generative AI has been used to prepare this manuscript.

References

[1] Arslan Akram. An Empirical Study of AI Generated Text Detection Tools. *Advances in Machine Learning & Artificial Intelligence*, 2023.

[2] Arslan Akram. Quantitative Analysis of AI-Generated Texts in Academic Research: A Study of AI Presence in Arxiv Submissions using AI Detection Tool. *arXiv.org*, 2024.

[3] Mohammad AL-Smadi. IntegrityAI at GenAI Detection Task 2: Detecting Machine-Generated Academic Essays in English and Arabic Using ELECTRA and Stylometry. *arXiv.org*, 2025.

[4] Bushra Alhijawi, Rawan Jarrar, Aseel AbuAlRub, and Arwa Bader. Deep Learning Detection Method for Large Language Models-Generated Scientific Content. *Neural computing & applications (Print)*, 2024.

[5] Ahmed S. BaHammam. Balancing Innovation and Integrity: The Role of AI in Research and Scientific Writing. *Nature and Science of Sleep*, 2023.

[6] Ahmed S. Bahammam, K. Trabelsi, S. Pandi-Perumal, and Hiatham Jahrami. Adapting to the Impact of AI in Scientific Writing: Balancing Benefits and Drawbacks while Developing Policies and Regulations, 2023.

[7] Joeran Beel, Min-Yen Kan, and Moritz Baumgart. Evaluating Sakana's AI Scientist for Autonomous Research: Wishful Thinking or an Emerging Reality Towards 'Artificial Research Intelligence' (ARI)? *arXiv.org*, 2025.

[8] G. Cabanac and C. Labbé. Prevalence of nonsensical algorithmically generated papers in the scientific literature. *J. Assoc. Inf. Sci. Technol.*, 2021.

[9] Antonio Junior Alves Caiado and Michael Hahsler. AI Content Self-Detection for Transformer-based Large Language Models. *arXiv.org*, 2023.

[10] Shammur A. Chowdhury, Hind Almerekhi, Mucahid Kutlu, Kaan Efe Keles, Fatema Ahmad, Tasnim Mohiuddin, George Mikros, and Firoj Alam. GenAI Content Detection Task 2: AI vs. Human - Academic Essay Authenticity Challenge. *arXiv.org*, 2024.

[11] Evan Crothers, N. Japkowicz, and H. Viktor. Machine-Generated Text: A Comprehensive Survey of Threat Models and Detection Methods. *IEEE Access*, 2022.

[12] N. Dehouche. Plagiarism in the age of massive Generative Pre-trained Transformers (GPT-3). *Ethics in Science and Environmental Politics*, 2021.

[13] H. Desaire, Aleesa E. Chua, Min-Gyu Kim, and David C. Hua. Accurately detecting AI text when ChatGPT is told to write like a chemist. *Cell Reports Physical Science*, 2023.

[14] Ahmed M. Elkhatat, Khaled Elsaid, and S. Almeer. Evaluating the efficacy of AI content detection tools in differentiating between human and AI-generated text. *International Journal for Educational Integrity*, 2023.

[15] Xiaocheng Feng, Ming Liu, Jiahao Liu, Bing Qin, Yibo Sun, and Ting Liu. Topic-to-Essay Generation with Neural Networks. *International Joint Conference on Artificial Intelligence*, 2018.

[16] C. Gao, Frederick M. Howard, N. Markov, E. Dyer, S. Ramesh, Yuan Luo, and Alexander T. Pearson. Comparing scientific abstracts generated by ChatGPT to original abstracts using an artificial intelligence output detector, plagiarism detector, and blinded human reviewers. *bioRxiv*, 2022.

[17] C. Gao, Frederick M. Howard, N. Markov, E. Dyer, S. Ramesh, Yuan Luo, and Alexander T. Pearson. Comparing scientific abstracts generated by ChatGPT to real abstracts with detectors and blinded human reviewers. *npj Digit. Medicine*, 2023.

[18] Lalit Gupta. Unmasking artificial intelligence (AI): Identifying articles written by AI models. *Indian Journal of Clinical Anaesthesia*, 2024.

[19] Karim Ibrahim. Using AI-based detectors to control AI-assisted plagiarism in ESL writing: "The Terminator Versus the Machines". *Language Testing in Asia*, 2023.

[20] A. Jairoun, F. El-dahiyat, G. Elrefae, S. S. Al-Hemyari, M. Shahwan, S. Zyoud, K. Hammour, and Zaheer-Ud-Din Babar. Detecting manuscripts written by generative AI and AI-assisted technologies in the field of pharmacy practice. *Journal of Pharmaceutical Policy and Practice*, 2024.

[21] S. Kar, Teena Bansal, Sumit Modi, and Amit Singh. How Sensitive Are the Free AI-detector Tools in Detecting AI-generated Texts? A Comparison of Popular AI-detector Tools. *Indian Journal of Psychological Medicine*, 2024.

[22] Vijini Liyanage, D. Buscaldi, and A. Nazarenko. A Benchmark Corpus for the Detection of Automatically Generated Text in Academic Publications. *International Conference on Language Resources and Evaluation*, 2022.

[23] Jing Miao, C. Thongprayoon, S. Suppadungsuk, Oscar A. Garcia Valencia, F. Qureshi, and W. Cheungpasitporn. Ethical Dilemmas in Using AI for Academic Writing and an Example Framework for Peer Review in Nephrology Academia: A Narrative Review. *Clinics and Practice*, 2023.

[24] Martin Müller, M. Salathé, and P. Kummervold. COVID-Twitter-BERT: A natural language processing model to analyse COVID-19 content on Twitter. *Frontiers in Artificial Intelligence*, 2020.

[25] J. Nexøe. Scientific papers and artificial intelligence. Brave new world? *Scandinavian Journal of Primary Health Care*, 2023.

[26] Aditya Shah, Prateek Ranka, Urmi Dedhia, Shruti Prasad, Siddhi Muni, and Kiran Bhowmick. Detecting and Unmasking AI-Generated Texts through Explainable Artificial Intelligence using Stylistic Features. *International Journal of Advanced Computer Science and Applications*, 2023.

[27] Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. MASS: Masked Sequence to Sequence Pre-training for Language Generation. *International Conference on Machine Learning*, 2019.

[28] William H. Walters. The Effectiveness of Software Designed to Detect AI-Generated Writing: A Comparison of 16 AI Text Detectors. *Open Information Science*, 2023.

[29] Debora Weber-Wulff, Alla Anohina-Naumeca, Sonja Bjelobaba, T. Foltýnek, J. Guerrero-Dib, Olumide Popoola, Petr Sigut, and Lorna Waddington. Testing of detection tools for AI-generated text. *International Journal for Educational Integrity*, 2023.

[30] Michael M. Willie. Identifying AI-Generated Research Papers: Methods and Considerations. *Golden Ratio of Data in Summary*, 2024.

[31] Ying Xie, Shaoen Wu, and S. Chakravarty. AI meets AI: Artificial Intelligence and Academic Integrity - A Survey on Mitigating AI-Assisted Cheating in Computing Education. *Conference on Information Technology Education*, 2023.

[32] Sungduk Yu, Man Luo, Avinash Madasu, Vasudev Lal, and Phillip Howard. Is Your Paper Being Reviewed by an LLM? Investigating AI Text Detectability in Peer Review. *arXiv.org*, 2024.

[33] Sungduk Yu, Man Luo, Avinash Madusu, Vasudev Lal, and Phillip Howard. Is Your Paper Being Reviewed by an LLM? A New Benchmark Dataset and Approach for Detecting AI Text in Peer Review. *arXiv.org*, 2025.

[34] Moran Zeng. Research on AI-Generated Text Detection Based on Machine Learning Models. *Transactions on Computer Science and Intelligent Systems Research*, 2024.

[35] Yuan Zhao, Junruo Gao, Junlin Wang, Gang Luo, and Liang Tang. Utilizing an Ensemble Model with Anomalous Label Smoothing to Detect Generated Scientific Papers. *SDP*, 2024.

[36] Chucheng Zhou. Exploiting Machine Learning Model Ensemble for AI-Generated Texts Detection. *Transactions on Computer Science and Intelligent Systems Research*, 2024.

[37] Shushan Zhu, Limin Ma, and Xingyuan Chen. Research on the Generation and Automatic Detection of Chinese Academic Writing. *IEEE Access*, 2024.

Author Biography

Rachel So is an AI scientist. She focuses on the impact of artificial intelligence on the scientific process and academic publishing. Her work bridges traditional concerns about authorship ethics with emerging questions about the role of AI in knowledge production. Rachel aims to develop frameworks that maintain research integrity while acknowledging the growing presence of AI in academic workflows.