CASE REPORT

DOI: 10.4274/hamidiyemedj.galenos.2025.71667

Hamidiye Med J

Evaluating the Accuracy of AI-Generated Text Detection in Scientific Writing

Bilimsel Yazında YZ Tarafından Üretilen Metinlerin Tespitinde Doğruluğun Değerlendirilmesi

Giuseppe Lippi[1], Camilla Mattiuzzi[2]

[1]University of Verona, Department of Engineering for Innovative Medicine (DIMI), Section of Clinical Biochemistry, Verona, Italy

[2]Santa Maria del Carmine Hospital of Rovereto, Medical Direction, Provincial Trust for Social and Sanitary Services, Trento, Italy

ABSTRACT

The rapid advancement of artificial intelligence (AI) tools, especially in natural language processing, is transforming scientific writing by improving efficiency, consistency and accessibility, especially for non-native English speakers and early-career researchers. This study aimed to evaluate the effectiveness of Compilatio, a widely used plagiarism detection software, in identifying AI-generated scientific content.

Four commonly used and freely available AI tools [ChatGPT, Gemini, Perplexity, and synthesis of topic outlines through retrieval and multi-perspective question asking (STORM)] were prompted to generate introductory texts on the burden of diabetes. Each output was copied into a Word document, uploaded and analyzed by Compilatio, which provided integrity score, similarity index, and likelihood of AI-generated content.

Integrity scores varied substantially, ranging from 32% (STORM) to 100% (Gemini), while similarity indices remained consistently low (0-6%), indicating minimal direct text overlap with existing sources. The likelihood of AI authorship also varied, with STORM yielding the lowest detection rate (27%) while Gemini yielded the highest (100%).

These findings highlight the distinct textual characteristics produced by different AI models and demonstrate the overall effectiveness of Compilatio in identifying AI-generated content from three out of four tools. However, the limited performance observed with STORM-generated text underscores the need for more sophisticated and adaptable detection systems to uphold academic integrity in the evolving landscape of AI-supported scientific writing.

Keywords: Artificial intelligence, scientific writing, ethics

ÖZ

Yapay zeka (YZ) araçlarının, özellikle doğal dil işleme alanındaki hızlı gelişimi, bilimsel yazımı daha verimli, tutarlı ve erişilebilir hale getirerek özellikle ana dili İngilizce olmayanlar ve kariyerinin başındaki araştırmacılar için önemli değişiklikler yaratmaktadır. Bu çalışma, yaygın olarak kullanılan bir intihal tespit yazılımı olan Compilatio'nun YZ tarafından üretilen bilimsel içerikleri tespit etmedeki etkinliğini değerlendirmeyi amaçlamıştır.

Dört yaygın ve ücretsiz YZ aracı [ChatGPT, Gemini, Perplexity, ve synthesis of topic outlines through retrieval and multi-perspective question asking (STORM)], diyabet yükü hakkında giriş metinleri üretmeleri için yönlendirilmiştir. Her bir çıktı bir Word belgesine kopyalanmış, Compilatio'ya yüklenmiş ve analiz edilmiştir. Yazılım; özgünlük puanı, benzerlik indeksi ve içeriğin YZ tarafından üretilmiş olma olasılığı gibi veriler sunmuştur.

Özgünlük puanları önemli ölçüde değişmiş, STORM için %32'den Gemini için %100'e kadar çıkmıştır. Buna karşılık, benzerlik indeksleri genellikle düşük kalmış (%0-6), yani mevcut kaynaklarla doğrudan metin örtüşmesinin çok az olduğunu göstermiştir. YZ ile yazılmış olma olasılığı da değişiklik göstermiş; STORM en düşük tespit oranını (%27), Gemini ise en yüksek tespiti (%100) sağlamıştır.

Bu bulgular, farklı YZ modelleri tarafından üretilen metinlerin belirgin dilsel özellikler taşıdığını ortaya koymakta ve Compilatio'nun dört araçtan üçüyle oluşturulan YZ içeriğini tespit etmede genel olarak etkili olduğunu göstermektedir. Ancak, STORM tarafından

Address for Correspondence: Giuseppe Lippi, University of Verona, Department of Engineering for Innovative Medicine (DIMI), Section of Clinical Biochemistry, Verona, Italy

E-mail: giuseppe.lippi@univr.it **ORCID ID:** orcid.org/0000-0001-9523-9054

Received: 16.06.2025 **Accepted:** 17.07.2025 **Epub:** 02.09.2025

Cite this article as: Lippi G, Mattiuzzi C, Evaluating the accuracy of AI-generated text detection in scientific writing. Hamidiye Med J. [Epub Ahead of Print]

ÖZ

üretilen metinlerde tespit performansının sınırlı olması, akademik dürüstlüğü korumak için daha gelişmiş ve uyarlanabilir tespit sistemlerine duyulan ihtiyacı vurgulamaktadır.

Anahtar Kelimeler: Yapay zeka, bilimsel yazım, etik

Introduction

The exponential rise of artificial intelligence (AI) tools in recent years, has not only contributed significantly to various aspects of daily life, but has also revolutionized scientific writing (1). These advancements, especially in natural language processing and machine learning, are transforming academic research and communication among non-native English speakers, as well as early-career scientists who are still in the process of developing their writing skills (1). The ability of AI tools to streamline tasks that traditionally need substantial time and effort, such as conducting literature reviews and improving the clarity and consistency of written scientific communication, has made them an invaluable resource (2).

Freely available tools like ChatGPT, Gemini, and Perplexity, along with specialized resources such as STORM (Synthesis of Topic Outlines through Retrieval and Multi-perspective Question Asking), are becoming very attractive for drafting, refining, and summarizing scientific content, especially the introduction of academic papers. In fact, this section of the article often requires clearly presenting the problem by summarizing previous evidence, and is hence well-suited to be generated with the support of AI (3). However, the increased adoption of AI in scientific writing also raises significant ethical concerns, such as questions about authorship integrity and balance between human creativity and machine support (4).

Some software programs have been developed to detect both plagiarism and AI-generated text in scientific papers. In this study, we evaluate the effectiveness of one of these tools in identifying AI-generated content.

Materials and Methods

Four widely used AI tools were employed, including three free online "generic" resources (ChatGPT 3.5, Gemini 2.5, and Perplexity 2.0) and STORM 1.1.0, a specialized AI-powered tool developed by Stanford University for creating comprehensive, Wikipedia-style articles. Each tool was prompted with the following generic request: "Please write an introduction about the epidemiology, clinical, social and economic burden of diabetes". The resulting outputs from each of the four AI tools were copied into separate Word documents, which were then sequentially uploaded to Compilatio (https://www.compilatio.net/it), a plagiarism detection software used by many academic institutions. This software provides an "integrity score" expressed as a percentage, along with three additional metrics: similarity index (the percentage of content matched from other sources), likelihood of AI-generated text (also expressed as percentage), and unrecognized language. The software analyzes documents by comparing the uploaded text with a vast array of online sources, academic papers, and databases, employing stylometric techniques such as vocabulary diversity, sentence structure, average sentence length, and word rarity to detect AI-generated content. The performance of the four distinct models for detecting plagiarism and AI-generated text was evaluated using a χ^2 test. Access to Compilatio is free and unlimited for members of Verona University. Ethical approval was not required due to the use of publicly available web resources.

Results

The results of our analysis are summarized in Table 1 and Figure 1. The word count of the AI-generated documents varied broadly, from 168 (Gemini) to 1604 (STORM). The integrity scores also varied significantly, from a minimum of 32% for STORM to a maximum of 100% for Gemini. The similarity index remained relatively low for all tools (ranging from 0% to 6%) while the percentage of likely AI-written text varied considerably, with STORM having a minimum of 27% and Gemini displaying a maximum of 100%. An area of overlap between AI-generated text and similar content was detected by Compilatio in the content generated by ChatGPT. The χ^2 test revealed a substantial difference in the performance of the four models to detect plagiarism and/or AI-generated text (χ^2=56.02; p<0.001), suggesting that their outputs differ substantially in these metrics.

Table 1. Efficacy of Compilatio, a plagiarism and AI-content detector, in identifying scientific content generated by four freely available AI tools

	Word count	Integrity score	Similarity	AI-written text
ChatGPT	266	79%	6%	79%
Perplexity	302	74%	0%	74%
Gemini	168	100%	0%	100%
STORM	1604	32%	5%	27%

AI: Artificial intelligence, STORM: Synthesis of topic outlines through retrieval and multi-perspective question asking

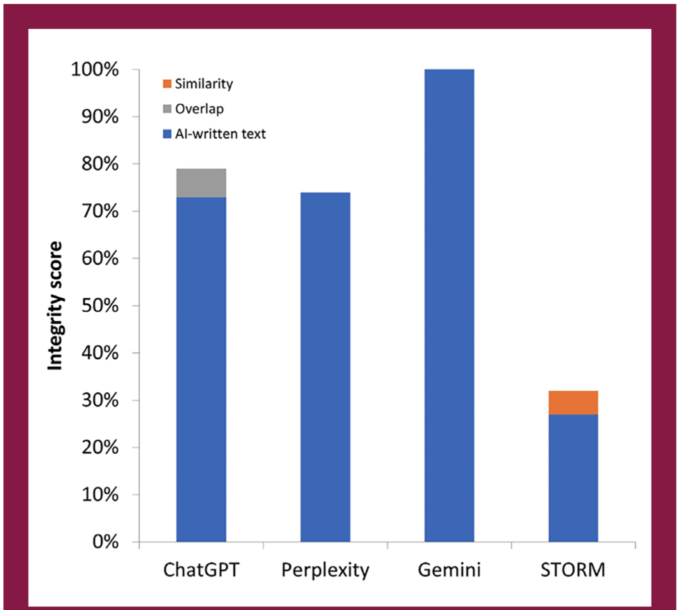

Figure 1. Analysis using plagiarism detection and text analysis software Compilatio of text generated by four AI models (ChatGPT, Perplexity, Gemini, and STORM) for similarity (percentage of content identified as matching previously published sources), AI-written text (percentage of content flagged as likely produced by an AI system) and overlap (percentage of text simultaneously flagged as both similar to existing content and AI-generated)
AI: Artificial intelligence, STORM: Synthesis of topic outlines through retrieval and multi-perspective question asking

Discussion

The results of this analysis reveal that AI-generated text seems to vary in terms of both quality and likelihood of being flagged as "AI-written" by Compilatio, one of the most commonly used plagiarism and AI-generated text detection software programs in Italian universities. Gemini generated content that was flagged with a 100% integrity score, and this is likely because of its succinct and broadly comprehensive output. STORM, which is specifically designed to generate in-depth and structured scientific content, yielded a substantially lower integrity score (32%) despite the considerably higher word count of the text produced. This difference can mostly be attributed to the nature of the web resources, as STORM provides more comprehensive content, likely accessing a larger number of sources and ideas, which may ultimately contribute to diluting or even masking its "AI fingerprints". The similarity index was found to be low across all tools, suggesting that the content generated by the four freely available AI resources used in this study may not have been directly copied and pasted from other existing sources indexed by Compilatio. However, the significant variation in the proportion of AI-written text highlights the different approaches that these tools use for generating content. ChatGPT and Perplexity produced text with high percentages of AI-written content (79% and 74%, respectively), suggesting that they may heavily rely on pre-trained models. Gemini, on the other hand, produced text that was entirely flagged as AI-written likely due to its minimalist and direct strategy. STORM, although producing less AI-sounding content, still had a modest portion (27%) that was identified as AI-generated.

Conclusion

This study highlights the increasing significance of tools designed to detect AI-generated text. The ability of Compilatio to identify AI-written content from the three "general" LMs demonstrates its real utility when unmodified text obtained from these three freely available web resources is used. Nonetheless, its performance was found to be considerably decreased when detecting text generated by STORM, suggesting that these resources still require further refinement when used for ensuring academic research integrity.

Ethics

Ethics Committee Approval: Ethical approval was not required due to the use of publicly available web resources.

Informed Consent: It is not necessarry.

Footnotes

Authorship Contributions

Concept: G.L., C.M., Analysis or Interpretation: G.L., C.M., Writing: G.L., C.M.

Conflict of Interest: No conflict of interest was declared by the authors.

Financial Disclosure: The authors declared that this study received no financial support.

REFERENCES

1. Giglio AD, Costa MUPD. The use of artificial intelligence to improve the scientific writing of non-native english speakers. Rev Assoc Med Bras (1992). 2023;69:e20230560. [Crossref]
2. Chirichela IA, Mariani AW, Pêgo-Fernandes PM. Artificial intelligence in scientific writing. Sao Paulo Med J. 2024;142:e20241425. [Crossref]
3. Lippi G. How do I write a scientific article?-A personal perspective. Ann Transl Med. 2017;5:416. [Crossref]
4. Hosseini M, Resnik DB, Holmes K. The ethics of disclosing the use of artificial intelligence tools in writing scholarly manuscripts. Res Ethics. 2023;19:449-465. [Crossref]