Review

The Role of Generative Artificial Intelligence for Scientific Writing: A Scoping Review of Empirical Evidence (2023-2026)

Jovan Shopovski *

Posted Date: 19 May 2026

doi: 10.20944/preprints202605.1200.v1

Keywords: generative artificial intelligence; scientific writing; ChatGPT; large language models; academic integrity; ethics of authorship; literature review; manuscript preparation

 doi:10.20944/preprints202605.1200.v1

Review

The Role of Generative Artificial Intelligence for Scientific Writing: A Scoping Review of Empirical Evidence (2023–2026)

Jovan Shopovski

Grigol Robakidze University, Georgia; jovanpraven@yahoo.com

Abstract

This paper examines the empirical evidence on the use of generative artificial intelligence (GenAI) in scientific writing. A search was conducted in Google Scholar and PubMed, followed by an analysis of the included studies, which was performed according to the academic field, AI tool, writing task, study design, and main findings. Following the PRISMA guide, this scoping review included 18 studies published between 1st January 2023 and 1st January 2026, representing the disciplines of medicine, education, dentistry, radiology, humanities, library, information science and cognitive science. The evidence base was dominated by studies on ChatGPT, making it the most empirically researched GenAI tool in this field. According to the studies reviewed, GenAI performed well on an array of measures (readability, fluency, and organization) and efficiency (the latter especially in terms of manuscript drafting, abstract writing, proposal development, and literature reviewing). However, the findings also disclosed several limitations, including incorrect or falsified references, inaccurate bibliographical metadata, shallow analysis, lack of originality, and insufficient methodological depth. Based on comparative evidence, newer model versions show improved coherence and reasoning and although improved with the newer GenAI versions, reference reliability still appears to be a recurring problem. Overall, GenAI can be a useful assistive tool for scientific writing; however, its usefulness is dependent upon human supervision and the task at hand, especially with regard to the accuracy of facts and their sources.

Keywords: generative artificial intelligence; scientific writing; ChatGPT; large language models; academic integrity; ethics of authorship; literature review; manuscript preparation

Introduction

Effective scientific writing is the cornerstone of academic research, allowing researchers to clarify their research objectives and effectively present their results (Lindsay, 2020). However, the main purpose of the process of scientific writing is to communicate the results to the reading audience (Gopen & Swan, 1990). It is a process that requires practice and focus on clarity (Heard, 2016). The value of scientific writing lies in recording data, supporting evidence-based ideas, and building on existing research (Okwemba, 2022). All of this makes the process of scientific writing and publishing significantly important for society.

Artificial Intelligence (AI) tools that fall outside the scope of Generative AI are already widely recognized within the scientific writing and publishing process and are extensively utilized in academia. They are especially beneficial for non-native English speakers by enhancing clarity, style, and coherence (Giglio & Costa, 2023). Moreover, AI is also widely used in detecting plagiarism, generating citations, and managing literature (Khalifa & Albadawy, 2024). The public access to Generative AI and the vast amount of data on which these new tools are trained, on the other hand, represents a new concept that is transforming society in many ways due to its unprecedented level of sophistication (Liang et al., 2024). Its ability to generate sophisticated content quickly has changed traditional academic practices and goes beyond the established AI usage in the field of scientific

 doi:10.20944/preprints202605.1200.v1

writing (Kaliyadan & Seetharam, 2023). Currently, Generative AI significantly impacts scientific writing across various fields, offering benefits such as idea organization, suggesting research gaps and questions, fostering literature reviews, generating research proposals and methods, as well as aiding in data management and analysis (Khalifa & Albadawy, 2024; Kaliyadan & Seetharam, 2023). The introduction of ChatGPT in November 2022 has gained significant attention in academic writing and its usage in academia in general. Alongside other Large Language Models (LLMs), which are integral to Generative AI and represent advanced AI tools, their usage in scientific writing is steadily growing (Liang et al., 2024). Furthermore, technological revolutions and advancements usually precede legal regulations. The absence of clear regulations and guidance for the ethical use of Generative AI is a reality. The need for responsible and transparent use opens many dilemmas and vigorous discussions (Muga, 2023; Vitente et al., 2023). Primarily, discussions and concerns have arisen regarding the potential inclusion of Generative AI (ChatGPT in particular) as the author of a paper. Some ethics organizations and leading publishers oppose such inclusion due to accountability issues. Despite this, ChatGPT has been included as a co-author in manuscripts, and these manuscripts have already received citations (Nazarovets & Teixeira da Silva, 2024). The threat to academic integrity is another ethical concern related to the advancement of Generative AI and its usage in academia. Concerns about plagiarism and misconduct, along with issues related to transparency and reporting, may arise from its usage (Lozic & Stular, 2023). Although Large Language Models, LLMs, generate scientific content that differs from human-generated, it is very difficult for the difference to be detected even by experts. Sophisticated AI tools for that purpose are required instead (Khera et al. 2025).

On the other hand, along with the growing level of Generative AI usage for research and writing, the exploration of potential and capabilities in current commercial Generative AI tools is also expanding. Scientific papers discussing the nuances of existing versions of LLMs and other Generative AI tools for scientific writing and publishing have already been published. However, a comprehensive review of literature summarizing the advantages and challenges of Generative AI in scientific writing is necessary to fill this research gap.

This article aims to address this gap by summarizing research findings and providing insights beneficial to stakeholders and crucial for shaping future exploration in Generative AI and scientific writing. This scoping literature review will map the current state of Generative AI tools used for scientific writing, offering a thorough analysis of studies in this domain and practical applications. It will conclude by evaluating reported functionalities and benefits, while also acknowledging the dilemmas and concerns associated with Generative AI usage in scientific writing.

The research questions of this scoping literature review are:

1. In which roles and aspects can Large Language Models (LLMs) be used in the process of improving the quality and efficiency of scientific writing?
2. What are the major challenges and ethical concerns that are associated with their use in the process of scientific writing?

Background

The traditional scientific writing process, which is widely accepted, involves some key steps such as selecting a topic, conducting a review of literature, and writing the article in a structured format (Grimm & Harvey, 2022). Manuscripts are usually prepared in the IMRAD format: Introduction, Methods, Results, and Discussion (Somashekhar, 2020). The key components of a research article are underscored, including the title, abstract, introduction, methods, results, discussion, and conclusion (Lunsford & Lunslord, 1996; Fischer & Zigmond, 2009). Some research focuses on evaluating the soundness of these components. It encompasses a theoretical base, problem statements, variable definition, population description, research design, test instruments, reliability and validity, result consistency, and recommendations for further research (Nielsen & Reilly, 1985). Finally, each manuscript, prior to publication, undergoes formatting and referencing to ensure adherence to specific style guidelines established by scientific journals (Watson, 2019). AI usage in

 doi:10.20944/preprints202605.1200.v1

scientific writing is a topic that attracts huge attention and opens debate. The potential of AI to facilitate diversity and increase efficiency in scientific communication is highlighted (Carobene et al., 2023; Khan et al., 2023). However, its usage raises concerns regarding the impact of AI on research integrity and the role of human researchers in the process.

AI-powered writing tools, such as Grammarly, Zotero, Mendeley, and others, have revolutionized how scientists approach the writing process, and the benefits of their usage are widely accepted and confirmed. These AI writing tools provide valuable assistance in various aspects of scientific writing, including grammar and syntax correction, citation management, and literature review organization (Razack et al., 2021). By using AI algorithms, these tools can analyze and understand the context of scientific writing, thereby offering real-time suggestions for improvements. However, the commercial usage of Generative AI is new, and its usage and capabilities go beyond these AI tools. The public access to Generative AI, particularly LLMs, is a relatively new topic that sparks interest and provides new opportunities and challenges simultaneously. After the release of ChatGPT, the usage of LLMs for generating content in scientific papers significantly increased, especially in the field of Computer Science (Liang et al., 2024).

Research studies have underlined the benefits of LLMs, including accelerating literature reviews and enhancing the writing process (AlSagri et al. 2024). The use of LLMs accelerates the manuscript output and reduces barriers for non-native English speakers (Kusumegi et al. 2025). They also highlight limitations, such as biases and ethical concerns (Muga, 2023; Boyko et al., 2023). While testing the ability of LLMs in scientific summarization tasks, it was evident that LLMs outperform humans in certain tasks. Nonetheless, they have limitations in generating long summaries and abstractive lay summaries (Fonseca & Cohen, 2024).

As Generative AI and LLMs continue to advance, their usage for scientific writing will increase. The growing ethical dilemmas and concerns will also foster the necessity for regulations and guidelines in this area. Several publishers have already established policies addressing the issue of Generative AI usage in manuscripts (Salimi & Saheb, 2023). Concurrently, the integration of Generative AI usage in writing scientific papers will advance, as well as the utilization of Generative AI in reviewing processes. Editors and reviewers will encounter and increasingly benefit from various opportunities for assessing submitted articles. However, they must remain vigilant about potential threats, such as biases and errors made by LLMs (Gilat & Cole, 2023). The issue of the inaccuracy of the information provided and analyzed by Generative AI has been present since the beginning of its usage in scientific writing. This applies to commercially available Generative AI tools to date (Dashti et al., 2023). Therefore, authors and reviewers should double-check the relevance of the information presented by Generative AI.

The studies in this area raise discussions regarding the ethical issues and challenges emanating from the progress of Generative AI and its steady usage in academia. Elali and Rachid (2023) examined the fabrication and plagiarism by the Generative AI content, highlighting the usage of the LLM tools to streamline the research process as well as the risk of undermining the legitimate work. Over the recent period, there has been a significant increase in studies analyzing the pros and cons of Generative AI usage in scientific writing. There is a clear need for a review of the existing results to establish a solid foundation for determining the strengths and weaknesses of current Generative AI tools and their usage for scientific writing. The main aim of this study is to fill this gap and summarize the findings of LLM usage in scientific writing, distinguishing between the benefits on one hand and the errors and challenges on the other. This scoping review of literature will also provide answers to which academic domains the LLM usage in scientific writing has been tested and whether it generates content with equal accuracy in social sciences, humanities, and natural and medical sciences. Furthermore, it will offer answers to which parts of the manuscript can be facilitated with the usage of Generative AI. Whether the existing versions of the LLM tools are efficient in creating abstracts, introductions, literature reviews, research methods, and conclusions. Finally, their language abilities in translation, editing, and proofreading of scientific content will be elaborated. On the other hand, it

 doi:10.20944/preprints202605.1200.v1

will underline the limitations of the commercially available LLM tools for scientific writing, the challenges that arise with Generative AI usage for scientific writing, and the ethical concerns.

Methods

Review Design

The research employed a scoping review approach to aim at mapping empirical findings regarding the application of Generative Artificial Intelligence (GenAI) to scientific writing. This method was chosen because of the rapid evolution of the field, its heterogeneity, and its interdisciplinary character extending into academic fields, spanning from medicine to engineering and education.

On the one hand, the article sought to define, classify, and generalize empirical findings on how GenAI tools can be used to handle diverse scientific writing tasks, such as the writing of a manuscript, the generation of an abstract, the leniency of language, summarization, and the provision of literature. Alongside mapping technical applications, the review was intended to encompass patterns pertinent to research integrity, such as accuracy, transparency, and reliability concerns about AI-assisted writing. The research employed a PRISMA-guided study selection procedure, which guaranteed transparency and reproducibility throughout the processes of identification, screening, evaluation of eligibility, and inclusion.

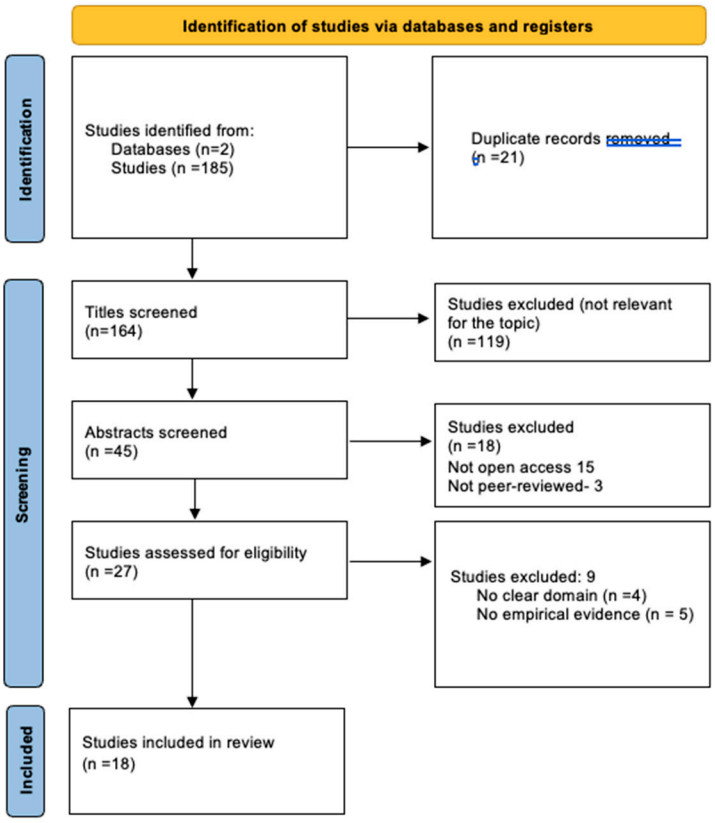

Figure 1. PRISMA flow diagram of study selection.

Search Strategy

Google Scholar and **PubMed**, two interdisciplinary search engines, were chosen because they offer a wide range of peer-reviewed studies not only in the fields of biomedical and social science but also in the field of technology.

 doi:10.20944/preprints202605.1200.v1

The search was narrowed down to **1 January 2023 to 1 January 2026**, the period when the widespread use of publicly available large language models (LLMs), including ChatGPT, became popular in academic life.

The keyword combinations that were used are as follows:

- "Generative AI" AND "academic writing"
- "LLM" OR "Large Language Models" AND "academic writing"
- "Generative AI" AND "scientific writing"
- "ChatGPT" AND "scientific writing"

These search terms were aimed at analyzing studies that have specifically focused on the use of GenAI tools in scientific and academic writing functions and not on the use of AI itself.

Eligibility Criteria

Studies were included if the following criteria were met:

1. Related to the applicability of the use of generative AI or large language models in scientific writing.
2. Delivered **empirical findings** such as experimental, comparative, evaluative, or quasi-experimental designs;
3. Were **peer-reviewed**;
4. Were published within the specified time period.

Studies were excluded if they:

- Were not peer-reviewed.
- Were not of specified academic discipline;
- Were not from an identifiable academic domain; or
- Were not addressing a writing-related function.

Study Selection

The first search retrieved 185 records. Following the elimination of 21 duplicate records, 164 studies remained for title screening. Title screening led to the retention of 45 studies for abstract review. During the abstract screening phase, 18 studies were removed because of unavailability, lack of peer review, or inadequate academic organization. This led to the selection of 27 eligible articles.

During the full-text review, **9 studies were excluded**, of which:

- 5 lacked a clearly defined academic domain; and
- 4 did not address a specific scientific writing function relevant to the review.

A total of 18 studies met all inclusion criteria and were included in the final synthesis.

Data Extraction

Data extraction used a review matrix designed specifically for this study. The following data were extracted for each article:

- study reference;
- academic field;
- AI tool and version (if specified);
- writing task assessed;
- study design;
- principal findings; and
- reported limitations.
- Further, to support descriptive analysis in the results section, the following data were also extracted:
- affiliated country (place of the study);
- number of co-authors; and

 doi:10.20944/preprints202605.1200.v1

- where published.

Data Synthesis

Due to the variability among studies in terms of disciplinary background of the study, research design and tasks assessed, a narrative synthesis was adopted.

The synthesis sought to identify common patterns in the studies:

- supported writing tasks;
- observed advantages, such as increased efficiency, clarity, organisation and language quality;
- reported drawbacks, especially regarding factual correctness, reference credibility, creativity and critical thinking; and
- differences between disciplines.

Furthermore, a descriptive synthesis was undertaken to explore the patterns of geographical location of studies, authorship and publishing outlets, which are also reported in the results.

3. Results

3.1. Characteristics of Included Studies

Table 1 provides an overview of the characteristics of studies, such as discipline, AI tool, writing task, study type, results and limitations.

Table 1. Characteristics of included studies.

Study	Field	AI tool	Task tested	Study design	Main finding
Altmae et al. (2023)	Reproductive medicine	ChatGPT-3.5	Manuscript drafting	Prompt-based illustrative experiment	Helped structure and draft text, but references were inaccurate
Athaluri et al. (2023)	Healthcare research	ChatGPT-3-based model	Research proposal generation	Analytical experiment	Produced structured proposals, but many references were fabricated
Kacena et al. (2024)	Musculoskeletal research	ChatGPT-4	Full review writing	Comparative experimental study	Reduced writing time and improved readability, but reference accuracy was poor
Lozic and Stular (2023)	Digital humanities	ChatGPT-3.5/4, Bard, Bing Chat, Claude-2, Aria	Scientific explanation and academic text generation	Comparative experimental evaluation	Generated fluent academic prose, but outputs were often derivative and factually inconsistent
Babl and Babl (2023)	Emergency medicine	ChatGPT (GPT-3)	Abstract generation	Prompt-based demonstration	Produced a well-structured abstract, but citation hallucination occurred

 doi:10.20944/preprints202605.1200.v1

Study	Field	AI tool	Task tested	Study design	Main finding
Donlon and Tiernan (2023)	Educational technology	ChatGPT-3.5	Academic paper generation	Exploratory case study	Generated coherent text and useful draft sections, with human refinement still needed
Hsu (2023)	Educational technology	ChatGPT-4	Short paper preparation	Exploratory case study	Supported brainstorming, outlining, drafting, and proofreading, but required human verification
Hwang et al. (2024)	Biomedical publishing	ChatGPT-3.5 and 4	Abstract generation from RCTs	Comparative experimental study	Generated readable abstracts, but reporting quality was lower than that of human abstracts
Mugaanyi et al. (2024)	Natural sciences/humanities	ChatGPT-3.5	Citation and reference generation	Cross-disciplinary evaluation	Citation relevance was moderate, but DOI accuracy was poor
Kung (2023)	Library and information science	Elicit	Literature review support	Functional test study	Supported paper discovery, summarization, and evidence organization
Jenko et al. (2024)	Musculoskeletal radiology	GPT-4 and the-literature.com	Literature review generation	Comparative evaluation	GPT-4 produced more relevant summaries, but neither tool captured all relevant literature
Dashti et al. (2025)	Dentistry	ChatGPT	Reference generation	Experimental verification study	Generated references were highly unreliable and required manual checking
Kumar et al. (2025)	Orthopedics/medical education	ChatGPT	Writing support for residents	Quasi-experimental study	Improved language, structure, and readability, but not higher-order research skills
Benichou (2026)	Medical scientific writing	GPT-3.5, GPT-4, GPT-4o, GPT-5	Full IMRaD article generation	Comparative experimental study	Newer models showed better structure and reasoning, with improved citation quality

 doi:10.20944/preprints202605.1200.v1

Study	Field	AI tool	Task tested	Study design	Main finding
Benichou (2023)	Medical scientific writing	ChatGPT	Generation of article sections	Comparative experimental test	Produced coherent and readable text quickly, but content was often superficial
Eymann et al. (2025)	Cognitive science	ChatGPT-4	Research proposal generation	Experimental comparison	Generated highly rated proposals, but originality and consistency remained uncertain
Pressman et al. (2025)	Plastic surgery	ChatGPT-4	Abstract generation	Experimental comparative study	AI-generated abstracts were preferred for clarity and writing quality
Hadan et al. (2024)	Human-computer interaction	Google Gemini	Detection and evaluation of AI-augmented writing	Reviewer survey experiment	Reviewers struggled to detect AI writing, and AI assistance did not reduce perceived quality

3.2. Yearly Distribution of Included Studies

The yearly distribution of the included studies confirms that empirical research on generative artificial intelligence in scientific writing is both recent and rapidly developing. Of the 18 studies included in this review, 8 were published in 2023, 5 in 2024, 4 in 2025, and 1 in 2026. This pattern indicates that scholarly attention intensified immediately following the widespread adoption of publicly accessible large language models such as ChatGPT, with the strongest concentration of studies appearing in the early phase of academic experimentation. The continued presence of studies across 2024–2026 further suggests that the field remains active, while gradually shifting from early exploratory work toward more comparative and evaluative assessments of specific tools, versions, and writing-related functions.

3.3. Geographical Distribution of Studies

The geographical origin of the included studies is presented in Figure 2.

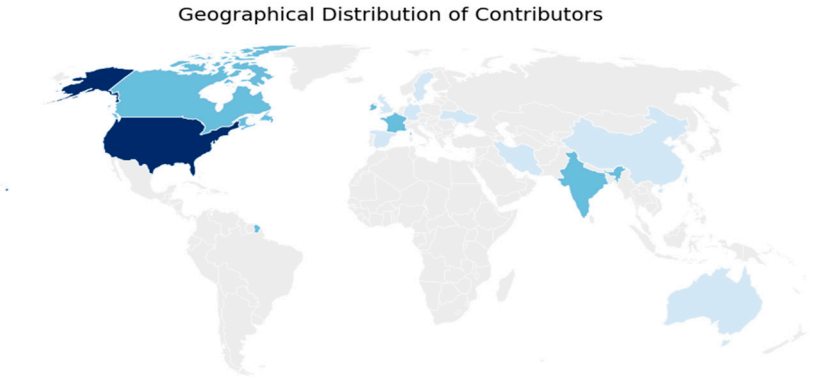

Figure 2. World map showing the geographical distribution of included studies.

 doi:10.20944/preprints202605.1200.v1

Created by Chat GPT 5.0 Plus Version

The distribution indicates that research on generative artificial intelligence in scientific writing is international in scope but unevenly distributed across countries. The largest concentration of contributions originates from the United States, followed by a smaller group of countries with multiple contributions, including India, France, Canada, and Ireland. Additional contributions are distributed across countries such as Spain, Sweden, Ukraine, Slovenia, Australia, Taiwan, China, Iran, Germany, and the United Kingdom, each represented by a smaller number of studies.

This pattern reflects a broad but asymmetrical global engagement, suggesting that empirical research in this field is currently concentrated in specific research environments while remaining accessible across multiple regions.

3.4. Authorship Patterns

The included studies demonstrate that authorship patterns are mainly collaborative. Author groups range from one to seven authors per study, with the majority of studies (12 out of 17) having two to six authors.

This suggests that the work on using generative AI for scientific writing tends to be conducted by small research teams, likely because of the multidisciplinary nature of the topic, which often requires a combination of expertise from the academic field, research methods and AI systems.

3.5. Publishing Outlets

The studies used in the article were found in a wide variety of scholarly journals and conference papers, including medical, technological, and interdisciplinary publications.

A large percentage of the papers were found in journals of **major international publishers**, including Elsevier, Springer, and Wiley, as well as specialized outlets such as the *Journal of Medical Internet Research* and *Computers in Human Behavior: Artificial Humans*.

The above distribution of publishing outlets suggests that research on generative AI in scientific writing lacks a unified discipline but is instead spread across a wide range of different fields, indicating its cross-cutting relevance and increasing adoption in the mainstream academic publishing sector.

3.6. Empirical Findings on the Use of Generative AI in Scientific Writing

Results report that generative artificial intelligence research in the area of scientific writing, which is based on empirical and quasi-empirical methods, has extended to a wide range of academic fields such as medicine, biomedical science, education, radiology, dentistry, humanities, library and information science, and cognitive science. The tasks analyzed were diversified, such as manuscript writing, abstract generation, research proposal writing, literature review generation or assistance, citation, proofreading, and idea generation.

In the reviewed articles, generative AI exhibited evident potential in assisting with various writing tasks, yet its effectiveness was uneven based on the characteristics of the task to be performed and the level of human supervision needed (Altmae et al., 2023; Donlon & Tiernan, 2023; Hsu, 2023; Kacena et al., 2024; Kumar et al., 2025).

Generative AI tools have demonstrated reliability in improving the formal and linguistic aspects of scientific writing. Some studies also showed that ChatGPT and its application could produce well-formed, grammatically accurate, and coherent academic text and can therefore be of high utility in writing essay sections, language optimization, and topic organization (Altmae et al., 2023; Donlon & Tiernan, 2023; Hsu, 2023; Benichou, 2023). ChatGPT showed pronounced improvement in writing technique, language accuracy, and structural organization in a quasi-experimental study of orthopedic residents, but was less creative and critical (Kumar et al., 2025). It was clear that comparing newer versions of ChatGPT to older ones showed improvement in reasoning and technical quality of content (Benichou, 2023, 2026). A comparative study of medical scientific writing

 doi:10.20944/preprints202605.1200.v1

established that AI-generated texts were frequently similar to human-generated texts in their level of readability and structure, and they were created significantly more quickly (Pressman et al., 2025).

AI is also quite helpful in generating abstracts. Under comparative and experimental research, AI-generated abstracts were mostly rated to be very readable and well-structured, yet they did not always achieve comparable quality and scientific thoroughness as human-written abstracts (Hwang et al., 2024; Pressman et al., 2025). The investigations conducted by Hwang et al. (2024) revealed that abstracts created by ChatGPT had high readability, particularly those generated by GPT-3.5, yet human-created abstracts scored much higher on the CONSORT-A checklist. Pressman et al. (2025) also discovered that ChatGPT-4-generated abstracts were better rated by evaluators on the quality and clarity of writing, but factual accuracy was not thoroughly evaluated in the study. These results indicate that AI could be exceptionally robust at presenting scientific content succinctly and in a refined manner, although the rigor of reporting could be lower compared to human-written text.

Generative AI also demonstrated significant support capabilities in the context of manuscript writing and proposal writing. Altmae et al. (2023) discovered that ChatGPT was helpful in outlining ideas, creating sections of a manuscript, and refining language quality, which could help speed up the writing process. Athaluri et al. (2023) indicated that ChatGPT was able to produce structured medical research proposals, and Hsu (2023) demonstrated that ChatGPT was able to assist in several steps of the manuscript preparation process, such as brainstorming, outlining, drafting, recommending potential research designs, and proofreading. ChatGPT-4 was found to perform better in general scoring of short research projects, particularly with structure, detail, and rationale justification; however, the results also revealed repetitive patterns and a lack of variation, with outputs appearing to be more creative than novel when performing a scientific creativity study (Eymann et al., 2025).

The literature on literature review support is more mixed regarding the findings; however, it points to perceived practical benefits.

It was discovered that AI-assisted tools like Elicit could aid literature discovery, summarization, extraction of key variables, and organization of references, allowing users to construct research matrices and speed up evidence-synthesis processes (Kung, 2023). GPT-4 demonstrated more effective performance in a comparative study based on musculoskeletal radiological research, as compared to a dedicated literature review tool, in generating meaningful summaries on various topics, yet neither did so reliably (Jenko et al., 2024). These findings indicate that AI applications can probably be useful in facilitating the early-stage process of review, particularly within the scoping and literature-organization phase, but the outputs cannot be regarded as complete and still require human input.

Although these merits are present, the strongest negative result of the reviewed literature is the consistent lack of reliability of AI-generated citations and references. False or fabricated references, false DOIs, non-existent publications, and misplaced bibliographic metadata were indicated across multiple fields (medicine, dentistry, biomedical writing, and cross-disciplinary citation assignments) multiple times (Altmae et al., 2023; Athaluri et al., 2023; Babl & Babl, 2023; Kacena et al., 2024; Mugaanyi et al., 2024; Dashti et al., 2025).

Athaluri et al. (2023) discovered that 28 references ChatGPT created during research proposal writing were entirely fabricated, and Dashti et al. (2025) discovered that all 75 references ChatGPT created in their journal-specific experiment did not correspond to real publications. Mugaanyi et al. (2024) identified that, although most of the citations were to actual publications, the success rate of DOIs was extremely low, particularly in humanities-related subjects. Kacena et al. (2024) also noted that nearly 70% of AI-generated review articles contained incorrect references. This similar tendency among studies is a strong indicator that citation generation is one of the most significant limitations of existing AI writing tools.

Another common limitation is related to form and lack of depth or thoroughness in the scientific text generated by AI. Although AI systems often create convincing and fluent prose, multiple studies have determined that the material was superficial, lacking critical analysis or methodological rigor

 doi:10.20944/preprints202605.1200.v1

(Benichou, 2023; Lozic & Stular, 2023; Kacena et al., 2024). ChatGPT-4 generates the best ratio of correct content among tested systems in humanities-oriented evaluations, but the results were still rated as derivative rather than original scholarly work (Lozic & Stular, 2023). Similarly, Kacena et al. (2024) found that AI-generated review papers were well written but needed to be fact-checked and edited. This was also indicated elsewhere by Kumar et al. (2025), who discovered that AI improved lower order writing skills but not higher order scientific skills.

Finally, new evidence also suggests that humans are not always able to distinguish between AI and human-generated writing, especially if they are assessing short texts and/or their writing quality. In the HCI study by Hadan et al. (2024), reviewers struggled to identify AI-generated text, and may have considered AI-enhanced text to be of acceptable quality. However, the human element (nuance, tone, personal opinion) was found in the same study to be highly valued by reviewers and to be more distinctive of human-generated writing.

Overall, the aforementioned literature suggests that generative AI may significantly improve writing speed, readability, and organisation. Nonetheless, such benefits are tempered by major drawbacks in the quality of references, the quality of scientific content, and the general accuracy of the information presented when used without human intervention.

4. Discussion

The findings of this updated review support the argument that a significant assistive role of generative artificial intelligence is possible in academic writing, but it is essentially a supportive (rather than independent) role. In the included studies, AI systems are consistently positive in the drafting, editing, structuring, summarising, and readability, and at the same time have historical limitations in the referencing, factual correctness, methodological correctness, and advanced scientific reasoning. The Janus pattern is one of the most robust findings among the empirical evidence.

The greatest strength of generative AI is its ability to accelerate the production of prose in academic style. Research has shown that AI can generate clear, readable, grammatically correct and logically coherent text, which can be on par with human-generated text (Donlon & Tiernan, 2023, 2026; Hsu, 2023, 2025; Benichou, 2023, 2026). These skills are remarkable, particularly when dealing with writing in limited structural formats such as abstracts, introductions and conclusions. Experimental studies also demonstrate that AI can improve writing skills in the educational context, especially related to grammar, structure and readability (Kumar et al., 2025). Taken together, these findings make generative AI a productivity-enhancing tool that could reduce language barriers and potentially improve productivity, particularly for early-career researchers and those who don't speak English as their first language.

The review also demonstrates that epistemic quality is not the same as writing quality. Reference hallucination and hallucination of bibliographic metadata are the most critical and most commonly reported issues. This can be seen in several domains and use cases, including the generation of proposals, abstracts, citations, and manuscripts (Altmae et al., 2023; Athaluri et al., 2023; Kacena et al., 2024; Mugaanyi et al., 2024; Dashti et al., 2025). The recurrence of fabricated or incorrect references in these different studies supports the hypothesis of this being a systematic failure of the current systems using LLMs. From an ethical perspective, this is a major disincentive to scholarly integrity because of the potential to deceive readers, reviewers and editors with plausible but incorrect references. So, human oversight is vital when using generative AI for scientific writing, with all sources needing to be checked by humans, to ensure human accountability for maintaining the integrity of the scholarly record rather than AI systems.

This is also highlighted in the review, when there is a clear difference between the short-term writing tasks and long-term scientific processes facilitated by AI. Generative AI is reliant on surface-level and organisational functions, including language editing, summarisation and organisation. However, they fall short on critical thinking, methodological decision-making, and interpretation (Lozic & Stular, 2023; Benichou, 2023; Kacena et al., 2024; Kumar et al., 2025). Qualitative reviews of

 doi:10.20944/preprints202605.1200.v1

AI output (even in studies where findings were deemed positive, e.g., Eymann et al., 2025) report that AI outputs were repetitive, had the usual methodological inconsistencies, and were not conceptually innovative. This suggests AI-generated content is not "new science" but is more likely to be a form of recombining existing science that does not inherently satisfy the ethical authorship, originality and contribution requirements.

Another insightful finding is that AI is not equally capable of all tasks. While AI can be used to write and summarise abstracts, it has been reported to be less complete and rigorous in its approach to reporting (Hwang et al., 2024; Pressman et al., 2025). Likewise, although literature review tools may increase efficiency in discovery and sorting, they are not systematic nor transparent in approach (Kung, 2023; Jenko et al., 2024). This suggests that ethical AI use is not just about whether AI is used, but how it is used, for what purposes and whether limitations are disclosed.

The findings also pose important concerns regarding bias and views on scientific quality. Researchers have shown reviewers may struggle to distinguish between AI and human compositions, and could perceive AI-assisted writing as readable (Hadan et al., 2024; Pressman et al., 2025). This can skew the peer-review process, with style overtaking substance. In these cases, AI can play a role in epistemic inflation, where high-quality writing masks weak or flawed research. This has implications for editors and may require more focus on research quality and strength of evidence, rather than writing quality.

A second consideration is transparency and disclosure. Given that AI-generated text is not always detectable, even by professional editors (Hadan et al., 2024), it is the responsibility of authors to disclose the use of generative AI to write. Non-disclosure could be seen as an issue of academic integrity, especially when AI is used extensively in manuscript preparation. However, the literature reviewed supports the opinion that AI is not an author, but a tool, consistent with the position of major publishers who believe that AI should not be cited as an author due to a lack of responsibility and accountability. But disclosure of AI use should be regarded as part of ethical research..

Authorship and responsibility go hand in hand. While AI can produce text, it cannot take responsibility for content, justify methodological decisions or engage in the peer review process. This is a problem of contribution and responsibility. Ethically, this supports the claim that AI can help to write but cannot be an author. Instead, its contribution should be recognised as technical support and human authors should retain intellectual ownership and accountability.

The other potential concern is over-reliance and skills erosion. The adoption of AI tools for writing tasks could result in a tendency among researchers, especially early-career researchers, to use AI as a crutch to generate writing structures and phrasing, rather than developing their own scientific writing and research skills. The studies reviewed indicate that AI can improve writing skills but not higher-level research skills (Kumar et al., 2025). This could pose an ethical issue in the future regarding the loss of expertise and autonomy in academic practice.

The review also shows a trend towards rapid improvement of AI tools, as newer versions show improvements in coherence, reasoning, and more accurate citation (Benichou, 2026). But these gains do not resolve underlying issues of veracity and interpretive capacity. Newer models may be more convincing, which increases the potential for errors to go unnoticed. Consequently, far from lessening human supervision, advances in technology heighten the need for critical assessment of AI-generated outputs.

Lastly, we must mention some limitations of the reviewed literature. Several studies have small sample sizes, are limited to one prompt, or specific disciplinary contexts and are explorative rather than comparative. Furthermore, because AI technology is evolving rapidly, study results may quickly become outdated. Nonetheless, while these studies have limitations, the consistency of findings across multiple studies suggests that generative AI is a powerful aid but not an independent replacement for human researchers' expertise.

 doi:10.20944/preprints202605.1200.v1

5. Conclusions

The findings from this review indicate that generative artificial intelligence has gained a seemingly beneficial role in scientific writing, but this is not uniform and highly specific to tasks. The most common positive findings in the reviewed studies relate to readability and language fluency, writing organisation, and writing efficiency. Generative AI tools were particularly adept at section writing, generation of abstracts, support with proposal writing, search for literature, and formalising academic text. At the same time, ChatGPT is the most researched generative AI application in the evidence base, as most studies are conducted on this application. This suggests that performance trajectories of the ChatGPT-based tools play an increasing role in our understanding of the use of generative AI in scientific writing.

But, as performance in organizational tasks suggests, high performance in these tasks does not imply high performance and reliability in the scientific tasks. One of the most common findings in the investigated studies was the continued presence of errors in AI-generated citations and references, such as invented sources, inaccurate metadata and misleading DOIs. Furthermore, several studies found that while AI-generated texts are often coherent, persuasive, and well-organised, they were often limited by shallowness, incompleteness, low originality and lack of precise methodological detail. The significance of these findings are that even complex scholarly language may be coupled with substantial epistemic problems.

The review also highlights how more recent versions of generative AI have resulted in improvements in coherence, reasoning, and technical quality. There is some evidence to support this assertion, as more recently developed models produce more realistic and well-formatted texts. However, these improvements don't solve the issues identified in the literature..

The use of unreliable references, the susceptibility to factual errors and lack of depth remain even in the face of improvements in fluency. This suggests technical advances should not be taken to demonstrate independent scientific reliability.

Overall, the results support a restrained but evidence-based conclusion: generative AI, especially as represented in the literature by ChatGPT, can substantially improve the efficiency and presentation of scientific writing, but it cannot yet be relied upon for autonomous scholarly production. Its practical value lies in augmentation rather than replacement. The evidence reviewed demonstrates the continued need for human oversight, particularly for source attribution, factuality, interpretability and methodological rigor.

Declaration of AI Use

The author used ChatGPT, GPT-5 version, developed by OpenAI, to assist with summarizing selected parts of the manuscript, generating ideas, and supporting image/figure generation. All outputs were reviewed, edited, and verified by the author, who takes full responsibility for the final content of the manuscript.

References

1. Altmae, S., Sola-Leyva, A., & Salumets, A. (2023). Artificial intelligence in scientific writing: A friend or a foe? *Reproductive BioMedicine Online, 47*(1), 3-9. https://doi.org/10.1016/j.rbmo.2023.04.009
2. AlSagri, H. S., Farhat, F., Sohail, S. S., & Saudagar, A. K. J. (2024). ChatGPT or Gemini: Who Makes the Better Scientific Writing Assistant?. *Journal of Academic Ethics*, 1-15
3. Athaluri, S. A., Manthena, S. V., Kesapragada, V. K. M., et al. (2023). Exploring the boundaries of reality: Investigating the phenomenon of artificial intelligence hallucination in scientific writing through ChatGPT references. *Cureus, 15*(4). https://doi.org/10.7759/cureus.37432
4. Atkinson, A. G., Lia, H., & Navarro, S. M. (2024). Advancing scientific writing with artificial intelligence: expanding the research toolkit. *Global Surgical Education-Journal of the Association for Surgical Education, 3*(1), 74.
5. Babl, F. E., & Babl, M. P. (2023). Generative artificial intelligence: Can ChatGPT write a quality abstract? *Emergency Medicine Australasia, 35*(5), 809-811. https://doi.org/10.1111/1742-6723.14233

 doi:10.20944/preprints202605.1200.v1

6. Boyko, J., Cohen, J., Fox, N., et al. (2023). An interdisciplinary outlook on large language models for scientific research. https://doi.org/10.48550/arXiv.2311.04929
7. Benichou, L. (2026). The role of using ChatGPT AI in writing medical scientific articles—Two years after. *Journal of Stomatology Oral and Maxillofacial Surgery*, 102751.https://doi.org/10.1016/j.jormas.2026.102751
8. Benichou, L. (2023). The role of using ChatGPT AI in writing medical scientific articles. *Journal of stomatology, oral and maxillofacial surgery*, *124*(5), 101456.https://doi.org/10.1016/j.jormas.2023.101456
9. Carobene, A., Padoan, A., Cabitza, F., et al. (2023). Rising adoption of artificial intelligence in scientific publishing: Evaluating the role, risks, and ethical implications in paper drafting and review process. *Clinical Chemistry and Laboratory Medicine (CCLM)*. https://doi.org/10.1515/cclm-2023-1136
10. Cheng, H. (2023). Challenges and limitations of ChatGPT and artificial intelligence for scientific research: A perspective from organic materials. AI, 4(2), 401-405. https://doi.org/10.3390/ai4020021
11. Dashti, M., Londono, J., Ghasemi, S., et al. (2023). How much can we rely on artificial intelligence chatbots such as the ChatGPT software program to assist with scientific writing? *The Journal of Prosthetic Dentistry*. https://doi.org/10.1016/j.prosdent.2023.05.023
12. Donlon, E., & Tiernan, P. (2023). Chatbots and citations: An experiment in academic writing with generative AI. *Irish Journal of Technology Enhanced Learning, 7*(2), 75-87. https://doi.org/10.22554/ijtel.v7i2.125
13. Elali, F. R., & Rachid, L. N. (2023). AI-generated research paper fabrication and plagiarism in the scientific community. *Patterns*, *4*(3). https://doi.org/10.1016/j.patter.2023.100706
14. Fischer, B. A., & Zigmond, M. J. (2009). Components of a research article. Retrieved from www.survival.pitt.edu
15. Eymann, V., Lachmann, T., Hekele, F., Ashok, A., Berns, K., & Czernochowski, D. (2025, October). Does ChatGPT know what robots are thinking?-Evaluation of human-created vs. AI-generated scientific ideas. In *Proceedings of the 36th Annual Conference of the European Association of Cognitive Ergonomics (EACE)* (pp. 1-5). https://dl.acm.org/doi/10.1145/3746175.3746208
16. Fonseca, M., & Cohen, S. B. (2024). Can large language model summarizers adapt to diverse scientific communication goals? https://doi.org/10.48550/arXiv.2401.10415
17. Gilat, R., & Cole, B. J. (2023). How will artificial intelligence affect scientific writing, reviewing, and editing? The future is here…. *Arthroscopy, 39*(5), 1119-1120. https://doi.org/10.1016/j.arthro.2023.01.014
18. Giglio, A. D., & Costa, M. U. (2023). The use of artificial intelligence to improve the scientific writing of non-native English speakers. *Revista da Associação Médica Brasileira, 69*. https://doi.org/10.1590/1806-9282.20230560
19. Gopen, G. D., & Swan, J. A. (1990). The science of scientific writing. *American Scientist, 78*(6), 550-558. Available at https://www.usenix.org/sites/default/files/gopen_and_swan_science_of_scientific_writing.pdf
20. Grimm, L. J., & Harvey, J. A. (2022). Practical steps to writing a scientific manuscript. *Journal of Breast Imaging, 4*(6), 640-648. https://doi.org/10.1093/jbi/wbac059
21. Heard, S. B. (2016). *The scientist's guide to writing: How to write more easily and effectively throughout your scientific career*. Princeton: Princeton University Press. https://doi.org/10.1515/9781400881147
22. Hsu, H. P. (2023). Can generative artificial intelligence write an academic journal article? Opportunities, challenges, and implications. *Irish Journal of Technology Enhanced Learning, 7*(2), 158-171. https://doi.org/10.22554/ijtel.v7i2.152
23. Hadan, H., Wang, D. M., Mogavi, R. H., Tu, J., Zhang-Kennedy, L., & Nacke, L. E. (2024). The great AI witch hunt: Reviewers' perception and (Mis) conception of generative AI in research writing. *Computers in Human Behavior: Artificial Humans*, *2*(2), https://doi.org/10.1016/j.chbah.2024.100095
24. Huang, J., & Tan, M. (2023). The role of ChatGPT in scientific communication: Writing better scientific review articles. *American Journal of Cancer Research, 13*(4), 1148. http://www.ncbi.nlm.nih.gov/pmc/articles/pmc10164801/
25. Hwang, T., Aggarwal, N., Khan, P. Z., et al. (2024). Can ChatGPT assist authors with abstract writing in medical journals? Evaluating the quality of scientific abstracts generated by ChatGPT and original abstracts. *PLoS ONE, 19*(2), e0297701. https://doi.org/10.1371/journal.pone.0297701

 doi:10.20944/preprints202605.1200.v1

26. Jenko, N., Ariyaratne, S., Jeys, L., et al. (2024). An evaluation of AI-generated literature reviews in musculoskeletal radiology. *The Surgeon*. https://doi.org/10.1016/j.surge.2023.12.005
27. Kacena, M. A., Plotkin, L. I., & Fehrenbacher, J. C. (2024). The use of artificial intelligence in writing scientific review articles. *Current Osteoporosis Reports, 1-7*. https://doi.org/10.1007/s11914-023-00852-0
28. Kaliyadan, F., & Seetharam, K. A. (2023). ChatGPT-Quo Vadis? *Indian Dermatology Online Journal, 14*(4), 457-458. https://doi.org/10.4103%2Fidoj.idoj_344_23
29. Khalifa, M., & Albadawy, M. (2024). Using artificial intelligence in academic writing and research: An essential productivity tool. *Computer Methods and Programs in Biomedicine Update, 100145*. https://doi.org/10.1016/j.cmpbup.2024.100145
30. Khan, N. A., Osmonaliev, K., & Sarwar, M. Z. (2023). Pushing the boundaries of scientific research with the use of artificial intelligence tools: Navigating risks and unleashing possibilities. *Nepal Journal of Epidemiology, 13*, 1258-1263. https://doi.org/10.3126/nje.v13i1.53721
31. Kim, J., & Peters, J. (2023). ChatGPT and its ethical implications for STEM research and higher education: A media discourse analysis. International Journal of STEM Education. https://doi.org/10.1186/s40594-023-00454-4
32. Kung, J. Y. (2023). Elicit. *The Journal of the Canadian Health Libraries Association, 44*(1), 15. https://doi.org/10.29173/jchla29657
33. Khera, R., Pedroso, A. F., Keloth, V. K., Xu, H., Silva, G. S., & Schwamm, L. H. (2025). Scientific Writing in the Era of Large Language Models: A Computational Analysis of AI-Versus Human-Created Content. *Stroke*, *56*(10), 3078-3083. 10.1161/STROKEAHA.125.051913
34. Kumar, R., Kumar, A., Selvam, A. A., Dhamu, I., & Balasubramanian, N. (2025). The impact of ChatGPT on orthopedic residents' scientific writing: A quasi-experimental study. *Journal of Clinical Orthopaedics and Trauma*, *69*, 103116.
35. Kusumegi, K., Yang, X., Ginsparg, P., de Vaan, M., Stuart, T., & Yin, Y. (2025). Scientific production in the era of large language models. *Science*, *390*(6779), 1240-1243. DOI: 10.1126/science.adw3000
36. Lindsay, D. M. (2020). Scientific writing = thinking in words. *Europhysics Letters, 143*. http://dx.doi.org/10.1071/9780643101579
37. Liang, W., Zhang, Y., Wu, Z., et al. (2024). Mapping the increasing use of LLMs in scientific papers. https://arxiv.org/pdf/2404.01268
38. Lozic, E., & Stular, B. (2023). Fluent but not factual: A comparative analysis of ChatGPT and other AI chatbots' proficiency and originality in scientific writing for humanities. *Future Internet, 15*(10), 336.
39. Lunsford, T. R., & Lunsford, B. R. (1996). How to critically read a journal research article. *JPO Journal of Prosthetics and Orthotics, 8*, 24-31. https://cdn.ymaws.com/www.oandp.org/resource/resmgr/docs/skc/journal-club/How_to_Critically_Read.pdf
40. Muga, G. (2023). Editorial—Artificial intelligence language models in scientific writing. *Europhysics Letters, 143*. https://ui.adsabs.harvard.edu/link_gateway/2023EL....14320000G/doi:10.1209/0295-5075/ace3ef
41. Mugaanyi, J., Cai, L., Cheng, S., et al. (2024). Evaluation of large language model performance and reliability for citations and references in scholarly writing: Cross-disciplinary study. *Journal of Medical Internet Research, 26*, e52935. https://doi.org/10.2196/52935
42. Nazarovets, S., & Teixeira da Silva, J. A. (2024). ChatGPT as an "author": Bibliometric analysis to assess the validity of authorship. *Accountability in Research, 1-11*. https://doi.org/10.1080/08989621.2024.2345713
43. Nielsen, E., & Reilly, P. L. (1985). A guide to understanding and evaluating research articles. *Gifted Child Quarterly, 29*, 90-92. https://doi.org/10.1177/001698628502900210
44. Okwemba, R. K. (2022). Introduction to scientific writing: A review. *International Journal of Scientific Research in Science and Technology*. https://doi.org/10.32628/IJSRST218631
45. Razack, H. I. A., Mathew, S. T., Saad, F. F. A., et al. (2021). Artificial intelligence-assisted tools for redefining the communication landscape of the scholarly world. *Science Editing, 8*(2), 134-144. https://doi.org/10.6087/kcse.244

 doi:10.20944/preprints202605.1200.v1

46. Pressman, S. M., Garcia, J. P., Borna, S., Gomez-Cabello, C. A., Haider, S. A., Haider, C. R., & Forte, A. J. (2025). Man versus machine: a comparative study of human and ChatGPT-generated abstracts in plastic surgery research. *Aesthetic Plastic Surgery*, *49*(17), 5013-5020. https://doi.org/10.1007/s00266-025-04836-6
47. Salimi, A., & Saheb, H. (2023). Large language models in ophthalmology scientific writing: Ethical considerations blurred lines or not at all? *American Journal of Ophthalmology*. https://doi.org/10.1016/j.ajo.2023.06.004
48. Smith, P., & Smith, L. (2023). This season's artificial intelligence (AI): Is today's AI really that different from the AI of the past? Some reflections and thoughts. *AI and Ethics, 1-4*. https://doi.org/10.1007/s43681-023-00388-0
49. Somashekhar, S. P. (2020). Art of scientific writing. *Indian Journal of Gynecologic Oncology, 18*, 1-3. https://doi.org/10.1007/s40944-020-00382-y
50. Vitente, A. C., Lazaro, R. T., Escuadra, C. J., et al. (2023). The use of artificial intelligence (AI)-assisted technologies in scientific discourse. *Philippine Journal of Physical Therapy*. https://doi.org/10.46409/002.HNUY6271
51. Watson, R. (2019). Avoiding desk rejection of a manuscript. *Nurse Author & Editor*. https://doi.org/10.1111/j.1750-4910.2019.tb00042.x