The Impact of AI Language Models on Scientific Writing and Scientific Peer Reviews: A Systematic Literature Review

Rand Alchokr
Otto-von-Guericke-University
Magdeburg, Germany
rand.alchokr@ovgu.de

Evelyn Starzew
Otto-von-Guericke-University
Magdeburg, Germany
evelyn.starzew@st.ovgu.de

Gunter Saake
Otto-von-Guericke-University
Magdeburg, Germany
saake@ovgu.de

Thomas Leich
Harz University & METOP GmbH
Weringerode & Magdeburg, Germany
tleich@hs-harz.de

Jacob Krüger
Eindhoven University of Technology
Eindhoven, The Netherlands
j.kruger@tue.nl

ABSTRACT

Recent advances in deep learning have led to the development of well-known AI language models, such as ChatGPT. Such models have gained widespread attention across various domains, including scientific research. In this context, discussions about the use of these models for writing and reviewing publications have started. Within this paper, we discuss the implications of integrating AI language models into the scientific writing process and provide a comprehensive overview of existing research on this topic. Therefore, we searched, describe, summarize, and organize existing research following systematic literature-review guidelines using the digital library Scopus. Since peer review is a crucial part of scientific research, we also focus on exploring the consequent impact of emerging models on the peer-reviewing process. Existing studies show that AI language models are used significantly in scientific writing. However, this usage requires guidelines and control to overcome potential challenges and problems.

CCS CONCEPTS

• **General and reference** → **General literature**.

KEYWORDS

Large Language Models, Scientific Writing, Peer Review, ChatGPT

ACM Reference Format:
Rand Alchokr, Evelyn Starzew, Gunter Saake, Thomas Leich, and Jacob Krüger. 2024. The Impact of AI Language Models on Scientific Writing and Scientific Peer Reviews: A Systematic Literature Review. In *The 2024 ACM/IEEE Joint Conference on Digital Libraries (JCDL '24), December 16–20, 2024, Hong Kong, China.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3677389.3702508

JCDL '24, December 16–20, 2024, Hong Kong, China

ACM ISBN 979-8-4007-1093-3/24/12
https://doi.org/10.1145/3677389.3702508

1 INTRODUCTION

The release of ChatGPT in 2022 received huge public attention [40]. ChatGPT allows its users to chat with software capable of generating human-like responses by incorporating a large knowledge base; and has fascinated users worldwide. Newspapers reported that it reached 100 million users just two months after its launch [17] and about the record-breaking 80 million user milestone [5]. ChatGPT results from recent developments in deep learning, which have led to the emergence of Large Language Models (LLMs).

The interest in LLMs can also be seen in scientific research, where the number of papers on this topic has strongly increased since [20, 35, 52]. LLMs have different capabilities depending on the targeted task. One possible task is text generation: these models could even be used to write a complete scientific article [52]. Consequently, controversies and discussions about the use of ChatGPT in scientific publications have spread rapidly [47]. Several studies assess the adequacy, opportunities, and challenges [3, 22, 36], while others warn against the use of such models for scientific work [44]. Moreover, negative hype regarding the use of ChatGPT for cheating, privacy risks, security, and potentially harmful content has formed [28]. A complementary application for LLMs in science is the peer-review process. Peer reviews aim to ensure high-quality research while upholding scientific standards [51]. However, there are many challenges around this process [4]. The question here is how do the advancements of LLMs impact peer reviews?

It is easy to lose track of all the different studies on LLMs in scientific writing and reviewing. Therefore, in this paper, we aim to describe, summarize, and organize the existing research on the use of LLMs for scientific writing and to explore the impact it has on peer-reviewing. Overall, our contribution is a structured literature review to answer two research questions (RQs):

- RQ_1 *How do AI language models contribute to scientific writing processes?*
- RQ_2 *What is the impact of AI language models on scientific peer reviewing?*

By answering these RQs, we hope to enable reasoning about the potential transformations of the peer-review process and to point out possibilities for improvements.

2 BACKGROUND

The basis of LLMs is language modeling (LM). The goal of LM is to model the likelihood of a specific word or sequence of words [52]. By now, four generations of language models exist: First, *Statistical Language Models* (SLMs) are based on statistical learning methods. Second, *Neural Language Models* (NLMs) are characterized by neural networks that predict the probability of word sequences. Third, *pre-trained language models* (PLMs) have introduced the paradigm of pre-training and fine-tuning to capture word representations depending on the context. These are then fine-tuned to work in specific downstream tasks. Lastly, *Large Language Models* (LLMs) have been developed, which are scaled PLMs to increase their abilities in solving tasks [38, 52]. LLMs are usually deep neural networks with a huge amount of parameters and are fundamentally built on the transformer architecture, which may also incorporate other structures [49, 52]. Such LLMs are pre-trained on vast amounts of text data, typically consisting of different publicly available textual datasets like Wikipedia articles, webpages, or books. The model captures the characteristics of the data including any correct and incorrect, toxic, biased, or harmful content [28, 52]. The breakthrough of LLMs can be attributed to the development of the transformer architecture, combined with the increased computational power and availability of huge amounts of training data [38].

ChatGPT is one of the most popular LLMs currently available. It is based on the generative pre-trained transformer GPT-3.5. GPT-4 was released in 2023 and has superior performance to earlier versions [52]. ChatGPT has a web-based browser interface or can be used as a mobile app. A user can ask questions in a dialogue field, which are being answered by generated text. These answers can then be detailed through further questions or remarks [28], which represents an optimization for conversations [52]. Due to these advancements, using LLMs for scientific writing and peer reviewing has become an important topic.

3 METHODOLOGY

To achieve our goal and answer our research questions, we employed a Systematic Literature Review (SLR) following the guidelines for software-engineering research proposed by Kitchenham et al. [25]. Using these guidelines, our methodology involves the steps we illustrate in Figure 1.

Literature Search. To answer RQ_1 and RQ_2, we applied the following search string on the Scopus digital library[1], which has a wide coverage of literature and enables easy and convenient search in the database [26, 43].

```
"((‘language model’ OR LLM OR ‘large language model’ OR
ChatGPT AND (‘peer review*’ OR ‘scientific writing’))"
```

We used inclusion and exclusion criteria to guide the selection of relevant studies. Specifically, selected studies should be peer-reviewed; published in a journal, workshop, or conference; and written in English. We assessed the quality of the corpus by answering established quality assessment questions [25].

Conduct. In Figure 1, we present the steps of our selection process during its conduct, including the number of remaining and excluded papers. We executed this process from April 2024 to May 2024. Finally, we included 30 papers as relevant primary studies.

[1]https://www.elsevier.com/products/scopus

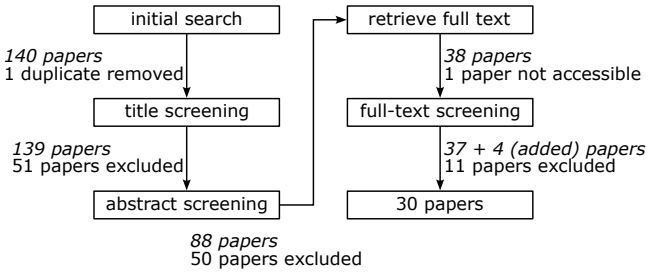

Figure 1: Flow chart of our study-selection process.

4 RESULTS

The 30 papers in our final corpus have been published from 2020 to 2024. Overall, 23 of them help answering RQ_1, whereas seven correspond to RQ_2. The papers belong to different domains: 18 out of 30 stem from the medical domain, five from computer science, and the rest from a variety of domains (e.g., education, neuroscience).

4.1 Contribution of LLMs to Scientific Writing

For RQ_1, we analyzed 23 papers, which we summarize in Table 1.

LLMs are Used for Scientific Writing. Liang et al. [32] show a steady increase in LLM usage for scientific writing, with the fastest and most significant growth in computer science. The authors conducted a large-scale systematic analysis of 950,000 papers published after the release of ChatGPT, measuring the fraction of generated text in the abstract and introduction. They emphasize that they measured not only the editing of text, but also substantial modifications. Similarly, LLMs can be used to improve language quality by editing human-written text [2]. Consequently, LLMs are already used extensively for scientific writing.

Use Cases of LLMs for Scientific Writing Differ. The analyzed papers differ in their understanding of scientific writing, meaning that they try to assess the ability to use LLMs for scientific writing by focusing on different aspects (cf. Table 1). We clustered the investigated use cases along three groups:

- **Generating a complete scientific paper.** Here, we further distinguished between two types of papers:
  - **Scientific articles.** Articles can be generated iteratively via prompts [15, 18], using data as a basis [34] and via question answering [27] or in a hybrid approach with a human-in-the-loop [11]. Májovský et al. [37] attempted to create a completely fraudulent scientific article via prompts.
  - **Literature reviews.** Literature reviews, such as SLRs, have been iteratively written by first generating an outline and then further describing the respective sections [46]. The text was generated by copying entire papers into the chat interface and prompting LLMs to screen the literature and extract relevant information [24] or by asking handcrafted questions across different rounds [9].
- **Generating parts of a scientific paper.** Dubinski et al. [14] as well as Kassem and Michahelles [23] examined the

use case of summarizing papers. Others have analyzed the more specific use case of generating abstracts based on the full text of a paper [29] based on data [6], title, and journal name [16] or iteratively via prompts [1]. Further use cases like using LLMs for introductions [1], reports [14], cover letters [13], and background sections [21] have been tested.
- **Miscellaneous.** The miscellaneous group consists of other use cases, such as, detecting AI-generated text in papers [30, 41], answering scientific questions [2, 39, 41], editing and fact-checking [23] or generating meta-reviews [7].

Neat Text: LLMs Can Generate Good-Looking Text. Research shows that LLMs are able generate "good-looking" scientific text that meets requirements for style and format [6, 7, 15, 18, 27, 34, 46]. This is especially true for cover letters, where no significant difference between human-written and AI-generated text was detected [13] or in sections like background, summaries [21] and abstracts [2] in which no novel knowledge is required. Actually, Gao et al. [16] showed that original abstracts tend to have a higher plagiarism score than generated abstracts. The generated texts seem to be legit and scientifically correct, leading to time-saving in the writing process. A statistically significant increase in efficiency has been shown by multiple studies [14, 15, 34].

Uncertain Authorship: AI-Generated Text is Difficult to Distinguish from Human-Written Text. Humans and detectors tend to struggle to determine the real authorship of a text sample [1, 16, 21, 29]. However, being able to determine authorship depends heavily on the choice of the text sample and the used detector. Therefore, high-quality detectors are crucial for transparent scientific writing [41]. In addition, Abani et al. [1] found that the ability to detect generated scientific texts worsens with the inspecting person having less knowledge about the specific research area. However, humans tend to perceive the generated texts to be superficial and vague [16]. Importantly, since the authors of a paper are responsible for the content, LLMs cannot be seen as authors [6, 12, 18, 19, 23, 34, 42].

Hallucinations in Generated Texts. LLMs like ChatGPT tend to hallucinate, which means that they can state incorrect information with high confidence. Hallucinations can be seen in factually inaccurate or wrong statements that may seem valid until closer inspection [1, 2, 9, 11, 33, 39, 46]. Especially, AI language models tend to hallucinate references. Either these references are completely made up [18, 27, 37, 41] or parts of the references are incorrect [14, 15, 34]. The difficulty in identifying hallucinations leads to the danger of misleading information being perceived as valid [46].

Human-Written Texts are Superior to Generated Ones. Current LLMs cannot always reach human-level quality [29]. Reasons for this may relate to superficiality in specific domains [27] and incorporated biases [2, 11]. This is also visible in the inability of LLMs to generate novel contributions and research questions [21, 33] or to think critically [33]. Furthermore, LLMs are not useful for fact-checking [24].

Outdated Information in Generated Texts. Generated texts do not always build on the most recent developments in a research field, but rather outdated information [2, 23, 39, 46].

Further Limitations: Guidelines and Ethical Concerns. Researchers have noted that ChatGPT does not work well in all areas of scientific writing. One example is conducting a literature review [27]. Even if ChatGPT is used for a literature review, it performs badly at screening and abstracting data [24]—and yields questionable completeness [9]. Additionally, the quality of responses given by LLMs depends on how well a prompt is written [1]. Still, Macdonald et al. [34] were able to show the ability of ChatGPT to self-correct errors after giving feedback.

Ethical pitfalls are an important concern related to using LLMs in scientific research. Disclosing the use of LLMs in the creation of papers is necessary to overcome this concern. Using such models can assist researchers in drafting high-quality scientific articles [13]. However, a lack of transparency puts reliability at risk [2, 41], which, in turn, led to discouragements of using LLMs for scientific writing [18]. Further problems include plagiarism [1, 6, 11], privacy concerns [11, 14, 27], and limitations regarding reproducibility [23, 29].

4.2 Impact of LLMs on Peer Reviews

To answer RQ_2, we analyzed seven primary studies.

Use Cases of LLMs for Peer Reviewing Vary. Liang et al. [30] investigated peer reviews at AI conferences and stated that approximately 15 % of them were written with significant contributions by LLMs. The seven studies we identified focused on three different use cases: writing, analyzing, and guiding peer reviews. Writing peer reviews can be further distinguished in writing as a first reviewer [8, 10, 31] or as a second reviewer [45]. Verharen [50] analyzed the used language and subjectivity of peer reviews with an LLM. Lastly, Su et al. [48] aimed to develop an LLM-based tool to support the writing of peer reviews. In addition to contributing to the writing, LLMs can be used to analyze existing human-written reviews to reveal trends and problems with fairness and bias. Therefore, LLMs can help reveal problems, understand certain phenomena, and improve existing processes.

Generated and Human-Written Reviews are Comparable. Liang et al. [31] conducted a large-scale empirical analysis of ChatGPT's ability to generate scientific feedback. They compared human-written and generated reviews of 3,000 papers and measured the overlap between both. The results indicate that the overlap is comparable to the overlap between two different human-written reviews. Moreover, the generated feedback was non-generic and perceived helpful by the original authors. These results were confirmed in a feasibility study [8]. The authors successfully generated valuable feedback on clarity, organization, and writing style, which showed a high level of agreement with reviews written by humans. Overall, this can lead to time savings for reviewers.

Divergence from Original Reviews. Subtle differences between human-written and generated reviews for complex articles have been identified [8]. Specifically, LLMs tend to focus on different aspects than humans [31]. These differences are also reflected in the inability of LLMs to analyze figures and images—contrary to humans [8]. Verharen [50] conducted a large-scale study on 500 peer reviews to analyze the subjectivity of the reviews with the help of ChatGPT and to simultaneously assess whether ChatGPT can perform language analysis of scientific texts. Although the majority of language used was polite, Verharen identified a gender

Table 1: Summary of the 23 primary studies related to RQ_1.

Ref	Use Case	Model	Neat Text	Uncertain Authorship	Halluci-nations	Human Superiority	Outdated Info	Guidelines, Transparency Need	Ethical Concerns
[1]	Abstract, Introduction, References	ChatGPT (05/2023)		●	●	●			●
[2]	Answering scientific questions	ChatGPT-3.5, ChatGPT-4	●		●	●	●	●	●
[6]	Abstract	GPT-4	●		●	●			●
[7]	Meta-Reviews	UniLM	●			●			
[9]	SLR	ChatGPT-4				●			●
[11]	Scientific article (Hybrid approach)	ChatGPT-4			●	●		●	●
[13]	Cover letters	ChatGPT-4	●			●		●	
[14]	Summaries reports	ChatGPT Jan 9 Version	●		●	●			●
[15]	Scientific article	ChatGPT PLUS (GPT-4)	●		●	●			●
[16]	Abstract	ChatGPT-3	●	●		●			●
[18]	Scientific article	ChatGPT-4	●		●	●			●
[21]	Background	GPT-3.5	●	●		●			●
[23]	Summary, Editing, Fact-Checking	ChatGPT (Feb13)				●	●	●	●
[24]	SLR	GPT-4				●			●
[27]	Scientific article	ChatGPT-3.5, Bard (May 2023)	●		●			●	●
[29]	Abstract	ChatGPT-3		●		●			●
[32]	Analysis of LLM in abstracts, introduction	Distributional LLM Quantification Framework							
[33]	Answering scientific questions	ChatGPT-3.5, ChatGPT-4, Bing, Bard, Claude 2, Aria			●	●			●
[34]	Scientific article	ChatGPT	●		●	●		●	●
[37]	Scientific article	ChatGPT-3			●	●			●
[39]	Answering scientific questions	ChatGPT (April 5th 2023), ChatSonic, New Bing, YouChat			●	●	●	●	●
[41]	Detection of AI generated scientific writing	ChatGPT		●	●	●		●	●
[46]	Review article	ChatGPT-4	●		●	●	●	●	●

bias. Female authors tend to receive less polite reviews than male authors. Moreover, they recognized high variability in the scoring of the same paper from different reviewers, which shows the possible subjectivity in peer reviews.

Different Models Disagree. Hallucinations and disagreements between different models represent another limitation. Researchers have tested the ability of different LLMs to generate reviews, and found hallucinations of titles as well as references [10, 45]. The results show that LLMs perform differently in such tasks and they disagreed many times. So, the outcomes of such reviews are inconsistent and depend on the model choice.

Further Limitations: Guidelines and Ethical Concerns. Saad et al. [45] discuss the poor ability of LLMs to score a paper and generate a correlated acceptance prediction. To test this, they prompted different versions of ChatGPT to give feedback, scores, and acceptance or rejection predictions on a paper. They argued that the models gave mostly positive results and were not able to identify manuscripts not meeting the editorial standards. This limitation could lead to inconsistent practices and ethical breaches. Therefore, researchers strongly recommend guidelines to define how to support decision-making, improve quality as well as transparency, and avoid biased responses [2, 8–11, 31, 45]. Overall, LLMs can guide the reviewing process, but are not able to replace human reviewers.

5 CONCLUSION

In this paper, we reported the results of an SLR on LLMs' ability in assisting scientific writing and peer reviewing, including their limitations and potential risks. LLMs can make a positive contribution to these processes, especially in cases where no novel contribution or critical thinking are required (e.g., summarizing, abstracts, language polishing). Consequently, we found a consensus in the literature that human skills in scientific writing are superior to LLMs. The models can still reduce the time needed, improve language quality, and help structure the writing process. Guidelines and standards are an important step to support the transparent and reliable use of LLMs in this direction.

REFERENCES

[1] S. Abani, H.A. Volk, S. De Decker, J. Fenn, C. Rusbridge, M. Charalambous, R. Goncalves, R. Gutierrez-Quintana, S. Loderstedt, T. Flegel, C. Ros, T.V. Klopmann, H.C. Schenk, M. Kornberg, N. Meyerhoff, A. Tipold, and J.N. Nessler. 2023. ChatGPT and scientific papers in veterinary neurology; is the genie out of the bottle? *Frontiers in Veterinary Science* 10 (2023). https://doi.org/10.3389/fvets.2023.1272755

[2] O. Abuyaman. 2023. Strengths and weaknesses of ChatGPT models for scientific writing about medical vitamin B12: Mixed methods study. *JMIR Formative Research* 7, 1 (2023). https://doi.org/10.2196/49459

[3] A. Aghemo, A. Forner, and L. Valenti. 2023. Should artificial intelligence-based language models be allowed in developing scientific manuscripts? A debate between ChatGPT and the editors of Liver International. *Liver International* 43, 5 (2023), 956–957. https://doi.org/10.1111/liv.15580

[4] R. Alchokr, J. Krüger, Y. Shakeel, G. Saake, and T. Leich. 2022. Peer-reviewing and submission dynamics around top software-engineering venues: A juniors' perspective. In *International Conference on Evaluation and Assessment in Software Engineering (EASE)*. ACM, 60–69. https://doi.org/10.1145/3530019.3530026

[5] H Anwar. 2024. *ChatGPT Celebrates Record 80M Daily User Milestone Amid Rumors Of A Possible Search Product Launch.* https://www.digitalinformationworld.com/2024/05/chatgpt-celebrates-record-80m-daily.html

[6] F.E. Babl and M.P. Babl. 2023. Generative artificial intelligence: can ChatGPT write a quality abstract? *EMA - Emergency Medicine Australasia* 35, 5 (2023), 809–811. https://doi.org/10.1111/1742-6723.14233

[7] C. Bhatia, T. Pradhan, and S. Pal. 2020. MetaGen: An academic meta-review generation system. In *SIGIR 2020 - Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (2020). 1653–1656. https://doi.org/10.1145/3397271.3401190

[8] S. Biswas, D. Dobaria, and H.L. Cohen. 2023. ChatGPT and the future of journal reviews: A feasibility study. *Yale Journal of Biology and Medicine* 96, 3 (2023), 415–420. https://doi.org/10.59249/SKDH9286

[9] A. Bond, D. Cilliers, F. Retief, R. Alberts, C. Roos, and J. Moolman. 2024. Using an artificial intelligence chatbot to critically review the scientific literature on the use of artificial intelligence in environmental impact assessment. *Impact Assessment and Project Appraisal* (2024). https://doi.org/10.1080/14615517.2024.2320591

[10] D. Carabantes, J.L. González-Geraldo, and G. Jover. 2023. ChatGPT could be the reviewer of your next scientific paper. Evidence on the limits of AI-assisted academic reviews. *Profesional de la Informacion* 32, 5 (2023). https://doi.org/10.3145/epi.2023.sep.16

[11] M. Cascella, J. Montomoli, V. Bellini, A. Ottaiano, M. Santorsola, F. Perri, F. Sabbatino, A. Vittori, and E.G. Bignami. 2023. Writing the paper "Unveiling artificial intelligence: an insight into ethics and applications in anesthesia" implementing the large language model ChatGPT: a qualitative study. *Journal of Medical Artificial Intelligence* 6 (2023). https://doi.org/10.21037/jmai-23-13

[12] I. Dergaa, K. Chamari, P. Zmijewski, and H.B. Saad. 2023. From human writing to artificial intelligence generated text: examining the prospects and potential threats of ChatGPT in academic writing. *Biology of Sport* 40, 2 (2023), 615–622. https://doi.org/10.5114/BIOLSPORT.2023.125623

[13] C.D. Deveci, J.J. Baker, B. Sikander, and J. Rosenberg. 2024. A comparison of cover letters written by ChatGPT-4 or humans. *Ugeskrift for Laeger* 186, 1 (2024), 1–1.

[14] D. Dubinski, S.-Y. Won, S. Trnovec, B. Behmanesh, P. Baumgarten, N. Dinc, J. Konczalla, A. Chan, J.D. Bernstock, T.M. Freiman, and F. Gessler. 2024. Leveraging artificial intelligence in neurosurgery—unveiling ChatGPT for neurosurgical discharge summaries and operative reports. *Acta Neurochirurgica* 166, 1 (2024). https://doi.org/10.1007/s00701-024-05908-3

[15] M. Elbadawi, H. Li, A.W. Basit, and S. Gaisford. 2024. The role of artificial intelligence in generating original scientific research. *International Journal of Pharmaceutics* 652 (2024). https://doi.org/10.1016/j.ijpharm.2023.123741

[16] C.A. Gao, F.M. Howard, N.S. Markov, E.C. Dyer, S. Ramesh, Y. Luo, and A.T. Pearson. 2023. Comparing scientific abstracts generated by ChatGPT to real abstracts with detectors and blinded human reviewers. *npj Digital Medicine* 6, 1 (2023), 75. https://doi.org/10.1038/s41746-023-00819-6

[17] The Guardian. 2023. *ChatGPT reaches 100 million users two months after launch.* https://www.theguardian.com/technology/2023/feb/02/chatgpt-100-million-users-open-ai-fastest-growing-app

[18] A. Guleria, K. Krishan, V. Sharma, and T. Kanchan. 2023. ChatGPT: ethical concerns and challenges in academics and research. *Journal of Infection in Developing Countries* 17, 9 (2023), 1292–1299. https://doi.org/10.3855/jidc.18738

[19] B.N. Hryciw, A.J.E. Seely, and K. Kyeremanteng. 2023. Guiding principles and proposed classification system for the responsible adoption of artificial intelligence in scientific writing in medicine. *Frontiers in Artificial Intelligence* 6 (2023). https://doi.org/10.3389/frai.2023.1283353

[20] J. Huang and M. Tan. 2023. The role of ChatGPT in scientific communication: writing better scientific review articles. *American Journal of Cancer Research* 13 4 (2023), 1148–1154. https://api.semanticscholar.org/CorpusID:258638310

[21] I.A. Huespe, J. Echeverri, A. Khalid, I. Carboni Bisso, C.G. Musso, S. Surani, V. Bansal, and R. Kashyap. 2023. Clinical research with large language models generated writing - clinical research with AI-assisted writing (CRAW) study. *Critical Care Explorations* 5, 10 (2023). https://doi.org/10.1097/CCE.0000000000000975

[22] E. Kasneci, K. Sessler, S. Küchemann, M. Bannert, D. Dementieva, F. Fischer, U. Gasser, G. Groh, S. Günnemann, E. Hüllermeier, S. Krusche, G. Kutyniok, T. Michaeli, C. Nerdel, J. Pfeffer, O. Poquet, M. Sailer, A. Schmidt, T. Seidel, M. Stadler, J. Weller, J. Kuhn, and G. Kasneci. 2023. ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences* 103 (2023). https://doi.org/10.1016/j.lindif.2023.102274

[23] K. Kassem and F. Michahelles. 2023. Etmachina: exploring the use of conversational agents such as ChatGPT in scientific writing. In *CEUR Workshop Proceedings* (2023), Vol. 3502.

[24] Q. Khraisha, S. Put, J. Kappenberg, A. Warraitch, and K. Hadfield. 2024. Can large language models replace humans in systematic reviews? Evaluating GPT-4's efficacy in screening and extracting data from peer-reviewed and grey literature in multiple languages. *Research Synthesis Methods* (2024). https://doi.org/10.1002/jrsm.1715

[25] B.A. Kitchenham, O. Pearl Brereton, D. Budgen, M. Turner, J. Bailey, and S. Linkman. 2009. Systematic literature reviews in software engineering – A systematic literature review. *Information and Software Technology* 51, 1 (2009), 7–15. https://doi.org/10.1016/j.infsof.2008.09.009

[26] J. Krüger, C. Lausberger, I. von Nostitz-Wallwitz, G. Saake, and T. Leich. 2020. Search. review. repeat? an empirical study of threats to replicating SLR searches. *Empirical Software Engineering* 25, 1 (2020), 627–677. https://doi.org/10.1007/s10664-019-09763-0

[27] A. Labouchère and W. Raffoul. 2024. ChatGPT and bard in plastic surgery: hype or hope? *Surgeries (Switzerland)* 5, 1 (2024), 37–48. https://doi.org/10.3390/surgeries5010006

[28] J. Lambert and M. Stevens. 2023. ChatGPT and generative AI technology: a mixed bag of concerns and new opportunities. *Computers in the Schools* (2023), 1–25. https://doi.org/10.1080/07380569.2023.2256710

[29] K.W. Lawrence, A.A. Habibi, S.A. Ward, C.M. Lajam, R. Schwarzkopf, and J.C. Rozell. 2024. Human versus artificial intelligence-generated arthroplasty literature: A single-blinded analysis of perceived communication, quality, and authorship source. *International Journal of Medical Robotics and Computer Assisted Surgery* 20, 1 (2024). https://doi.org/10.1002/rcs.2621

[30] W. Liang, Z. Izzo, Y. Zhang, H. Lepp, H. Cao, X. Zhao, L. Chen, H. Ye, S. Liu, Z. Huang, D.A. McFarland, and J.Y. Zou. 2024. Monitoring AI-modified content at scale: a case study on the impact of ChatGPT on AI conference peer reviews. arXiv:2403.07183 [cs] http://arxiv.org/abs/2403.07183

[31] W. Liang, Y. Zhang, H. Cao, B. Wang, D. Ding, X. Yang, K. Vodrahalli, S. He, D. Smith, Y. Yin, D. McFarland, and J. Zou. 2023. Can large language models

provide useful feedback on research papers? A large-scale empirical analysis. arXiv:2310.01783 [cs] http://arxiv.org/abs/2310.01783

[32] W. Liang, Y. Zhang, Z. Wu, H. Lepp, W. Ji, X. Zhao, H. Cao, S. Liu, S. He, Z. Huang, D. Yang, C. Potts, C.D. Manning, and J.Y. Zou. 2024. Mapping the increasing use of LLMs in scientific papers. arXiv:2404.01268 [cs] http://arxiv.org/abs/2404.01268

[33] E. Lozić and B. Štular. 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 (2023). https://doi.org/10.3390/fi15100336

[34] C. Macdonald, D. Adeloye, A. Sheikh, and I. Rudan. 2023. Can ChatGPT draft a research article? An example of population-level vaccine effectiveness analysis. *Journal of Global Health* 13 (2023). https://doi.org/10.7189/JOGH.13.01003

[35] R. Megawati, H. Listiani, N. Pranoto, M. Akobiarek, and Ruth S. 2023. Role of GPT chat in writing scientific articles: a systematic literature review. *Jurnal Penelitian Pendidikan IPA* 9 (11 2023), 1078–1084. https://doi.org/10.29303/jppipa.v9i11.5559

[36] J.G. Meyer, R.J. Urbanowicz, P.C.N. Martin, K. O'Connor, R. Li, P.-C. Peng, T.J. Bright, N. Tatonetti, K.J. Won, G. Gonzalez-Hernandez, and J.H. Moore. 2023. ChatGPT and large language models in academia: opportunities and challenges. *BioData Mining* 16, 1 (2023). https://doi.org/10.1186/s13040-023-00339-9

[37] M. Májovský, M. Černý, M. Kasal, M. Komarc, and D. Netuka. 2023. Artificial intelligence can generate fraudulent but authentic-looking scientific medical articles: pandora's box has been opened. *Journal of Medical Internet Research* 25 (2023). https://doi.org/10.2196/46924

[38] H. Naveed, A.U. Khan, S. Qiu, M. Saqib, S. Anwar, M. Usman, N. Akhtar, N. Barnes, and A.l Mian. 2024. A comprehensive overview of large language models. arXiv:2307.06435 [cs] http://arxiv.org/abs/2307.06435

[39] D. Negrini and G. Lippi. 2023. Generative artificial intelligence in (laboratory) medicine: friend or foe? *Biochimica Clinica* 47, 3 (2023), 259–265. https://doi.org/10.19186/BC_2023.025

[40] OpenAI. 2024. ChatGPT (October 2023 version). https://openai.com. Accessed: 2024-10-01.

[41] A. Pawlicka, M. Pawlicki, R. Kozik, and M. Choraś. 2024. *The rise of AI-powered writing: how ChatGPT is revolutionizing scientific communication for better or for worse.* Communications in Computer and Information Science, Vol. 2014 CCIS. https://doi.org/10.1007/978-981-97-0903-8_30 Pages: 327.

[42] M. Perkins and J. Roe. 2024. Academic publisher guidelines on AI usage: a ChatGPT supported thematic analysis. *F1000Research* 12 (2024). https://doi.org/10.12688/f1000research.142411.2

[43] R. Pranckutė. 2021. Web of science (WoS) and Scopus: the titans of bibliographic information in today's academic world. *Publications* 9, 1 (2021). https://doi.org/10.3390/publications9010012

[44] K. Quach. 2024. *Researchers warned against using AI to peer review academic papers.* https://www.semafor.com/article/05/08/2024/researchers-warned-against-using-ai-to-peer-review-academic-papers

[45] A. Saad, N. Jenko, S. Ariyaratne, N. Birch, K.P. Iyengar, A.M. Davies, R. Vaishya, and R. Botchu. 2024. Exploring the potential of ChatGPT in the peer review process: an observational study. *Diabetes and Metabolic Syndrome: Clinical Research and Reviews* 18, 2 (2024). https://doi.org/10.1016/j.dsx.2024.102946

[46] M. Safrai and K.E. Orwig. 2024. Utilizing artificial intelligence in academic writing: an in-depth evaluation of a scientific review on fertility preservation written by ChatGPT-4. *Journal of Assisted Reproduction and Genetics* (2024). https://doi.org/10.1007/s10815-024-03089-7

[47] C. Stokel-Walker. 2024. *AI Chatbots Have Thoroughly Infiltrated Scientific Publishing.* https://www.scientificamerican.com/article/chatbots-have-thoroughly-infiltrated-scientific-publishing/

[48] X. Su, T. Wambsganss, R. Rietsche, S.P. Neshaei, and T. Käser. 2023. Reviewriter: AI-generated instructions for peer review writing. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics* (2023). 57–71.

[49] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, and I. Polosukhin. 2023. Attention is all you need. arXiv:1706.03762 [cs] http://arxiv.org/abs/1706.03762

[50] J.P. Verharen. 2023. ChatGPT identifies gender disparities in scientific peer review. *eLife* 12 (2023). https://doi.org/10.7554/eLife.90230

[51] M.L. Voight and B.J. Hoogenboom. 2012. Publishing your work in a journal: understanding the peer review process. *International Journal of Sports Physical Therapy* 7, 5 (2012), 452–460.

[52] W.X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong, Y. Du, C. Yang, Y. Chen, Z. Chen, J. Jiang, R. Ren, Y. Li, X. Tang, Z. Liu, P. Liu, J.-Y. Nie, and J.-R. Wen. [n. d.]. A survey of large language models. arXiv:2303.18223 [cs] http://arxiv.org/abs/2303.18223