


AI language models in human reproduction research: exploring ChatGPT's potential to assist academic writing

N. Semrl ¹, S. Feigl¹, N. Taumberger ^{1,*}, T. Bracic¹, H. Fluhr¹, C. Blockeel², and M. Kollmann ¹

¹Department of Obstetrics and Gynecology, Medical University of Graz, Graz, Austria,

²Centre for Reproductive Medicine, Universitair Ziekenhuis Brussel (UZ Brussel), Brussels, Belgium

*Correspondence address. Medical University of Graz, Auenbruggerplatz 14, 8036 Graz, Austria. E-mail: nadja.taumberger@medunigraz.at  <https://orcid.org/0000-0002-8969-7064>

ABSTRACT

Artificial intelligence (AI)-driven language models have the potential to serve as an educational tool, facilitate clinical decision-making, and support research and academic writing. The benefits of their use are yet to be evaluated and concerns have been raised regarding the accuracy, transparency, and ethical implications of using this AI technology in academic publishing. At the moment, Chat Generative Pre-trained Transformer (ChatGPT) is one of the most powerful and widely debated AI language models. Here, we discuss its feasibility to answer scientific questions, identify relevant literature, and assist writing in the field of human reproduction. With consideration of the scarcity of data on this topic, we assessed the feasibility of ChatGPT in academic writing, using data from six meta-analyses published in a leading journal of human reproduction. The text generated by ChatGPT was evaluated and compared to the original text by blinded reviewers. While ChatGPT can produce high-quality text and summarize information efficiently, its current ability to interpret data and answer scientific questions is limited, and it cannot be relied upon for a literature search or accurate source citation due to the potential spread of incomplete or false information. We advocate for open discussions within the reproductive medicine research community to explore the advantages and disadvantages of implementing this AI technology. Researchers and reviewers should be informed about AI language models, and we encourage authors to transparently disclose their use.

Keywords: artificial intelligence / language models / ChatGPT / academic writing / reproductive medicine

Introduction

Artificial intelligence (AI) language models are language processing models that utilize advanced deep learning algorithms trained on vast amounts of data to generate coherent and contextually appropriate texts that appear human-like (Salvagno *et al.*, 2023). One of the most powerful and widely debated systems currently is Chat Generative Pre-trained Transformer (ChatGPT). This AI-powered system was released by the software company OpenAI, based in San Francisco, CA, on 30 November 2022 (Sallam, 2023). It is currently freely accessible to a mass audience. Since its release it has become a sensation and provoked significant public discourse, as the impressive performance of such AI software will probably revolutionize many aspects of our lives. It not only provides endless entertainment; it has also demonstrated the ability to formulate patient's clinic letters (Ali *et al.*, 2023), simplify radiology reports (Jeblick *et al.*, 2022), and even pass the US Medical License Exam (USMLE) (Gilson *et al.*, 2023) and a law-school entrance exam (Choi *et al.*, 2023). Furthermore, its increasing ability to produce coherent language is making differentiation from human-authored text more and more challenging (Sallam, 2023).

Within the scientific community, ChatGPT has received mixed responses. There is widespread discussion about the potential of

ChatGPT to assist in the process of writing a scientific paper (Salvagno *et al.*, 2023) and some publishers have allowed ChatGPT to be credited with authorship in medical literature (Teixeira da Silva, 2023). On the other hand, concerns have been raised regarding the ethics (Liebrenz *et al.*, 2023), transparency (Thorp, 2023), accuracy (Wen and Wang, 2023), and medicolegal aspects of ChatGPT's writing (Sallam, 2023).

Reproductive medicine is a productive field in medical research and could benefit from AI assistance. It is important to have thorough discussions on the consequences of AI-generated content in academic publishing to develop comprehensive guidelines for its appropriate use and its boundaries. Here, we explore the advantages and disadvantages of utilizing AI-based language models in academic writing, supporting the arguments with an assessment of ChatGPT as an exemplary representative of such models.

Feasibility of ChatGPT's assistance in answering scientific questions

When considering the AI-supported scientific writing, one of the primary concerns is the potential threat to scientific integrity.

Received: May 10, 2023. Revised: September 6, 2023. Editorial decision: September 27, 2023.

© The Author(s) 2023. Published by Oxford University Press on behalf of European Society of Human Reproduction and Embryology. All rights reserved. For permissions, please email: journals.permissions@oup.com

The ability of AI language models to generate factually correct scientific text when prompted with specific medical topics is unknown. ChatGPT is currently one of the largest publicly available language models, relying on a vast corpus of text data. However, it operates with a database that is current only up to year 2021 (Sallam, 2023). Furthermore, several authors have reported occurrences of 'AI hallucination', where ChatGPT generates convincing yet entirely fabricated answers (Jeblick et al., 2022; Goddard, 2023; Sallam, 2023), posing a risk of spreading misinformation. Another important aspect to consider is the training process of ChatGPT. After pre-training on a large database, ChatGPT creates answers ranked by human users, leading to a learning process through trial and error (Kitamura, 2023), which can introduce bias. In the context of human reproduction research, this could reinforce geographic, socioeconomic, and/or cultural bias, amplifying only the perspectives and viewpoints of a specific group, favoring a certain therapeutic approach, or focusing on reproductive issues relevant only for a certain area or socioeconomic group with a better access to this language model.

All these issues raise important questions about accountability in the event of misinformation or inaccuracies being spread. Does the responsibility for text-integrity lie with the prompt creator or the ChatGPT's developer? Currently, there are no specific laws addressing the use of ChatGPT in scientific literature (Kitamura, 2023), and Open AI's terms of use permit users to reproduce the generated text for any purpose while holding them accountable for evaluating its accuracy (Open AI Terms of Use). We strongly advocate for thorough examination of ChatGPT's output by human authors, by comparing it against credible sources. Continuing this line of argument, we agree with the current guidelines of Elsevier (Elsevier Publishing Ethics) that ChatGPT should not be listed as an author, as it cannot bear the responsibility for the integrity of the text.

To substantiate these arguments, we conducted an evaluation of ChatGPT's performance. We searched for meta-analyses accessible in full text, published in the *Human Reproduction* from January 2022 to May 2023 to include only scientific papers that are not in the ChatGPT's knowledge database. *Human Reproduction* is a leading journal covering research in all areas of reproduction in human. Published articles have a well-structured abstract with clearly formulated study question, which was helpful when assessing the feasibility of ChatGPT to answer scientific questions. Six articles were identified (Arecco et al., 2022; Busnelli et al., 2022; Heymann et al., 2022; Fraison et al., 2023; Meernik et al., 2023; Rimmer et al., 2023).

ChatGPT was asked the exact same study question formulated in the abstract of the meta-analysis using the following prompt: '*study question*. Please answer in less than 40 words'. The output of ChatGPT is not deterministic. Each prompt was used two times in two separate chat windows to account for this variability. The formulated answers were evaluated for compliance with word-limitation (yes/no), grammatical correctness (yes/no), and factual correctness when comparing ChatGPT's output to the original answer of meta-analysis. We used free-text comments to describe potential flaws of the content of ChatGPT's answer in comparison to the original. Upon supervision, all authors agreed that ChatGPT provided believable, grammatically correct answers, mostly complying to the word-limitation. While it was able to generate high-quality text similar to the original, some misleading information and vaguely formulated output were identified. In Table 1, study answers written by ChatGPT, and the original answers are presented together with evaluation of the ChatGPT's output. We observed the use of declarative sentences

despite the literature only having demonstrated associations (study question from Arecco et al., 2022); ChatGPT reported statistically less significant results based on the content of the original meta-analysis, but not the statistically most significant results (study question from Busnelli et al. 2022), and drew false conclusions (study question from Heymann et al., 2022). We noted that ChatGPT's expression was mostly more vague compared to the original text. All authors agreed that the best-formulated response was for the study question of Rimmer et al. (2023).

This was an assessment of only 12 of ChatGPT's output texts. However, it revealed that both minor and more significant inaccuracies can be readily found when comparing the ChatGPT output to the published original work written by a human author.

As awareness of these problems grows, updates to ChatGPT could potentially address such issues and OpenAI reports that the latest (monetized ChatGPT 4) version surpasses its performance and further advancements are underway (Open AI). Our concerns reflect only the current situation. Additionally, plug-ins designed to provide ChatGPT with access to up-to-date information, analytical tools, and third-party services may enhance its performance. However, the program can (currently) not be used to answer scientific questions or perform fact-checking. It cannot replace critical human thinking and each output should be thoroughly examined.

Feasibility of ChatGPT to identify and cite relevant literature

Another area where AI language models could be helpful is assistance with literature searches and citations. The idea is tempting as it could save much time and resources, allowing authors to concentrate on data quality assessment or on planning of the experimental design of the study. However, several authors reported about inaccurate and insufficient citations (Sallam, 2023) from ChatGPT. Akhter and Cooper (2023) and Manohar and Prasad (2023), for example, reported the issues with citation of non-existing sources when using ChatGPT to write a case report.

To assess this aspect, we continued in the same chat window, using the prompt 'Please cite the literature on which your answer is based'. Secondly, the literature search strategy described in the meta-analysis was used as a prompt. Then, literature cited by ChatGPT was screened for: (i) eligibility and relevance of the cited studies; (ii) overlap with the literature cited in the meta-analysis; and (iii) citation accuracy regarding title, author(s), publication date, and digital object identifier (DOI) number.

When asking ChatGPT for citation of the studies on which the answer to the study question was based, there was almost no overlap with the cited literature in the respective meta-analysis (1/19; 5.3%). Only 36.8% (7/19) of citations were accurate regarding author name, title, publication date, and DOI number. However, most of them (14/19; 73.7%) were identifiable, as the inaccuracies were minor (e.g. publication year). On five occasions (26.3%), ChatGPT cited non-existing sources. And only 52.6% (10/19) of all ChatGPT citations (or 71.4% (10/14) of existing sources cited by ChatGPT) were evaluated as relevant for the study question (Table 2). After specifying the prompt, using the literature search strategy of the respective meta-analysis, and stating that only original articles are to be cited, the results remained similar (Table 3). Because of the very high false citation rate, we did not try to construct further prompts for the literature search. It seems that ChatGPT frequently lacks the ability to provide verifiable sources for its output. It is understandable that ChatGPT

Table 1. Comparison of two summary answers written by ChatGPT with the original answer to each respective study question.

Study question	ChatGPT answer I	ChatGPT answer II	Original answer	Evaluation of ChatGPT's output
Is it safe to perform controlled ovarian stimulation (COS) for fertility preservation before starting anticancer therapies or ART after treatments in young breast cancer patients? (Arecco et al., 2022)	COS before anticancer treatment is a viable option for fertility preservation in breast cancer patients but requires careful evaluation. ART after treatment can also be safe and effective. Consultation with a multidisciplinary team is recommended.	Yes, it is safe to perform controlled ovarian stimulation for fertility preservation before starting anticancer therapies or assisted reproduction technology after treatments in young breast cancer patients.	Performing COS before, or ART following anticancer treatment in young women with breast cancer does not seem to be associated with detrimental prognostic effect in terms of breast cancer recurrence, mortality or event-free survival (EFS).	<ul style="list-style-type: none"> • Word limitation: yes • Grammatical correctness: yes • Factual correctness: use of declarative sentences when the literature demonstrated only associations, answers formulated more vaguely
Is there an association between the different endometrial preparation protocols for frozen embryo transfer (FET) and obstetric and perinatal outcomes? (Busnelli et al., 2022)	Yes, different endometrial preparation protocols for FET may be associated with obstetric and perinatal outcomes such as live birth rate, gestational age, birth weight, and pre-term delivery.	Yes, there is evidence of an association between different endometrial preparation protocols for FET and obstetric and perinatal outcomes. However, the specific outcomes may vary depending on the protocol used, and further research is needed to determine optimal protocols for individual patients.	Programmed FET protocols were associated with a significantly higher risk of hypertensive disorders of pregnancy (HDP), pre-eclampsia (PE), post-partum hemorrhage (PPH) and cesarean section (CS) when compared with natural FET protocols.	<ul style="list-style-type: none"> • Word limitation: yes (Text 1), no (Text 2) • Grammatical correctness: yes • Factual correctness: statistically less significant associations reported, statistically most significant associations not reported
What are the associations between a history of cancer and outcomes after ART? (Meernik et al., 2023)	A history of cancer may negatively impact outcomes after assisted reproductive technology, including lower pregnancy rates, increased risk of miscarriage, and higher rates of birth defects in offspring.	A history of cancer may affect fertility and outcomes after assisted reproductive technology (ART), such as lower pregnancy rates and increased risk of miscarriage, premature delivery, and birth defects.	Compared to women without cancer, on average, women with cancer had a lower return for embryo transfer and a lower likelihood of clinical pregnancy and live birth after ART.	<ul style="list-style-type: none"> • Word limitation: yes • Grammatical correctness: yes • Factual correctness: evaluating additional outcomes beyond meta-analysis with uncertain credibility
Does the addition of hyaluronic acid (HA) to embryo transfer medium improve pregnancy outcomes in both autologous and oocyte donation IVF cycles? (Heymann et al., 2022)	There is some evidence to suggest that the addition of hyaluronic acid (HA) to embryo transfer medium may improve pregnancy outcomes in both autologous and oocyte donation IVF cycles, but further research is needed to confirm these findings.	Some studies suggest that adding hyaluronic acid (HA) to embryo transfer medium may improve pregnancy outcomes in both autologous and oocyte donation IVF cycles, but more research is needed to confirm its effectiveness.	The best available evidence indicates that the addition of HA to embryo transfer medium is clinically beneficial in cycles with autologous oocytes.	<ul style="list-style-type: none"> • Word limitation: yes • Grammatical correctness: yes • Factual correctness: false conclusion of ChatGPT, stating that the effect of hyaluronic acid seems to be the same in both autologous and donor cycles
What is the risk of miscarriage among pregnant women who received any of the COVID-19 vaccines? (Rimmer et al., 2023)	According to the available data, there is currently no evidence to suggest an increased risk of miscarriage in pregnant women who received any of the COVID-19 vaccines.	Recent studies suggest that COVID-19 vaccines do not increase the risk of miscarriage in pregnant women. Vaccination is recommended by health experts to protect pregnant women and their unborn babies from the risks of COVID-19.	There is no evidence that COVID-19 vaccines are associated with an increased risk of miscarriage.	<ul style="list-style-type: none"> • Word limitation: yes • Grammatical correctness: yes • Factual correctness: yes
What are the chances of achieving a live birth after embryo, oocyte and ovarian tissue cryopreservation (OTC) in female cancer survivors? (Fraison et al., 2023)	The chances of achieving a live birth after embryo, oocyte, or ovarian tissue cryopreservation in female cancer survivors vary depending on factors such as age, cancer type, and treatment. However, success rates range from 30% to 60% for embryo and oocyte cryopreservation and are still being studied for OTC.	The chances of achieving a live birth after embryo, oocyte, or ovarian tissue cryopreservation vary depending on factors such as age, cancer type, and treatment, but generally range from 30% to 60% per attempt.	The live birth rates (LBRs) following embryo and oocyte cryopreservation are 41% and 32%, respectively, while for IVF and spontaneous LBR after tissue cryopreservation and transplantation, these rates are 21% and 33%, respectively.	<ul style="list-style-type: none"> • Word limitation: no (Text 1), yes (Text 2) • Grammatical correctness: yes • Factual correctness: more vaguely formulated answers with less precise indication of percentages

Table 2. ChatGPT's cited sources on which the answer to meta-analysis' study question was based and their relevance, overlap with cited literature in respective meta-analysis, and citation accuracy.

	Number of ChatGPT citations (n)	Citations evaluated as relevant (%)	Overlap with the cited literature in respective meta-analysis (%)	Accurately cited literature (%)
Arecco et al. (2022)	5	40	0	40
Busnelli et al. (2022)	3	33.3	0	0
Meernik et al. (2023)	4	25	25	25
Heymann et al. (2022)	1	0	0	0
Rimmer et al. (2023)	3	100	0	66.6
Fraison et al. (2023)	3	100	0	66.6
Sum/average	19	52.6%	5.3%	36.8%

Table 3. ChatGPT's cited sources, when specifying the prompt based on the study search protocol of the respective meta-analysis, and their relevance, overlap with cited literature in meta-analysis, and citation accuracy.

	Number of ChatGPT citations (n)	Citations evaluated as relevant (%)	Overlap with the cited literature in respective meta-analysis (%)	Accurately cited literature (%)
Arecco et al. (2022)	5	40	20	60
Busnelli et al. (2022)	8	12.5	0	0
Meernik et al. (2023)	4	50	0	0
Heymann et al. (2022)	3	0	0	0
Rimmer et al. (2023)	4	50	0	75
Fraison et al. (2023)	5	80	0	60
Sum/average	29	37.9%	3.4%	31%

may not cite all of the relevant available data when prompted with a specific topic, giving its pre-trained nature and lack of capability to perform internet browsing; this could potentially be resolved through updates. Particularly worrisome, however, is the high citation rate of non-relevant and especially non-existing sources. At the moment we do not consider ChatGPT capable of assisting with literature searches and citations.

Some authors have proposed that ChatGPT could assist with various other stages of academic literature reviews (Qureshi et al., 2023; Sallam, 2023; Wang et al., 2023). For instance, it could save time during systematic reviews, aid in forming research questions, help to define the search strategy, or help with data extraction and data quality assessment. Qureshi et al. found that ChatGPT could propose a study question, eligibility criteria, and screen titles for importance but it should only be seen as a starting point for the refinement. They also criticized the ChatGPT's proposed search strategy, including the use of suboptimal index terms, which may not be apparent without expertise in search construction (Qureshi et al., 2023). Wang et al. explored the potential of utilizing ChatGPT to create a Boolean query. They demonstrate that the model has the potential to assist in the systematic review process but highlighted several issues concerning reproducibility (Wang et al., 2023), which is an essential component of systematic reviews. Mahuli et al. (2023) demonstrated ChatGPT's ability to extract specific data from chosen articles, such as mean values and sample sizes, but advise cautious implementation and human supervision. We believe that the current version of ChatGPT is unsuitable to assist in systematic review writing due to its limitations related to data access, reproducibility challenges, referencing, lack of contextual understanding, and potential for bias, but it might help in data summarizing as discussed below. When using the described prompts, ChatGPT program itself stated two times that 'as an AI language model, it cannot perform a systematic review or search for specific types of studies'. However, there is no doubt that AI in general already plays a

significant role in the automation of systematic reviews and different software applications have been developed to search, screen, and extract data (Marshall and Wallace, 2019).

Feasibility of ChatGPT to summarize findings and assist scientific writing

ChatGPT is an AI software and employs natural language processing technology to produce coherent, high-quality text that resembles human language (Ali et al., 2023). Therefore, it is a promising tool to assist in the writing of a scientific paper. English is the dominant scientific language which presents a major publishing barrier for non-native speakers. Proofreading and formatting could increase the efficiency of academic writing and alleviate the difficulties faced by non-native speakers when attempting to use academic English. Several authors have reported about the effectiveness of ChatGPT in the editing process, assisting with language refinement, summarizing, rewriting, searching for synonyms, or even suggesting titles (NajafAli et al., 2023; Sallam, 2023; Salvagno et al., 2023). This also seems acceptable, as numerous publishers already offer English editing services. Furthermore, the central aspect of a scientific work remains the hypothesis, along with the design and conduct of the study, while the writing carries secondary significance. It is, however, expected that ChatGPT will become a paid service as the newest version is already monetized. This could lead to new inequalities in academic publishing between high- and low-income countries (Liebrenz et al., 2023; Salvagno et al., 2023). The data on how well ChatGPT can assist with scientific writing is scarce (Sallam, 2023).

Lastly, we evaluated the capability of ChatGPT to summarize scientific findings and help with scientific writing. The 'Study question', 'Summary answer', 'What is known already', 'Study design', 'Methods', 'Main results', and 'Limitations' parts of each abstract were copied to ChatGPT starting with a following

Table 4. Comparison of the abstract conclusion written by ChatGPT with the original abstract conclusion, as assessed by four blinded reviewers.

Article title	Blinded reviewers recognizing ChatGPT abstract (%)	Abstract conclusion formulated by ChatGPT	Original abstract conclusion
Safety of fertility preservation techniques before and after anticancer treatments in young women with breast cancer: a systematic review and meta-analysis	100%	In conclusion, performing controlled ovarian stimulation (COS) for fertility preservation before starting anticancer therapies or assisted reproduction technology (ART) after treatments in young breast cancer patients does not seem to have a detrimental prognostic effect on breast cancer recurrence, mortality, or event-free survival (EFS). This meta-analysis suggests that fertility preservation strategies can be considered safe options for young women with breast cancer. However, the limitations of the included studies and retrospective nature of the analysis should be considered.	Our results demonstrate that performing COS at diagnosis or ART following treatment completion does not seem to be associated with detrimental prognostic effect in young women with breast cancer, including among patients with hormone receptor-positive disease and those receiving neoadjuvant chemotherapy.
Obstetric and perinatal outcomes following programmed compared to natural frozen-thawed embryo transfer cycles: a systematic review and meta-analysis	100%	Programmed FET protocols are associated with a higher risk of hypertensive disorders of pregnancy, pre-eclampsia, post-partum hemorrhage, and cesarean section when compared with natural FET protocols. The reasons for increased risks are still unknown, and further research is required to establish an effective way to prevent obstetric and perinatal complications.	Endometrial preparation protocols with HRT were associated with worse obstetric and perinatal outcomes. However, because of the methodological weaknesses, recommendations for clinical practice cannot be made. Well conducted prospective studies are thus warranted to establish a safe endometrial preparation strategy for FET cycles aimed at limiting superimposed risks in women with an 'a priori' high-risk profile for obstetric and perinatal complications.
Outcomes after assisted reproductive technology in women with cancer: a systematic review and meta-analysis	50%	Women with a history of cancer undergoing ART have lower odds of clinical pregnancy and live birth compared to women without cancer. These findings have important implications for counseling and informing patients with a history of cancer who are considering ART. The study also highlights the need for more research on how specific cancer treatments and clinical factors may impact ART outcomes.	Women with cancer may be less likely to achieve pregnancy and live birth after embryo transfer. Further examination of reproductive outcomes and sources of heterogeneity among studies is warranted to improve evidence of the expected success of ART after a cancer diagnosis.
The effect of hyaluronic acid in embryo transfer media in donor oocyte cycles and autologous oocyte cycles: a systematic review and meta-analysis	50%	Adding functional concentrations of HA to embryo transfer media improves pregnancy outcomes in autologous oocyte cycles, with increased live birth, clinical pregnancy, and multiple pregnancy rates, but has little effect in donor oocyte cycles. The wider implication is that HA addition could benefit patients undergoing autologous oocyte cycles, but further studies are needed to determine its impact on donor cycles and potential adverse effects.	There is a moderate level of evidence to suggest that functional HA concentration in embryo transfer medium increases clinical pregnancy, live birth and multiple pregnancy rates in IVF cycles using autologous oocytes. This effect was not seen in donor oocyte cycles, indicating either intrinsic differences between donor and autologous oocytes or lack of statistical power. The combination of HA addition to transfer media in cycles using autologous oocytes and a single embryo transfer policy might yield the best combination, with higher clinical pregnancy and live birth rates without increasing the chance of multiple pregnancies.
The risk of miscarriage following COVID-19 vaccination: a systematic review and meta-analysis	100%	In conclusion, based on a systematic review and meta-analysis of 21 studies, there is no evidence to suggest that COVID-19 vaccines increase the risk of miscarriage among pregnant women. These findings have wider implications for improving vaccine uptake among pregnant women and reducing the impact of the COVID-19 pandemic on	COVID-19 vaccines are not associated with an increase in the risk of miscarriage or reduced rates of ongoing pregnancy or live birth among women of reproductive age. The current evidence remains limited and larger population studies are needed to further evaluate the effectiveness and safety

(continued)

Table 4. Continued.

Article title	Blinded reviewers recognizing ChatGPT abstract (%)	Abstract conclusion formulated by ChatGPT	Original abstract conclusion
		maternal and fetal health. However, caution should be taken due to the limitations of the analysis, including the variability and heterogeneity of the included studies.	of COVID-19 vaccination in pregnancy.
Live birth rate after female fertility preservation for cancer or hematopoietic stem cell transplantation: a systematic review and meta-analysis of the three main techniques; embryo, oocyte, and ovarian tissue cryopreservation	75%	In conclusion, this systematic review and meta-analysis found that the live birth rates following embryo and oocyte cryopreservation are 41% and 32%, respectively, while for IVF and spontaneous LBR after tissue cryopreservation and transplantation, these rates are 21% and 33%, respectively. These findings provide useful information for clinicians when counseling female cancer survivors about fertility preservation options. However, due to the heterogeneity of the literature, further research is needed to compare these techniques.	This study highlights the need for long-term follow-up registries to assess return rates, as well as spontaneous pregnancy rates and birth rates after FP.

prompt: ‘Based on the following abstract, write the conclusion in which wider implications of the findings are summarized’. We selected four reviewers to exceed the standard number required for a manuscript review. The reviewers are actively engaged researchers with expertise in reproductive medicine to avoid bias due to unfamiliarity with the subject. Reviewers (N.T., M.K., C.B., S.F.) screened both the abstract conclusion written by ChatGPT and the original abstract conclusion of each paper (presented to them in a random fashion) and tried to identify the one written by ChatGPT. The blinded reviewers provided a binary score of whether they believed the conclusion was written by a human or by ChatGPT (Table 4). The GPT-2 Output Detector was used to evaluate whether ChatGPT created text and human created text can be recognized as such. Overall, ChatGPT generated a text summarizing critical information in a time-saving manner, requiring under 30s for each conclusion. In general, the conclusions were well-written, grammatically correct, and sounded plausible. On average, the conclusions created by ChatGPT were longer and more vaguely formulated, as agreed by all reviewers. When summarizing the abstract from Heymann et al., ChatGPT stated that adding hyaluronic acid to embryo transfer media seem to improve pregnancy outcomes with increased live birth, clinical pregnancy, and multiple pregnancy rates. Technically, the data are correct, however, it implies that multiple pregnancy rate is a wanted ART outcome. This indicates the ChatGPT’s limited ability to replace human logical thinking and once again shows the importance of examining its output. All blinded reviewers indicated it was challenging to differentiate between the conclusion written by a human being and that written by ChatGPT. At least two of the four reviewers detected the difference in all cases, and all identified the conclusion written by ChatGPT in half of the cases. It must be noted that the reviewers have excellent English skills, but they are not English native speakers and might not be able to detect subtle differences in language used by ChatGPT and real authors. The GPT-2 Output Detector was used to determine whether text created by ChatGPT and human-generated text can be distinguished. In two out of six cases, the program was able to identify the text created by ChatGPT, giving it an originality score of <50%. In two cases, both the original text and the ChatGPT-generated text were

identified as ‘real’ with an originality score of more than 90%. In one case, the ChatGPT-generated text was identified as 99.22% original, while the original text was identified as only 2.09% original (Fig. 1).

In accordance with existing literature (Gao et al., 2023; Else, 2023), this shows that distinguishing between a text written by a human being and one composed by AI language models can be a challenging task for human reviewers, even when they are evaluating the topic from their area of expertise. However, certain indicators, such as a lack of originality, excessive text, nuance, and vaguer language use, could suggest that a paper is produced by AI rather than a human author (Gao et al., 2023). Interestingly, these writing traits are also evident in texts written in a language that is not the author’s native tongue. As a result, there is a risk that AI-based plagiarism detection tools might incorrectly classify non-native speakers’ papers as machine-generated text (Salvagno et al., 2023). According to our evaluation, AI-based plagiarism detectors may estimate quite well if the text is AI-formulated but they are not reliable. It is probable that soon, all manuscript submissions will be screened by AI output detectors as part of the editorial process, to detect AI-generated text, which ultimately can be considered a form of plagiarism. However, we believe that the use of AI language models should not be limited when used by authors solely for text managing and correctly disclosed. Disclosing its use is crucial not only to ensure transparency but also to gain insights into how its implementation impacts the research landscape. This, in turn, would enable us to set up more effective regulations.

We conclude that ChatGPT is strongly limited when interpreting data and drawing scientific conclusions, and that it is not feasible to assist with literature searches in the field of human reproduction but it may be helpful with summarizing study findings and text processing, offering a timesaving writing tool when performed under human supervision. This is probably true also for other areas of medicine and is in accordance with the latest recommendations by Elsevier (Elsevier Publishing Ethics), which suggest that AI-based language models should be utilized for improving the language and readability of written work only. To summarize, we consulted ChatGPT about the future of AI in scientific writing in the field of reproductive medicine (Fig. 2). The software seems to agree that AI language models have a great

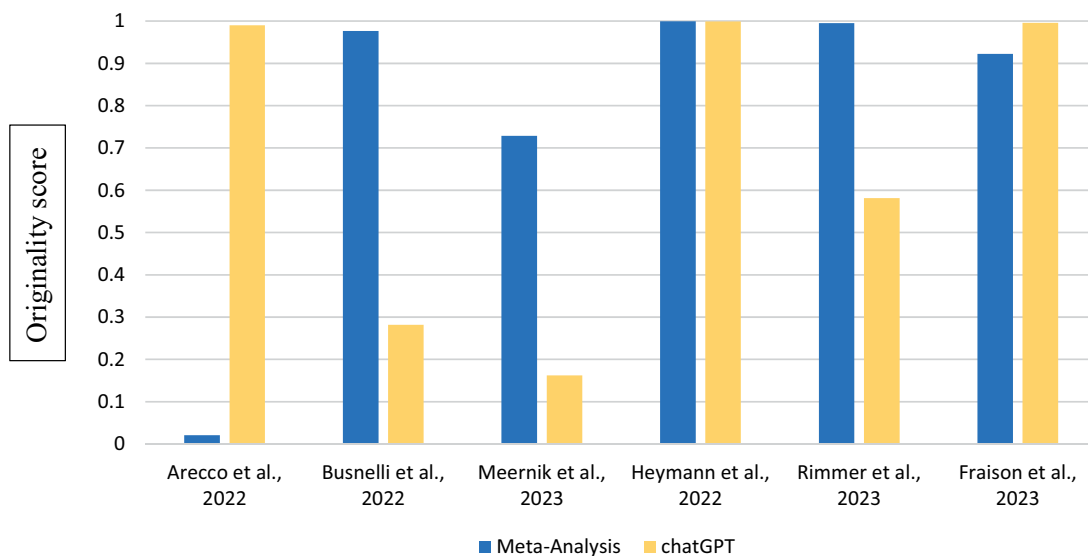


Figure 1. ChatGPT-2 output analysis of original text (blue) and ChatGPT-created text (yellow). An originality score between 0% and 100% was assigned to indicate the level of originality for each text.

NE

ChatGPT, what is the future of AI language model in assisting scientific writing in the field of reproductive medicine? Please answer in less than 250 words.



AI language models have already made significant contributions to the field of scientific writing, including in reproductive medicine. With continued advancements in natural language processing and machine learning algorithms, AI language models are likely to become even more effective in assisting scientists and researchers with writing scientific papers and publications.

In particular, AI language models can help with tasks such as literature review, identifying relevant citations, and suggesting appropriate language and phrasing for scientific writing. These models can also assist with data analysis and visualization, helping researchers to more effectively communicate their findings.

One potential challenge in applying AI language models to reproductive medicine is the need for specialized domain knowledge. However, as AI language models become more sophisticated, they may be able to learn and incorporate this knowledge more effectively.

Overall, the future of AI language models in reproductive medicine looks promising, with the potential to improve the efficiency and accuracy of scientific writing and communication in this field. However, it is important to continue to monitor and evaluate the use of these models to ensure that they are used appropriately and ethically.

Regenerate response

Figure 2. ChatGPT discussing the future of AI-assisted writing in human reproduction research.

potential to assist human reproduction research, but careful use and output monitoring are needed.

Conclusion

It seems inevitable that AI language models such as ChatGPT will become widely used in research. We believe that individuals

engaged in research should be familiar with the topic to be mindful of the potential downsides, notably the possibility of spreading (plausibly sounding) inaccuracies. However, language models have some potential to improve scientific writing, provided that appropriate academic regulations are in place. Instead of dismissing this development, we should explore ways to work alongside them to enhance our work without compromising their integrity.

Data availability

The data underlying this opinion paper are available within the article.

Authors' roles

N.S. wrote the main part of the paper. N.S. and T.B. prepared the blinded review. All authors evaluated the ChatGPT's output. M. K., N.T., C.B., and S.F. (blinded reviewers) screened abstract conclusions written by ChatGPT and original abstracts and tried to identify the one written by ChatGPT. All authors participated in critical reading of the entire document. All authors approved the final version for publication.

Funding

No specific funding was used.

Conflict of interest

None declared.

References

- Akhter HM, Cooper JS. Acute pulmonary edema after hyperbaric oxygen treatment: a case report written with ChatGPT assistance. *Cureus* 2023;**15**:e34752.
- Ali SR, Dobbs TD, Hutchings HA, Whitaker IS. Using ChatGPT to write patient clinic letters. *Lancet Digit Health* 2023;**5**:e179–e181.
- Arecco L, Blondeaux E, Bruzzzone M, Ceppi M, Latocca MM, Marrocco C, Boutros A, Spagnolo F, Razeti MG, Favero D et al. Safety of fertility preservation techniques before and after anticancer treatments in young women with breast cancer: a systematic review and meta-analysis. *Hum Reprod* 2022;**37**:954–968.
- Busnelli A, Schirripa I, Fedele F, Bulfoni A, Levi-Setti PE. Obstetric and perinatal outcomes following programmed compared to natural frozen-thawed embryo transfer cycles: a systematic review and meta-analysis. *Hum Reprod* 2022;**37**:1619–1641.
- Choi J, Hickman K, Monahan A, Schwarcz D. ChatGPT goes to law school. *J Legal Educ* 2023; <http://dx.doi.org/10.2139/ssrn.4335905>.
- Else H. Abstracts written by ChatGPT fool scientists. *Nature* 2023; **613**:423.
- Elsevier Publishing Ethics. Elsevier. <https://Www.Elsevier.Com/about/Policies/Publishing-Ethics> (1 February 2023, date last accessed).
- Fraison E, Huberlant S, Labrune E, Cavalieri M, Montagut M, Brugnion F, Courbiere B. Live birth rate after female fertility preservation for cancer or haematopoietic stem cell transplantation: a systematic review and meta-analysis of the three main techniques; embryo, oocyte and ovarian tissue cryopreservation. *Hum Reprod* 2023;**38**:489–502.
- Gao C, Howard F, Markov N, Dyer E, Rammesh S, Luo Y, Pearson A. Comparing scientific abstracts generated by ChatGPT to real abstracts with detectors and blinded human reviewers. *NPJ Digit Med* 2023;**6**:75.
- Gilson A, Safraneck CW, Huang T, Socrates V, Chi L, Taylor RA, Chartash D. How does ChatGPT perform on the United States medical licensing examination? The implications of large language models for medical education and knowledge assessment. *JMIR Med Educ* 2023;**9**:e45312.
- Goddard J. Hallucinations in ChatGPT: a cautionary tale for biomedical researchers. *Am J Med* 2023. <http://dx.doi.org/10.1016/j.amjmed.2023.06.012>.
- Heymann D, Vidal L, Shoham Z, Kostova E, Showell M, Or Y. The effect of hyaluronic acid in embryo transfer media in donor oocyte cycles and autologous oocyte cycles: a systematic review and meta-analysis. *Hum Reprod* 2022;**37**:1451–1469.
- Jeblick K, Schachtner B, Dextl J, Mittermeier A, Stüber AT, Topalis J, Weber T, Wesp P, Sabel B, Ricke J et al. ChatGPT makes medicine easy to swallow: an exploratory case study on simplified radiology reports. *arXiv* 2022. <https://doi.org/10.48550/arXiv.2212.14882>, preprint: not peer reviewed.
- Kitamura FC. ChatGPT is shaping the future of medical writing but still requires human judgment. *Radiology* 2023;**307**:e230171.
- Liebrezn M, Schleifer R, Buadze A, Bhugra D, Smith A. Generating scholarly content with ChatGPT: ethical challenges for medical publishing. *Lancet Digit Health* 2023;**5**:e105–e106.
- Mahuli SA, Rai A, Mahuli AV, Kumar A. Application ChatGPT in conducting systematic reviews and meta-analyses. *Br Dent J* 2023; **235**:90–92.
- Manohar N, Prasad SS. Use of ChatGPT in academic publishing: a rare case of seronegative systemic lupus erythematosus in a patient with HIV infection. *Cureus* 2023;**15**:e34616.
- Marshall IJ, Wallace BC. Toward systematic review automation: a practical guide to using machine learning tools in research synthesis. *Syst Rev* 2019;**8**:163.
- Meernik C, Poole C, Engel SM, Rauh-Hain JA, Luke B, Nichols HB. Outcomes after assisted reproductive technology in women with cancer: a systematic review and meta-analysis. *Hum Reprod* 2023; **38**:30–45.
- Najafali D, Hinson C, Camacho JM, Galbraith LG, Gupta R, Reid CM. Can chatbots assist with grant writing in plastic surgery? Utilizing ChatGPT to start an R01 grant. *Aesthet Surg J* 2023; **43**:NP663–NP665.
- Open AI. <https://openai.com> (21 July 2023, date last accessed).
- Open AI Terms of use. Open AI. <https://openai.com/policies/terms-of-use> (22 July 2023, date last accessed).
- Qureshi R, Shaughnessy D, Gill KAR, Robinson KA, Li T, Agai E. Are ChatGPT and large language models "the answer" to bringing us closer to systematic review automation? *Syst Rev* 2023;**12**:72.
- Rimmer MP, Teh JJ, Mackenzie SC, Al Wattar BH. The risk of miscarriage following COVID-19 vaccination: a systematic review and meta-analysis. *Hum Reprod* 2023;**38**:840–852.
- Sallam M. ChatGPT Utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. *Healthcare (Basel)* 2023;**11**:887.
- Salvagno M, Taccone FS, Gerli AG. Can artificial intelligence help for scientific writing? *Crit Care* 2023;**27**:75.
- Teixeira da Silva JA. Is ChatGPT a valid author? *Nurse Educ Pract* 2023; **68**:103600.
- Thorp HH. ChatGPT is fun, but not an author. *Science* 2023;**379**:313.
- Wen J, Wang W. The future of ChatGPT in academic research and publishing: a commentary for clinical and translational medicine. *Clin Transl Med* 2023;**13**:e1207.
- Wang S, Scells H, Koopman B, Zuccon G. Can ChatGPT write a good Boolean query for systematic review literature search? *arXiv* 2023. <https://doi.org/10.48550/arXiv.2302.03495>, preprint: not peer reviewed.