European Journal of Internal Medicine xxx (xxxx) xxx

Contents lists available at ScienceDirect

European Journal of Internal Medicine

journal homepage: www.elsevier.com/locate/ejim

Artificial intelligence in scientific medical writing: Legitimate and deceptive uses and ethical concerns

Davide Ramoni [a,1], Cosimo Sgura [a,1], Luca Liberale [a,b], Fabrizio Montecucco [a,b], John P.A. Ioannidis [c], Federico Carbone [a,b,*]

[a] Department of Internal Medicine, University of Genoa, 6 viale Benedetto XV, 16132 Genoa, Italy
[b] IRCCS Ospedale Policlinico San Martino, Genoa – Italian Cardiovascular Network, Largo Rosanna Benzi 10, 16132, Genoa, Italy
[c] Departments of Medicine of Epidemiology and Population Health of Biomedical Science and of Statistics and Meta-Research Innovation Center at Stanford (METRICS), Stanford University, Stanford CA 94305, USA

ARTICLE INFO

Keywords:
Artificial intelligence
Chatbots
ChatGPT
Medical writing
Natural language understanding
Large language models

ABSTRACT

The debate surrounding the integration of artificial intelligence (AI) into scientific writing has already attracted significant interest in medical and life sciences. While AI can undoubtedly expedite the process of manuscript creation and correction, it raises several criticisms. The crossover between AI and health sciences is relatively recent, but the use of AI tools among physicians and other scientists who work in the life sciences is growing very fast. Within this whirlwind, it is becoming essential to realize where we are heading and what the limits are, including an ethical perspective.

Modern conversational AIs exhibit a context awareness that enables them to understand and remember any conversation beyond any predefined script. Even more impressively, they can learn and adapt as they engage with a growing volume of human language input. They all share neural networks as background mathematical models and differ from old chatbots for their use of a specific network architecture called transformer model [1]. Some of them exceed 100 terabytes (TB) (e.g., Bloom, LaMDA) or even 500 TB (e.g., Megatron-Turing NLG) of text data, the 4.0 version of ChatGPT (GPT-4) was trained with nearly 45 TB, but stays updated by the internet connection and may integrate with different plugins that enhance its functionality, making it multimodal.

1. Large language models and medical domain knowledge

The ghost-writer potential of ChatGPT and related AI tools has been widely discussed in many editorials and essays. Yet, general AI models to-date have failed to fully utilize medical language and this still places some limits to applications in medical science and healthcare [2]. Open-domain question answering firstly attracted the interest of the natural language processing (NLP) community and gave rise to a subset of sources that includes large-scale, multi-subject, and multi-choice datasets for medical domain question answering. They combine question answering datasets spanning professional medical exams, medical research, and consumer queries (Table 1).

The evolution towards the use of large language models (LLM) in NLP applications has broadened the scope of data used for training and inference. Once an LLM is trained, the AI can be deployed across various domains for practical purposes: text generation, translation, content summary, rewriting content, classification and categorization, sentiment analysis, and conversational AI/chatbots.

Relatedly, the instruction-tuned variant of Med-PaLM achieved an early significant milestone by successfully obtaining a "passing score" on the United States Medical License Exam (USMLE) [2]. It demonstrated an accuracy of 67.6 % on MedQA (USMLE) and consistently outperformed previous benchmarks on MedMCQA, PubMedQA, and Massive Multitask Language Understanding (MMLU) clinical topics. The release of Med-PaLM 2 by Google® elevated the score to 86.5 % in MedQA dataset, marking a remarkable 19 % improvement over the preceding LLM (Table 2).

Nevertheless, such a strong performance still leaves notable gaps. The adherence of Med-PALM2 answers to scientific consensus was later judged quite low by a panel of clinicians (61.9 %) making it inappropriate for use in the safety-critical medical domain [2]. Moreover, there is some concern that many materials related to the questions relevant to

* Corresponding author at: First Clinic of Internal Medicine, Department of Internal Medicine, University of Genoa, 6 viale Benedetto XV, 16132 Genoa, Italy.
E-mail address: federico.carbone@unige.it (F. Carbone).
[1] The authors equally contribute to as first authors of this manuscript.

https://doi.org/10.1016/j.ejim.2024.07.012
Received 6 May 2024; Received in revised form 14 June 2024; Accepted 10 July 2024

Table 1
Evolution over time of some common question answering datasets in the medical field.

Model	Year	Language	Format	Q/A dataset size
LiveQA [1]	2017	English	Dataset of consumers questions about medical knowledge	Training/ test 634/ 104
MedicationQA [2]	2017	English	Dataset of consumers questions about medication	674
PubMedQA [3] https://pubmedqa.github.io/	2019	English	Question/answer pairs with yes/no/ maybe using the corresponding abstracts	1 k expert-annotated 61.2k unlabelled 211.3k artificially generated
MedQA [4]	2021	English Chinese (simplified/ traditional)	Medical board exams in US, mainland China, and Taiwan	12,723 34,251, and 14,123
MedMCQA [5]	2022	English	Medical board exams in India	194K
MMLU [6]	2020	English	Benchmark that covers a broad spectrum of questions including 57 domains	1,212
HealthSearchQA [7]	2022	English	Dataset of consumers questions about medical conditions and their associated symptoms	3,173
MultiMedQA [7]	2023	English	Pool of Medication QA, LiveQA, PubMedQA, MMLU, MedMCQA, MedQA	–

The table lists the format and size of single-question answering datasets used for testing large language models (LLMs) on the US Medical Licensing Examination. To address their intrinsic limitations, MultiMedQA finally pooled them into a single question-answering benchmark, used for testing the performance of LLMs in a recent comparative study [7]. In this study, Flan-PaLM exceeded the state-of-the-art performance of other LLMs (i.e., PubMedGPT, DRAGON, BioLinkBERT, PubMedBERT, GPT_Neo) on MultiMedQA. Information on the language, format, and size of each dataset has been retrieved from reference articles included in the aforementioned study.

[1] Abacha AB AE, Pinter Y, Demmer-Fushman D. Overview of the Medical Question Answering Task at TREC 2017 LiveQA. Text REtrieval Conference (TREC) 2017. 2017.
[2] Abacha AB, Mrabet Y, Sharp M, Goodwin TR, Shooshan SE, Demner-Fushman D. Bridging the Gap Between Consumers' Medication Questions and Trusted Answers. Studies in health technology and informatics. 2019;264:25-9.
[3] Jin Q DB, Liu Z, Cohen WW, Lu X. ubMedQA: A Dataset for Biomedical Research Question Answering. arXiv:190906146. 2019.
[4] Jin D PE, Oufattole N, Weng WH, Fang H, Szolovitz P. What Disease Does This Patient Have? A Large-Scale Open Domain Question Answering Dataset from Medical Exams. Appl Sci. 2021;11.
[5] Pal A UK, Sankarasubbu M. MedMCQA: A Large-scale Multi-Subject Multi-Choice Dataset for Medical domain Question Answering. arXiv:220314371. 2022.
[6] Hendrycks D BC, Basart S, Zou A, Mazeika M, Song D, Steinhardt J. Measuring Massive Multitask Language Understanding. arXiv:200903300. 2021.
[7] Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. Nature. 2023;620:172-80.

Table 2
Performance of large language models in US Medical License Examination (MedQA).

Model	Year	Format and results	Dataset size
PubmedBERT [1]	2020	Pre-training LLM on unlabelled text Accuracy: 38.1%	100M
DRAGON [2]	2022	pre-trained LLM from PubMed abstracts Accuracy: 47.5%	360M
BioLinkBERT [3]	2022	Pre-training LLM on Stanford CRFM Accuracy: 45.1%	BERT tiny (4.4M), BERT base (110M), BERT large (340M)
Galactica [4]	2022	48M scientific articles, textbooks and websites. Withdrawn due to massive release of fake, fraudulent and/or plagiarised scientific papers. Accuracy: 44.4%	120B
BioMedLM (formerly PubMed GPT) https://hai.stanford.edu/news/stanford-crfm-introduces-pubmedgpt-27b	2022	LLM trained on biomedical literature Accuracy: 50.3%	2.7B
GPT-Neo https://www.eleutther.ai/artifacts/gpt-neo	2020	LLM similar to BioMedLM, but not domain-specific Accuracy: 33.3%	2.7B
Med-PaLM [5]	2022	LLM with strong performance in answering medical questions, combined with effective instruction prompt tuning Accuracy: 67.6%	540B
BioGPT [6]	2022	A domain-specific generative transformer language model pre-trained on large scale biomedical No exact data on accuracy	355 B
Med-PaLM 2 [7] https://sites.research.google/med-palm/	2023	LLM that has been trained on a massive dataset of medical text and code, was shown to perform at an "expert" level on USMLE, an over 19% improvement from Med-PaLM's previous performance. Still under development Accuracy: 86.5%	Not yet declared
MedAlpaca [8]	2023	Pre-training LLM on a high-quality collection of medical text data Accuracy: 47.3%	Various: 7B, 13B, 33B, 65B

CRFM, Center for Research on Foundation Models; LLM, Large language models; USMLE, US Medical License Examination. The searches of studies were conducted until February 2024 in the PubMed database and the arXiv.org website. The Mesh terms identified for the PubMed search were: "large language model" and "medical question answering". The same terms were also applied on the arXiv.org website. When appropriate, the free text terms have been truncated in order to include alternative word endings. The search result was limited to articles that were written in English as well as articles published from 2020. The database searches were complemented with manual review of the reference lists of relevant articles.

[1] Tinn R, Cheng H, Gu Y, Usuyama N, Liu X, Naumann T, et al. Fine-tuning large neural language models for biomedical natural language processing. Patterns. 2023;4:100729.
[2] Yasunaga M BA, Ren H, Zhang H, Manning CD, Liang P,Leskovec J. Deep Bidirectional Language-Knowledge Graph Pretraining. arXiv:221009338. 2022.
[3] Yasunaga M LJ, Liang P. LinkBERT: Pretraining Language Models with Document Links. arXiv:220315827. 2022.
[4] Taylor R KM, Cucurull G, Scialom T, Hartshorn A, Saravia E, Poulton E, Kerkez V, Stojnic R. Galactica: A Large Language Model for Science.

arXiv:221109085. 2022.
[5] Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. Nature. 2023;620:172-80.
[6] Luo R SL, Xia Y, Qin T, Zhang S, Poon H, Liu T-Y. BioGPT: Generative Pre-trained Transformer for Biomedical Text Generation and Mining. arXiv:221010341. 2022.
[7] Singhal K TT, Gottweis J, Sayres R, Wulczyn E, Hou L, Clark K, Pfohl S, Cole-Lewis H, Neal D, Schaekermann M, Wang A, Amin M, Lachgar S, Mansfield P, Prakash S, Green B, Dominowska E, Aguera y Arcas B, Tomasev N, Liu Y, Wong R, Semturs C, Mahdavi S, Barral J, Webster D, Corrado GS, Matias Y, Azizi S, Karthikesalingam A, Natarajan V. Towards Expert-Level Medical Question Answering with Large Language Models. arXiv:230509617 2023.
[8] Han T AL, Papaioannou J-M, Grundmann P, Oberhauser T, Löser A, Truhn D, Bressem KK. MedAlpaca – An Open-Source Collection of Medical Conversational AI Models and Training Data. arXiv:230408247. 2023.

these types of exams have been spilled on the internet, therefore LLMs may be aware of them, thus explaining the high performance. Regardless, further genuine improvements are very likely, but their real-world validity and clinical relevance will need to be carefully vetted.

2. Large language models' applications to scientific writing

2.1. Publish or perish and negative consequences fuelled by AI tools

Under pressure, many scientists may gamble on the "publish or perish" culture, inadvertently contributing to the growth of research waste and low-quality papers with spurious or even misleading contributions. At the extreme of the spectrum, they may even become costumers of paper mills that sell authorship in fake papers [3]. The phenotype of scientists who may resort to AI to build spurious CVs with poor-quality or even fraudulent papers needs to be carefully studied. In different countries, settings, and microenvironments, these phenotypes may be different. Corrupt environments in medicine and science are probably common worldwide. Alternatively, even scientists at prestigious institutions may be involved for a variety of reasons in this conundrum.

2.2. Legitimate applications

The ever-expanding volume of scientific literature makes it challenging to keep up with the latest research trends. This is where the AI input may legitimately help. PubMed itself employs different algorithms to enhance its information retrieval process [4]. Additionally, external AI tools could streamline the process of summarizing research papers from PubMed with application programming interfaces.

Besides ChatGPT, a plethora of tools have now been developed that may assist medical research and writing in daily practice. The research assistants IRIS.AI and Scite.ai can support the classification of research workplaces through the categorization of scientific articles in a visual map or by tracking the total number of citations for a given article and distinguishing between those in favour and those against. The accuracy and exact benefits of such tools need careful validation. Given that AI tools theoretically excel in absorbing existing text rather than creating new ideas, it may be expected that a key use may be in improving the efficiency of reviewing the literature. However, it is rather unclear if AI tools can yet produce high quality reviews with rigorous evidence-based medicine standards.

Offering support in statistical analysis is another highly touted application for AI chatbots, given the widely recognized lack of statistical resources and statistical literacy across medicine and biomedical research and the limited number of available methodologists and expert biostatisticians. Indeed, LLMs may generate streamlined code and algorithms even for the most complex statistical analyses. E.g., ChatGPT excels in producing practical R language code, being supported by a growing package collection that may enhance accessibility and workflow, especially for newcomers. (e.g., air, OpenAIR, RTutor, CodeLingo, askgpt, gptstudio, gpttools, gptchatteR packages) [5]. However, caution is warranted, as code generation errors can occur and still necessitate basic expertise. In the long run, it is unknown whether exposing larger numbers of medical staff and medical scientists to advanced statistical and algorithmic options will make things better or worse. A pessimistic scenario is that this will lead to more widespread errors, as fancy models will be used and mostly misused by people who are not properly trained in them and do not understand their premises, assumptions, and limitations of use.

Finally, text translation is another legitimate purpose of LLM use. Essentially, leveraging AI tools to develop the writer rather than focusing solely on the writing may offer non-native English speakers transformative learning opportunities, independent of their career stages and linguistic backgrounds. A real-time feedback on clarity, coherence, and organization may allow writers to understand not only what but 'why' of changes when applied to the early stage drafting [6]. AI can also align the manuscript with the journal's guidelines and support the peer-review process in a time-saving manner.

2.3. Deceptive applications

The breakthroughs in AI text generation have the potential to undermine medical research due to their (mis)use, which can be harmful through malevolent intent or culpable negligence. These threats posed by LLMs are subtle and then far from trivial. Paper mills are an increasingly recognized threat for the medical literature and synthetic fake scientific papers may be alarmingly common with an estimated figure of 24 % in the field of medicine [7].

Concurrently, AI may also help in identifying such fake, paper mill products. GPT-2 Output Detector retains an excellent discriminating value for fake detection (AUC 0.94), better than current plagiarism-detection website (i.e., Plagiarism Detector, iThenticate) and blinded human reviewers [8]. Aware of this problem, the International Association of Scientific, Technical, and Medical Publishers (STM) has introduced the Integrity Hub, a robust cloud-based environment to scrutinize submitted articles by leveraging shared data and experiences. To intensify the fight against paper mills, STM is also developing its own AI detection software consolidating resources from similar tools available elsewhere (https://www.stm-assoc.org/stm-integrity-hub/). While GPT-2 output detector (https://openai-openai-detector.hf.space/) and Copyleaks already works to generally address authenticity (https://copyleaks.com/), ownership, and AI content detection, some publishers such as Taylor & Francis are employing also their own tool. Regarding the features to consider, both author-, manuscript and journal-related features should be taken into account (Table 3).

Similarly, the use of conversational AI for research grant writing is another slippery field. Beyond scientific paper drafting, AI can significantly shorten the time needed for "dead documents" that heavily burden grant applications [9]. It is then not surprisingly that AI use in writing would cover 25 % of manuscript writing and 15 % of research grant writing, as indicated by a 2023 survey in Nature [9].

3. Ethical concerns

The exponential growth in the use of chatbots and the emergence of a dual-use dilemma raise concerns about the evolving quality of medical research in both the short and long term. Medical research was already infested with a large share of waste even before the advent of advanced AI tools [10].

The first issue is related to basic technical issues: these tools still often fail to provide correct replies to questions and can generate fraudulent – albeit superficially seemingly high-quality – medical articles. Unlike the human brain, AI is currently unable to articulate complex inductive reasoning, which is essential for coherent thinking. The focus on explainability in AI is then a key aspect for improving modern machine learning algorithms. Inductive Logic Programming (ILP) has

Table 3
Summary of warnings/ risk factors from items related to paper mills and related unethical phenomena.

WARNING OR RISK FACTOR	Comment
At risk countries (A)	China, Iran, and Russia are most prominent producers that have been detected, but location may diversify over time; authors who buy authorship in paper mills may come from almost any country
ORCID ID lacking/ incomplete (A)	Not very specific criterion since many legitimate papers still lack ORCID
Suspicious e-mail address of authors (A, M)	E.g. not institutional, or weird institutions
Suspicious reviewers' e-mail address (A, M)	Same as above
Not consistent references (M)	Use of references/citations is very poor in the vast majority of the scientific literature, but with paper mills sometimes misuse may reach egregious levels; AI tools may not necessarily improve this, since they are trained on flawed previous papers with flawed use of references.
Plagiarism (M)	Classic plagiarism is detectable with standard software; more difficult with paraphrasing
Weird phrasing and terminology (M)	Forced rephrasing and misuse of scientific terms
Image quality (M)	Routine checking of images for signs of problems may reveal many problems that may be related to papermills or may be independent to them
Severe methodological weaknesses(M)	Not specific to paper mills, most published papers are weak methodologically
Manuscript formatting and style (M, J)	Unusual formatting and style in submission or requested by the journal
Non-provision of raw data (M, J)	Not specific, since most scientific papers unfortunately do not provide raw data, but some paper mill or other fraudulent studies may overtly seem implausible (e.g. data volume exceeds capacity of team/center)
Authorship change after acceptance (M, J)	Not specific to paper mills, but probably far more common, since authors are added even at such a late stage, some paper mills advertise that they sell authorship after acceptance
Detection of AI-generated material (M, J)	See text, methods of detection improve over time, but they are far from perfect, risk of false negatives and false positives
Too fast review time (J)	May herald a problem with the legitimacy of the journal and/or its ability to reject paper mill papers and other fraudulent or very low-quality work; extreme cases abound (e.g. highjacked sites of previously legitimate journals)
Similarities in reviewer comments (J)	Not specific, given the generally poor quality of peer review in most journals, but extreme similarity should raise concerns

A: author-related, M: manuscript-related, J: journal-related.
Table adapted from Dadkhah M, Oermann MH, Hegedus M, Raman R, David LD. Detection of fake papers in the era of artificial intelligence. Diagnosis. 2023;10:390–7.

the potential to effectively leverage abductive reasoning and generate first-order clausal theories. ILP and its variants are then expected to overcome some of the current drawbacks and foster user trust [11]. It is unclear though whether they will alleviate or exacerbate the ethical dilemmas.

Addressing the inclusion of AI in authorship is another current and relevant question. When ChatGPT was listed as a co-author for the first time, this tongue-in-cheek choice sparked a lively discussion in the scientific community. The journal ultimately published a corrigendum excluding AI from the author list [12]. In line, practically all leading scientific journals now claim AI doesn't fulfil the necessary criteria to be recognized as authors [13,14], whereas the International Committee of Medical Journal Editors recently stressed 'accountability' among the authorship criteria, thus formally excluding AI from any authorship (https://www.icmje.org/recommendations/).

What might happen in the longer term is not very predictable. The massive and largely unregulated use of AI in medical writing may even lead to a progressive homogenization of text resulting in a loss of originality. This could be particularly dangerous for the new generations of scientists. The volume of published papers may grow, validity may decrease, and ability to detect problems with validity may dwindle. Scientists may be trapped in a mesh of voluminous trivia and waste.

On the other hand, embracing the benefits of AI may relieve scientists, physicians, and physician-scientists from several burdensome and repetitive tasks. This might ultimately improve their productivity in terms of both scientific advance and patient care, contributing to the overall well-being of healthcare systems and our communities. Moreover, AI tools may have beneficial uses across a large array of research practices (besides writing) and facilitate progress in medicine and other scientific fields (for a review see ref [15]).

4. Conclusion

The advance of AI in human life is pervasive and likely unstoppable. We are already living in a world where the border between human and computer-generated content is difficult to draw, as never before. The question is not whether to accept it or not, but how to manage it. Not for the first time, science is facing a dual-use dilemma. To deal with it fairly, we will need to develop a defensible account of why and how much openness matters, as well as establish the trade-offs between openness and the associated harms.

Relatedly, a range of regulatory measures may be needed. The question arises of whether to prefer self-regulation by the AI community, external government, other stakeholders (e.g., journals, publishers, scientific communities, universities, and more) or a combination of multiple players. There is still room to develop an ethical framework before this window closes. Any work done in this space should be deeply and critically discussed to find the right way to manage the potential of AI, taking into account also its emerging capabilities and trying to anticipate also the impact of their further growth.

Conflict of interest

Luca Liberale - Payment or honoraria for lectures, presentations, speakers bureaus, manuscript writing or educational events: Daiichi-Sankyo. Patents planned, issued or pending: co-inventor on the international patent WO/2020/226,993 filed in April 2020. There are no other conflicts of interest.

Acknowledgments

Work supported by #NEXTGENERATIONEU (NGEU) and funded by the Ministry of University and Research (MUR), National Recovery and Resilience Plan (NRRP), project MNESYS (PE0000006) – A Multiscale integrated approach to the study of the nervous system in health and disease (DN. 1553 11.10.2022); recipient Federico Carbone and Fabrizio Montecucco. This research was also funded by a grant from the Rete Cardiologica of Italian Ministry of Health RCR- 2022- 23682288) to Prof. F. Montecucco.

References

[1] Briganti G. How ChatGPT works: a mini review. European archives of oto-rhino-laryngology: official journal of the European Federation of Oto-Rhino-Laryngological Societies. 2023.
[2] Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. Nature 2023;620:172–80.
[3] Sanderson K. Science's fake-paper problem: high-profile effort will tackle paper mills. Nature 2024;626:17–8.
[4] Kiester L, Turp C. Artificial intelligence behind the scenes: PubMed's Best Match algorithm. Journal of the Medical Library Association: JMLA 2022;110:15–22.
[5] Macdonald C, Adeloye D, Sheikh A, Rudan I. Can ChatGPT draft a research article? An example of population-level vaccine effectiveness analysis. Journal of global health 2023;13:01003.
[6] Ingley SJ, Pack A. Leveraging AI tools to develop the writer rather than the writing. Trends in ecology & evolution 2023;38:785–7.

[7] Brainard J. New tools show promise for tackling paper mills. Science 2023;380:568–9.
[8] Gao CA, Howard FM, Markov NS, Dyer EC, Ramesh S, Luo Y, et al. Comparing scientific abstracts generated by ChatGPT to real abstracts with detectors and blinded human reviewers. NPJ digital medicine 2023;6:75.
[9] Parrilla JM. ChatGPT use shows that the grant-application system is broken. Nature 2023;623:443.
[10] Ioannidis JP. Why most published research findings are false. PLoS Med 2005;2:e124.
[11] Zhang Z, Yilmaz L, Liu B. A Critical Review of Inductive Logic Programming Techniques for Explainable AI. IEEE transactions on neural networks and learning systems 2023.
[12] O'Connor S. Corrigendum to "Open artificial intelligence platforms in nursing education: Tools for academic progress or abuse?". Nurse Educ. Pract. 2023;66:103537. Nurse education in practice. 2023;67:103572.
[13] Stokel-Walker C. ChatGPT listed as author on research papers: many scientists disapprove. Nature 2023;613:620–1.
[14] Tools such as ChatGPT threaten transparent science; here are our ground rules for their use. Nature 2023;613:612.
[15] Telenti A, Auli M, Hie BL, Maher C, Saria S, Ioannidis JPA. Large language models for science and medicine. European journal of clinical investigation 2024:e14183.