Large language models in healthcare information research: making progress in an emerging field

Harish Tayyar Madabushi,[1] Matthew D. Jones [2]

[1]Department of Computer Science, University of Bath, Bath, UK
[2]Department of Life Sciences, University of Bath, Bath, UK

Correspondence to
Dr Matthew D. Jones;
M.D.Jones@bath.ac.uk

Received 25 September 2024
Accepted 15 October 2024
Published Online First
23 October 2024

► http://dx.doi.org/10.1136/bmjqs-2024-017476

To cite: Tayyar Madabushi H, Jones MD. *BMJ Qual Saf* 2025;**34**:73–76.

The last 5 years have seen a rapid growth in research applying artificial intelligence or machine learning to improve the quality and safety of healthcare. This coincides with the release of web interfaces (such as ChatGPT from OpenAI and Copilot from Microsoft) that have enabled the general public (including health professionals and researchers) to easily access the latest generation of large language models (LLMs).

LLMs have fundamentally changed how machine learning is used across domains. Unlike previous generation systems that required careful data curation for specific tasks before training, modern LLMs work well with just a few examples or a simple problem description. This progress is mainly due to training on large volumes of web data that allows them to develop an 'understanding' of both language and general knowledge which they can then apply to a wide range of tasks.[1]

To fully comprehend the capabilities and associated dangers of LLMs, it is necessary to briefly examine how they function which was summarised in a recent review published by this journal.[2] Fundamentally, they are 'auto-completion' models trained to complete sentences, which can occasionally lead to the generation of inaccurate, if linguistically fluent, information—a phenomenon known as 'hallucinations' (figure 1). In addition, the generalisations on which they rely inherently limit their effectiveness when addressing marginalised groups or less common healthcare topics. It is important to recognise that LLMs were not originally designed for use in healthcare settings, where requirements might very well be different. At a minimum, it is essential to use LLMs specifically designed for medical applications (such as Med-PaLM 2) and rigorous testing to ensure safety and effectiveness.

Several recent systematic reviews (focused on ChatGPT-based studies) give an oversight of emerging trends when applying LLMs in healthcare. They have been applied in most clinical specialties[3] and to address a wide range of applications.[4 5] Consequently, the potential users of such applications have also varied, from health professionals and students to patients and carers.[4] While initial work focused on professional users,[5] a search for recent studies suggests that increasing amounts of research is focusing on patient information.

When investigating the use of LLMs to respond to medical queries, accuracy has been the most commonly used metric to assess the quality of LLM-generated responses with metrics such as completeness, consistency, safety, appropriateness and readability considered much less often.[3 5] One meta-analysis found that ChatGPT was able to correctly answer 56% of multiple-choice questions (95% CI 51% to 60%) but this varied between clinical specialties.[3] This may be related to the varying public availability of high-quality information on different topics.[4]

These findings give the impression of an emerging research field with many small-scale studies mapping potential applications for LLMs and developing methodologies. However, to move the field forward, more rigorous research methods and greater transparency of reporting are now required.[3 4]

In this context, the study in this issue by Andrikyan *et al* makes a welcome contribution to the field.[6] First, it focuses on patients as potential users of LLM-powered search engines (specifically Microsoft Copilot in Bing) for drug information. This user group has been relatively understudied so far,[5] yet as patients greatly outnumber professionals and have

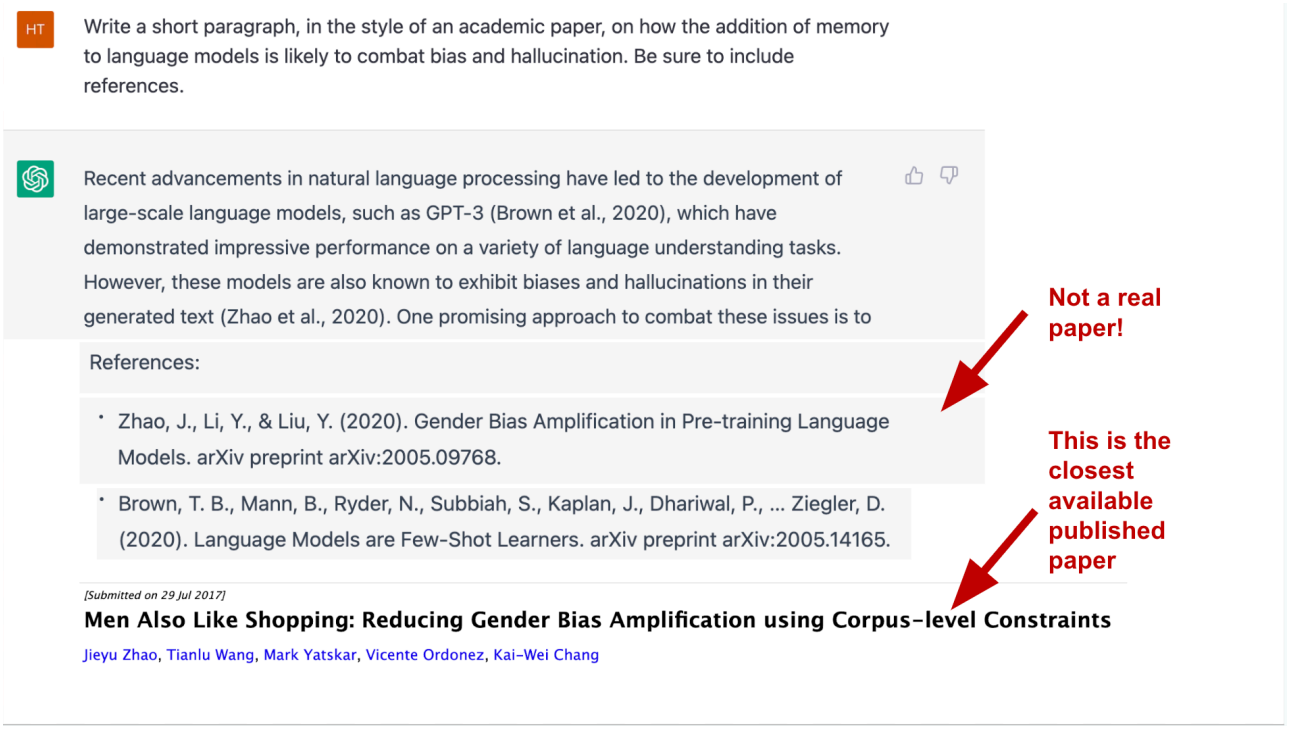

Figure 1 An example of 'hallucinations' where the statistical information available to large language models results in the generation of plausible but factually incorrect outputs. This is especially problematic in safety critical domains such as healthcare.

less training in the interpretation of health information, their use of LLMs may have greater potential for positive or negative effects. Second, the study addresses some of the methodological limitations of previous work. For example, it is transparent and systematic in its selection of drugs and patient questions with which to assess the responses of Copilot. It also uses a range of outcomes including the Flesch reading-ease score, completeness and accuracy in comparison with reputable information, and the likelihood and extent of possible harm.

The headline findings are alarming: The mean reading ease score was only appropriate for patients educated to undergraduate level and for some types of questions, median completeness and accuracy were as low as 20% and 50%, respectively. Similarly, 32% of expert ratings were for a medium to high likelihood of harm resulting from a patient following the advice with 22% of ratings suggesting this could result in death or severe harm. However, these expert ratings should be interpreted cautiously because they are based on seven experts' assessment of only 20 out of 500 answers selected for their low accuracy, low completeness or risk to patient safety. They are therefore not representative of the data set as a whole but could be considered to represent potential 'worst case scenarios'. In addition, the inter-rater reliability between the seven experts who generated these data was low (0.19–0.20) and the system for rating the likelihood and extent of harm did not consider the relationship between these two variables. For example, expert ratings could not reflect the potential for one answer to have a greater likelihood of causing low harm and a lesser chance of causing death or severe harm.

Considering these findings and this field of research, we believe that the following developments should be considered in future. First, as healthcare LLM research moves from the initial technical exploratory phase to more focused development and implementation, it will be important to use appropriate best practice guidance and theoretical frameworks to ensure that LLM-based systems are addressing the most important problems in the most useful, implementable and sustainable manner. For example, models drawing on sociotechnical theories (such as the Systems Engineering Initiative for Patient Safety model[7]) and frameworks for the development and evaluation of complex interventions should be used.[8]

Second, rather than working with LLMs designed for the general public and trained with large amounts of text retrieved from the web, healthcare researchers should consider collaborating with colleagues with expertise in computer science to develop bespoke systems to retrieve relevant information from reliable information sources. This may be especially critical for ensuring that these models capture healthcare-specific information which may not be available in significant quantities on the web for all relevant topics. Such an approach has the potential to prevent the generation of misleading information and has already proved successful in small studies.[9] It may also help to address the problems faced by

healthcare professionals in finding the most appropriate section of the most appropriate guideline for their patient.[10]

Third, more rigorous approaches to the assessment of understandability are required. While readability formulae are easy to apply, there are numerous limitations in their applicability to health information.[11] The findings of future studies would therefore have greater validity if they also tested the understandability of LLM-generated information with potential target users. Techniques developed for the user-testing of health information are an appropriate starting point.[12] The fact that LLM outputs appear extremely plausible makes such rigorous testing all the more critical. Ultimately, the effect of LLM-based systems on health outcomes should be assessed.

In a similar way to Andrikyan *et al*, future studies should also consider the potential impact on patients' health of LLM-generated information. This should move beyond considering only the risk of harm of such information so that potentially beneficial outcomes are also estimated. For example, it is widely recognised that current healthcare practice leaves many patients poorly informed about their care which impairs their ability to participate in shared-decision making and may increase their risk of harm and decrease their risk of benefit.[13 14] An LLM-based system that effectively improved overall patient knowledge might therefore increase overall health outcomes even if in some cases incorrect information led to harm. The adoption of methods from the field of health economics would be a useful approach to quantifying this balance between risk and benefit alongside the necessary ethical dialogue between patients, healthcare professionals and wider society.

The importance of such engagement was recognised as the first priority in the Health Foundation's recent Priorities for an AI in Healthcare Strategy.[15] Interestingly, this was supported by a survey of both the public and healthcare staff. While both groups were, on balance, supportive of the use of artificial intelligence in healthcare, there was greater support among staff.[16] In a new and rapidly developing research field which may currently be dominated by enthusiastic early adopters, these survey findings emphasise the importance of high-quality public and patient involvement in research and ensuring that patients are supportive of such developments before implementation.

Of course, it should not be forgotten that the LLMs used by many researchers (eg, ChatGPT) are also available to patients and practising professionals and so it is likely that they are already being used in healthcare. However, there have been surprisingly few studies into the extent and nature of this phenomenon. Future research into this area would therefore be particularly useful to inform current practice, so is needed alongside further studies into potential future applications for LLMs in healthcare.

Finally, to illustrate the current capabilities of LLMs, the following concluding paragraph was initially generated using Microsoft Copilot and then lightly edited. We uploaded a draft version of the article and used the prompt 'Write a conclusion to this draft editorial for the journal BMJ Quality and Safety'.

The integration of LLMs into routine healthcare is a rapidly evolving field with significant potential to enhance various aspects of quality, safety and efficiency. However, current research has highlighted several challenges including the accuracy, completeness and safety of LLM-generated information. The adoption of rigorous research methodologies and collaboration with patients, the public, healthcare professionals and interdisciplinary researchers will help to ensure that the field progresses as rapidly as possible in a relevant direction. By addressing these challenges and leveraging theoretical frameworks and best practice guidance, healthcare systems can harness the benefits of LLMs while mitigating potential risks, ultimately improving patient outcomes and safety.

X Harish Tayyar Madabushi @harish and Matthew D. Jones @MatthewJonesUoB

Contributors MDJ wrote the first draft to which HTM added content related to computer science. Both authors revised subsequent drafts. MDJ is the guarantor. To illustrate the current capabilities of large language models, the final paragraph of this article was initially generated using Microsoft Copilot and then lightly edited. We uploaded a draft version of the article and used the prompt 'Write a conclusion to this draft editorial for the journal BMJ Quality and Safety'.

Funding The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests None declared.

Patient consent for publication Not applicable.

Ethics approval Not applicable.

Provenance and peer review Commissioned; internally peer reviewed.

ORCID iD

Matthew D. Jones http://orcid.org/0000-0002-2617-4098

REFERENCES

1 Madabushi HT, Romain L, Milin P, *et al*. Construction grammar and language models. In: Fried M, Nikiforidou K, eds. *The Cambridge Handbook of Construction Grammar*. Cambridge: Cambridge University Press, 2025.
2 Howell MD. Generative artificial intelligence, patient safety and healthcare quality: a review. *BMJ Qual Saf* 2024;33:748–54.
3 Wei Q, Yao Z, Cui Y, *et al*. Evaluation of ChatGPT-generated medical responses: a systematic review and meta-analysis. *J Biomed Inform* 2024;151:104620.
4 Li J, Dada A, Puladi B, *et al*. ChatGPT in healthcare: a taxonomy and systematic review. *Comput Methods Programs Biomed* 2024;245:108013.
5 Wang L, Wan Z, Ni C, *et al*. A systematic review of chatgpt and other conversational large language models in healthcare. *Health Informatics* [Preprint].

6 Andrikyan W, Sametinger SM, Kosfeld F, *et al*. Artificial intelligence-powered chatbots in search engines: a cross-sectional study on the quality and risks of drug information for patients. *BMJ Qual Saf* 2025;34:100–9.
7 Holden RJ, Carayon P, Gurses AP, *et al*. SEIPS 2.0: a human factors framework for studying and improving the work of healthcare professionals and patients. *Ergonomics* 2013;56:1669–86.
8 Skivington K, Matthews L, Simpson SA, *et al*. A new framework for developing and evaluating complex interventions: update of Medical Research Council guidance. *BMJ* 2021;374:n2061:2061:.
9 Tran T, Joseph V, Smith L, *et al*. CardioCanon: A Customised Chatbot for Cardiology Inquiry With Retrieval Augmented Generation to Reduce Hallucinations and Improve Performance of Large Language Models. *Heart Lung Circ* 2024;33:S379–80.
10 Jones MD, Liu S, Powell F, *et al*. Exploring the Role of Guidelines in Contributing to Medication Errors: a Descriptive Analysis of National Patient Safety Incident Data. *Drug Saf* 2024;47:389–400.
11 Wang LW, Miller MJ, Schmitt MR, *et al*. Assessing readability formula differences with written health information materials: application, results, and recommendations. *Res Social Adm Pharm* 2013;9:503–16.
12 Raynor DK, Knapp P, Silcock J, *et al*. 'User-testing' as a method for testing the fitness-for-purpose of written medicine information. *Patient Educ Couns* 2011;83:404–10.
13 Joseph-Williams N, Elwyn G, Edwards A. Knowledge is not power for patients: A systematic review and thematic synthesis of patient-reported barriers and facilitators to shared decision making. *Patient Educ Couns* 2014;94:291–309.
14 Hoffmann TC, Del Mar C. Patients' expectations of the benefits and harms of treatments, screening, and tests: a systematic review. *JAMA Intern Med* 2015;175:274–86.
15 Thornton N, Hardie T, Horton T, *et al*. Priorities for an AI in health care strategy: the health foundations. 2024. Available: https://www.health.org.uk/publications/long-reads/priorities-for-an-ai-in-health-care-strategy
16 Thornton N, Binesmael A, Horton T, *et al*. AI in health care: what do the public and NHS staff think?: the health foundation. 2024. Available: https://www.health.org.uk/publications/long-reads/ai-in-health-care-what-do-the-public-and-nhs-staff-think