


Large language models for reducing clinicians' documentation burden

Kirk Roberts

 Check for updates

Evaluation of a clinical summarization method based on GPT-4 suggests that such models might reduce the documentation burden on clinicians – but prospective evaluation with high-priority tasks will be the true test of its potential.

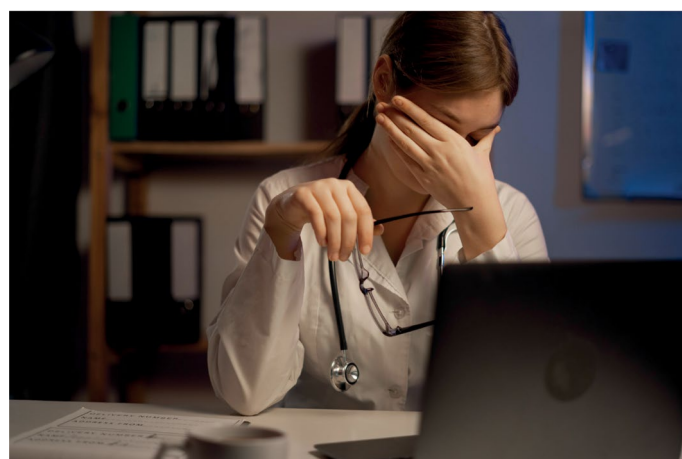
It has long been assumed that artificial intelligence (AI) tools would be adopted by clinicians once AI researchers could provide rigorous evidence of the safety and efficacy of such tools. Instead, some of the first major applications of AI in widespread clinical use^{1,2} are occurring without rigorous evidence and in spite of the concerns of AI researchers and others about the safety of these tools^{3,4}.

To be clear, the excitement over this technology in clinical medicine is not the result of improvements in AI alone: outdated regulatory requirements, narrowly focused health-system priorities, and user-unfriendly electronic health records have all contributed to burnout as clinicians are forced to choose between administrative priorities, duty to patients, and their own health⁵. Amidst this burgeoning crisis of clinician burnout, AI offers a potential means of deliverance. As a case in point, in this issue of *Nature Medicine*, Van Veen et al.⁶ show that AI-generated clinical summaries can reduce large amounts of patient data to feasible quantities, resulting in reduced clinician stress, more time for patient care, and possibly even fewer documentation errors. The question remains, however, whether the existing state of AI can truly offer a deliverance from these burdens without adversely influencing patient care.

The pivotal AI technology used by Van Veen et al.⁶ – large language models (LLMs) – represents the latest step in an evolution towards larger and more interconnected models (generally following the transformer architecture⁷), using massive amounts of data to train these models via methods such as self-supervised pre-training, supervised fine-tuning, and reinforcement learning with human feedback. Yet this latest evolutionary step has been far from incremental in its impact, exemplified by the Generative Pre-trained Transformer (GPT) line of models from OpenAI. These models possess remarkable potential for customization to suit a wide variety of tasks, including complex tasks in the medical domain.

Within this context, Van Veen et al.⁶ present a robust evaluation of LLMs for clinical text summarization. They evaluate eight LLM models across six clinical summarization data sets, along with two adaptation strategies to improve model performance. The result of these experiments is that OpenAI's GPT-4 model – using an in-context learning approach (that is, 'few-shot' learning with limited annotated examples) for adaptation – generally performed best.

The authors then carried out a head-to-head comparison of their adapted GPT-4 model and the 'human' summary (as derived from the



benchmark), with ten physicians rating both summaries in terms of completeness, correctness, and conciseness. In this regard, the LLM was largely equal to or better than the human-authored summary. This is an impressive result, not least because of the diversity of the datasets evaluated (encompassing radiology reports, progress notes, patient questions, and doctor–patient dialogues). Of additional importance, OpenAI's GPT-4 application programming interface (API) and the relatively straightforward prompting should be easy to replicate, allowing such approaches to be widely implemented (the authors also make their code freely available).

Before health systems rush to integrate such technology into clinical practice, however, it is important to put this work into context. The evaluations of Van Veen et al.⁶ were conducted on retrospective data sets, not in prospective scenarios in which the generated summaries were for actual patients of the clinicians involved in the user study. Furthermore, although the datasets were diverse, they do not necessarily represent high-priority targets to improve clinical workflows. For example, in the radiology datasets, the task was to generate the 'Impression' section (the key high-level information) of a radiology report, given the information in the report's 'Findings' section. This is a common task for the development of natural language processing (NLP) methodology, not least because it is easy to acquire large amounts of training data (the three datasets used by Van Veen et al.⁶ represent approximately 200,000 reports), but also because carrying out this task well requires the summarization model to identify, store, prioritize, and generate the key findings. Although useful for the development of NLP methods in both summarization⁸ and other tasks⁹, however, this is not of great practical importance; once a human radiologist has read an image and recorded all potentially relevant information in the 'Findings' section, writing the 'Impression' section is a low-burden task and unlikely to require automation on its own. The fact that the GPT-4-based approach did not require more than a few examples of this clinical summarization

task to perform at a human level suggests that it is indeed a major advance over prior NLP approaches. But the task is still, at best, a proxy for more clinically relevant summarization applications.

The existing state of LLMs for reducing clinical burdens through summarization provides sufficient impetus for prospective trials. Notably, this should include clinicians' use of summarization tools involving their own patients to assess not just the quality of the summary, but also whether a clinician would actually trust such a tool in practice, and how this would affect patient outcomes.

Much has been made of these models' tendency to 'hallucinate' (which, in this context, would involve including unsubstantiated information in the summary), but in summarization tasks, omission is a problem as well. This is especially concerning when it comes to real-world clinical data sets that are likely to include far more patient data than provided to GPT-4 for summarization by Van Veen et al.⁶, including hundreds of clinical notes as well as structured data and perhaps images. Many summarization tasks, furthermore, are guided by user input, wherein the clinician may provide a specific aspect of the patient's medical history to summarize (for example, their history of heart failure). This often requires incorporation of information-retrieval (search) methods, to which LLMs on their own are not necessarily well suited¹⁰.

It is well known that there are a wide variety of ethical considerations regarding the use of LLMs for clinical purposes. Not only do the models contain biases that reflect historical human biases, but the best-performing models (such as those from OpenAI) are far from transparent: we do not know which data they were built upon, which makes it even more difficult to assess how equitable their impact on society will be. Such proprietary APIs also potentially expose patient data, raising privacy concerns¹¹.

Finally, the success of LLMs across a wide variety of language tasks should force us to consider a fundamental, yet provocative, question: how much of traditional clinical documentation is even necessary? Given the ability of AI systems to understand patient-provider conversations, and given that AI methods will continue to advance in this

way, perhaps the primary form of clinical documentation should be the patient-provider interaction itself, be it in the form of a text transcript, audio, or even a video recording of a clinical encounter. AI-generated documentation should not be limited to the technology that was available at the time of the encounter, but should be generated from the best model available at the time the information is needed. In this way, instead of using AI tools as a direct replacement for traditional documentation, perhaps AI could obviate the need for most clinical documentation entirely. There are regulatory and administrative hurdles to this goal, of course, but these hurdles represent the same sources of burnout that have driven interest in AI-based summarization in the first place. Perhaps the best use of AI will be to make it obvious that clinical documentation requirements are not only burdensome, but outdated. Then AI could truly offer the deliverance hoped for by so many.

Kirk Roberts  

McWilliams School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX, USA.

 e-mail: Kirk.Roberts@uth.tmc.edu

Published online: 01 April 2024

References

1. Palmer, K. & Ross, C. *STAT* <https://go.nature.com/4a6c7D8> (21 February 2024).
2. Taylor, J. *Guardian* <https://go.nature.com/3TwNTMT> (27 July 2023).
3. Zhang, G. et al. Preprint at <https://arxiv.org/abs/2311.11211> (2023).
4. Duffourc, M. & Gerke, S. *J. Am. Med. Assoc.* **330**, 313–314 (2023).
5. National Academy of Medicine. *Taking Action Against Clinician Burnout: A Systems Approach to Professional Well-Being* (National Academies Press, 2019).
6. Van Veen, D. et al. *Nat. Med.* <https://doi.org/10.1038/s41591-024-02855-5> (2024).
7. Vaswani, A. et al. In *Adv. Neural Inf. Process. Syst.* Vol 30 (2017).
8. Zhang, Y. et al. In *Int. Workshop Health Text Mining Inf. Anal.* 204–213 (2018).
9. Soni, S. et al. In *Proc. 13th Language Res. Eval. Conf.* 6250–6259 (2022).
10. Hersh, W. *J. Am. Med. Inform. Assoc.* <https://doi.org/10.1093/jamia/ocae014> (2024).
11. Canter, G. P. & Packel, E. A. *J. Am. Med. Assoc.* **330**, 311–312 (2023).

Competing interests

The author declares no competing interests.