Invited Commentary | **Health Informatics**

Enhancing Health Care Communication With Large Language Models—The Role, Challenges, and Future Directions

Charumathi Raghu Subramanian, MD; Daniel A. Yang, MD; Raman Khanna, MD

Large language models (LLMs), a subset of artificial intelligence (AI) models trained on massive data sets that can synthesize and generate human-like text, have attracted tremendous interest in health care.[1] Researchers and health care professionals are actively exploring ways to integrate the use of LLMs for clinical administration tasks, clinical note generation, diagnostic support, and other activities. In addition to the Generative Pre-trained Transformer (GPT), there are other general purpose models, such as bidirectional encoder representations from transformers, and pathways language model, that have fine-tuned variants for medical and biomedical domains. There are also models trained exclusively on clinical text corpora, such as GatorTron and NYUTron, that have shown early promise in projecting hospital length of stay, in-hospital mortality, and hospital readmissions.[2]

Effective patient communication is critical in health care, and given the significant role that written text plays in this context, text-based AI models present numerous possibilities for enhancing patient communication.[3] Organizational health literacy is the degree to which health care organizations implement strategies to make it easier for patients to understand health information, navigate the health care system, and manage their health.[4] Central to achieving organizational health literacy is the enhancement of patient-oriented written communication. Studies have demonstrated that improving the readability of such communications is positively associated with patient outcomes,[5] and LLMs could be a potential tool here.

In this context, the study by Zaretsky and team[6] addresses an important task: using an LLM, GPT-4 (OpenAI), to transform hospital discharge summaries into a format that is readable for patients. Zaretsky et al[6] reviewed existing summaries and patient preferences to determine which elements should be included or excluded in the new format, processed the original summaries to retain only the specific patient-relevant elements (eg, removing billing codes), and inputted the processed summary into the LLM with instructions, modified through prompt engineering, to produce a concise, 1-page, patient-readable summary. These revised summaries were then assessed for their readability and understandability coupled with the balancing metrics of accuracy and completeness. Zaretsky et al[6] found that the LLM-generated summaries were shorter, more readable, and more understandable than the original discharge summaries. More than half of discharge summaries (54 of 100 summaries [54%]) were transformed successfully into a patient-readable format with top ratings in accuracy; yet 18 summaries were flagged for potential safety risks. Furthermore, among the 46 reviews flagged as less accurate, 24 (52%) were due to omissions and 4 (9%) were due to hallucinations.[6] These findings improve on readability scores and rates of hallucinations found in other patient education studies using LLMs.[7,8]

Given the principle of primum non nocere, the potential safety risk that Zaretsky et al[6] found is concerning. It is perhaps understandable that omissions occurred, given strict word restrictions to fit the discharge summary into a single page. It is conceivable that a longer discharge summary could have reduced rates of omission. While not evaluated in this study,[6] it is also possible that human clinicians might also omit key clinical elements in a brief discharge summary.

More concerning, despite their relative rarity, was the presence of hallucinations, particularly given the high confidence with which they were presented.[6] While the impact of LLM-generated hallucinations in health care remains unclear, hallucinations that are patient facing are more concerning than those facing clinicians, who may be able to better identify and correct them.

Inaccuracies noted in this study by Zaretsky et al,[6] such as mentioning nonexistent infections or chest pain, not only pose safety risks but also threaten the trust between patients and health care practitioners.

In addition to these safety concerns, the practical challenges of implementing such a tool in routine clinical practice must be considered. Manual processing of the discharge summaries to remove unnecessary elements is unlikely to be feasible on a busy clinical service, although there are technical solutions that could automate some or all of this work. More substantively, Zaretsky et al[6] underwent an extensive 6-week process of prompt engineering, and it is unclear the extent to which their work could be reusable directly or whether it would require additional reengineering at other institutions with different practice patterns and disease burdens.

With that said, Zaretsky and colleagues[6] should be commended for this early study, which highlights a novel use case of LLMs in health care to generate net new text to enhance patient education. Other studies have assessed LLM performance in replacing writing that clinicians would need to otherwise generate (eg, clinical notes, claim denial appeal letters); the study by Zaretsky et al[6] evaluates LLM performance for writing that clinicians would not usually, and perhaps rarely, have time to write for their patients. Using technology to create new products rather than just replace existing ones is an important step in realizing the gains of new digital technologies.[9] As Zaretsky et al[6] find, LLMs may not be ready for widespread unsupervised use to generate patient-facing discharge summaries, given real safety risks and formidable, if solvable, technological and workflow barriers, but perhaps in the near future, with better safety profiles, more automated inputs and outputs, and strict clinician oversight, they may become important tools in enhancing health care communication.

REFERENCES

1. Bommasani R, Hudson DA, Adeli E, et al. On the Opportunities and Risks of Foundation Models. *arXiv*. Preprint posted online July 12, 2022. http://arxiv.org/abs/2108.07258

2. Jiang LY, Liu XC, Nejatian NP, et al. Health system-scale language models are all-purpose prediction engines. *Nature*. 2023;619(7969):357-362. doi:10.1038/s41586-023-06160-y

3. Clusmann J, Kolbinger FR, Muti HS, et al. The future landscape of large language models in medicine. *Commun Med (Lond)*. 2023;3(1):141. doi:10.1038/s43856-023-00370-1

4. Brega AG, Hamer MK, Albright K, et al. Organizational health literacy: quality improvement measures with expert consensus. *Health Lit Res Pract*. 2019;3(2):e127-e146. doi:10.3928/24748307-20190503-01

5. Kaper MS, Sixsmith J, Reijneveld SA, de Winter AF. Outcomes and critical factors for successful implementation of organizational health literacy interventions: a scoping review. *Int J Environ Res Public Health*. 2021;18 (22):11906. doi:10.3390/ijerph182211906

6. Zaretsky J, Kim JM, Baskharoun S, et al. Generative artificial intelligence to transform inpatient discharge summaries to patient-friendly language and format. *JAMA Netw Open*. 2024;7(3):e240357. doi:10.1001/jamanetworkopen.2024.0357

7. Hung YC, Chaker SC, Sigel M, Saad M, Slater ED. Comparison of patient education materials generated by chat generative pre-trained transformer versus experts: an innovative way to increase readability of patient education materials. *Ann Plast Surg*. 2023;91(4):409-412. doi:10.1097/SAP.0000000000003634

8. Moons P, Van Bulck L. Using ChatGPT and Google Bard to improve the readability of written patient information: A proof-of-concept. *Eur J Cardiovasc Nurs*. 2023;zvad087. doi:10.1093/eurjcn/zvad087

9. Brynjolfsson E, McAfee A. *The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies*. W. W. Norton & Co; 2016.