MAYO CLINIC PROCEEDINGS: DIGITAL HEALTH

COMMENTARY

A Framework for Critically Assessing ChatGPT and Other Large Language Artificial Intelligence Model Applications in Health Care

Jonathan Ilicki, MD, MSc, MA

From the Innovation Team, Platform24, Sweden

Large language models (LLMs) are pretrained artificial intelligence (AI) algorithms that can interpret text and generate human-like text in real time.[1] Recent studies on LLMs (eg, PaLM and GPT-3.5) have found near-human performance on medical examinations and many possible future applications have been discussed, such as drafting discharge summaries, answering consultations, or generating lists of differential diagnosis.[1-7] Moreover, ChatGPT can already automatically generate empathetic responses to patients and seemingly genuine scientific abstracts.[8,9] The number of possible applications is likely to continue to increase, especially with large public investments in health care AI and as LLMs' capabilities continue to improve.[10,11]

However, it can be challenging for clinicians with limited technical understanding to assess the feasibility of such applications.[12] This could result in focusing on unrealistic applications or neglecting promising ones. Frameworks have been developed to assist in assessing AI applications for specific domains, eg, for radiology.[13] However, no such framework exists for LLM applications. Therefore, this article posits a simple framework for nontechnical health care professionals for assessing the feasibility of potential LLM applications in health care.

The framework consists of the following 4 steps:

1. Determine the main source of health care data that the LLM uses
2. Determine the intended recipient of the LLM's output
3. Combine the answers from (1) and (2) to identify a category
4. Assess fundamental limitations for that category

Step 1: Determine the Main Source of Health Care Data

Determine whether the health care data that the LLM will use to reply to prompts comes from patients (eg, health data or medical records), health care providers (eg, information on procedures, medications, research or organizational information, such as opening hours), or payers (eg, information on reimbursement of procedures).

Step 2: Determine Intended Recipient

Determine whether the main reader of the output of LLM is a patient, provider, or payer.

Step 3: Identify Category

Combine the answers from steps 1 and 2 in the LLM feasibility framework (Table 1) to identify which category the solution belongs to.

Step 4: Assess Fundamental Limitations

Table 2 describes categories and corresponding fundamental limitations, which can be used to assess the feasibility of a specific application. LLMs can be used in many different ways and are developing rapidly. However, some limitations are intrinsic to the AI model itself, which can be seen in the literature to date, and these are unlikely to change despite the rapid development. In brief, these limitations are as follows:

i. **Lack of understanding**: LLMs lack a human-like understanding of the real-world phenomena that words describe and only process their semantic representation. This

TABLE 1. LLM Feasibility Framework: Matrix for Determining Category of LLM Application

	Main source of health care data		
	Using patient data…	Using provider data…	Using payer data…
Main recipient of output	… to highly automate summaries or explanations of…		
Patients Adapting output (see examples) to, eg., individual patients' health literacy, medical history, and current medications	**Category 1** Example: Patient's own medical records (eg, discharge notes, laboratory results, investigations)	**Category 2** Example: Provider information (eg, medications, treatments, preoperative processes)	**Category 3** Example: Payer information (eg, coverage, explanation of health care system, available providers)
Providers Adapting output (see examples) to, eg., providers' specific clinical context, resources, or inquiry	**Category 2** Example: Pertinent patient information (eg, from medical records, laboratory results)	**Category 2** Example: Relevant medical information (eg, merging local or international guidelines, research)	**Category 3** Example: Relevant payer information (eg, reimbursement, quality measures, or coverage)
Payers Adapting output (see examples) to, eg, payers' specific rules on coverage, reimbursement, or quality measures	**Category 2** Example: Relevant population data (eg, aggregate statistics from free text medical records)	**Category 3** Example: Relevant provider information (eg, quality, efficiency or cost of providers/ pathways)	**Category 3** Example: Improving existing internal knowledge management systems

LLM, large language model.

lack of understanding is highlighted by unpredictable illogical errors in reasoning in recent LLM studies.[2,5,14] This lack of real-world understanding limits the extent to which LLMs can act autonomously without oversight and creates the need of control mechanisms to ensure the appropriateness of the output.

ii. Lack of predictability: LLMs run the risk of creating "hallucinations" (text responses that are either nonsensical or unfaithful to the content they should use) and errors that are difficult to predict, which can entail patient risks.[1,4,15] Manufacturers must ensure that a medical LLM software performs in a safe and predictable manner according to relevant legislation (eg, the Medical Device Regulation in Europe). Guaranteeing that a LLM does not create any hallucination is challenging. This risk can be partially mitigated by, for example, letting a clinician assess the output before it is acted on (to identify errors) or by forcing the LLM to reference external sources for the statements in its output (to allow comparisons with the original data that the output is based on).

iii. Lack of empathy: Even if LLMs can generate seemingly empathetic responses, they cannot experience emotions or empathize with a patient when providing emotional support.[16] Moreover, people may not perceive empathy as genuine when coming from an algorithm.[17] This may change over time but is currently a limitation in, for example, using unsupervised LLM output to provide patients with sensitive information.

The framework could be applied as follows: Imagine a LLM application that aims to improve patient adherence by adapting generic medication information (provider data) to patients (patient recipient) and different levels of health literacy. The combination of data and recipient places the application in category 2, and therefore it would be important to understand how the application addresses the lack of understanding and predictability by LLMs.

This framework has several limitations. First, it is not exhaustive but is designed as a simple heuristic for an initial understanding of what fundamental limitations a LLM application may have. This framework does not replace a comprehensive assessment, which is needed before clinical implementation. Such

Mayo Clin Proc Digital Health ■ June 2023;1(2):185-188 ■ https://doi.org/10.1016/j.mcpdig.2023.03.006
www.mcpiqojournal.org

TABLE 2. LLM Feasibility Framework: Limitations relevant for each category

Category	Example of healthcare data used	Fundamental limitations relevant for category: Lack of understanding	Lack of predictability	Lack of empathy
1: Output without clinical supervision	- Patient health data: e.g. medical records, blood results, patient reported outcome measures, data from wearables	✓	✓	✓
2: Supervised output which can impact clinical decisions	- Patient health data (as above) - Generic provider data: information about e.g. medications, treatments, procedures, research - Specific provider data: information about e.g. clinicians, opening hours, services provided	✓	✓	
3: Administrative output	- Provider information (generic/specific as above) - Payer data: administrative data, process measures, reimbursement, costs	✓		

LLM, Large language model.

an assessment will include several important aspects, such as interpretability of models (to what extent one can understand why they produce a certain result), which LLM is used, if the training data are sufficiently representative and of high quality, and whether the model has been fine-tuned to medical data. Second, it can only be used to identify potentially impractical solutions but not to confirm the feasibility of solutions. Last, despite incorporating limitations that seem fundamental, these may change as LLMs and social norms develop.

Conclusion

Notwithstanding the abovementioned limitations, this framework aims to aid nontechnical health care professionals in critically assessing emerging LLM applications and ensuring their development into clinically safe and useful tools. LLMs have great potential to improve many parts of health care, but more research is needed to understand their performance, safety, and effect on health care systems. Finally, when addressing novel emerging technologies, keep Amara's law in mind: "We tend to overestimate the effect of a technology in the short run and underestimate the effect in the long run."[18]

POTENTIAL COMPETING INTERESTS

Dr Ilicki is employed by Platform24 which develops software for health care. Platform24 does not currently develop or use any LLM in its software. J.I. also serves on the boards of Hypocampus and Geras Solutions, which do not currently develop or use any LLM in their software. The author reports no competing interests.

Abbreviations and Acronyms: **AI,** artificial intelligence; **LLM,** large language model

Grant Support: This study did not receive any financial support.

Correspondence: Address to Jonathan Ilicki, MD, MSc, MA, Platform24, Västra Järnvägsgatan 7, 111 64 Stockholm, Sweden (j.ilicki@gmail.com).

ORCID
Jonathan Ilicki: https://orcid.org/0000-0002-6514-8554

REFERENCES

1. Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. Preprint. Posted online December 26, 2022. arXiv. 2212.13138. https://doi.org/10.48550/arXiv.2212.13138
2. Liévin V, Egeberg Hother C, Winther O. Can large language models reason about medical questions? Preprint. Posted online July 17, 2022. arXiv. 2207.08143. https://doi.org/10.48550/arXiv.2207.08143
3. Patel SB, Lam K. ChatGPT: the future of discharge summaries? *Lancet Digit Health*. 2023;5(3):e107-e108.
4. Hirosawa T, Harada Y, Yokose M, Sakamoto T, Kawamura R, Shimizu T. Diagnostic accuracy of differential-diagnosis lists generated by generative pretrained transformer 3 chatbot for clinical vignettes with common chief complaints: a pilot Study. *Int J Environ Res Public Health*. 2023;20(4):3378.
5. Gilson A, Safranek CW, Huang T, et al. 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.
6. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. *PLOS Digit Health*. 2023;2(2):e0000198.
7. Howard A, Hope W, Gerada A. ChatGPT and antimicrobial advice: the end of the consulting infection doctor? *Lancet Infect Dis*. 2023;23(4):405-406.
8. Else H. Abstracts written by ChatGPT fool scientists. *Nature*. 2023;613(7944):423.
9. Goode J. A mental health tech company ran an AI experiment on real users. Nothing's stopping apps from conducting more. NBC News. https://www.nbcnews.com/tech/internet/chatgpt-ai-experiment-mental-health-tech-app-koko-rcna65110. Accessed January 19, 2023.
10. Health Secretary announces £250 million investment in artificial intelligence.Gov.UK. https://www.gov.uk/government/news/health-secretary-announces-250-million-investment-in-artificial-intelligence. Accessed January 19, 2023.
11. Rosemain M, Rose M. France to spend $1.8 billion on AI to compete with U.S., China. Reuters Techonology News. https://www.reuters.com/article/us-france-tech-idUSKBN1H51XP. Accessed January 19, 2023.
12. Chen M, Zhang B, Cai Z, et al. Acceptance of clinical artificial intelligence among physicians and medical students: a systematic review with cross-sectional survey. *Front Med (Lausanne)*. 2022;9:990604.
13. Omoumi P, Ducarouge A, Tournier A, et al. To buy or not to buy-evaluating commercial AI solutions in radiology (the ECLAIR guidelines). *Eur Radiol*. 2021;31(6):3786-3796.
14. Rao A, Kim J, Kamineni M, Pang M, Lie W, Succi MD. Evaluating ChatGPT as an adjunct for radiologic decision-making. Preprint. Posted online February 7, 2023. medRxiv. 23285399. https://doi.org/10.1101/2023.02.02.23285399
15. Ji Z, Lee N, Frieske R, et al. Survey of hallucination in natural language generation. Preprint. Posted online February 8, 2022. arXiv. 2202.03629. https://doi.org/10.1145/3571730
16. Montemayor C, Halpern J, Fairweather A. In principle obstacles for empathic AI: why we can't replace human empathy in healthcare. *AI Soc*. 2022;37(4):1353-1359.
17. Morris RR, Kouddous K, Kshirsagar R, Schueller SM. Towards an artificially empathic conversational agent for mental health applications: system design and user perceptions. *J Med Internet Res*. 2018;20(6):e10148.
18. Amara R. 1925-2017, American futurologist. In: Ratcliffe S, ed. Oxford Essential Quotations.*5th ed*. Oxford University Press; 2016.