Evaluating the Application of Large Language Models in Clinical Research Contexts

Roy H. Perlis, MD, MSc; Stephan D. Fihn, MD, MPH

The development of humanlike, interactive systems has been the quest of artificial intelligence (AI) research for decades. The seemingly abrupt advent of readily accessible, large language models (LLMs) has been greeted with great, and in some cases irrational, enthusiasm, accompanied by breakneck efforts to implement this new technology in myriad aspects of daily life, including medicine. LLMs are being applied in a wide array of clinical and medical educational settings, among them responding to patients' questions, generating notes from clinical encounters, creating and answering test questions, assisting with diagnoses, and guiding or engaging in therapeutic interactions. As is the case with other applications of AI in medicine, such as image processing and clinical risk prediction, there is the possibility of great benefit, as well as legitimate concern about unrecognized suboptimal or erroneous performance and even potential harm. Accordingly, research in this area is essential and needs to be conducted at a pace that parallels the rapidity with which this technology is evolving. At JAMA Network Open, we currently receive dozens of manuscripts evaluating various applications of LLM every month.

In proposing how this type of technology might be evaluated, nearly 75 years ago Alan Turing enumerated some key criteria that remain valid today. He recommended using a question-answer format that permits the system to demonstrate the use of natural language, reason, knowledge, and learning. (He also thought such systems should exhibit humanlike emotion.) Much more recently, an abundance of guidelines for the application of AI in biomedical research have emerged. Reassuringly, many of them focus on principles that will be familiar to any clinical investigator, including validity, fairness, and transparency.

In an effort to guide authors seeking to publish on clinical use of LLM, the editors of JAMA Network Open have developed the following criteria for submissions:

1. The application being evaluated must be clinically meaningful. An ideal report addresses a breadth of related experimental conditions relevant to clinical settings and not a single, narrow area.
2. Experimental conditions must be described with sufficient detail to allow replication by a knowledgeable user. All LLMs have multiple tunable parameters, such as temperature, that will affect their output, even if chatbots do not necessarily expose these parameters. Moreover, the version of the model itself can change over time and needs to be specified.
3. Multiple replicates are necessary to understand the stability of responses. Because text generation is probabilistic, the same prompt may yield responses that are quite different from run to run, so it is important to repeat the same request multiple times to estimate effects or associations and some measure of variability. Authors are also encouraged to consider the extent to which alternate prompts may change outputs, because these can be useful sensitivity analyses in establishing the robustness of results.
4. Comparison of LLM versions are generally better suited to other journals. Although comparisons between models are a sine qua non of many AI articles, authors should consider that most nonspecialists are not interested in how much better a new model is compared with an old one, particularly given the rapid rate at which models are being supplanted. Studies evaluating the most current versions of an LLM model are of greatest interest.
5. Beyond basic indices of accuracy, characteristics of incorrect or flawed responses should be described. As with biomarkers, the value of accuracy depends critically on the context in which a technology is applied and the consequences of error. So, for example, a mostly correct response to a case presentation that recommends a suboptimal but reasonable treatment may have very

Open Access. This is an open access article distributed under the terms of the CC-BY License.
different consequences from a very-incorrect one that recommends a contraindicated treatment. We encourage efforts to understand how and when models report wrong results, with examples of such responses.

6. Bias and fairness should be assessed.4 As with other forms of AI, LLMs trained on the breadth of the internet are susceptible to providing biased responses. Despite efforts to censor such responses or fine-tune training to minimize them, authors should consider whether modifying prompts to reflect individual characteristics yields different results.

7. Confidentiality must be protected. Among the most promising applications of LLMs may be summarization and interpretation of clinical data.1 As has been pointed out in the context of manuscript and grant reviews,5,6 simply uploading such data to a third party represents a breach of confidentiality. There are ways that appropriate protections can be maintained (eg, by hosting LLMs within an institutional firewall), but they require additional thought and investment of resources. Manuscripts that apply LLMs to clinical data must report how confidentiality was maintained.

8. As Turing advised,2 whenever possible, authors should aim to compare LLM results to some established expert reference, rather than simply describing model output. In accordance with other journal policies regarding data sharing, making these reference standards available as part of publication is strongly encouraged. If that standard comes from clinicians7 or other human participants, authors must consider carefully whether review by a human participants committee is required; in many cases, the individuals establishing a reference standard may be considered research participants. Even if the investigator believes a study is likely to be considered exempt, this is a determination made by the institutional review board.

As with any new technology, we anticipate that standards will evolve as the strengths and limitations of generative AI are better understood. Modern medicine entails a team effort; learning to incorporate this powerful but challenging new team member will require the sort of evidence that JAMA Network Open hopes to publish in the years to come.
REFERENCES


