Radiology

REVIEWS AND COMMENTARY • EDITORIAL

Feasibility and Prospect of Privacy-preserving Large Language Models in Radiology

Wenli Cai, PhD

Wenli Cai is an assistant professor of radiology at Massachusetts General Hospital and Harvard Medical School. His research interests focus on quantitative image analysis, medical image processing, computer-aided diagnosis, radiomics, machine learning, and their applications in clinical oncology.

Recent advances in large language models (LLMs), exemplified by the success of models like ChatGPT and GPT-4 developed by OpenAI, have had a profound impact on the health care and medical domains (1). In radiology, institutions are actively exploring the adoption of this technology in clinical settings to streamline and automate various clinical text processing tasks aimed at assisting radiologists in image interpretation and report generation (2).

However, the use of ChatGPT in clinical domains requires patient data to be transmitted to OpenAI's external server (the third party) through either a chat window or an application programming interface. This situation inevitably raises concerns regarding data security and patient privacy (3). To address these concerns, privacy-preserving LLMs (PP-LLMs) are emerging and have received substantial attention. PP-LLMs adopt a range of measures to ensure the protection of patient privacy and data security (4), including *(a)* encrypting or encoding users' prompts or queries before they are transmitted to a model; *(b)* maintaining confidentiality of any information related to the private data, prompts, and queries that the model receives and processes; and *(c)* enabling models to be cloned, compressed, and then deployed locally without sending any data to third parties. This ensures that patient data and the model's operations remain in full control of the local user, with no access possible by external parties.

In this issue of *Radiology*, Mukherjee et al (5) investigated the feasibility of using a locally deployed Vicuna-13B, a prototype PP-LLM, for labeling key findings in chest radiography reports. In contrast to the closed-source proprietary LLM family of GPT, Vicuna is one of the derivatives of LLaMA, an open-source LLM developed by Meta. Vicuna was fine-tuned using the training code from Alpaca, another member of the LLaMA series, and 70 000 user-shared ChatGPT conversations. Preliminary evaluation showed that Vicuna achieved over 90% of ChatGPT's response quality in user preference tests (6), despite having fewer parameters (13 billion) than either GPT-3.5 (175 billion) or GPT-4 (approximately 1.7 trillion).

In the study by Mukherjee et al (5), the performance of Vicuna-13B in the task of labeling 13 specific findings in chest radiography reports was compared with two well-established labeling tools, CheXpert and CheXbert. Two radiograph data sets were used, including 3269 free-text radiology reports from the publicly available MIMIC-CXR data set (7) and 25 596 reports from the National Institutes of Health (NIH) ChestX-ray14 data set (which are not publicly available) (8). It is worth mentioning that CheXpert and CheXbert were extensively trained for labeling radiography reports on data sets consisting of 222 750 reports from MIMIC-CXR and 78 506 reports from the NIH data set, respectively. In contrast, Vicuna-13B was used without any specific training for the labeling of chest radiography reports and without any form of fine-tuning.

Mukherjee et al (5) designed two tasks and two prompts to assess the agreement between Vicuna outputs and those from CheXpert and CheXbert.

Task 1 involved the direct labeling of 13 possible findings in a radiography report by assigning each finding to either a positive (value of 1), negative (0), not mentioned (NA), or unsure (−1) category.

Task 2 was a simplified version of task 1; it mapped all findings previously categorized as "not mentioned" or "unsure" to the "negative" category.

Two different prompts were designed to instruct Vicuna to generate the desired outputs.

Prompt 1 was a single-step prompt that instructed the model to generate a structured radiography labeling report by directly exporting each of the 13 findings (0, 1, −1, or NA for task 1 and 0 or 1 for task 2).

Prompt 2 was a multistep prompt that used a rule-based interactive prompting strategy to guide the model in a label-by-label manner in answering three yes-or-no questions to determine whether the finding should be classified as present, absent, or unsure/not mentioned.

When using prompt 1, Vicuna outputs for the 13 findings in both data sets showed, on average, poor agreement with outputs from CheXpert and CheXbert for task 1 (κ median, −0.48 to −0.40) and moderate agreement with outputs from the two labelers for task 2 (κ median, 0.46–0.56). When prompt 2 was used, Vicuna outputs in both data sets showed, on average, fair to moderate agreement with outputs from CheXpert and CheXbert

From the Department of Radiology, Massachusetts General Hospital and Harvard Medical School, 399 Revolution Dr, 13W44, Sommerville, MA 02145. Received September 1, 2023; revision requested and received September 7; accepted September 8. **Address correspondence to** the author (email: *dr.wlcai@gmail.com*).

Conflicts of interest are listed at the end of this article.

See also the article by Mukherjee et al in this issue.

for task 1 (κ median, 0.31–0.41) and moderate to substantial agreement with outputs from the two labelers for task 2 (κ median, 0.52–0.64). Compared with prompt 1, Vicuna outputs with prompt 2 showed better agreement with outputs from the two labelers in both data sets for task 1 ($P < .001$) and in the MIMIC-CXR data set for task 2 ($P = .02$). The authors also carried out a human evaluation study in which a random subset of 100 reports from the NIH data set were manually reviewed and annotated by a senior radiologist. For task 2, Vicuna with prompt 2 performed on par with CheXpert or CheXbert for nine of the 11 findings with more than a single true-positive finding in the subset of examinations.

The impacts of this study by Mukherjee et al (5) lie not only in its evaluation of the ability of general-purpose LLMs to label radiography reports, but more importantly, in the feasibility of locally deploying LLMs to protect patient privacy and data security. Indeed, to my knowledge, this study is one of the first investigations of a PP-LLM in a clinical setting.

First, the study showed the natural language processing capabilities of LLMs can be used for clinical tasks in radiology. As mentioned before, both CheXpert and CheXbert underwent substantial training for the special task of labeling radiography reports. This machine learning process requires substantial human efforts to collect and label large data sets, followed by training and testing of the model. In contrast, general-purpose LLMs such as ChatGPT or Vicuna are pretrained with extensive collections of text sourced from the internet. As such, the model itself has no specific knowledge of labeling radiography reports. With minimal effort of prompt coding, these pretrained LLMs could be readily applied to various tasks in radiology, such as the radiology report labeling demonstrated in their study, as well as radiology report generation, image annotation interpretation, clinical decision support, and even cancer screening and detection (9). LLMs offer the potential of a streamlined alternative to the traditional labor-intensive machine learning processes for domain-specific tasks.

Second, Vicuna, being an open-source LLM, supports local deployment within an institution. Unlike with ChatGPT, this local deployment strategy avoids the transmission of patient data to third-party servers. The institution intranet and firewall protect patient health information from being disclosed to and accessed by unauthorized entities. In addition, Vicuna has a substantially smaller number of parameters than GPT models do, which makes local deployment more feasible when computational resources are limited. Thus, the main advantages of model accessibility (open source), compactness (low number of model parameters), and security (local deployment), along with a performance quality similar to that of GPT models, make Vicuna one viable option for developing and deploying PP-LLMs in clinical settings, where the compliance with patient privacy and data security is a top priority.

Third, the study by Mukherjee et al (5) highlights the importance of structuring a specific and meaningful prompt in an LLM to capture the semantic differences in the context of a task. All it needs is a good prompt that is contextual, succinct, and informative to instruct the LLM through a specific task. Vicuna instructed by prompt 2, a multistep interactive prompt, improved the LLM's ability to label chest radiography reports without the need for any additional training. While the performance achieved in their feasibility study might not be optimal for immediate clinical implementation, it is expected that the model could be enhanced further through advanced prompt engineering techniques. For example, chain-of-thought prompting enables complex reasoning using few-shot exemplars, where the reasoning process is explicitly outlined, which could be used to build a well-structured and contextually rich framework for Vicuna (10).

While local deployment of an open-source LLM offers great potential for implementation of a PP-LLM, there are also considerations to bear in mind. For instance, the hardware costs for LLMs can be high, and models require ongoing maintenance and development efforts to ensure they continue to perform accurately. Additionally, appropriate regulation of access control, suitable methods for user authentication, and adherence to any data compliance requirements (such as potential data leakage or privacy breaches) when improving a model with queries involving patient data must be rigorously examined before implementation.

Interest in PP-LLMs is growing, and these models are expected to become one of the hottest research topics of generative AI in health care. The study by Mukherjee et al (5) contributes to this emerging field by demonstrating the feasibility of implementing a PP-LLM through local deployment in clinical settings, which may effectively address concerns about patient privacy and data security when using LLMs for clinical tasks.

Disclosures of conflicts of interest: W.C. Support from the National Institutes of Health/National Cancer Institute (grant no. R42CA189637) and the Children's Tumor Foundation (grant no. CTF-2021-10-02); stock in IQ Medical Imaging.

References

1. Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. N Engl J Med 2023;388(13):1233–1239.
2. Elkassem AA, Smith AD. Potential use cases for ChatGPT in radiology reporting. AJR Am J Roentgenol 2023;221(3):373–376.
3. Kanter GP, Packel EA. Health care privacy risks of AI chatbots. JAMA 2023;330(4):311–312.
4. Raeini M. Privacy-preserving large language models (PPLLMs). http://dx.doi.org/10.2139/ssrn.4512071. Posted July 24, 2023. Accessed August 25, 2023.
5. Mukherjee P, Hou B, Lanfredi RB, Summers RM. Feasibility of using the privacy-preserving large language model Vicuna for labeling radiology reports. Radiology 2023;e231147. Published online October 10, 2023.
6. Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality. The Large Model Systems Organization. https://lmsys.org/blog/2023-03-30-vicuna/. Published March 30, 2023. Accessed August 25, 2023.
7. Johnson AEW, Pollard TJ, Berkowitz SJ, et al. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. Sci Data 2019;6(1):317.
8. Wang X, Peng Y, Lu L, et al. ChestX-Ray8: hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR): IEEE, 2017.
9. Shen Y, Heacock L, Elias J, et al. ChatGPT and other large language models are double-edged swords. Radiology 2023;307(2):e230163.
10. Ott S, Hebenstreit K, Liévin V, et al. ThoughtSource: a central hub for large language model reasoning data. Sci Data 2023;10(1):528.