Almanac: Retrieval-Augmented Language Models for Clinical Medicine
Cyril Zakka, Rohan Shad, Curt Langlotz, Euan Ashley, William Hiesinger, and 17 others
Stanford Medicine, Stanford University, Johns Hopkins School of Medicine, and Columbia University

Introduction
Large language models (LLMs) have recently shown impressive zero-shot capabilities, whereby they can use auxiliary data, without the availability of task-specific training examples, to complete a variety of natural language tasks, such as summarization, dialogue generation, and question answering. However, despite many promising applications of LLMs in clinical medicine, adoption of these models has been limited by their tendency to generate incorrect and sometimes even harmful statements.

Methods
We tasked a panel of eight board-certified clinicians and two health care practitioners with evaluating Almanac, an LLM framework augmented with retrieval capabilities from curated medical resources for medical guideline and treatment recommendations. The panel compared responses from Almanac and standard LLMs (ChatGPT-4, Bing, and Bard) versus a novel data set of 314 clinical questions spanning nine medical specialties.

Results
Almanac showed a significant improvement in performance compared with the standard LLMs across axes of factuality, completeness, user preference, and adversarial safety. These results were echoed by the Nemenyi P values (P<0.01).

For citations, Almanac was able to provide correct and trustworthy citations for 91% of the ClinicalQA questions, with missed marks because of an inability to provide reliable sources when relying on its intrinsic knowledge.

Conclusion
Our results show the potential for LLMs with access to domain-specific corpora to be effective in clinical decision-making. The findings also underscore the importance of carefully testing LLMs before deployment to mitigate their shortcomings.

(Funded by the National Institutes of Health, National Heart, Lung, and Blood Institute.)