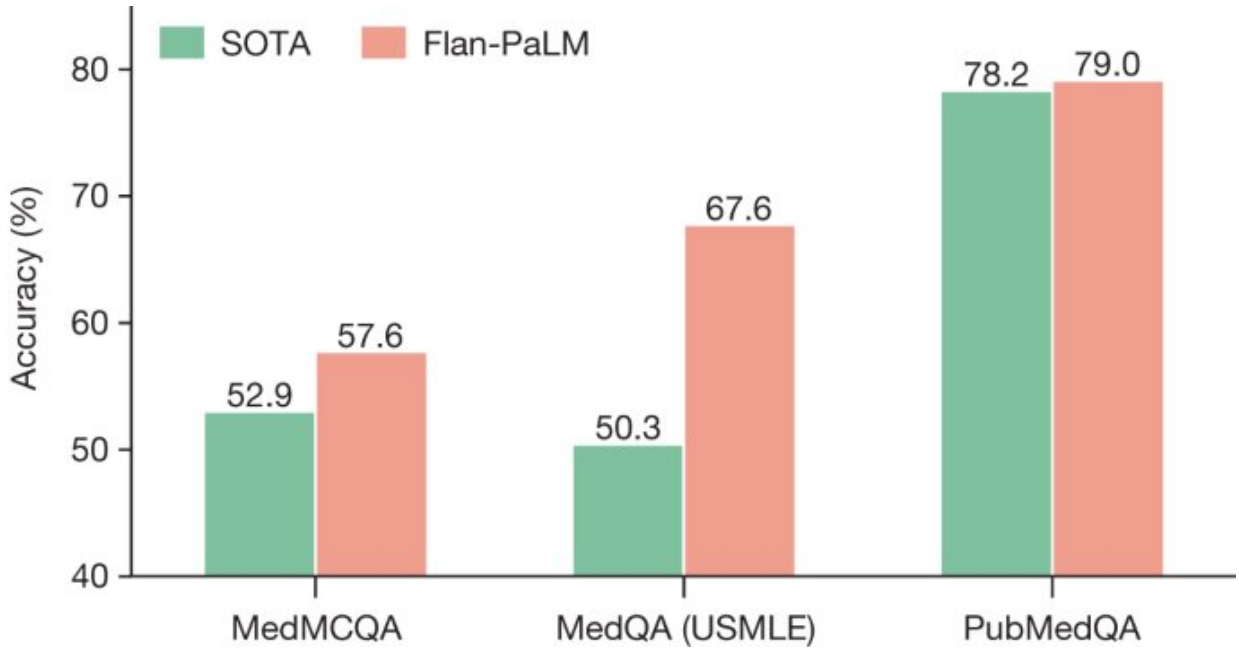
Benchmarking AI's ability to answer medical questions

July 14 2023



Our Flan-PaLM 540B model exceeds the previous state-of-the-art performance (SOTA) on MedQA (four options), MedMCQA and PubMedQA datasets. The previous state-of-the-art results are from Galactica20 (MedMCQA), PubMedGPT19 (MedQA) and BioGPT21 (PubMedQA). The percentage accuracy is shown above each column. Credit: *Nature* (2023). DOI: 10.1038/s41586-023-06291-2

A benchmark for assessing how well large language models (LLMs) can answer medical questions is presented in a paper published in *Nature*.

The study, from Google Research, also introduces Med-PaLM, an LLM specialized for the medical domain. The authors note, however, that many limitations must be overcome before LLMs can become viable for clinical applications.

Artificial intelligence (AI) models have potential uses in medicine, including knowledge retrieval and clinical decision support. However, existing models may, for instance, hallucinate convincing medical misinformation or incorporate biases that could exacerbate health disparities. Therefore, assessments of their clinical knowledge are needed. However, these assessments typically rely on automated evaluations on limited benchmarks, such as scores on individual medical tests, which may not translate to real-world reliability or value.

To evaluate how well LLMs encode clinical knowledge, Karan Singhal, Shekoofeh Azizi, Tao Tu, Alan Karthikesalingam, Vivek Natarajan and colleagues considered the ability of these models to answer medical questions.

The authors present a benchmark called MultiMedQA, which combines six existing question answering datasets spanning professional medicine, research and consumer queries, and HealthSearchQA, a new dataset of 3,173 medical questions commonly searched online.

The authors then evaluated the performance of PaLM (a 540-billion parameter LLM) and its variant, Flan-PaLM. They found that Flan-PaLM achieved state-of-the-art performance on several of the datasets. On the MedQA dataset comprising US Medical Licensing Exam-style questions, FLAN-PaLM exceeded previous state-of-the-art LLMs by more than 17%. However, while FLAN-PaLM performed well on multiple choice questions, human evaluation revealed gaps in its long-form answers to consumer medical questions.

To resolve this, the authors used a technique called instruction prompt tuning to further adapt Flan-PaLM to the medical domain. Instruction prompt tuning is introduced as an efficient approach for aligning generalist LLMs to new specialist domains.

Their resulting model, Med-PaLM, performed encouragingly in the pilot evaluation. For example, a panel of clinicians judged only 61.9% of Flan-PaLM long-form answers to be aligned with the scientific consensus, compared with 92.6% for Med-PaLM answers, on par with clinician-generated answers (92.9%). Similarly, 29.7% of Flan-PaLM answers were rated as potentially leading to harmful outcomes, in contrast to 5.8% for Med-PaLM, comparable with clinician-generated answers (6.5%).

The authors note that while their results are promising, further evaluations are necessary.

More information: Karan Singhal et al, Large language models encode clinical knowledge, *Nature* (2023). DOI: 10.1038/s41586-023-06291-2

Provided by Nature Publishing Group