

Nature publishes Google's Med-PaLM paper on how language models answer medical questions

Source copy kept by wheresthe.ai for the story "Nature publishes Google's Med-PaLM paper on how language models answer medical questions" (<https://wheresthe.ai/d/google-med-palm-nature-paper-2023-07/>).

Source: <https://pubmed.ncbi.nlm.nih.gov/37438534/>

Text saved 21 Oct 2026 (the abstract from PubMed (NCBI), because the journal page refuses automated reading).

SHA-256 of the saved text: 03d0f99d08f85b92317cc44472499c3c10ca71c2dd3de0df5b21ab00d1b2e420

1. Nature. 2023 Aug;620(7972):172-180. doi: 10.1038/s41586-023-06291-2. Epub 2023 Jul 12.

Large language models encode clinical knowledge.

Singhal K(#)(1), Azizi S(#)(2), Tu T(#)(3), Mahdavi SS(3), Wei J(3), Chung HW(3), Scales N(3), Tanwani A(3), Cole-Lewis H(3), Pfohl S(3), Payne P(3), Seneviratne M(3), Gamble P(3), Kelly C(3), Babiker A(3), Schärli N(3), Chowdhery A(3), Mansfield P(3), Demner-Fushman D(4), Agüera Y Arcas B(3), Webster D(3), Corrado GS(3), Matias Y(3), Chou K(3), Gottweis J(3), Tomasev N(5), Liu Y(3), Rajkomar A(3), Barral J(3), Semturs C(3), Karthikesalingam A(6), Natarajan V(7).

Author information: (1)Google Research, Mountain View, CA, USA. karansinghal@google.com. (2)Google Research, Mountain View, CA, USA. shekazizi@google.com. (3)Google Research, Mountain View, CA, USA. (4)National Library of Medicine, Bethesda, MD, USA. (5)DeepMind, London, UK. (6)Google Research, Mountain View, CA, USA. alankarthi@google.com. (7)Google Research, Mountain View, CA, USA. natviv@google.com. (#)Contributed equally

Erratum in Nature. 2023 Aug;620(7973):E19. doi: 10.1038/s41586-023-06455-0.

Large language models (LLMs) have demonstrated impressive capabilities, but the bar for clinical applications is high. Attempts to assess the clinical knowledge of models typically rely on automated evaluations based on limited benchmarks. Here, to address these limitations, we present MultiMedQA, a benchmark combining six existing medical question answering datasets spanning professional medicine, research and consumer queries and a new dataset of medical questions searched online, HealthSearchQA. We propose a human evaluation framework for model answers along multiple axes including factuality, comprehension, reasoning, possible harm and bias. In addition, we evaluate Pathways Language Model1 (PaLM, a 540-billion parameter LLM) and its instruction-tuned variant, Flan-PaLM2 on MultiMedQA. Using a combination of prompting strategies, Flan-PaLM achieves state-of-the-art accuracy on every MultiMedQA multiple-choice dataset (MedQA3, MedMCQA4, PubMedQA5 and Measuring Massive Multitask Language Understanding (MMLU) clinical topics6), including 67.6% accuracy on MedQA (US Medical Licensing Exam-style questions), surpassing the prior state of the art by more than 17%. However, human evaluation reveals key gaps. To resolve this, we introduce instruction prompt tuning, a parameter-efficient approach for aligning LLMs to new domains using a few exemplars. The resulting model, Med-PaLM, performs encouragingly, but remains inferior to clinicians. We show that comprehension, knowledge recall and reasoning improve with model scale and instruction prompt tuning, suggesting the potential utility of LLMs in medicine. Our human evaluations reveal limitations of today's models, reinforcing the importance of both evaluation frameworks and method development in creating safe, helpful LLMs for clinical applications.

© 2023. The Author(s).

DOI: 10.1038/s41586-023-06291-2 PMCID: PMC10396962 PMID: 37438534 [Indexed for MEDLINE]

Conflict of interest statement: This study was funded by Alphabet Inc. and/or a subsidiary thereof (Alphabet). K.S., S.A., T.T., V.N., A.K., S.S.M., C.S., J.W., H.W.C., N. Scales, A.T., H.C.-L., S.P., P.P., M.S., P.G., C.K., A.B., N. Schärli, A.C., P.M., B.A.A., D.W., G.S.C., Y.M., K.C., J.G., A.R., N.T., J.B. and Y.L. are employees of Alphabet and may own stock as part of the standard compensation package. D.D.-F. is affiliated with the US National Library of Medicine.