

Google says Med-PaLM 2 scores 85.4% on US medical licensing-style questions

Source copy kept by wheresthe.ai for the story “Google says Med-PaLM 2 scores 85.4% on US medical licensing-style questions” (<https://wheresthe.ai/d/google-med-palm-2-announced-2023-03/>).

Source: <https://sites.research.google/med-palm/>

Text saved 18 Oct 2026 (the page as published).

SHA-256 of the saved text: 8f2c5103bd4845aa8b27339c4f79b71e9047cc67d0872b7c4ab6cbcf1fc4284c

Med-PaLM

A large language model from Google Research, designed for the medical domain.

Watch video

Introduction

Med-PaLM is a large language model (LLM) designed to provide high quality answers to medical questions. Our second version, Med-PaLM 2, is one of the research models that powers MedLM – a family of foundation models fine-tuned for the healthcare industry. MedLM is now available to Google Cloud customers who have been exploring a range of applications, from basic tasks to complex workflows. Med-PaLM harnesses the power of Google’s large language models, which we have aligned to the medical domain and evaluated using medical exams, medical research, and consumer queries. Our first version of Med-PaLM, preprinted in late 2022 and published in Nature in July 2023, was the first AI system to surpass the pass mark (>60%) in the U.S. Medical Licensing Examination (USMLE) style questions. Med-PaLM also generates accurate, helpful long-form answers to consumer health questions, as judged by panels of physicians and users. We introduced Med-PaLM 2 at Google Health’s annual event, The Check Up, in March 2023. Med-PaLM 2 was the first to reach human expert level on answering USMLE-style questions. According to physicians, the model’s long-form answers to consumer medical questions improved substantially.

Quick links

[Med-PaLM Nature Paper](#)

[Med-PaLM 2 Preprint](#)

[The Keyword](#)

[The Check Up](#)

[Google Cloud: Med-PaLM 2](#)

[Google Cloud: MedLM](#)

[Share](#)

[Copy link](#)

×

[play](#) [silent looping video](#) [pause](#) [silent looping video](#)

[unmute video](#) [mute video](#)

Med-PaLM 2 reached 86.5% accuracy on the MedQA medical exam benchmark in research. Since medicine is inherently multimodal, we have also introduced research on a multimodal version of Med-PaLM, called Med-PaLM M. We are also exploring a wide range of other techniques to build medical AI systems that can bring information together from a wide range of data modalities.

Medical question–answering: a grand challenge for AI

Progress in AI over the last decade has enabled it to play an increasingly important role in healthcare and medicine. Breakthroughs such as the Transformer have enabled LLMs – such as PaLM – and other large models

to scale to billions of parameters letting generative AI move beyond the limited pattern-spotting of earlier AIs and into the creation of novel expressions of content, from speech to scientific modeling. Developing AI that can answer medical questions accurately has been a long-standing challenge with several research advances over the past few decades. While the topic is broad, answering USMLE-style questions has recently emerged as a popular benchmark for evaluating medical question answering performance.

Above is an example USMLE-style question. You are presented with a vignette containing a description of the patient, symptoms, and medications. Answering the question accurately requires the reader to understand symptoms, examine findings from a patient's tests, perform complex reasoning about the likely diagnosis, and ultimately, pick the right answer for what disease, test, or treatment is most appropriate. In short, a combination of medical comprehension, knowledge retrieval, and reasoning is necessary to do well. It takes years of training for clinicians to be able to accurately and consistently answer these questions. The generation capabilities of large language models also enable them to produce long-form answers to consumer medical questions. However, ensuring model responses are accurate, safe, and helpful has been a crucial research challenge, especially in this safety-critical domain.

In a pairwise study, Med-PaLM 2 answers were preferred to physician answers across eight of nine axes considered.

Evaluating answer quality

We assessed Med-PaLM and Med-PaLM 2 against a benchmark we call 'MultiMedQA', which combines seven question answering datasets spanning professional medical exams, medical research, and consumer queries. Med-PaLM was the first AI system to obtain a passing score on USMLE-style questions from the MedQA dataset, with an accuracy of 67.6%. Med-PaLM 2 improves on this further with state of the art performance of 86.5%. Importantly, in this work we go beyond multiple-choice accuracy to measure and improve model capabilities in medical question answering. Our model's long-form answers were tested against several criteria — including scientific factuality, precision, medical consensus, reasoning, bias, and likelihood of possible harm — which were evaluated by clinicians and non-clinicians from a range of backgrounds and countries. Both Med-PaLM and Med-PaLM 2 performed encouragingly across three datasets of consumer medical questions. In a pairwise study, Med-PaLM 2 answers were preferred to physician answers across eight of nine axes considered.

Check out how Med-PaLM 2 answers medical questions

*Examples only. Med-PaLM 2 is currently being evaluated to ensure safe and responsible use.

Check out how Med-PaLM 2 answers medical questions

*Examples only. Med-PaLM 2 is currently being evaluated to ensure safe and responsible use.

Extending Med-PaLM 2 beyond language

The practice of medicine is inherently multi-modal and incorporates information from images, electronic health records, sensors, wearables, genomics and more. We believe AI systems that leverage these data at scale using self-supervised learning with careful consideration of privacy, safety and health equity will be the foundation of the next generation of medical AI systems that scale world-class healthcare to everyone. Building on the "PaLM-E" vision-language model, we designed a multimodal version of Med-PaLM, called Med-PaLM M. This system can synthesize and communicate information from images like chest X-rays, mammograms, and more to help doctors provide better patient care. Within scope are several modalities alongside language: dermatology, retina, radiology (3D and 2D), pathology, health records and genomics. We're excited to explore how this technology can benefit clinicians in the future.

play silent looping video pause silent looping video

unmute video mute video

*Example only. This image reflects early exploration of Med-PaLM M's future capabilities.

Limitations

While Med-PaLM 2 reached state-of-the-art performance on several multiple-choice medical question answering benchmarks, and our human evaluation shows answers compare favorably to physician answers across several clinically important axes, we know that more work needs to be done to ensure these models are safely and

effectively deployed. Careful consideration will need to be given to the ethical deployment of this technology including rigorous quality assessment in different clinical settings with guardrails to mitigate against risks. For example, the potential harms of using a LLM for diagnosing or treating an illness are much greater than using a LLM for information about a disease or medication. Additional research will be needed to assess LLMs used in healthcare for homogenization and amplification of biases and security vulnerabilities inherited from base models. We dive into many important areas for further research in our Med-PaLM and Med-PaLM 2 papers.

In the press

Article

AI Chatbots Can Diagnose Medical Conditions at Home. How Good Are They?

Scientific American

Article

Google's working on an updated version of its medical A.I. that can answer health questions

CNBC

Article

Google AI Performs at 'Expert' Level on U.S. Medical Licensing Exam

Med Page Today

Article

Google's AI is best yet at answering medical and health questions

New Scientist

Article

A bioethicist and a professor of medicine on regulating AI in health care

The Economist

Article

Are AI doctors on the horizon?

Advisory Board

Article

Google will let health care customers test its generative AI model, ramping up rivalry with GPT-4

STAT

Article

Google to offer limited access to medical LLM

Mobi Health News

Article

How Tech Leaders Compete In The Battle Of Healthcare AI

Forbes

Blog

MedLM: generative AI fine-tuned for the healthcare industry

Google Cloud Blog

Acknowledgements

Med-PaLM research: Karan Singhal*, Shekoofeh Azizi*, Tao Tu*, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Abubakr Babiker, Yu-Han Liu, Nathanael Schärli, Aakanksha Chowdhery, Philip Mansfield, Dina Demner-Fushman, Blaise Agüera y Arcas, Dale Webster, Greg S. Corrado, Yossi Matias, Katherine Chou, Juraj Gottweis, Nenad Tomasev, Yun Liu, Alvin Rajkomar, Joelle Barral, Christopher Semturs, Alan Karthikesalingam**, and Vivek Natarajan** Med-PaLM 2 research: Karan Singhal*, Tao Tu*, Juraj Gottweis*, Rory Sayres*, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, Mike Schaekermann, Amy Wang, Mohamed Amin, Sami Lachgar, Philip Mansfield, Sushant Prakash, Bradley Green, Ewa Dominowska, Blaise Agüera y Arcas, Nenad Tomasev, Yun Liu, Renee Wong, Christopher Semturs, S. Sara Mahdavi, Joelle Barral, Dale Webster, Greg S. Corrado, Yossi Matias, Shekoofeh Azizi**, Alan Karthikesalingam**, Vivek Natarajan** Med-PaLM M research: Tao Tu*, Shekoofeh Azizi*, Danny Driess, Mike Schaekermann, Mohamed Amin, Pi-Chuan Chang, Andrew Carroll, Chuck Lau, Ryutaro Tanno, Ira Ktena, Basil Mustafa, Aakanksha Chowdhery, Yun Liu, Simon Kornblith, David Fleet, Philip Mansfield, Sushant Prakash, Renee Wong, Sunny Virmani, Christopher Semturs, S. Sara Mahdavi, Bradley Green, Ewa Dominowska, Blaise Agüera y Arcas, Joelle Barral, Dale Webster, Greg S. Corrado, Yossi Matias, Karan Singhal, Pete Florence, Alan Karthikesalingam** and Vivek Natarajan** * - equal contributions ** - equal leadership Additional contributors: Preeti Singh, Kavita Kulkarni, Jonas Kemp, Anna Iurchenko, Lauren Winer, Will Vaughan, Le Hou, Jimmy Hu, Yuan Liu, Jonathan Krause, John Guilyard, Divya Pandya. We thank Michael Howell, Boris Babenko, Naama Hammel, Cameron Chen, Basil Mustafa, David Fleet, Douglas Eck, Simon Kornblith, Fayruz Kibria, Gordon Turner, Lisa Lehmann, Ivor Horn, Maggie Shiels, Shravya Shetty, Jukka Zitting, Evan Rappaport, Lucy Marples, Viknesh Sounderajah, Ali Connell, Jan Freyberg, Dave Steiner, Cian Hughes, Brett Hatfield, SiWai Man, Gary Parakkal, Sudhanshu Sharma, Megan Jones-Bell, Susan Thomas, Martin Ho, Sushant Prakash, Bradley Green, Ewa Dominowska, Frederick Liu, Kate Weber, Annisah Um'rani, Laura Culp, and Xuezhi Wang for their assistance, insights, and feedback during our research. We are also grateful to Yossi Matias, Karen DeSalvo, Zoubin Ghahramani, James Manyika, and Jeff Dean for their support throughout this project.

[Discover more](#)

[Blog](#)

[A responsible path to generative AI in healthcare](#)

[Blog](#)

[Our latest health AI research updates](#)

[Blog](#)

[The next generation of AI for developers and Google Workspace](#)

[blog](#)

[Google Research, 2022 & beyond: Health](#)

[Blog](#)

[Google Health + blog](#)

× < >

Output of 'what foods to avoid if you have rosacea?'

Output of CT scan question