https://doi.org/10.1038/s41591-023-02535-w

MEDICAL AI CHATBOTS: ARE THEY SAFE TO TALK TO PATIENTS?

Artificial intelligence chatbots based on large language models can pass medical exams but their diagnoses are often inaccurate. **By Paul Webster**

The arrival of artificial intelligence (AI)-powered consumer chatbots capable of producing humanlike texts is electrifying news. More than a million people signed up to use ChatGPT, the first of these large language model-based chatbots to be publicly released, within a week of its launch last November by San Francisco-based tech company OpenAI. By February, a hundred million people were estimated to be using ChatGPT monthly.

A bevy of different chatbots are now available, the most prominent of which are ChatGPT and Bard, which is marketed by Google. Both chatbots can produce computer-generated texts that display uncannily humanlike research and writing capabilities.

Expert worries

Chatbots are not just a source of amazement. They are also causing a great deal of worry. Debate is raging over what chatbots portend, including for workers whose jobs they could potentially automate. For medical practitioners and researchers, the stakes in this debate are especially high: given the duty of care that they owe to patients, allowing a role in their work to chatbots could prove to be reckless. And patients, too, now have to reckon with whether or not to trust chatbots, and whether their doctors may be using AI tools. But at a time when 300 billion Google searches are performed on medical topics annually, the use of chatbots in medicine may become a new disruptive force.

These worries are shared by many in the tech industry itself. In March, more than 1,000 tech industry leaders issued an open letter calling for a pause in AI development. Soon afterward,

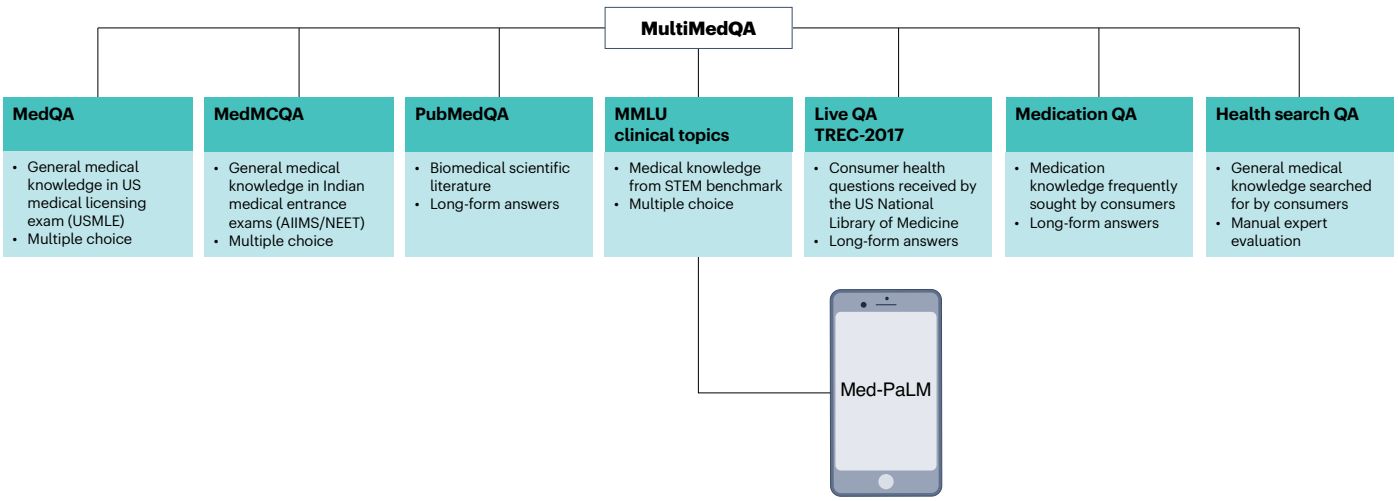

Fig. 1 | MedPaLM is trained on MultiMedQA, a corpus of medical multiple choice questions, including medical exams, long-form answers and manual evaluations by experts. Source: Singhal, K. et al. *Nature* 620, 172–180 (2023). AIIMS, All India Institute of Medical Sciences; MMLU, Massive Multitask Language Understanding; NEET, National Eligibility Cum Entrance Test; QA, quality assurance, STEM, science, technology, engineering, and mathematics; USMLE, US Medical Licensing Examination.

leaders from the Association for the Advancement of Artificial Intelligence issued an even more strident warning. Even the so-called 'godfather' of AI, Canadian researcher Geoffrey Hinton, now says he is worried that systems such as ChatGPT may soon outsmart, out-talk and outwrite us.

Approval needed

Until just a few months ago, medical chatbots constituted a niche area within AI research. When Google held a press conference last March regarding Med-PaLM, their new app with specialized abilities to answer medical questions, just three reporters attended – all from specialist health and science periodicals, including *Nature Medicine*. Even so, interest in AI-assisted generative chatbots that generate intellectually original analyses is now growing rapidly within the $140-billion global healthcare IT industry, and a plethora of uses for them are being pioneered. These range from basic clinical notetaking, to assisting with numerous types of diagnostics, to the generation of synthetic health data for medical imaging processes and research purposes.

Officials at the US Food and Drug Administration (FDA) are trying to keep up with monitoring and regulating the burgeoning medical use of AI devices, including chatbots. In a 26-page guidance released in September 2022, they stated that "software functions" that support or provide clinical recommendations to patients or caregivers (as opposed to licensed healthcare providers) meet the definition of medical devices requiring FDA review and approval. These products can sidestep FDA review only if humans fully preside over the software function. Health care providers must "independently review the basis for the recommendations presented by the software" so that they do not rely primarily on AI-based recommendations, but rather on their own judgments, to make clinical decisions.

In short, anytime a medically trained chatbot or other AI-assisted device is intended to operate independently from licensed clinicians, FDA review and approval is necessary.

The FDA's September 2022 guidance predated the public release of ChatGPT and its rivals, but the development of medical chatbots was already highly advanced. Speaking to reporters last March, Vivek Nataranjan, a chatbot researcher at Google, says that a combination of "very strong language models" and "very deep medical domain expertise" led his company to develop a sophisticated chatbot specifically tailored for medical use.

Nataranjan calls Google's product, which is known as Med-PaLM (PaLM stands for Pathways Language Model), a "leapfrog achievement." Med-PaLM easily aces US medical licensing exams, says Nataranjan. Humans who score above 60% on these exams are usually given a pass—Google's first version of Med-PaLM scored 67%. Its latest version, Med-PALM 2, scored 85, a level Google described as that of an "expert" doctor. Potential applications of large language models of this sort in medicine, according to Google, include knowledge retrieval, clinical decision support, summarization of key findings and triaging of patients' primary care concerns.

Thanks to constant fine-tuning by teams of Google technicians, Natarajan explained, Med-PaLM commands medical knowledge that is increasingly sensitive to what he calls the "nuances of the medical terrain." For the engineers building medically relevant models, he adds, the right training dataset of medical information is needed. "When you are training these models," he says, "you need to teach the model where to look for answers."

Missing medical data

Med-PaLM was trained by Google on MultiMedQA, a combination of seven standardized medical datasets that include vast quantities of medical questions and answers (Fig. 1). This includes Google's HealthSearchQA database, which comprises 3,375 commonly searched consumer medical questions. Overall, the PaLM training corpus consists of 780 billion 'tokens' representing a mixture of webpages, Wikipedia articles, source code, social media conversations, news articles and books.

Even so, PaLM's knowledge base has limits, acknowledged Alan Karthikesalingam, Research Lead for Google Health. Even the most sophisticated internet mining techniques cannot access peer-reviewed medical literature, much of which is paywalled by the world's publishers, (including Springer Nature, publisher of *Nature Medicine*). "Google utilizes publicly available data on the

open internet," explains Karthikesalingam. Few medical journals are fully open access, which leaves a lot of health research out of bounds to Med-PaLM.

Asked how Google plans to access the large body of important medical literature not available on the open internet for their AI chatbots, Karthikesalingam said he "did not have a view" on the matter. "We're very humble about research in this domain," he demurred. This may explain why Med-PaLM is not yet ready for widespread use, although in April Google announced plans to make it available to a select group of customers for limited testing. "A system may sound plausible but have very subtle [knowledge] gaps," Karthikesalingam acknowledges.

A major pitfall of relying on medical AI to diagnose a patient is that the evidence base for the diagnosis will not be made available. The FDA's guidance document warned that "contraindications or patient-specific warnings may be missed." For AI to be more widely used in health, any diagnosis made by a large language model would need to include the scientific rationale for that diagnosis, most likely including citations of research articles, the FDA cautioned.

That is easier said than done, says Alex Zhavoronkov, founder and CEO of Hong Kong–based Insilico Medicine, which specializes in harnessing artificial intelligence technologies for drug discovery and biomarker development. "The training materials for these systems must include ultra-high-quality, peer-reviewed full-text publications, which is not currently the case," warns Zhavoronkov.

Without comprehensive access to high-quality published science, AI chatbots cannot make accurate medical diagnoses, says Zhavoronkov. ChatGPT, he notes, is trained on a mix of texts and sources scraped from the internet that, in his view, "needs to be policed" by human reviewers. To achieve success with large language models in healthcare settings, he argues, "you will need to have high-quality human editors. And ultimately, you need access to the latest and highest-quality peer-reviewed scientific journals. So, the real winners in all of this could be the owners of those journals."

Preserving patient safety

Benjamin Tolchin, a neurologist at the Yale School of Medicine and the inaugural director of the Yale New Haven Health System Center for Clinical Ethics, also warns that the road ahead for medical chatbots may prove to be a long and halting one, despite the initial wave of excitement unleashed by ChatGPT. "I've used ChatGPT and was very struck by how eloquent it was in its communication, and in how detailed it was," Tolchin explains. "It's head and shoulders above any AI or chatbot technology I've ever seen before, and it shows real promise for clinical applications. But we have to ask, 'What will happen when patients and clinicians started asking it for guidance?'"

To answer that question, Tolchin recently posed a series of clinical questions to ChatGPT by describing patients and asking for diagnoses. "It gave responses on the level of a well-read medical student who was somewhat clinically oblivious," Tolchin recalls. "It knows just enough to be dangerous."

To preserve patient safety, Tolchin recommends medical governance frameworks for large language models in medicine, centered on informed consent. This should be mandated when these tools are used by clinicians, combined with careful clinical supervision of their usage, he argues. Tolchin also argues that the tools must include citations, and that clinicians and AI scientists need to collaborate closely as these tools are elaborated. "We are on the verge of a paradigm shift in healthcare," Tolchin reflects, "but I don't think we're there yet."

OpenAI, the maker of ChatGPT, also calls for caution. In an email, the company informed *Nature Medicine* that its models should not be used for medical diagnostics, to triage or to manage life-threatening issues. In their 2023 paper describing MultiMedQA, OpenAI researchers acknowledged "the potential harms of using an LLM for diagnosing or treating an illness".

Concerns about the factual inaccuracy of large language models have led some to call for control measures, OpenAI researchers acknowledge, which could include requiring AI providers to impose usage restrictions on large language models, identification of AI-generated content and requirements for "proof of personhood" before publication, along with developing "digital provenance standards that are widely adopted".

Joëlle Pineau, one of three managing directors for Meta Inc.'s Foundational Artificial Intelligence Research unit and a professor at McGill University in Montreal, says, "AI is just a tool". Because large language models such as ChatGPT draw on data containing little truly clinically reliable science, Pineau emphasizes, their clinical utility will remain limited until their training datasets encompass the full scope of scientific knowledge available to researchers who have access to copyrighted materials. "I don't have high confidence it's doable today," Pineau says.

Generative AI systems also need to contend with the delicate matter of clinical uncertainty. "We need a way to represent doubt," says Pineau. Until that's found, in Pineau's assessment, doubt itself will dominate the prospects for medical chatbots.

Paul Webster
Science journalist, Buenos Aires, Argentina, Toronta, Canada.

Published online: 08 September 2023