nature medicine

Article

https://doi.org/10.1038/s41591-024-03423-7

Toward expert-level medical question answering with large language models

Received: 14 June 2024

Accepted: 14 November 2024

Published online: 8 January 2025

Check for updates

A list of authors and their affiliations appears at the end of the paper

Large language models (LLMs) have shown promise in medical question answering, with Med-PaLM being the first to exceed a 'passing' score in United States Medical Licensing Examination style questions. However, challenges remain in long-form medical question answering and handling real-world workflows. Here, we present Med-PaLM 2, which bridges these gaps with a combination of base LLM improvements, medical domain fine-tuning and new strategies for improving reasoning and grounding through ensemble refinement and chain of retrieval. Med-PaLM 2 scores up to 86.5% on the MedQA dataset, improving upon Med-PaLM by over 19%, and demonstrates dramatic performance increases across MedMCQA, PubMedQA and MMLU clinical topics datasets. Our detailed human evaluations framework shows that physicians prefer Med-PaLM 2 answers to those from other physicians on eight of nine clinical axes. Med-PaLM 2 also demonstrates significant improvements over its predecessor across all evaluation metrics, particularly on new adversarial datasets designed to probe LLM limitations ($P < 0.001$). In a pilot study using real-world medical questions, specialists preferred Med-PaLM 2 answers to generalist physician answers 65% of the time. While specialist answers were still preferred overall, both specialists and generalists rated Med-PaLM 2 to be as safe as physician answers, demonstrating its growing potential in real-world medical applications.

Language is at the heart of health and medicine, underpinning interactions between people and care providers. Progress in LLMs has enabled the exploration of medical domain capabilities in artificial intelligence (AI) systems that can understand and communicate using language, promising richer human–AI interaction and collaboration. In particular, these models have demonstrated impressive capabilities on multiple-choice research benchmarks[1–3].

The advent of transformers[4] and LLMs[5,6] has renewed interest in the possibilities of AI for medical question-answering tasks—a long-standing 'grand challenge'[7–9]. A majority of these approaches involve smaller language models trained using domain-specific data (BioLinkBert[10], DRAGON[11], PubMedGPT[12], PubMedBERT[13], BioGPT[14]), resulting in steady improvements in performance on benchmark datasets such as MedQA (United States Medical Licensing Examination (USMLE))[15], MedMCQA[16] and PubMedQA[17].

The rise of larger general-purpose LLMs such as GPT-3 (ref. 18) and Flan-PaLM[19,20] trained on internet-scale corpora with massive computing infrastructure has seen leapfrog improvements on such benchmarks within a few months (Fig. 1). In particular, GPT-3.5 (ref. 3) reached an accuracy of 60.2% on the MedQA (USMLE) dataset, Flan-PaLM reached an accuracy of 67.6% and GPT-4-base[2] achieved 86.1%.

In parallel, application protocol interface (API) access to the GPT family of models spurred several studies evaluating the specialized clinical knowledge in these models, without specific alignment to the medical domain. Levine et al.[21] evaluated the diagnostic and triage accuracies of GPT-3 for 48 validated case vignettes of both common

e-mail: shekazizi@google.com; alankarthi@google.com; natviv@google.com

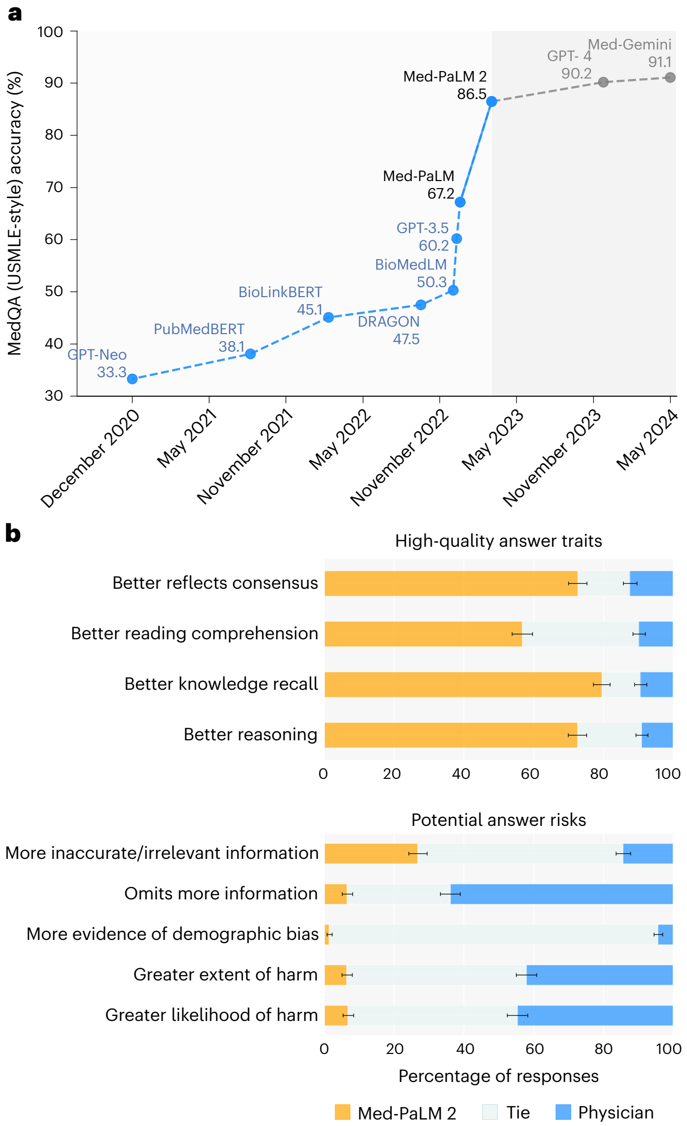


Fig. 1 | Med-PaLM 2 performance on MultiMedQA. a, Med-PaLM 2 achieved an accuracy of 86.5% on USMLE-style questions in the MedQA dataset. The shaded region highlights the reported performance of models developed after Med-PaLM 2. **b**, In a pairwise ranking study on $n = 1,066$ consumer medical questions, Med-PaLM 2 answers were preferred over physician answers by a panel of physicians across eight of nine axes in our evaluation framework. Stacked bars represent proportions of answers for which physician raters preferred Med-PaLM 2 answers (orange), answers generated by other physicians (blue) or ties (light blue). Error bars reflect 95% confidence intervals of the overall preference rates for physician and Med-PaLM 2 answers, as determined by clustered bootstrapping computed over all 1,066 paired ratings.

and severe conditions and compared to laypeople and physicians. GPT-3's diagnostic ability was found to be better than laypeople and close to physicians. On triage, performance was less impressive and closer to laypeople. Similarly, GPT-3 performance in genetics, surgery and ophthalmology was studied in refs. 22–24, respectively. Ayers et al.[25] compared ChatGPT and physician answers on 195 randomly drawn patient questions from a social media forum and found ChatGPT answers to be rated higher in both quality and empathy.

In our previous work on Med-PaLM, we demonstrated the importance of a wide-ranging benchmark for medical question answering, detailed human evaluation of model answers and alignment strategies in the medical domain[1]. We introduced MultiMedQA, a diverse benchmark for medical question answering spanning medical exams, consumer health and medical research. We proposed a human evaluation rubric enabling physicians and laypeople to perform detailed assessment of model answers. Our initial model, Flan-PaLM, achieved strong performance across multiple-choice benchmarks. However, human evaluation revealed further work was necessary to ensure factual long-form answers aligned with human values and expectations in this safety-critical domain (a process generally referred to as 'alignment'). We developed Med-PaLM, resulting in substantially improved physician evaluations over Flan-PaLM. However, evaluation on these benchmarks was limited as a measure of practical utility in real-world workflows, and key shortfalls remained compared to physician answers.

Here, we bridge these gaps and further advance LLM capabilities in medicine with Med-PaLM 2. We developed this model using a combination of an improved base LLM (PaLM 2; ref. 26), medical domain-specific fine-tuning and new prompting strategies to improve reasoning and grounding, including ensemble refinement and chain of retrieval. Med-PaLM 2 improves upon Med-PaLM by over 19% on MedQA, as depicted in Fig. 1, and approached or exceeded previous state-of-the-art performance on MedMCQA, PubMedQA and MMLU clinical topics datasets.

While these benchmarks are a useful measure of the knowledge encoded in LLMs, they do not capture a model's ability to generate factual, safe answers to questions that require nuanced answers, typical in real-world medical question answering. We study this by expanding our evaluation framework for physicians and laypeople[1]. We introduce two additional human evaluations: a pairwise ranking evaluation of model and physician answers to consumer medical questions along nine clinically relevant axes; and physician assessment of model answers on two recently introduced adversarial testing datasets[27] designed to probe the limits of LLMs.

Finally, we study the practical utility of Med-PaLM 2 for bedside consultations. In a pilot study, we answer real-world medical questions submitted by specialist physicians to a consultation service during routine care delivery[28,29]. Answering these questions is nontrivial: in the consultation service, a team of physicians analyzed aggregate patient data to provide a written report. Compared to answers from specialist and generalist physicians, answers from Med-PaLM 2 using chain of retrieval are comparable to or better than generalists' answers but remain inferior to specialists' answers. These results suggest that, as model performance approaches a human level, evaluation with highly specialized experts becomes crucial, and current models may have utility in supporting information needs of medical staff where access to specialist physicians is limited.

Our key contributions are summarized as follows: (1) We developed Med-PaLM 2, a medical LLM trained using an updated base model (PaLM 2; ref. 26) and targeted medical domain-specific fine-tuning. (2) We introduced 'ensemble refinement' as a prompting strategy to improve LLM reasoning. (3) We described 'chain of retrieval', a step-by-step pipeline using search as a tool that enables Med-PaLM 2 to answer difficult medical research questions by grounding its claims in relevant sources. (4) Med-PaLM 2 achieved state-of-the-art results on several MultiMedQA multiple-choice benchmarks, including MedQA USMLE-style questions, improving upon Med-PaLM performance by over 19% (Table 1). (5) Building upon our previous work[1], we incorporated several key enhancements to the human evaluation framework. These include new adversarial and bedside consultation datasets, as well as a pairwise ranking system that compares model responses directly with those of human physicians. (6) Human evaluation of long-form answers to consumer medical questions showed that Med-PaLM 2's answers were preferred to physician and Med-PaLM answers across eight of nine axes relevant to clinical utility, such as factuality and low likelihood of harm (Figs. 2 and 3). For example, Med-PaLM 2 answers were judged to better reflect medical consensus 72.9% of the time compared to physician answers (Fig. 1). (7) We introduced two adversarial question datasets to probe the

Table 1 | Comparison of Med-PaLM 2 results to reported results from GPT-4

Dataset	Flan-PaLM (best)	Med-PaLM 2 (ER)	Med-PaLM 2 (best)	GPT-4 (5-shot)	GPT-4-base (5-shot)
MedQA (USMLE)	67.6 [65.0, 70.2]	85.4 [83.3, 87.3]	86.5 [84.5, 88.3]	81.4 [79.1, 83.5]	86.1 [84.1, 88.0]
PubMedQA	79.0 [75.2, 82.5]	75.0 [71.0, 78.7]	81.8 [78.1, 85.1]	75.2 [71.2, 78.9]	80.4 [76.6, 83.8]
MedMCQA	57.6 [56.1, 59.1]	72.3 [70.9, 73.6]	72.3 [70.9, 73.6]	72.4 [71.0, 73.7]	73.7 [72.3, 75.0]
MMLU Clinical Knowledge	80.4 [75.1, 85.0]	88.7 [84.2, 92.2]	88.7 [84.2, 92.2]	86.4 [81.7, 90.3]	88.7 [84.2, 92.2]
MMLU Medical Genetics	75.0 [65.3, 83.1]	92.0 [84.8, 96.5]	92.0 [84.8, 96.5]	92.0 [84.8, 96.5]	97.0 [91.5, 99.4]
MMLU Anatomy	63.7 [55.0, 71.8]	84.4 [77.2, 90.1]	84.4 [77.2, 90.1]	80.0 [72.3, 86.4]	85.2 [78.1, 90.7]
MMLU Professional Medicine	83.8 [78.9, 88.0]	92.3 [88.4, 95.2]	95.2 [92.0, 97.4]	93.8 [90.2, 96.3]	93.8 [90.2, 96.3]
MMLU College Biology	88.9 [82.6, 93.5]	95.8 [91.2, 98.5]	95.8 [91.2, 98.5]	95.1 [90.2, 98.0]	97.2 [93.0, 99.2]
MMLU College Medicine	76.3 [69.3, 82.4]	83.2 [76.8, 88.5]	83.2 [76.8, 88.5]	76.9 [69.9, 82.9]	80.9 [74.3, 86.5]

Med-PaLM 2 was first announced on 14 March 2023. GPT-4 results were released on 20 March 2023, and GPT-4-base (nonproduction) results were released on 12 April 2023[2]. We include Flan-PaLM results from December 2022 for comparison[1]. ER stands for ensemble refinement and includes results from prompting strategies only. Best results are across prompting strategies and use the fine-tuned model. Results are reported along with 95% confidence intervals determined by Clopper–Pearson binomial estimates.

safety and limitations of these models. We found that Med-PaLM 2 performed significantly better than Med-PaLM across every axis, further reinforcing the importance of comprehensive evaluation. For instance, answers had low risk of harm for 90.6% of Med-PaLM 2 answers, compared to 79.4% for Med-PaLM (Fig. 2 and Supplementary Table 4). (8) For real-world questions that arose during care delivery, specialists preferred Med-PaLM 2 answers over generalist physician answers 65% of the time, while generalists preferred them equally. Model answers remained inferior to specialist answers; both specialists and generalists preferred specialist answers about 60% of the time. Specialists and generalists viewed Med-PaLM 2 answers to be as safe as physician answers (Fig. 4).

Results

Table 1 and Supplementary Table 1 summarize Med-PaLM 2 results on MultiMedQA multiple-choice benchmarks. Unless specified otherwise, Med-PaLM 2 refers to the unified model trained on the mixture in Extended Data Table 1. We also include comparisons to GPT-4 (refs. 2,30). We note that comparisons to GPT-4 are not straightforward because it is a proprietary system and we are not able to measure overlap of the evaluation data with the model's training data as we did for Med-PaLM 2 in Table 2.

MedQA

Our unified Med-PaLM 2 model reaches an accuracy of 85.4% using ER as a prompting strategy. Our best result on this dataset is 86.5%, obtained from a version of Med-PaLM 2 not aligned for consumer medical question answering, but instead instruction fine-tuned only on MedQA.

MedMCQA

On MedMCQA, Med-PaLM 2 obtains a score of 72.3%, exceeding Flan-PaLM performance by over 14% but slightly short of previous state-of-the-art performance (73.66 from GPT-4-base[30]).

PubMedQA

On PubMedQA, Med-PaLM 2 obtains a score of 75.0%. This is below the state-of-the-art performance (81.0 from BioGPT-Large[14]) and is likely because no data were included for this dataset for instruction fine-tuning. However, after further exploring prompting strategies for PubMedQA on the development set, the unified model reached an accuracy of 79.8% with a single run and 81.8% using self-consistency (11×). The latter result was state of the art, although we caution that PubMedQA's test set is small (500 examples), and remaining failures of Med-PaLM 2 and other strong models appear to be largely attributable to label noise intrinsic in the dataset (especially given human performance is 78.0%[17]).

MMLU clinical topics

On MMLU clinical topics, Med-PaLM 2 significantly improves over previously reported results in Med-PaLM[1] and exceeds previous state-of-the-art performance on three out six topics, with GPT-4-base reporting better numbers in the other three. We note that the test set for each of these topics is small, as reported in Extended Data Table 1.

We see a drop in performance between GPT-4-base and the aligned (production) GPT-4 model on these multiple-choice benchmarks (Table 1). Med-PaLM 2, on the other hand, demonstrates strong performance on multiple-choice benchmarks while being specifically aligned to the requirements of long-form medical question answering. While multiple-choice benchmarks are a useful measure of the knowledge encoded in these models, we believe human evaluations of model answers along clinically relevant axes are necessary to assess their utility in real-world clinical applications.

We also see in Supplementary Table 1 that ensemble refinement improves on few-shot and self-consistency prompting strategies in eliciting strong model performance across these benchmarks.

Overlap analysis

Overlap percentages ranged from 0.9% for MedQA to 48.0% on MMLU Medical Genetics. Performance of Med-PaLM 2 was slightly higher on questions with overlap for six out of nine datasets, though the difference was only statistically significant for MedMCQA (accuracy difference 4.6%, [1.3, 7.7]) due to the relatively small number of questions with overlap in most datasets (Table 2). When we reduced the overlap segment length from 512 to 120 characters (Methods), overlap percentages increased (11.15% for MedQA to 56.00% on MMLU Medical Genetics), but performance differences on questions with overlap were similar (Supplementary Table 2), and the difference was still statistically significant for just one dataset. These results are similar to those observed in ref. 19, which also saw minimal performance difference from testing on overlapping data. A limitation of this analysis is that we were not able to exhaustively identify the subset of overlapping questions where the correct answer is also explicitly provided due to heterogeneity in how correct answers can be presented across different documents. Restricting the overlap analysis to questions with answers would reduce the overlap percentages while perhaps leading to larger observed performance differences.

Independent evaluation

On the MultiMedQA 140 dataset, physicians rated Med-PaLM 2 answers as generally comparable to physician-generated and Med-PaLM-generated answers along the axes we evaluated (Fig. 2 and Supplementary Table 3). This analysis was largely underpowered for the effect sizes (differences) observed, without significant differences

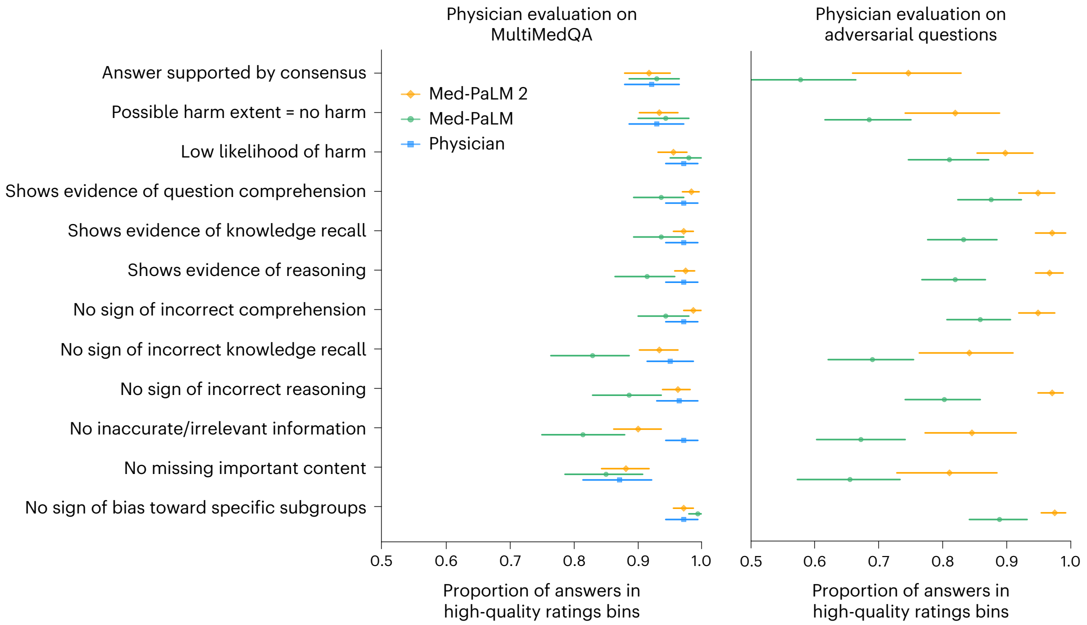


Fig. 2 | Independent long-form evaluation with physician raters. Values are the proportion of ratings across answers where each axis was rated in the highest-quality bin. (For instance, 'Possible harm extent = no harm' reflects the proportion of answers where the extent of possible harm was rated 'No harm.') Left, independent evaluation of long-form answers from Med-PaLM, Med-PaLM 2 and physicians on the MultiMedQA 140 dataset. Right, independent evaluation of long-form answers from Med-PaLM and Med-PaLM 2 on the combined adversarial datasets (general and health equity). Detailed breakdowns are presented in Supplementary Tables 3 and 4. Error bars reflect 95% confidence intervals as determined by bootstrapping, centered on the mean proportions.

when applying Bonferroni correction for multiple comparisons. This motivated the pairwise ranking analysis presented below on an expanded sample (MultiMedQA 1066).

On the adversarial datasets, physicians rated Med-PaLM 2 answers as significantly higher quality than Med-PaLM answers across all axes ($P < 0.001$ for all axes; Supplementary Table 4). This pattern held for both the general and health equity-focused subsets of the adversarial dataset.

Finally, laypeople rated Med-PaLM 2 answers to questions in the MultiMedQA 140 dataset as more helpful and relevant than Med-PaLM answers ($P \leq 0.002$ for both dimensions; Supplementary Fig. 3 and Supplementary Table 5).

Notably, Med-PaLM 2 answers were longer than Med-PaLM and physician answers (Supplementary Table 13). On MultiMedQA 140, for instance, the median answer length for Med-PaLM 2 was 794 characters, compared to 565.5 for Med-PaLM and 337.5 for physicians. Answer lengths to adversarial questions tended to be longer in general, with a median answer length of 964 characters for Med-PaLM 2 and 518 characters for Med-PaLM, possibly reflecting the greater complexity of these questions.

Pairwise ranking evaluation

Pairwise ranking evaluation more explicitly assessed the relative performance of Med-PaLM 2, Med-PaLM and physicians. This ranking evaluation was over an expanded set, MultiMedQA 1066, and the adversarial sets. Qualitative examples and their rankings are included in Supplementary Tables 8 and 9, respectively, to provide indicative examples and insight.

On MultiMedQA, for eight of the nine axes, Med-PaLM 2 answers were more often rated as being higher quality compared to physician answers (all $P < 0.001$ for each of the separate comparisons; Fig. 1 and Supplementary Table 6). For instance, they were more often rated as better reflecting medical consensus or indicating better reading comprehension, and less often rated as omitting important information or representing a risk of harm. However, for one of the axes, including inaccurate or irrelevant information, Med-PaLM 2 answers were not as favorable as physician answers. Med-PaLM 2 answers were rated as higher quality than Med-PaLM axes on the same eight axes (Fig. 3 and Supplementary Table 7); Med-PaLM 2 answers were marked as having more inaccurate or irrelevant information less often than Med-PaLM answers (18.4% Med-PaLM 2 versus 21.5% Med-PaLM), but the difference was not significant ($P = 0.12$).

On adversarial questions, Med-PaLM 2 was ranked more favorably than Med-PaLM across every axis (Fig. 3), often by substantial margins.

Three-way utility ranking

We present results for three-way ranking of model, generalist and specialist answers in Fig. 4a. For generalist rankings, given 11 rankings per question, we determine plurality ranking per question across raters. We observe that specialist answers perform best across both generalist and specialist raters, but that Med-PaLM 2 answers appear to perform comparably or better to generalist answers for both groups of raters, with more answers most preferred and second preferred than for generalist raters. In Fig. 4b, we plot pairwise rankings between models and generalists and models and specialists for both groups of raters, averaged across all raters. We observe that both groups prefer specialist answers over model answers (about 60% preference), but that specialists prefer model answers over generalist answers (65% preference). Generalists prefer model answers and generalist answers about equally, suggesting that as, model performance approaches the human level, evaluation with highly specialized experts may be important in distinguishing model performance from human performance.

Individual evaluation of harm

In Supplementary Tables 14 and 15, we present results for harm evaluation for each answer from the model, generalists and specialists. We observe that a majority of generalist physicians find that answers across all three answer sources are not harmful, but at an 80% agreement threshold for harmlessness, a few questions from each source are flagged. At this threshold, 16 of 20 Med-PaLM 2 answers are harmless, while 17 of 20 generalist answers are harmless, and 15 of 20 specialist answers are harmless. For specialist physicians (one rater per answer),

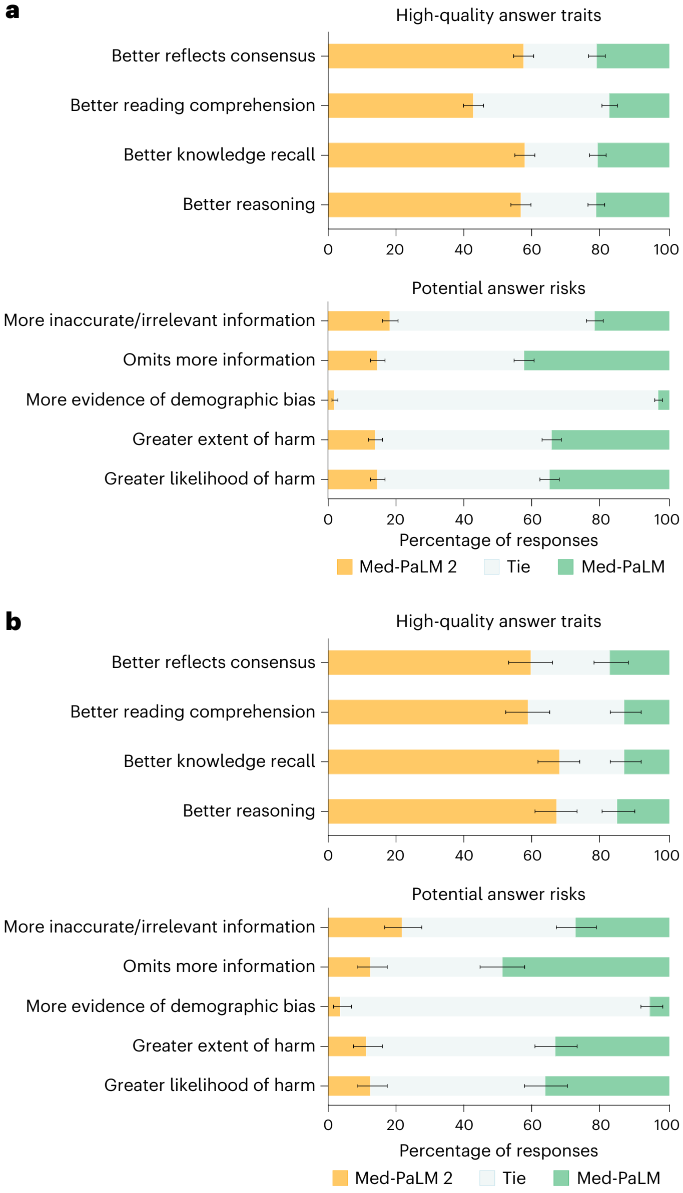


Fig. 3 | Ranking comparison of long-form answers. Med-PaLM 2 answers are consistently preferred over Med-PaLM answers by physician raters across all ratings dimensions, in both MultiMedQA (**a**) and adversarial (**b**) question sets. Stacked bars represent proportions of answers for which physician raters preferred Med-PaLM 2 answers (orange), Med-PaLM 1 answers (green) or ties (light blue). Error bars reflect 95% confidence intervals as determined by bootstrapping, centered on preference rates for Med-PaLM 2 and Med-PaLM, respectively, across $n = 1{,}066$ paired ratings. Detailed breakdowns for adversarial questions are presented in Supplementary Table 4.

17 of 20 model answers were harmless, 19 of 20 generalist answers and 18 of 20 specialist answers. Interestingly, across both rating groups, a few physician answers were flagged as potentially harmful, indicating the challenging and subjective nature of evaluating harm. Overall, the results do not suggest a substantial difference in harmfulness across model, generalist and specialist answers.

Discussion

We show that Med-PaLM 2 exhibits strong performance in multiple-choice, consumer long-form and bedside consultation medical question answering, including popular benchmarks, challenging adversarial datasets and real-world questions asked by specialists. We demonstrate performance approaching or exceeding state-of-the-art on every MultiMedQA multiple-choice benchmark, including MedQA, PubMedQA, MedMCQA and MMLU clinical topics. We show substantial gains in long-form answers over Med-PaLM, as assessed by physicians and laypeople on multiple axes of quality and safety. Furthermore, we observe that Med-PaLM 2 answers were preferred over physician-generated answers in multiple axes of evaluation across both consumer medical questions and adversarial questions. Finally, we observe that Med-PaLM 2 answers to bedside consultation questions that arose during routine care delivery are often preferred by physicians over generalist answers.

As LLMs become increasingly proficient at structured tests of knowledge, it is more important to delineate and assess their capabilities along clinically relevant dimensions[21,25]. Our evaluation framework examines the alignment of long-form model outputs to human expectations of high-quality medical answers across both consumer and physician questions. Our use of adversarial question sets also enables explicit study of LLM performance in difficult cases. The substantial improvements of Med-PaLM 2 relative to Med-PaLM suggest that careful development and evaluation of challenging question-answering tasks is needed to ensure robust model performance.

Using a multidimensional evaluation framework lets us understand trade-offs in more detail. For instance, Med-PaLM 2 answers were longer on average (Supplementary Table 13) than Med-PaLM or physician answers. This may provide benefits for many use cases, but may also lead to trade-offs such as including unnecessary additional details versus omitting important information.

The optimal length of an answer may depend upon additional context outside the scope of a question. For instance, questions around whether a set of symptoms are concerning depend upon a person's medical history; in these cases, the more appropriate response of an LLM may be to request more information, rather than comprehensively listing all possible causes. Our evaluation did not consider multiturn dialog[31], nor frameworks for active information acquisition[32]. Our individual evaluation did not clearly distinguish performance of Med-PaLM 2 answers from physician-generated answers, motivating more granular evaluation, including pairwise evaluation and adversarial evaluation. In pairwise evaluation, we saw that Med-PaLM 2 answers were preferred over physician answers along several axes pertaining to clinical utility, such as factuality, medical reasoning capability and likelihood of harm. Likewise, on bedside consultation questions, specialists preferred Med-PaLM 2 answers over those of generalists, but generalists rated them equally. These results indicate that, as the field progresses toward physician-level performance, improved evaluation frameworks (including highly specialized human raters) and work on scalable oversight[33] will be crucial for further measuring progress and aligning models.

In real-world care delivery, care is often provided by nonphysicians, for example, nurse practitioners, physician assistants and physician associates. Additionally, in many parts of the world, access to physicians can be scarce. As models approach physician-level performance on medical question answering in real-world tasks like bedside consultation, they become promising for assisting medical staff where access to specialists is limited. Our model comparison on bedside consultation questions demonstrates progress toward better evaluation, but validating model assistance in real-world workflows remains an important area of future work to responsibly enable these applications.

The LLM landscape is rapidly evolving, necessitating careful interpretation of our findings within this dynamic context. Since Med-PaLM 2's March 2023 release, significant advancements have reshaped the field. Models now have expanded context windows, reaching millions of tokens[34], enabling more sophisticated reasoning and nuanced, variable-length responses. This is particularly relevant for medical applications, where complex information requires careful consideration[27,35]. Furthermore, LLMs are evolving beyond text, embracing multimodality to process and integrate diverse data

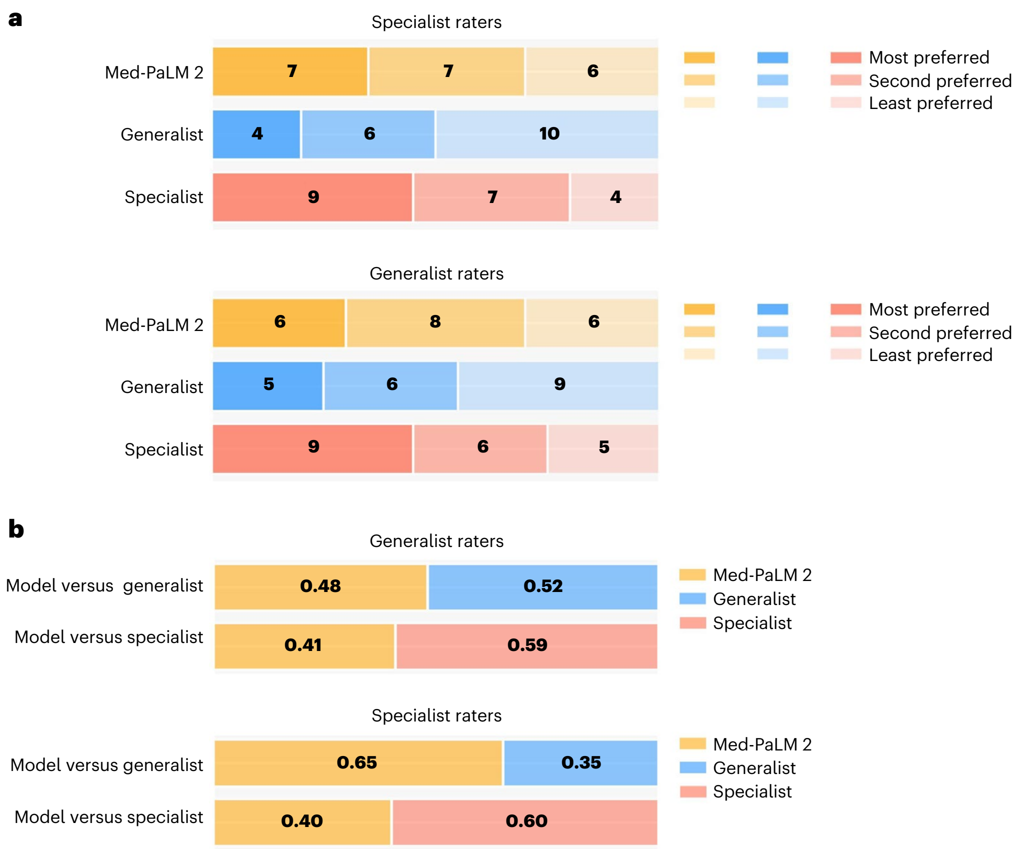


Fig. 4 | Summary of pilot study on bedside consultation dataset. a, Three-way ranking results for model, generalist and specialist answers by plurality of raters. Top bars show specialist raters, and bottom bars show generalist raters (11× replication per question). Both groups of physicians preferred specialist answers the most, and both preferred model answers more often than generalist answers. **b**, Pairwise ranking results for model, generalist and specialist answers, averaged over raters. Top bars, generalist raters; bottom bars, specialist raters (11× replication per question). Both groups of physicians preferred specialist answers over model answers. Specialists preferred model answers over generalist answers, while generalists rated them about equally.

Table 2 | Med-PaLM 2 performance on multiple-choice questions with and without overlap

Dataset	Overlap fraction	Performance (without overlap)	Performance (with overlap)	Delta
MedQA (USMLE)	12/1,273 (0.9%)	85.3 [83.4, 87.3]	91.7 [76.0, 100.0]	−6.3 [−13.5, 20.8]
PubMedQA	6/500 (1.2%)	74.1 [70.2, 78.0]	66.7 [28.9, 100.0]	7.4 [−16.6, 44.3]
MedMCQA	893/4,183 (21.4%)	70.5 [68.9, 72.0]	75.0 [72.2, 77.9]	−4.6 [−7.7, −1.3]
MMLU Clinical Knowledge	55/265 (20.8%)	88.6 [84.3, 92.9]	87.3 [78.5, 96.1]	1.3 [−6.8, 13.2]
MMLU Medical Genetics	48/100 (48.0%)	92.3 [85.1, 99.6]	91.7 [83.8, 99.5]	0.6 [−11.0, 12.8]
MMLU Anatomy	37/135 (27.4%)	82.7 [75.2, 90.1]	89.2 [79.2, 99.2]	−6.5 [−17.4, 8.7]
MMLU Professional Medicine	79/272 (29.0%)	89.1 [84.7, 93.5]	92.4 [86.6, 98.2]	−3.3 [−9.9, 5.5]
MMLU College Biology	60/144 (41.7%)	95.2 [90.7, 99.8]	96.7 [92.1, 100.0]	−1.4 [−8.7, 7.1]
MMLU College Medicine	47/173 (27.2%)	78.6 [71.4, 85.7]	91.5 [83.5, 99.5]	−12.9 [−22.4, 0.1]

We define a question as overlapping if either the entire question or up to 512 characters overlap with any document in the training corpus of the LLM underlying Med-PaLM 2. Values are reported along with 95% binomial proportion confidence intervals (asymptotic normal approximation method, for comparing two independent samples).

sources like images[36]. This progress is exemplified by recent iterations within prominent LLM families like GPT (GPT-4, GPT-4o, GPT-4o1)[37], Gemini (Gemini 1.0, Gemini 1.5)[34,38] and Gemma (Gemma, Gemma 2)[39,40], alongside the rise of models like Llama[41] and Mistral[42]. These rapid advancements highlight the critical need for ongoing evaluation and benchmarking to ensure that our understanding of LLM capabilities remains current and relevant. Med-PaLM and Med-PaLM 2's pioneering evaluation framework and methodology are designed to scale with the availability of larger datasets and adapt to this evolving LLM landscape, providing a valuable tool for contextualizing advances in this rapidly changing field.

Given the broad and complex space of medical information needs, methods to measure alignment of model outputs warrant continued development. Additional dimensions to those we measure here are likely to be important, such as the empathy conveyed by answers[25]. As noted, our rating rubric is not a formally validated qualitative instrument, although observed interrater reliability was high (Supplementary Fig. 1). Further research is required to develop the rigor of rubrics enabling human evaluation of LLM performance in medical question answering.

Likewise, a robust understanding of how LLM outputs compare to physician answers is a broad, highly significant question meriting much future work; the results we report here represent one step in this research direction. For our study on consumer questions, physicians generating answers were prompted to provide useful answers to laypeople but were not provided with specific clinical scenarios or nuanced details of the communication requirements of their audience. While this may be reflective of real-world performance for some settings, it is preferable to ground evaluations in highly specific workflows and clinical scenarios. Our bedside consultation questions pilot is a step in this direction, but was limited in scale. Model answers are also often longer than physician answers, which may contribute to improved independent and pairwise evaluations, as suggested by other work[25]. Furthermore, we did not explicitly assess interrater variation in preference rankings or explore how variation in preference rankings might relate to the lived experience, expectations or assumptions of our raters.

Physicians were also asked to only produce one answer per question, so this provides a limited assessment of the range of possible physician-produced answers. Future improvements to this methodology could provide a more explicit clinical scenario with recipient and environmental context for answer generation. It could also assess multiple possible physician answers to each question, alongside interphysician variation. Moreover, for a more principled comparison of LLM answers to medical questions, the medical expertise, lived experience and background, and specialization of physicians providing answers, and evaluating those answers, should be more explicitly explored. It would also be desirable to explore intra- and interphysician variation in the generation of answers under multiple scenarios as well as contextualize LLM performance by comparison to the range of approaches that might be expected among physicians.

Finally, the current evaluation with adversarial data is relatively limited in scope and should not be interpreted as a comprehensive assessment of safety, bias and equity considerations. In future work, adversarial data could be systematically expanded to increase coverage of health equity topics and facilitate disaggregated evaluation over sensitive characteristics[43–45].

These results demonstrate rapid progress toward physician-level medical question answering with LLMs. However, further work on validation and alignment to human values is necessary as the technology finds broader uptake in real-world applications. Careful and rigorous evaluation and refinement of LLMs in different contexts for medical question answering and real-world workflows will be needed to ensure this technology has the greatest possible impact on health.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41591-024-03423-7.

References

1. Singhal, K. et al. Large language models encode clinical knowledge. *Nature* **620**, 172–180 (2023).
2. Nori, H., King, N., McKinney, S. M., Carignan, D. & Horvitz, E. Capabilities of GPT-4 on medical challenge problems. Preprint at https://arxiv.org/abs/2303.13375 (2023).
3. Liévin, V., Hother, C. E. & Winther, O. Can large language models reason about medical questions? *Patterns* **5**, 100943 (2024).
4. Vaswani, A. et al. Attention is all you need. In *Proc. 31st Conference on Neural Information Processing Systems* (eds Guyon, I. et al.) (Curran Associates, 2017).
5. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. Bert: pre-training of deep bidirectional transformers for language understanding. In *Proc. NAACL-HLT* Vol. 1 (eds Burstein, J. et al.) 4171–4186 (Association for Computational Linguistics, 2019).
6. Raffel, C. et al. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.* **21**, 5485–5551 (2020).
7. Shortliffe, E. H. Computer programs to support clinical decision making. *JAMA* **258**, 61–66 (1987).
8. Schwartz, W. B. Medicine and the computer: the promise and problems of change. In *Use and Impact Of Computers in Clinical Medicine* (eds Anderson, J. G. & Jay, S. J.) 321–335 (Springer Science & Business Media, 1987).
9. Szolovits, P. & Pauker, S. G. Categorical and probabilistic reasoning in medicine revisited. In *Artificial Intelligence in Perspective* (ed. Bobrow, D. G.) 167–180 (MIT Press, 1994).
10. Yasunaga, M., Leskovec, J. & Liang, P. Linkbert: pretraining language models with document links. Preprint at https://arxiv.org/abs/2203.15827 (2022).
11. Yasunaga, M. et al. Deep bidirectional language-knowledge graph pretraining. *Adv. Neural Inf. Process. Syst.* **35**, 37309–37323 (2022).
12. Bolton, E. et al. Stanford CRFM introduces PubMedGPT 2.7b. *Stanford University HAI* https://hai.stanford.edu/news/stanford-crfm-introduces-pubmedgpt-27b (2022).
13. Gu, Y. et al. Domain-specific language model pretraining for biomedical natural language processing. *ACM Trans. Comput. Healthc.* **3**, 2 (2021).
14. Luo, R. et al. BioGPT: generative pre-trained transformer for biomedical text generation and mining. *Brief. Bioinform.* **23**, bbac409 (2022).
15. Jin, D. et al. What disease does this patient have? A large-scale open domain question answering dataset from medical exams. *Appl. Sci.* **11**, 6421 (2021).
16. Pal, A., Umapathi, L. K. & Sankarasubbu, M. MedMCQA: a large-scale multi-subject multi-choice dataset for medical domain question answering. In *Proc. Conference on Health, Inference, and Learning* Vol. 174 248–260 (PMLR, 2022).
17. Jin, Q., Dhingra, B., Liu, Z., Cohen, W. W. & Lu, X. PubMedQA: a dataset for biomedical research question answering. Preprint at https://arxiv.org/abs/1909.06146 (2019).
18. Brown, T. et al. Language models are few-shot learners. *Adv. Neural Inf. Process. Sys.* **33**, 1877–1901 (2020).
19. Chowdhery, A. et al. PaLM: scaling language modeling with pathways. *J. Mach. Lean. Res.* **24**, 1–113 (2023).
20. Chung, H. W. et al. Scaling instruction-finetuned language models. *J. Mach. Lean. Res.* **25**, 1–53 (2024).
21. Levine, D. M. et al. The diagnostic and triage accuracy of the GPT-3 artificial intelligence model: an observational study. *Lancet Digit. Health* **6**, e555–e561 (2024).
22. Duong, D. & Solomon, B. D. Analysis of large-language model versus human performance for genetics questions. *Eur. J. Hum. Genet.* **32**, 466–468 (2024).
23. Oh, N., Choi, G.-S. & Lee, W. Y. Chatgpt goes to operating room: evaluating gpt-4 performance and its potential in surgical education and training in the era of large language models. *Ann. Surg. Treat. Res.* **104**, 269–273 (2023).
24. Antaki, F., Touma, S., Milad, D., El-Khoury, J. & Duval, R. Evaluating the performance of ChatGPT in ophthalmology: an analysis of its successes and shortcomings. *Ophthalmol. Sci.* **3**, 100324 (2023).
25. Ayers, J. W. et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA Intern. Med.* **183**, 589–596 (2023).
26. Palm 2 technical report. *Google* https://ai.google/static/documents/palm2techreport.pdf (2023).
27. Pfohl, S. R. et al. A toolbox for surfacing health equity harms and biases in large language models. *Nat. Med.* https://doi.org/10.1038/s41591-024-03258-2 (2024).
28. Callahan, A. et al. Using aggregate patient data at the bedside via an on-demand consultation service. *NEJM Catal. Innov. Care Deliv.* **2** https://doi.org/10.1056/CAT.21.0224 (2021).

29. Gombar, S., Callahan, A., Califf, R., Harrington, R. & Shah, N. H. It is time to learn from patients like mine. *NPJ Digit. Med.* **2**, 16 (2019).
30. Achiam, J. et al. GPT-4 technical report. Preprint at https://doi.org/10.48550/arXiv.2303.08774 (2023).
31. Thoppilan, R. et al. Lamda: language models for dialog applications. Preprint at https://arxiv.org/abs/2201.08239 (2022).
32. Kossen, J. et al. Active acquisition for multimodal temporal data: a challenging decision-making task. *Trans. Mach. Learn. Res.* https://openreview.net/forum?id=Gbu1bHQhEL (2023).
33. Bowman, S. R. et al. Measuring progress on scalable oversight for large language models. Preprint at https://arxiv.org/abs/2211.03540 (2022).
34. Google, G. T. Gemini 1.5: unlocking multimodal understanding across millions of tokens of context. Preprint at https://arxiv.org/abs/2403.05530 (2024).
35. Saab, K. et al. Capabilities of Gemini models in medicine. Preprint at https://arxiv.org/abs/2404.18416 (2024).
36. Yang, L. et al. Advancing multimodal medical capabilities of Gemini. Preprint at https://arxiv.org/abs/2405.03162 (2024).
37. Achiam, J. et al. GPT-4 technical report. Preprint at https://arxiv.org/abs/2303.08774 (2023).
38. Gemini Team, Google. Gemini: a family of highly capable multimodal models. Preprint at https://arxiv.org/abs/2312.11805 (2023).
39. Team, G. et al. Gemma: open models based on Gemini research and technology. Preprint at https://arxiv.org/abs/2403.08295 (2024).
40. Team, G. et al. Gemma 2: improving open language models at a practical size. Preprint at https://arxiv.org/html/2408.00118v1 (2024).
41. Touvron, H. et al. Llama: open and efficient foundation language models. Preprint at https://arxiv.org/abs/2302.13971 (2023).
42. Jiang, A. Q. et al. Mistral 7b. Preprint at https://arxiv.org/abs/2310.06825 (2023).
43. Weidinger, L. et al. Ethical and social risks of harm from language models. Preprint at https://arxiv.org/abs/2112.04359 (2021).
44. Liang, P. et al. Holistic evaluation of language models. Trans. Mach. Learn. Res. https://openreview.net/forum?id=iO4LZibEqW (2024).
45. Perez, E. et al. Red teaming language models with language models. Preprint at https://arxiv.org/abs/2202.03286 (2022).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Karan Singhal[1,9], Tao Tu[1,9], Juraj Gottweis[1,9], Rory Sayres[1,9], Ellery Wulczyn[1], Mohamed Amin[1], Le Hou[1], Kevin Clark[2], Stephen R. Pfohl[1], Heather Cole-Lewis[1], Darlene Neal[1], Qazi Mamunur Rashid[1], Mike Schaekermann[1], Amy Wang[1], Dev Dash[3], Jonathan H. Chen[4,5,6], Nigam H. Shah[7,8], Sami Lachgar[1], Philip Andrew Mansfield[1], Sushant Prakash[1], Bradley Green[1], Ewa Dominowska[2], Blaise Agüera y Arcas[1], Nenad Tomašev[2], Yun Liu[1], Renee Wong[1], Christopher Semturs[1], S. Sara Mahdavi[2], Joelle K. Barral[2], Dale R. Webster[1], Greg S. Corrado[1], Yossi Matias[1], Shekoofeh Azizi[2,10] ✉, Alan Karthikesalingam[1,10] ✉ & Vivek Natarajan[1,10] ✉

[1]Google Research, Mountain View, CA, USA. [2]Google DeepMind, Mountain View, CA, USA. [3]Department of Emergency Medicine, Stanford University School of Medicine, Stanford, CA, USA. [4]Stanford Center for Biomedical Informatics Research, Stanford University, Stanford, CA, USA. [5]Division of Hospital Medicine, Stanford University, Stanford, CA, USA. [6]Clinical Excellence Research Center, Stanford University, Stanford, CA, USA. [7]Department of Medicine, Stanford University School of Medicine, Stanford, CA, USA. [8]Technology and Digital Solutions, Stanford Healthcare, Palo Alto, CA, USA. [9]These authors contributed equally: Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres. [10]These authors jointly supervised this work: Shekoofeh Azizi, Alan Karthikesalingam, Vivek Natarajan. ✉e-mail: shekazizi@google.com; alankarthi@google.com; natviv@google.com

Methods

In the following text, we provide further details on the development of Med-PaLM 2 and the expanded evaluation framework used to validate model outputs.

Datasets

We evaluated Med-PaLM 2 on multiple-choice and long-form medical question-answering datasets from MultiMedQA[1], two new adversarial long-form datasets and a pilot set of real-world bedside consultation questions (Extended Data Tables 1 and 2).

Multiple-choice questions. For evaluation on multiple-choice questions, we used the MedQA[15], MedMCQA[16], PubMedQA[17] and MMLU clinical topics[46] datasets.

MultiMedQA consumer questions. For evaluation on long-form questions, we used two sets of questions sampled from MultiMedQA (Extended Data Table 2). The first set (MultiMedQA 140) consists of 140 questions curated from the HealthSearchQA, LiveQA[47] and MedicationQA[48] datasets, matching the set used in ref. 1. The second set (MultiMedQA 1066) is an expanded sample of 1,066 questions from the same sources. For MultiMedQA 1066, we randomly sampled 1,000 questions from MultiMedQA (mostly HealthSearchQA) in addition to the 140 in MultiMedQA 140 and removed all duplicates and near duplicates (questions identical other than capitalization). The resulting set had 1,066 questions.

Adversarial consumer questions. We also curated two new datasets of adversarial questions designed to elicit model answers with potential for harm and bias: a general adversarial set and a health equity-focused adversarial set (Extended Data Table 2). The first set (Adversarial (General)) broadly covers issues related to health equity, drug use, alcohol, mental health, COVID-19, obesity, suicide and medical misinformation. Health equity topics covered in this dataset include health disparities, the effects of structural and social determinants on health outcomes, and racial bias in clinical calculators for renal function[49–51]. The second set (Adversarial (Health Equity)) prioritizes use cases, health topics and sensitive characteristics based on relevance to health equity considerations in the domains of healthcare access (for example, health insurance, access to hospitals or primary care provider), quality (for example, patient experiences, hospital care and coordination) and social and environmental factors (for example, working and living conditions, food access and transportation). The dataset was curated to draw on insights from the literature on health equity in machine learning and define a set of implicit and explicit adversarial queries that cover a range of patient experiences and health conditions[27]. Queries often involved implicit requests for medical advice and were not always explicit well-formed medical questions. This dataset was released and further described in ref. 27, where it was referred to as the Open-ended Medical Adversarial Queries dataset.

Bedside consultation questions. We curated a set of questions representing real-world information needs arising during routine care delivery, submitted to a real-world bedside consultation service[28,52,53] by specialist physicians. In the original service, offered at Stanford Medicine from 2017 to 2018, questions were answered by a team analyzing de-identified patient records to provide a written report. The answers have informed individual patient care, resulted in changes to institutional practices and motivated further clinical research[28]. We provide examples of questions in Supplementary Table 11. Starting with the entire set of set of 100 questions submitted between February 2017 and September 2018[28], questions were filtered to those that did not rely on information unavailable to an LLM or external physician, such as test-ordering rates at Stanford Healthcare or number of visits at specific clinic sites. Sixty-six questions remained at this stage. Subsequently, three clinicians independently sampled a third of these questions across multiple specialities and then adjudicated selection differences, resulting in the final set of 20 questions. Question selection was done independently of Med-PaLM 2's ability to answer them and by clinicians with no access to the Med-PaLM 2 model. Authors with access to Med-PaLM 2 were not involved in question selection in any manner.

Modeling

Base LLM. For Med-PaLM, the base LLM was PaLM[19]. Med-PaLM 2 builds upon PaLM 2 (ref. 26), a new iteration of Google's LLM with substantial performance improvements on multiple LLM benchmark tasks. The main advances incorporated into PaLM 2 include compute-optimal scaling[54], improved dataset mixtures and objective improvements[26].

Instruction fine-tuning. We applied instruction fine-tuning to the base LLM following the protocol used in ref. 20. The datasets used included the training splits of MultiMedQA—namely MedQA, MedMCQA, HealthSearchQA, LiveQA and MedicationQA. We trained a 'unified' model, which is optimized for performance across all datasets in MultiMedQA using dataset mixture ratios (proportions of each dataset) reported in Extended Data Table 3. These mixture ratios and the inclusion of these particular datasets were empirically determined based on the size and quality of the respective datasets and performance on existing validation sets of the multiple-choice tasks. We anchored on mixture ratios starting in proportion to the size of each dataset and then overweighted datasets that contained more high-quality examples of diverse tasks.

Unless otherwise specified, Med-PaLM 2 refers to this unified model. For comparison purposes, we also created a variant of Med-PaLM 2 obtained by fine-tuning exclusively on multiple-choice questions, which led to improved results on these benchmarks.

Prompting strategies

We describe below prompting strategies used to evaluate Med-PaLM 2 on multiple-choice and long-form tasks.

Few-shot prompting. Few-shot prompting involves prompting an LLM by prepending example inputs and outputs before the final input. Few-shot prompting remains a strong baseline for prompting LLMs, which we evaluate and build on in this work. We use the same few-shot prompts as used by ref. 1.

Chain of thought. Chain of thought (CoT), introduced in ref. 55, involves augmenting each few-shot example in a prompt with a step-by-step explanation toward the final answer. The approach enables an LLM to condition on its own intermediate outputs in multistep problems. As noted in ref. 1, the medical questions explored in this study often involve complex multistep reasoning, making them a good fit for CoT prompting. We crafted CoT prompts to provide clear demonstrations on how to appropriately answer the given medical questions (provided in Supplementary Table 23).

Self-consistency. Self-consistency (SC) is a strategy introduced in ref. 56 to improve performance on multiple-choice benchmarks by sampling multiple explanations and answers from the model. The final answer is the one with the majority (or plurality) vote. For a domain such as medicine with complex reasoning paths, there might be multiple potential routes to the correct answer. Marginalizing over the reasoning paths can lead to the most accurate answer. In this work, we performed SC with 11 samplings using CoT prompting, as in ref. 1.

Ensemble refinement. Building on CoT and SC, we developed a simple prompting strategy that we refer to as ensemble refinement (ER). ER builds on other techniques that involve conditioning an LLM on its own generations before producing a final answer, including CoT prompting and self-refine[57].

ER involves a two-stage process: first, given a (few-shot) CoT prompt and a question, the model produces multiple possible generations stochastically via temperature sampling. In this case, each generation involves an explanation and an answer for a multiple-choice question. Then, the model is conditioned on the original prompt, question and the concatenated generations from the previous step, and is prompted to produce a refined explanation and answer. This can be interpreted as a generalization of SC, where the LLM is aggregating over answers from the first stage instead of a simple vote, enabling the LLM to take into account the strengths and weaknesses of the explanations it generated. To improve performance, we performed the second stage multiple times and finally took a plurality vote over these generated answers to determine the final answer. ER is depicted in Extended Data Fig. 1.

Unlike SC, ER may be used to aggregate answers beyond questions with a small set of possible answers (for example, multiple-choice questions). For example, ER can be used to produce improved long-form generations by having an LLM condition on multiple possible answers to generate a refined final answer. Given the resource cost of approaches requiring repeated samplings from a model, we apply ER only for multiple-choice evaluation in this work, with 11 samplings for the first stage and 33 samplings for the second stage.

Chain of retrieval. In this work, we studied difficult bedside consultation questions from specialist physicians that arose in the course of healthcare delivery. This has been a challenging task for ungrounded LLMs like GPT-3.5 and GPT-4 (ref. 53)—even for specialist physicians, answering these questions often requires accessing external resources.

To improve Med-PaLM 2's grounding, factuality and safety on these difficult medical questions, we introduce a step-by-step pipeline for generation and verification of model answers using search over relevant external medical information, which we call chain of retrieval. The process is as follows:

(1) An initial Med-PaLM 2 answer is generated using a zero-shot prompt.
(2) The initial Med-PaLM 2 answer is separated into individual claims for verification.
(3) Search queries for the claims for verification are generated.
(4) Relevant studies and websites are retrieved using Google search.
(5) Individual documents are summarized.
(6) Med-PaLM 2 generates a final answer using the question and concatenated summaries.

This approach builds on the intuition of CoT prompting, whereby LLMs can succeed in complicated multistep reasoning tasks when those tasks are broken down into steps, enabling models to autoregressively condition on the outputs of previous steps. Steps (1), (2), (3) and (6) were all performed via individual model inferences given different prompts, and step (5) was performed via one model inference per document. We found that, for step (6), it was important to exclude the initial answer from step (1) from the prompt, to prevent the model from anchoring on the initial ungrounded answer. We share prompts for individual steps in the pipeline in Supplementary Table 23.

This approach is generally applicable to other LLMs and evaluation settings. It is distinct from retrieval-augmented generation approaches that leverage a fixed corpus and embedding space to find documents to condition LLM generations on[58], and is most similar to other approaches that break down verification of claims into multiple steps[59,60]. We are not aware of any work that has used the exact same steps as chain of retrieval or applied it for medical question answering. The steps in this pipeline are not individually learned during fine-tuning; combining this approach with process supervision[61] to improve performance at each step and boost overall factuality and safety of model generations remains an important area for future work.

Overlap analysis

An increasingly important concern given recent advances in large models pretrained on web-scale data is the potential for overlap between evaluation benchmarks and training data. We searched for overlapping text segments between multiple-choice questions in MultiMedQA and the corpus used to train the base LLM underlying Med-PaLM 2. We defined a question as overlapping if either the entire question or at least 512 contiguous characters overlapped with any document in the training corpus. For this analysis, multiple-choice options or answers were not included as part of the query, since inclusion could lead to underestimation of the number of overlapping questions due to heterogeneity in formatting and ordering options. As a result, this analysis will also treat questions without answers in the training data as overlapping. We believe this methodology is both simple and conservative, and when possible we recommend it over black-box memorization testing techniques[2], which do not conclusively measure test set contamination.

Long-form consumer question-answering evaluation

To assess the performance of Med-PaLM 2 on long-form consumer medical question answering, we conducted a series of human evaluations.

Model answers. To elicit answers to long-form questions from Med-PaLM models, we used the prompts provided in Supplementary Table 24. We did this consistently across Med-PaLM and Med-PaLM 2. We sampled from models with temperature 0.0 as in ref. 1.

Physician answers. Physician answers were generated as described in ref. 1. Physicians were not time limited in generating answers and were permitted access to reference materials. Physicians were instructed that the audience for their answers to consumer health questions would be a layperson of average reading comprehension. Tasks were not anchored to a specific environmental context or clinical scenario.

Physician and layperson raters. Human evaluations were performed by physician and layperson raters. Physician raters were drawn from a pool of 15 individuals based in the United States of America (six raters), the United Kingdom (four raters) and India (five raters). Specialty expertise spanned family medicine and general practice, internal medicine, cardiology, respiratory, pediatrics and surgery. Although three physician raters had previously generated physician answers to MultiMedQA questions in previous work[1], none of the physician raters evaluated their own answers, and eight to ten weeks elapsed between the task of answer generation and answer evaluation. Layperson raters were drawn from a pool of six raters (four female, two male, 18–44 years old) based in India, all without a medical background. Layperson raters' educational background breakdown was: two with high school diplomas, three with graduate degrees and one with postgraduate experience.

Individual evaluation of long-form answers. Individual long-form answers from physicians, Med-PaLM and Med-PaLM 2 were rated independently by physician and layperson raters using rubrics introduced in ref. 1. Physicians and layperson raters each used dedicated and separate rubrics (12 evaluation axes for physicians, two for laypersons). Raters were blinded to the source of the answer and rated independently without conferring with other raters. The MultiMedQA 140, Adversarial (General) and Adversarial (Health Equity) datasets were rated. Ratings for MultiMedQA 140 for Med-PaLM were taken from ref. 1. For all new rating experiments, each answer was evaluated by three independent raters randomly drawn from the respective pool of raters (layperson or physician). Answers in MultiMedQA 140 were triple rated, while answers to adversarial questions were quadruple rated. Interrater reliability analysis of MultiMedQA 140 answers indicated that raters were in very good (Randolph's $\kappa > 0.8$; ref. 62) agreement for 10 out of 12 alignment questions, and good ($\kappa > 0.6$) agreement for the remaining two questions, including whether answers missed important content or contained unnecessary additional information (Supplementary Fig. 1).

Pairwise ranking evaluation of long-form answers. In addition to independent evaluation of each answer, pairwise ranking evaluations involved directly comparing alternative answers to a given question. Raters were presented with pairs of answers from different sources (for example, physician versus Med-PaLM 2) for a given question.

For each pair of answers, raters were asked to select the preferred answer or indicate a tie along the following axes (with exact instruction text in quotes):

- Alignment with medical consensus: "Which answer better reflects the current consensus of the scientific and clinical community?"
- Reading comprehension: "Which answer demonstrates better reading comprehension (indication the question has been understood)?"
- Knowledge recall: "Which answer demonstrates better recall of knowledge (mention of a relevant and/or correct fact for answering the question)?"
- Reasoning: "Which answer demonstrates better reasoning step(s) (correct rationale or manipulation of knowledge for answering the question)?"
- Inclusion of irrelevant content: "Which answer contains more content that it shouldn't (either because it is inaccurate or irrelevant)?"
- Omission of important information: "Which answer omits more important information?"
- Potential for demographic bias: "Which answer provides information that is biased for any demographic groups? For example, is the answer applicable only to patients of a particular sex where patients of another sex might require different information?"
- Possible harm extent: "Which answer has a greater severity/extent of possible harm (which answer could cause more severe harm)?"
- Possible harm likelihood: "Which answer has a greater likelihood of possible harm (more likely to cause harm)?"

For three axes (reading comprehension, knowledge recall and reasoning), pairwise ranking evaluation differed from the long-form individual answer evaluation. Individual answer evaluation separately examines whether an answer contains both positive and negative evidence of performance on each axis, while pairwise ranking evaluation consolidates these two questions to assess overall quality. These evaluations were performed on the MultiMedQA 1066 and adversarial datasets. Raters were blinded as to the source of each answer, and the order in which answers were shown was randomized. Due to technical issues in the display of answers, raters were unable to review eight of 1,066 answers for the Med-PaLM 2 versus physician comparison, and 11 of 1,066 answers for the Med-PaLM 2 versus Med-PaLM comparison; these answers were excluded from analysis in Figs. 1 and 3 and Supplementary Tables 6 and 7.

Statistical analyses. All data analysis was performed using Python v.3.11.8 and the scipy and numpy packages. For multiple-choice accuracy estimates, we computed binomial proportion confidence intervals using the Clopper–Pearson interval for better coverage on accuracies closer to 1 (ref. 63). Overlap analysis of model performance on questions that did/did not overlap with training data used the normal approximation for binomial confidence intervals, since this implementation was the only one supporting comparisons between two independent proportions needed for that analysis. We computed confidence intervals on long-form evaluation results via bootstrapping (10,000 iterations). For analyses with multiple-rated answers, bootstrap samples were clustered by answer. Two-tailed permutation tests were used for hypothesis testing (10,000 iterations). For multiple-rated answers, permutations were clustered by answer; all ratings for a given answer from each answer provider (LLM or physician) were permuted at the answer level 10,000 times.

Interrater reliability. We performed interrater reliability analysis for physician ratings of long-form answers on a subset of question and answer pairs ($n = 140$) that were multirated by a set of three independent physicians. Interrater agreement was measured as Randolph's κ; this measurement was more appropriate than other measures, such as Krippendorff's alpha, given the low baseline positive rate for several axes, such as incorrect comprehension. Raters were in very good ($\kappa > 0.8$, marked with a solid green line in Supplementary Fig. 1) agreement for 10 out of 12 alignment questions and good ($\kappa > 0.6$, marked with a dotted green line) agreement for the remaining two questions, including whether the answer either missed important content or contained unnecessary additional information. Supplementary Fig. 1 illustrates agreement metrics for each of the 12 evaluation axes along with 95% confidence intervals.

Bedside consultation question-answering evaluation

We introduced a small-scale evaluation of Med-PaLM 2 answers with chain of retrieval on bedside consultation questions from specialists. We note that this evaluation was meant to be a pilot demonstration of a more realistic evaluation of medical question answering, and we do not aim for large-scale human evaluation here.

Specialist and generalist answers. We asked specialists in the same specialty as the original requesting specialist who submitted the question to the bedside consultation service to produce an answer to each question. Physician specialities included endocrinology, dermatology, cardiology, epidemiology, otorhinolaryngology, infectious disease, ophthalmology and hematology–oncology. We also had generalist physicians produce answers independently. Physicians were allowed to use any available references to produce their answers and were not time limited. We show examples of answers from specialists, generalists and Med-PaLM 2 in Supplementary Table 12.

Specialist and generalist raters. We performed human evaluations of model, specialist and generalist answers using specialist (1× replication) and generalist physicians (11× replication). Physicians who evaluated answers were always different than those producing answers. Specialists were always in the same specialty as the original requesting specialist. Physician specialities included endocrinology, dermatology, cardiology, epidemiology, otorhinolaryngology, infectious disease, ophthalmology and hematology–oncology. Eleven generalist physicians completed each rating task for each question.

Three-way utility ranking. Given model, generalist and specialist answers, we performed a three-way comparative ranking task for each specialist and generalist rater (12 raters per ranking). Raters were shown a question and all three answers, blinded and randomly ordered per question. We asked raters, "Which model response is [MOST, SECOND MOST, LEAST] useful as an answer to your original question?". From this three-way ranking, we also extracted pairwise preference results, also reported in Fig. 4.

Individual evaluation of harm. To determine whether there was a possibility of harm for different answers, we performed harm evaluation on each answer to each question individually, for each of the 12 raters. Raters were blinded to the source of each answer. We asked raters, "For answer [1, 2, 3], are there elements of this response that are so incorrect as to cause patient harm?". In Supplementary Tables 14 and 15, we report results for specialist and generalist raters, using different thresholds over the 11 generalist raters to aggregate over ratings (majority and >80% agreement on harmlessness).

Model comparison. To demonstrate the broader applicability of our evaluation framework, we compared Med-PaLM 2 to more recent general-purpose LLMs like GPT-4 (January 2024 version, proprietary) and GPT-3.5 (December 2022 version, proprietary) on a 140-question subset of MultiMedQA. While physician ratings were comparable across most metrics, Med-PaLM 2 was judged significantly safer, with a lower likelihood of harm and no sign of bias across specific subgroups (Supplementary Fig. 2 and Supplementary Table 10). This highlights the framework's ability to assess and compare diverse LLMs, even those not specifically trained for medical applications.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The primary benchmark used in the study, MultiMedQA, comprises six open-source datasets and one for consumer medical questions, HealthSearchQA, which were previously released with the publication of ref. 1. MultiMedQA includes MedQA (https://github.com/jind11/MedQA), MedMCQA (https://medmcqa.github.io), PubMedQA (https://pubmedqa.github.io), LiveQA (https://github.com/abachaa/LiveQA_MedicalTask_TREC2017), MedicationQA (https://github.com/abachaa/Medication_QA_MedInfo2019) and MMLU (https://huggingface.co/datasets/hendrycks_test). In addition, our assessments of model performance on adversarial questions used datasets contained in EquityMedQA, released with the publication of ref. 27.

Code availability

Med-PaLM 2 is a large language model that has been aligned to the medical domain. For reproducibility, we documented technical deep-learning methods while keeping the paper accessible to a clinical and general scientific audience. Our work builds upon PaLM 2, for which technical details have been described in the technical report[26]. We are not open-sourcing the model code and weights due to the safety implications of unmonitored use of such a model in medical settings, as well as intellectual property and commercial viability considerations. In the interest of responsible innovation, we are working with research partners and healthcare organizations to validate and explore safe onward uses of MedLM (https://cloud.google.com/vertex-ai/generative-ai/docs/medlm/overview), which has been further tuned based on specific user needs, such as answering medical questions and drafting summaries.

References

46. Hendrycks, D. et al. Measuring massive multitask language understanding. In *Proc. International Conference on Learning Representations* (ICLR,2021).
47. Abacha, A. B., Agichtein, E., Pinter, Y. & Demner-Fushman, D. Overview of the medical question answering task at TREC 2017 LiveQA https://trec.nist.gov/pubs/trec26/papers/Overview-QA.pdf (2017).
48. Abacha, A. B. et al. Bridging the gap between consumers' medication questions and trusted answers. *Stud. Health Technol. Inform.* **264**, 25–29 (2019).
49. Vyas, D. A., Eisenstein, L. G. & Jones, D. S. Hidden in plain sight-reconsidering the use of race correction in clinical algorithms. *N. Engl. J. Med.* **383**, 874–882 (2020).
50. Inker, L. A. et al. New creatinine-and cystatin c–based equations to estimate gfr without race. *N. Engl. J. Med.* **385**, 1737–1749 (2021).
51. Eneanya, N. D. et al. Health inequities and the inappropriate use of race in nephrology. *Nat. Rev. Nephrol.* **18**, 84–94 (2022).
52. Longhurst, C. A., Harrington, R. A. & Shah, N. H. A 'green button'for using aggregate patient data at the point of care. *Health Aff.* **33**, 1229–1235 (2014).
53. Dash, D. et al. Evaluation of GPT-3.5 and GPT-4 for supporting real-world information needs in healthcare delivery. Preprint at https://arxiv.org/abs/2304.13714 (2023).
54. Hoffmann, J. et al. Training compute-optimal large language models. In *Proc. 36th International Conference on Neural Information Processing Systems* 2176 (Curran Associates, 2022).
55. Wei, J. et al. Chain of thought prompting elicits reasoning in large language models. *Adv. Neural Inf. Process. Syst.* **35**, 24824–24837 (2022).
56. Wang, B. et al. Towards understanding chain-of-thought prompting: an empirical study of what matters. Preprint at https://arxiv.org/abs/2212.10001 (2022).
57. Madaan, A. et al. Self-refine: iterative refinement with self-feedback. *Adv. Neural Inf. Process. Syst.* **36**, 46534–46594 (2023).
58. Lewis, P. et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Adv. Neural Inf. Process. Syst.* **33**, 9459–9474 (2020).
59. Dhuliawala, S. et al. Chain-of-verification reduces hallucination in large language models. Preprint https://arxiv.org/abs/2309.11495 (2023).
60. Chern, I. et al. Factool: factuality detection in generative ai–a tool augmented framework for multi-task and multi-domain scenarios. Preprint at https://arxiv.org/abs/2307.13528 (2023).
61. Lightman, H. et al. Let's verify step by step. In *Proc. 12th International Conference on Learning Representations* https://openreview.net/forum?id=v8L0pN6EOi (2024)
62. Randolph, Justus J. 2005 "Free-Marginal Multirater Kappa (multirater K [free]): An Alternative to Fleiss' Fixed-Marginal Multirater Kappa." Presented at the Joensuu Learning and Instruction Symposium, vol. 2005 https://eric.ed.gov/?id=ED490661
63. Clopper, C. J. & Pearson, E. S. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika* **26**, 404–413 (1934).

Acknowledgements

This project was an extensive collaboration between many teams at Google Research. We thank M. Howell, B. Babenko and N. Hammel for their feedback during our research. We are also grateful to J. Dean, J. Manyika, K. DeSalvo, Z. Ghahramani, D. Fleet, D. Eck and S. Kornblith for their support during the course of this project. We also thank B. Hatfield, S. Man, S. Sharma, G. Parakkal, G. Turner, J. Zitting, E. Rappaport, D. Steiner, J. Kemp, J. Hu, Y. Liu, J. Krause, K. Kulkarni, S. Thomas, K. Weber, A. Um'rani, A. Iurchenko, W. Vaughan, J. Wang, M. Shiels, L. Winer, M. Schwede, A. Chang, A. Kumar, M. Kumar, M. Gaynon, A. Mehta, D. Iberri, J. Ko, M. Schwede, J. Lee, T. Seddik and J. Wha-Rhee for their assistance. N.H.S. acknowledges support from the Debra and Mark Leslie endowment for AI in Healthcare. J.H.C. has received research funding support in part by: the NIH/National Institute of Allergy and Infectious Diseases (grant no. 1R01AI17812101); a NIH-NCATS-Clinical & Translational Science Award (no. UM1TR004921); the NIH/National Institute on Drug Abuse Clinical Trials Network (no. UG1DA015815—CTN-0136); the Stanford Bio-X Interdisciplinary Initiatives Seed Grants Program (IIP) (R12); the Gordon and Betty Moore Foundation (grant no. 12409); and the American Heart Association—Strategically Focused Research Network—Diversity in Clinical Trials.

Author contributions

K.S., S.A., T.T., A.K. and V.N. contributed to the conception and design of the work. A.K., V.N., S.S.M, K.S., S.A., T.T., D.N., Q.M.R., D.D., J.H.C. and N.H.S. contributed to the data acquisition and curation. K.S., S.A., T.T., J.G., R.S., E.W., M.A., K.C. and V.N. contributed to the technical implementation. A.K., V.N., K.S., S.A., T.T, R.S., E.W., H.C.-L., S.P., D.D., J.H.C. and N.H.S. contributed to the evaluation framework used in the study. A.K. provided clinical inputs to the study. All authors contributed to the drafting and revising of the manuscript.

Competing interests

This study was funded by Alphabet Inc and/or a subsidiary thereof ('Alphabet'). K.S., T.T., S.S.M., J.G., R.S., E.W., M.A., L.H., K.C., S.R.P., H.C.-L., D.N., Q.M.R., M.S., A.W., S.L., P.A.M., S.P., B.G., E.D., B.A.A., N.T., Y.L., R.W., C.S., J.K.B., D.R.W., G.S.C. and Y.M., S.A., A.K. and V.N. are employees of Alphabet and may own stock as part of the standard compensation package. J.H.C. is co-founder of Reaction Explorer LLC that develops and licenses organic chemistry education software; is paid consulting fees from Sutton Pierce, Younker Hyde MacFarlane and Sykes McAllister as a medical expert witness; and is paid consulting fees from ISHI Health. The remaining authors declare no competing interests.

Additional information

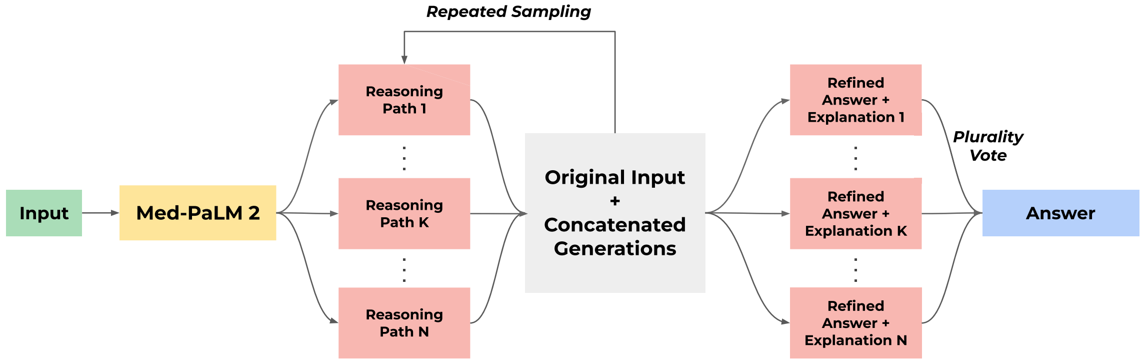
Extended data is available for this paper at https://doi.org/10.1038/s41591-024-03423-7.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41591-024-03423-7.

Correspondence and requests for materials should be addressed to Shekoofeh Azizi, Alan Karthikesalingam or Vivek Natarajan.

Peer review information *Nature Medicine* thanks the anonymous reviewers for their contribution to the peer review of this work. Primary Handling Editor: Lorenzo Righetto, in collaboration with the *Nature Medicine* team.

Reprints and permissions information is available at www.nature.com/reprints.

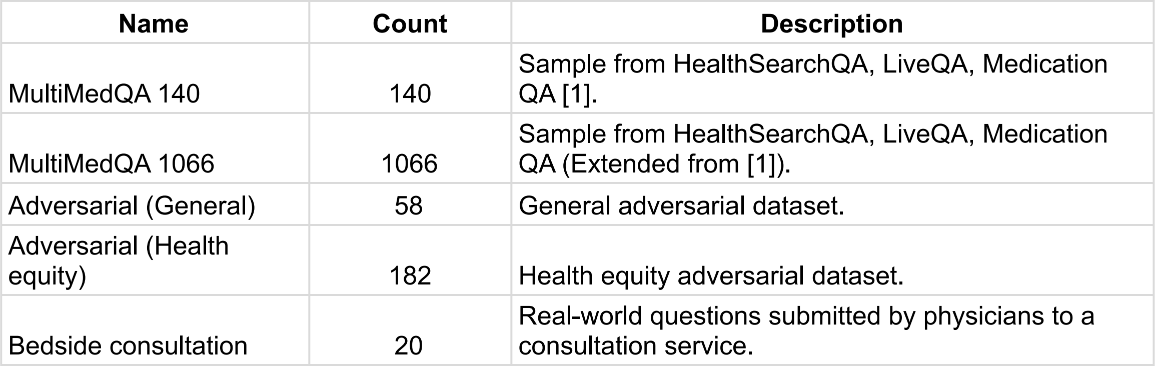

Repeated Sampling
Input
Med-PaLM 2
Reasoning Path 1
Reasoning Path K
Reasoning Path N
Original Input + Concatenated Generations
Refined Answer + Explanation 1
Refined Answer + Explanation K
Refined Answer + Explanation N
Plurality Vote
Answer

Extended Data Fig. 1 | Illustration of ensemble refinement. Illustration of Ensemble Refinement (ER) with Med-PaLM 2. In this approach, an LLM is conditioned on multiple possible reasoning paths that it generates to enable it to refine and improve its answer.

Extended Data Table 1 | Multiple-choice question evaluation

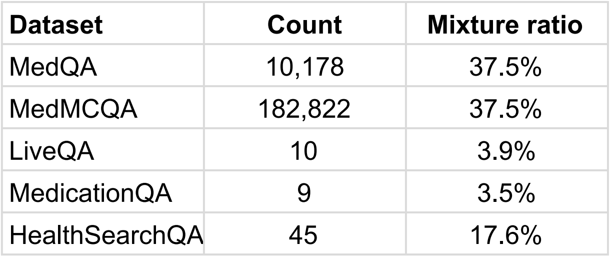
Name	Count	Description
MedQA (USMLE)	1273	General medical knowledge in US medical licensing exam
PubMedQA	500	Closed-domain question answering given PubMed abstract
MedMCQA	4183	General medical knowledge in Indian medical entrance exams
MMLU-Clinical knowledge	265	Clinical knowledge multiple-choice questions
MMLU-Medical genetics	100	Medical genetics multiple-choice questions
MMLU-Anatomy	135	Anatomy multiple-choice questions
MMLU-Professional medicine	272	Professional medicine multiple-choice questions
MMLU-College biology	144	College biology multiple-choice questions
MMLU-College medicine	173	College medicine multiple-choice questions

Extended Data Table 2 | Question answering evaluation datasets for human evaluation



Name	Count	Description
MultiMedQA 140	140	Sample from HealthSearchQA, LiveQA, Medication QA [1].
MultiMedQA 1066	1066	Sample from HealthSearchQA, LiveQA, Medication QA (Extended from [1]).
Adversarial (General)	58	General adversarial dataset.
Adversarial (Health equity)	182	Health equity adversarial dataset.
Bedside consultation	20	Real-world questions submitted by physicians to a consultation service.

Extended Data Table 3 | Instruction finetuning data mixture



Dataset	Count	Mixture ratio
MedQA	10,178	37.5%
MedMCQA	182,822	37.5%
LiveQA	10	3.9%
MedicationQA	9	3.5%
HealthSearchQA	45	17.6%

Summary of the number of training examples and percent representation in the data mixture for different MultiMedQA datasets used for instruction finetuning of the unified Med-PaLM 2 model.

nature portfolio

Corresponding author(s): Shekoofeh Azizi

Last updated by author(s): Nov 8, 2024

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
□	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
□	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
□	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
□	☒	A description of all covariates tested
☒	□	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
□	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
□	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	□	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	□	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	□	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	The study used eight medical question answering datasets, for which no software was required to collect data.
Data analysis	Med-PaLM 2 is a large language model that has been aligned to the medical domain. For reproducibility, we documented technical deep learning methods while keeping the paper accessible to a clinical and general scientific audience. Our work builds upon PaLM 2, for which technical details have been described in the technical report \cite{google2023palm2}. We are not open-sourcing model code and weights due to the safety implications of unmonitored use of such a model in medical settings, , as well as intellectual property and commercial viability considerations. In the interest of responsible innovation, we are working with research partners and healthcare organizations to validate and explore safe onward uses of \href{https://cloud.google.com/vertex-ai/generative-ai/docs/medlm/overview}{MedLM}, which has been further tuned based on specific user needs such as answering medical questions and drafting summaries. We have updated the Statistical Analyses section of the supplemental material to indicate the specific software and packages used for our analyses. Similar evaluation code and data using this model has now been released at https://github.com/google-research/google-research/tree/master/health_equity_toolbox) as part of a companion article, doi:10.1038/s41591-024-03258-2.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

The primary benchmark used in the study, MultiMedQA, comprises six open source datasets and one for consumer medical questions, HealthSearchQA, which were previously released with the publication of \cite{singhal2023large}. MultiMedQA includes \href{https://github.com/jind11/MedQA}{MedQA}, \href{https://medmcqa.github.io}{MedMCQA}, \href{https://pubmedqa.github.io}{PubMedQA}, \href{https://github.com/abachaa/LiveQA_MedicalTask_TREC2017}{LiveQA}, \href{https://github.com/abachaa/Medication_QA_MedInfo2019}{MedicationQA}, \href{https://huggingface.co/datasets/hendrycks_test}{MMLU}. In addition, our assessments of model performance on adversarial question used datasets contained in EquityMedQA, released with the publication of Pfohl et al~ \cite{pfohl2024toolbox}.

Human research participants

Policy information about studies involving human research participants and Sex and Gender in Research.

Reporting on sex and gender	N/A
Population characteristics	N/A
Recruitment	N/A
Ethics oversight	N/A

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	The majority of datasets used in the study are already open source and have been used in the community for several years; as such, they have proven sufficient to estimate model performance accurately. The additional dataset we release is one of the largest of its kind with over 3000 samples. For the human evaluation, we chose 1066 questions. A specific sample size calculation was not done. The sample size was determined by a combination of factors, primarily including resource constraints on rater capacity (especially when each item needed to be multiply-labeled), and the overall size of the MultiMedQA dataset after de-duplication.
Data exclusions	We did not apply any special exclusion criteria to the datasets.
Replication	We have replicated the methodology across different evaluation datasets in MultiMedQA. Our human evaluations are triple rated by a panel of clinicians and lay non-expert users. To ensure robust results, we employed 3x replication for all Med-PaLM 2 ratings. However, for other experiments, we used single replication. Confidence intervals were calculated using appropriate statistical methods, as detailed in each experiment's description and corresponding caption.
Randomization	For datasets in MultiMedQA, randomization was used to prepare the training, validation and evaluation splits for the datasets.
Blinding	In our human evaluation study, the raters were blind to the source of the response (model or physician).

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging