

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 (<i>n</i>) 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 <i>Only common tests should be described solely by name; describe more complex techniques in the Methods section.</i>
☐	☒ 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. <i>F</i> , <i>t</i> , <i>r</i>) with confidence intervals, effect sizes, degrees of freedom and <i>P</i> value noted <i>Give P values as exact values whenever suitable.</i>
☒	☐ 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 <i>d</i> , Pearson's <i>r</i>), 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 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](https://www.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