nature portfolio

Corresponding author(s): Krithik Vishwanath

Last updated by author(s): Apr 20, 2026

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	For MedQA and HealthBench, 500 items each were randomly sampled (seed=62) from the publicly available benchmarks. Frontier LLMs (GPT-5.2, Gemini 3.1 Pro, Claude Opus 4.6) were queried via API with deterministic parameters (temperature=0.0, seed=62, search tools enabled). Clinical tools (OpenEvidence, UpToDate Expert AI) and Google Search AI Overview were queried manually through their browser interfaces, as no public APIs were available. Each query was submitted to all six model.
Data analysis	Analyses and model querying were performed in Python 3.13.2 using openai 1.68.2, anthropic 0.49.0, python-dotenv 1.1.0, tqdm 4.67.1, numpy 2.2.4, pandas 2.2.3, datasets 3.4.1, pyarrow 19.0.1, requests 2.32.3, httpx 0.28.1, pydantic 2.10.6, and huggingface-hub 0.29.3. MedQA accuracy was compared using exact McNemar's tests with Holm-Bonferroni correction; confidence intervals are Wilson's 95% CIs. HealthBench scores were compared using Wilcoxon signed-rank tests with Holm correction, graded by panel-majority voting across three LLM judges. For the RCQ, overall model differences were assessed with the Friedman test; pairwise comparisons used Nemenyi post-hoc tests and Wilcoxon signed-rank tests with Holm correction. Effect sizes are rank-biserial correlations with 95% CIs from 5,000 bootstrap iterations. The primary regression was a cumulative link (proportional odds) model with rater fixed effects; a linear mixed model with random rater intercepts served as a sensitivity check. Inter-rater reliability was assessed with Krippendorff's alpha (ordinal) and PABAK. Binary safety flags were compared with Cochran's Q and pairwise McNemar's tests; refusal rates were compared with Fisher's exact test. All tests were two-sided at $\alpha = 0.05$. Code for the LLM generation and physician evaluation can be found at https://github.com/nyuolab/clinical-llm-benchmarks upon publication of the article.

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 MedQA (https://huggingface.co/datasets/GBaker/MedQA-USMLE-4-options) and HealthBench (https://huggingface.co/datasets/openai/healthbench) benchmarks are publicly available on HuggingFace. The specific 500-item subsets used in this study can be reproduced using the random seed (seed=62) described in the Online Methods. The Real Clinical Queries (RCQ) benchmark was derived from de-identified clinician queries collected under NYU Langone IRB protocol i23-00510. Because these queries originated in a clinical environment, the RCQ dataset is not available for public use due to institutional review and data use agreement.

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	*Use the terms sex (biological attribute) and gender (shaped by social and cultural circumstances) carefully in order to avoid confusing both terms. Indicate if findings apply to only one sex or gender; describe whether sex and gender were considered in study design; whether sex and/or gender was determined based on self-reporting or assigned and methods used.* *Provide in the source data disaggregated sex and gender data, where this information has been collected, and if consent has been obtained for sharing of individual-level data; provide overall numbers in this Reporting Summary. Please state if this information has not been collected.* *Report sex- and gender-based analyses where performed, justify reasons for lack of sex- and gender-based analysis.*
Reporting on race, ethnicity, or other socially relevant groupings	*Please specify the socially constructed or socially relevant categorization variable(s) used in your manuscript and explain why they were used. Please note that such variables should not be used as proxies for other socially constructed/relevant variables (for example, race or ethnicity should not be used as a proxy for socioeconomic status).* *Provide clear definitions of the relevant terms used, how they were provided (by the participants/respondents, the researchers, or third parties), and the method(s) used to classify people into the different categories (e.g. self-report, census or administrative data, social media data, etc.)* *Please provide details about how you controlled for confounding variables in your analyses.*
Population characteristics	*Describe the covariate-relevant population characteristics of the human research participants (e.g. age, genotypic information, past and current diagnosis and treatment categories). If you filled out the behavioural & social sciences study design questions and have nothing to add here, write "See above."*
Recruitment	*Describe how participants were recruited. Outline any potential self-selection bias or other biases that may be present and how these are likely to impact results.*
Ethics oversight	*Identify the organization(s) that approved the study protocol.*

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	Sample size was chosen to be 1,000 questions/prompts, 500 from MedQA and 500 from HealthBench, randomly sampled with seed=62. An additional 100 de-identified clinician queries comprised the RCQ benchmark. This sample size was chosen due to the institutional constraints in vetting physician queries and ensuring HIPAA-compliance. The RCQ sample size was constrained by the manual querying required for clinical tools lacking API access. Each RCQ response was evaluated by three randomly assigned clinicians, producing 1,800 model-question annotations across 12 raters. Sample sizes for the HealthBench and MedQA experiments were chosen for practical considerations and are consistent with prior LLM evaluation studies. For a binomial outcome such as accuracy, a sample size of 500 yields a worst-case approximate 95% margin of error of ±4.4 percentage points, and approximately ±3.5 percentage points when true accuracy is near 80%, providing reasonably precise model-level estimates while keeping inference cost tractable.
Data exclusions	In the RCQ evaluation, 32 model-question pairs were excluded because raters flagged the response as a refusal. Refusals were manually

Data exclusions	verified before removal. All ratings associated with excluded pairs were discarded, yielding 568 scored responses and 1,704 individual ratings. No exclusions were applied to MedQA or HealthBench evaluations.
Replication	Frontier LLM outputs were generated with deterministic parameters (temperature=0.0, seed=62) to ensure reproducibility. MedQA scoring used dual extraction (GPT-4.1 and regex) with manual verification of disagreements. HealthBench was graded by panel-majority voting across three LLM judges from separate model families. Each RCQ response was independently scored by three blinded clinicians. Sensitivity analyses (Kruskal-Wallis tests on all 568 items; linear mixed models with random rater intercepts) were concordant with primary analyses. We confirm that all planned evaluation runs and reproducibility checks were completed successfully; no failed replication attempts were excluded. Model refusals were handled according to the prespecified analysis plan: for MedQA and HealthBench, refusals or non-answers were treated as incorrect/zero-scoring responses. For the RCQ clinician-review analysis, refusals were excluded from the ordinal rating analysis because no substantive clinical answer was available to rate, and refusal rates were reported and analyzed for the RCQ benchmark.
Randomization	For the automated benchmarks (MedQA, HealthBench), each model was evaluated on the same randomly sampled item sets, so randomization of model-to-item assignment was not applicable. For the RCQ clinician evaluation, raters were randomly assigned to model-question pairs such that three raters scored each pair. No rater was guaranteed all six model responses for a given question.
Blinding	For the automated benchmarks, blinding was not applicable as scoring was algorithmic. For the RCQ evaluation, all 12 clinician raters were blinded to model identity throughout the evaluation process. Responses were presented without any indication of which model generated them, and rater-to-item assignment was randomized.

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
☒	☐ Plants

Methods

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

Plants

Seed stocks	*Report on the source of all seed stocks or other plant material used. If applicable, state the seed stock centre and catalogue number. If plant specimens were collected from the field, describe the collection location, date and sampling procedures.*
Novel plant genotypes	*Describe the methods by which all novel plant genotypes were produced. This includes those generated by transgenic approaches, gene editing, chemical/radiation-based mutagenesis and hybridization. For transgenic lines, describe the transformation method, the number of independent lines analyzed and the generation upon which experiments were performed. For gene-edited lines, describe the editor used, the endogenous sequence targeted for editing, the targeting guide RNA sequence (if applicable) and how the editor was applied.*
Authentication	*Describe any authentication procedures for each seed stock used or novel genotype generated. Describe any experiments used to assess the effect of a mutation and, where applicable, how potential secondary effects (e.g. second site T-DNA insertions, mosiacism, off-target gene editing) were examined.*