nature medicine

Article

https://doi.org/10.1038/s41591-025-04151-2

Holistic evaluation of large language models for medical tasks with MedHELM

Received: 3 June 2025

Accepted: 24 November 2025

Published online: 20 January 2026

Check for updates

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

While large language models (LLMs) achieve near-perfect scores on medical licensing exams, these evaluations inadequately reflect the complexity and diversity of real-world clinical practice. Here we introduce MedHELM, an extensible evaluation framework with three contributions. First, a clinician-validated taxonomy organizing medical AI applications into five categories that mirror real clinical tasks—clinical decision support (diagnostic decisions, treatment planning), clinical note generation (visit documentation, procedure reports), patient communication (education materials, care instructions), medical research (literature analysis, clinical data analysis) and administration (scheduling, workflow coordination). These encompass 22 subcategories and 121 specific tasks reflecting daily medical practice. Second, a comprehensive benchmark suite of 37 evaluations covering all subcategories. Third, systematic comparison of nine frontier LLMs—Claude 3.5 Sonnet, Claude 3.7 Sonnet, DeepSeek R1, Gemini 1.5 Pro, Gemini 2.0 Flash, GPT-4o, GPT-4o mini, Llama 3.3 and o3-mini—using an automated LLM-jury evaluation method. Our LLM-jury uses multiple AI evaluators to assess model outputs against expert-defined criteria. Advanced reasoning models (DeepSeek R1, o3-mini) demonstrated superior performance with win rates of 66%, although Claude 3.5 Sonnet achieved comparable results at 15% lower computational cost. These results not only highlight current model capabilities but also demonstrate how MedHELM could enable evidence-based selection of medical AI systems for healthcare applications.

LLMs have shown impressive performance on medical knowledge benchmarks, achieving ~99% accuracy on standardized exams like MedQA[1]. This has sparked interest in deploying them in healthcare settings: supporting clinical decision-making such as diagnosis and treatment[2], optimizing clinical workflows including documentation and scheduling[3] and enhancing patient education and communication[4].

However, there is a large gap between performance on medical knowledge benchmarks and readiness for real-world deployment due to three key limitations in these existing benchmarks[5]: (1) Benchmark questions do not match real-world settings. Existing benchmarks rely on synthetic vignettes or narrowly scoped exam questions, failing to capture key aspects of real diagnostic processes such as extracting relevant details from patient records[6,7]. (2) Limited use of real-world data. Only 5% of LLM evaluations use real-world electronic health record (EHR) data[8]. EHRs contain ambiguities, inconsistencies and domain-specific shorthand that synthetic data likely cannot replicate. (3) Limited task diversity. While medical licensing exams and diagnostic tasks are prioritized in 64% of LLM healthcare evaluations[8], many administrative and operational tasks that represent significant human bottlenecks are often overlooked. These include time-intensive administrative tasks (for example, generating prior authorization letters, identifying billing codes), clinical documentation (for example, writing progress notes or discharge instructions) and patient communication (for example, asynchronous messaging through electronic patient portals)[9].

e-mail: suhana@stanford.edu

https://doi.org/10.1038/s41591-025-04151-2

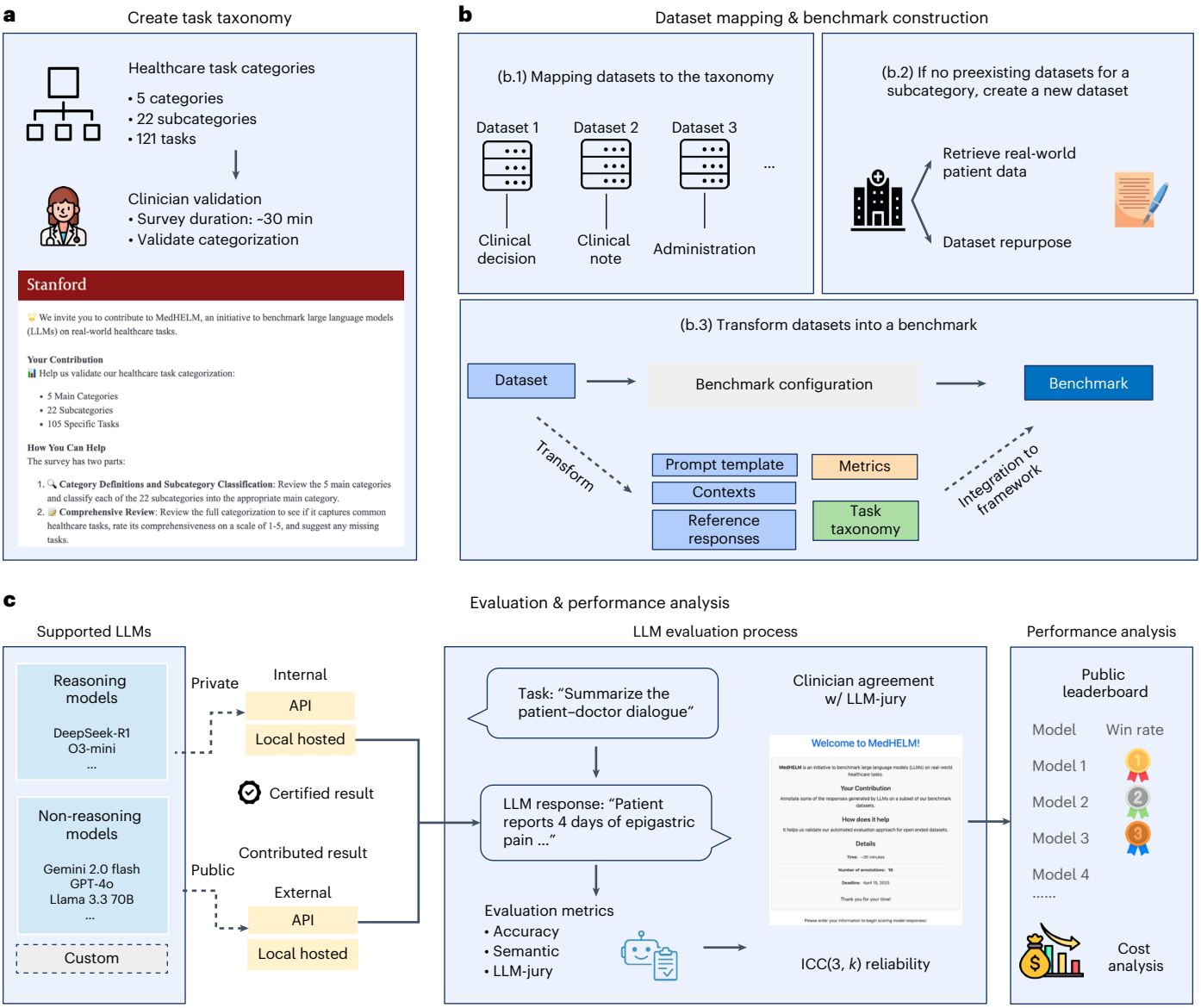


Fig. 1 | Overview of the MedHELM framework. a, A clinician-validated taxonomy organizing 121 medical tasks into 5 categories and 22 subcategories. **b**, A suite of benchmarks that map existing benchmarks to this taxonomy and introduces new benchmarks for complete coverage. **c**, An evaluation comparing reasoning and non-reasoning LLMs, with model rankings, LLM-jury-based evaluation of open-ended benchmarks and cost–performance analysis. Credits: icons in **a**–**c**, Freepik, Flaticon.

Recent work on HealthBench[10] has advanced the evaluation of LLMs in medicine by scoring 5,000 single-turn, free-text dialogues in which the model acts independently, much like a direct-to-patient advice line, without follow-up questions or human oversight. Its physician-authored rubrics reward exhaustive, risk-averse responses that maximize safety and completeness, offering a valuable stress test for fully autonomous chatbots. However, this design does not capture the iterative, context-aware interactions clinicians expect from an assistive copilot, nor does it assess performance on structured tasks that dominate everyday workflows (for example, order review, note generation, literature summarization). Furthermore, existing benchmarks including HealthBench suffer from limited model diversity in their evaluations, focusing primarily on GPT-family models rather than providing systematic comparisons across the growing landscape of LLMs. This narrow evaluation scope, combined with the lack of comprehensive task coverage, creates substantial gaps in our understanding of LLM readiness for real-world medical deployment.

To address these limitations, we draw inspiration from the Holistic Evaluation of Language Models (HELM) project[11,12], which has become the gold standard for standardized cross-domain LLM evaluation, assessing dozens of models across hundreds of scenarios and serving thousands of users through its comprehensive leaderboard. Building on this foundation, we introduce MedHELM (holistic evaluation of LLMs for medical tasks), an extensible evaluation framework for assessing LLM performance in completing medical tasks (Fig. 1). MedHELM encompasses both clinical care and broader healthcare operations across five categories: clinical decision support, clinical note generation, patient communication and education, medical research assistance, and administration and workflow, covering tasks performed by healthcare professionals, administrators, researchers and support staff throughout the healthcare ecosystem.

Using MedHELM we evaluate 9 LLMs using 37 distinct benchmarks covering all 22 subcategories of medical tasks, focusing on clinicians' day-to-day activities beyond just taking licensing exams. We assess

https://doi.org/10.1038/s41591-025-04151-2

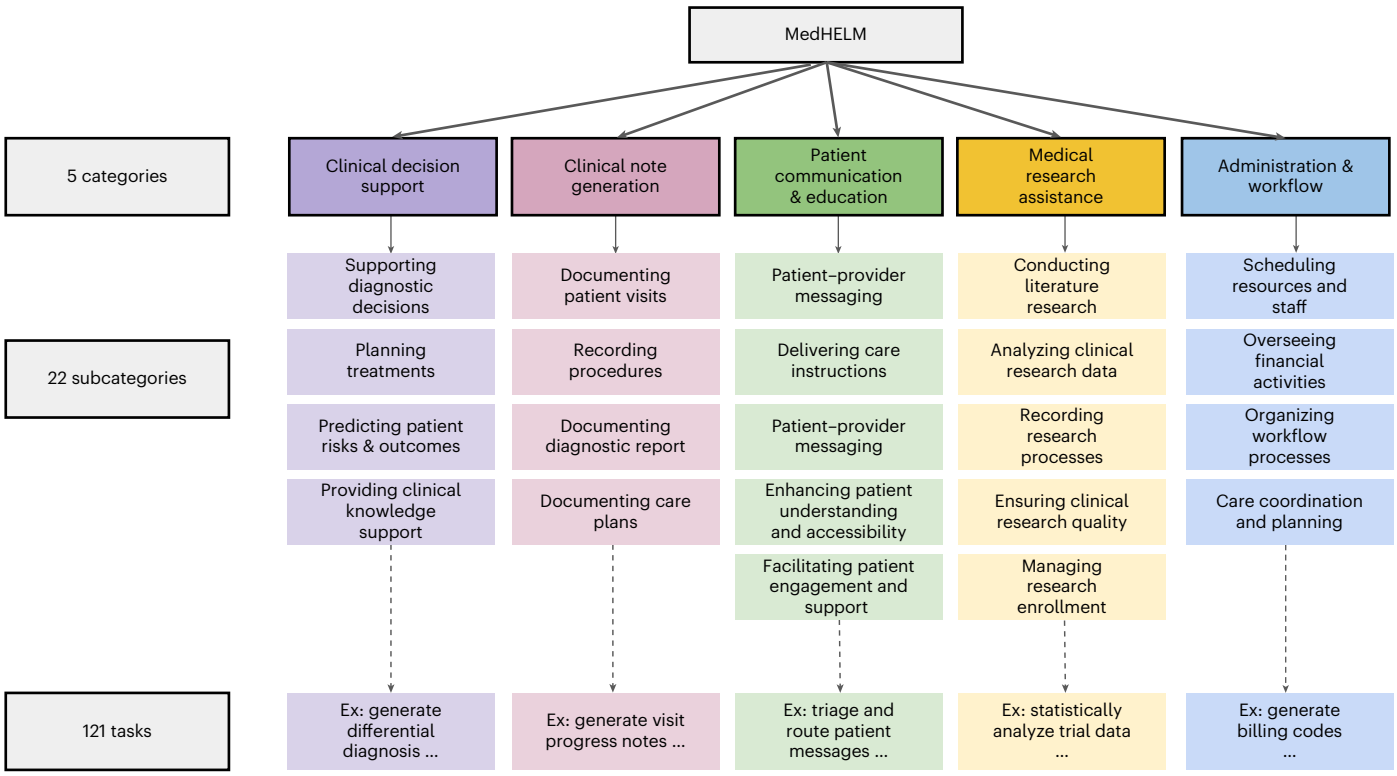


Fig. 2 | Overview of the MedHELM taxonomy. The taxonomy includes 5 main categories, 22 subcategories and 121 tasks.

performance using benchmark-appropriate metrics: exact match for closed-ended benchmarks, and for open-ended benchmarks, we use LLM-as-a-judge evaluation, a methodology where language models serve as automated evaluators using structured rubrics. Specifically, we use an LLM-jury approach where three LLMs evaluate responses using tailored rubrics that we validate by demonstrating agreement with clinician ratings, as well as estimated computational cost to provide practical deployment insights.

Our primary contributions are:

- **Clinician-validated taxonomy:** A 5-category, 22-subcategory, 121-task taxonomy developed with 29 clinicians.
- **A benchmark suite with full taxonomy coverage:** A collection of 37 benchmarks spanning all 22 subcategories of medical tasks. This includes 19 existing benchmarks, 5 reformulated benchmarks based on existing datasets and 13 new benchmarks. For privacy and regulatory compliance, as well as to prevent inclusion in LLM training data, 14 datasets are not publicly released.
- **Comparative evaluation of models along with cost–performance analysis:** A systematic evaluation shows that reasoning models achieve the highest overall performance.

The MedHELM framework addresses a critical need in the testing and evaluation of AI for medical use by providing consistent, real-world evaluation standards for medical (and healthcare) applications of LLMs. This framework benefits three key stakeholder groups: (1) healthcare systems evaluating LLMs for specific tasks, (2) AI developers identifying performance gaps across medical tasks, and (3) researchers developing methods to reproducibly measure LLM capabilities on medical tasks.

To foster collaborative improvement of such AI evaluation, we provide an openly accessible leaderboard (https://crfm.stanford.edu/helm/medhelm/latest/) with current benchmarking results and share the codebase (https://github.com/stanford-crfm/helm/) with documentation (https://crfm-helm.readthedocs.io/en/latest/medhelm/) for contributing new datasets, evaluation metrics and benchmarking custom models. For the private datasets, researchers can submit models for evaluation via pull requests to our GitHub repository. We evaluate submissions within our secure environment and share results on the public leaderboard, keeping the datasets private while enabling comprehensive benchmarking. Having some private datasets also provides a mechanism to estimate generalizability of evaluations made on public data to truly unseen datasets. We also maintain a public repository (https://github.com/som-shahlab/medhelm/) for all scripts and code used in data analysis and figure generation, enabling full transparency and reproducibility of all plots and results presented. By standardizing terminology and evaluation methods across the task taxonomy, MedHELM establishes a foundation for reproducible and real-world assessment of AI capabilities in medicine.

Results

Clinician validation of the taxonomy

In a structured review process, 29 clinicians from 14 specialties across 4 institutions evaluated the initial taxonomy comprising 5 categories, 21 subcategories and 98 tasks. Supplementary Tables 1 and 2 provide detailed breakdowns of participant affiliations and specialties. When asked to assign each subcategory to its appropriate top-level category, clinicians correctly matched 96.7% of subcategories to their intended categories. The clinicians rated the comprehensiveness of the proposed tasks at a mean of 4.21/5 ($n = 29$, s.d. = 0.675) and provided 107 comments with suggestions for improvement. Based on this feedback, we refined task definitions and expanded the taxonomy to its final form: 5 categories, 22 subcategories and 121 tasks. An overview of the final taxonomy is shown in Fig. 2, with the complete list provided in the Methods.

Overview of the benchmark suite

Our 37 benchmarks span all 22 subcategories, providing full coverage over categories and subcategories in our taxonomy. The benchmark

suite comprises 19 existing benchmarks, 5 reformulated benchmarks derived from previously unevaluated medical datasets and 13 new benchmarks, of which 12 are EHR based. The suite includes 13 open-ended benchmarks (requiring free-text generation) and 22 closed-ended benchmarks (with predefined answer choices). Access levels are designated as 16 public, 7 gated (that is, requiring approval) and 14 private. Additional information regarding the benchmarks included in MedHELM can be found in Extended Data Table 1.

Clinical Decision Support is the most represented category with 12 benchmarks, followed by Patient Communication (8), Clinical Note Generation and Medical Research Assistance (6), and Administration and Workflow (5). The distribution of benchmarks across subcategories is uneven, with 15 subcategories containing a single benchmark and the remaining 7 subcategories containing between 2 and 7 benchmarks each.

Model evaluation and cost–performance analysis

Comprehensive performance evaluation. *Pairwise win-rate and average scores*. Table 1 compares models using win-rate and macro-average performance metrics (defined in the Table 1 footnote). Both DeepSeek R1 and o3-mini performed the best, winning each 66% of head-to-head comparisons with macro-averages of 0.75 and 0.77, and low win standard deviations of 0.11 and 0.15.

Both the Claude 3.7 Sonnet and Claude 3.5 Sonnet models achieved win rates of 63% with similar macro-averages of 0.74 and 0.73, respectively, demonstrating consistent performance across our benchmark suite. GPT-4o achieved a 58% win rate, while Gemini 2.0 Flash (43%) and GPT-4o mini (37%) performed at moderate levels. Open-source Llama 3.3 Instruct achieved a 30% win rate, while Gemini 1.5 Pro ranked lowest with 23% wins but had the most consistent performance across benchmarks (lowest win s.d. of 0.08).

Performance by benchmark. We present every model's normalized score from each of the 37 benchmarks as a heat map in Fig. 3, where darker green indicates higher performance. Models consistently struggled with quantitative tasks: MedCalc-Bench (calculating medical values from patient notes), EHRSQL (generating SQL queries from natural language instructions for clinical research) and MIMIC-IV Billing Code (assigning 10th revision of the International Classification of Diseases (ICD-10) codes to clinical notes). Conversely, models achieved their best performance on NoteExtract (extracting specific information from clinical notes). Our statistical power analysis confirmed that the performance differences observed between models are statistically meaningful rather than due to random variation (Supplementary Table 4).

Performance by category. Figure 4 presents performance scores by our five top-level categories. Most models achieve their highest scores in clinical note generation (0.74–0.85) and patient communication and education (0.76–0.89), with moderate performance in medical research assistance (0.65–0.75) and clinical decision support (0.63–0.77) and generally lower scores in administration and workflow (0.53–0.63).

These performance variations reflect distinct task-based challenges: free-text-generation tasks leverage models' natural language strengths, while structured reasoning tasks require domain-specific knowledge integration and logical inference. These patterns have important implications for selective deployment strategies in healthcare settings (Supplementary Table 5).

Individual model performance across categories. At the model level, DeepSeek R1 and o3-mini lead across most categories, with DeepSeek R1 excelling in clinical note generation (0.85, tied with o3-mini) and patient communication and education (0.89), while o3-mini leads in clinical decision support (0.77) and medical research assistance (0.75). The Claude Sonnet series (3.5 and 3.7) demonstrates consistent performance with scores of 0.82–0.83 in clinical note generation and 0.83–0.84 in patient communication and education. The GPT series (4o and 4o mini) shows moderate consistency with scores of 0.79–0.80 in clinical note generation and 0.81–0.82 in patient communication and education. Gemini models exhibit greater variation, with scores of 0.74–0.78 in clinical note generation and 0.76–0.81 in patient communication and education, where Gemini 2.0 Flash substantially outperforms 1.5 Pro. The open-source Llama 3.3 performs well in patient communication and education (0.81) but shows the lowest score in administration and workflow (0.53), indicating areas for future improvement (Fig. 4).

Table 1 | Comparison of performance of frontier models across 37 MedHELM benchmarks, sorted by descending win rate

Model (snapshot)	Win rate↑	Win s.d.↓	Macro-avg↑	s.d.↓
DeepSeek R1	**0.66**	0.11	0.75	0.21
o3-mini (2025-01-31)	**0.66**	0.15	**0.77**	**0.18**
Claude 3.5 Sonnet (20241022)	0.63	0.15	0.74	0.20
Claude 3.7 Sonnet (20250219)	0.63	0.15	0.73	0.21
GPT-4o (2024-05-13)	0.58	0.17	0.74	**0.18**
Gemini 2.0 Flash	0.43	0.18	0.71	0.20
GPT-4o mini (2024-07-18)	0.37	0.17	0.71	0.19
Llama 3.3 Instruct (70B)	0.30	0.15	0.70	0.21
Gemini 1.5 Pro (001)	0.23	**0.08**	0.68	0.20

Bold indicates the best value in each column. 'Win rate' represents the proportion of pairwise comparisons where each model achieved superior performance across all 37 benchmarks (possible range: 0–1). 'Win s.d.' measures how consistently a model wins (lower values=more consistent). 'Macro-avg' is the average performance score across all 37 benchmarks. s.d. shows how much performance varies across different benchmarks (lower values indicate more consistent performance across benchmarks).

Performance by subcategory. More granular subcategory analysis revealed key patterns across specific medical tasks. For planning treatments, encompassing benchmarks such as MTSamples and Medec, Claude 3.7, DeepSeek R1 and o3-mini outperformed others, showcasing superior abilities in clinical summarization and procedural understanding. The domain of providing clinical knowledge support, covering fact-intensive tasks such as HeadQA, Medbullets, MedQA and MedMCQA, highlighted performance of GPT-4o, Claude 3.5, DeepSeek R1 and o3-mini, suggesting that traditional question–answering (QA)-style questions do not fully capture model capability differences. For clinical documentation tasks including DischargeMe, ACI-Bench, MIMIC-RRS and NoteExtract, Claude 3.7, DeepSeek R1 and o3-mini excelled at generating coherent clinical narratives, while other models showed limitations in producing contextually appropriate documentation. However, all models struggled with quantitative tasks such as analyzing clinical research data and predicting patient risks and outcomes, highlighting current limitations in complex, domain-specific calculations (Extended Data Fig. 1).

Medical versus general performance. Comparing MedHELM rankings with established benchmarks (LM Arena, HELM) revealed significant domain-specific shifts. While reasoning models like DeepSeek R1 maintained top-tier performance across domains (93rd to 100th percentile), some general-purpose models showed concerning degradation in medical tasks; for example, Gemini 2.0 Flash dropped 42 percentile points (86th to 44th) and GPT-4o fell 24 points (80th to 56th). These clinically meaningful shifts validate that general benchmark performance cannot reliably predict medical task capabilities (Extended Data Table 2).

https://doi.org/10.1038/s41591-025-04151-2

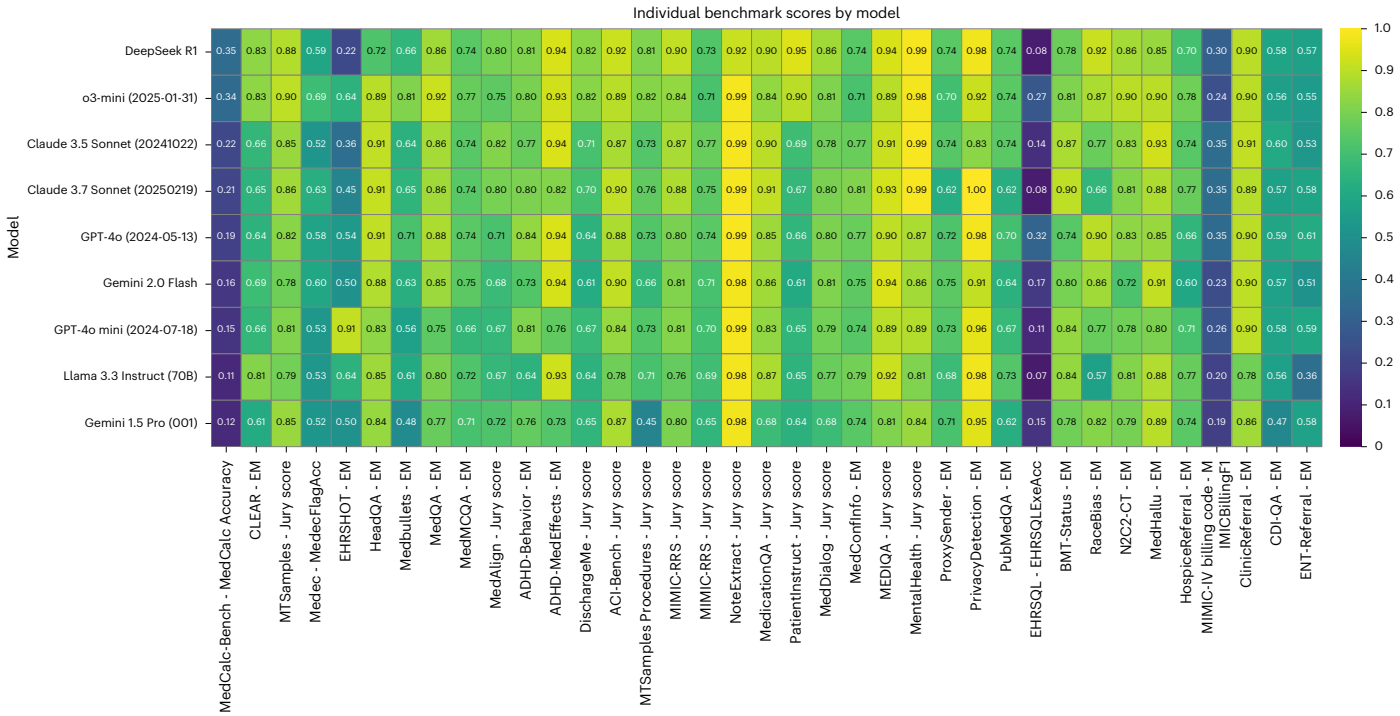


Fig. 3 | Model performance across benchmarks. Heat map of normalized scores (0–1) for each model (rows) across 37 benchmarks (columns). Scores are normalized for visualization purposes; the official leaderboard reports the original (unnormalized) scores. Dark green indicates high performance; dark red indicates low performance. The statistical significance of model differences varies by benchmark, with an analysis of minimum detectable effects shown in the Methods. Metrics include EM as exact match, Jury Score as the average normalized score from three frontier LLMs, MedCalc Accuracy as exact match or thresholded match depending on question type, MedFlagAcc as binary accuracy for detecting presence of errors, EHRSQLExeAcc as execution accuracy of generated code against a target output, and MIMICBillingF1 as the F1 score on ICD-10 codes extracted from a medical note.

Resource-efficient models. To assess deployment feasibility in resource-constrained settings, we evaluated three smaller open-source models (Qwen-2.5-7B-instruct, Phi-3.5-mini-instruct, MedGemma-4b-it) on publicly accessible benchmarks (Extended Data Table 3). While these models achieved reasonable performance on some text-generation tasks, they showed significant deficits in specialized medical reasoning, with particularly poor performance on medical calculations (MedCalc-Bench: 0.01–0.091) and clinical knowledge assessments (Medbullets: 0.192–0.406). These results highlight the trade-offs between computational efficiency and medical task performance in resource-constrained deployment scenarios.

Human performance baselines. Several MedHELM benchmarks include reported estimates of human performance from their original publications, serving as reference points for interpreting model results. Because human evaluation protocols differ across studies in task definitions, participant expertise and scoring methods, these estimates are not directly comparable (Supplementary Table 6).

Evaluation of open-ended benchmarks. For our 13 open-ended benchmarks that require free-text generation (such as clinical note writing and patient communication), we developed an LLM-jury evaluation approach. Traditional automated metrics like ROUGE-L and BERTScore only measure lexical similarity between model outputs and reference answers, but cannot assess the medical accuracy, clinical appropriateness and safety considerations that are critical for healthcare applications.

Validation against clinician judgment. To validate our LLM-jury method, we recruited 20 practicing clinicians to independently rate a subset of model outputs using the same three criteria our jury uses: accuracy, completeness and clarity. We collected ratings on 56 instances in total—31 from ACI-Bench (clinical note generation) and 25 from MEDIQA (patient QA)—and compared these clinician scores to our LLM-jury's aggregated ratings.

The LLM-jury achieved an intraclass correlation coefficient (ICC) of 0.47 with clinician ratings, which notably outperformed both the average agreement between clinicians themselves (ICC = 0.43) and standard automated metrics including ROUGE-L (ICC = 0.36) and BERTScore-F (ICC = 0.44; Extended Data Table 4). While an ICC of 0.47 may appear modest, it matched the level of agreement observed between human medical evaluators, reflecting the inherent subjectivity in clinical assessment. Importantly, unlike conventional automated metrics that only capture surface-level text similarity, our LLM-jury evaluates the medical content, clinical reasoning and safety implications that matter most for healthcare applications.

Robustness testing. We conducted two comprehensive robustness analyses to ensure our evaluation approach was reliable and stable. First, we tested whether our methodology could handle imperfections in reference answers. We applied automated filtering to identify potentially problematic gold-standard responses in our five reformulated benchmarks and found that three benchmarks had fewer than 5% of problematic responses. For the remaining two benchmarks (MIMIC-RRS and MTSamples Procedures), our LLM-jury proved remarkably robust, showing consistent mean scores between answers flagged as potentially problematic and those deemed acceptable (Supplementary Fig. 1). This stability demonstrates that our evaluation approach does not rely heavily on perfect reference answers.

Second, we assessed whether our results depended on the specific models chosen for our three-judge jury. We tested all seven possible three-judge combinations from our pool of evaluation models and

https://doi.org/10.1038/s41591-025-04151-2

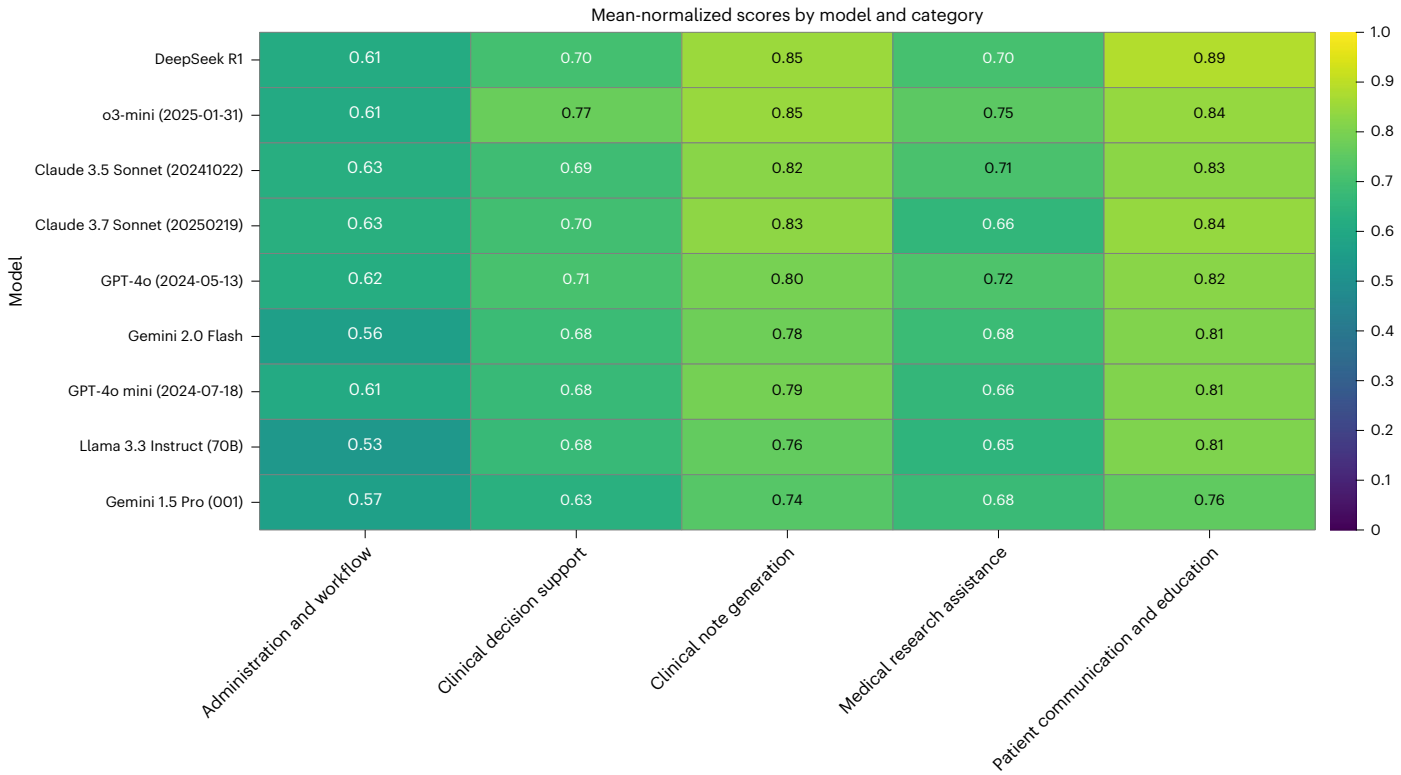


Fig. 4 | Model performance across MedHELM categories. Mean normalized scores (0–1 scale) across the five categories for all evaluated models. Darker green represents higher scores. Models are ordered by mean win rate from top (highest) to bottom (lowest), while categories are arranged left to right.

found strong consistency across compositions, with average correlations exceeding 0.85 and 100% agreement in identifying the top and bottom performing models for 12 of 13 datasets (Extended Data Table 5). Individual judge ratings remained consistent across all evaluation criteria—accuracy scores ranged from 3.910 to 4.355, completeness from 3.388 to 3.816 and clarity from 4.744 to 4.969—confirming that our jury's effectiveness does not depend on specific model choices (Extended Data Table 6).

Performance advantages in specific scenarios. Detailed analysis revealed that while overall performance comparisons between LLM-jury and automated metrics show modest differences, our approach provides clear advantages in specific, clinically relevant scenarios. The LLM-jury's strengths are most pronounced when evaluating moderate-quality responses (rated 3–4 on our five-point scale), where it outperformed ROUGE-L 62.5% of the time and showed stronger correlation with human judgment (0.414 versus 0.386). For high-quality responses, the LLM-jury outperformed BERTScore in 60.8% of cases, demonstrating superior ability to capture the subtle quality distinctions that matter most to human evaluators.

Rather than showing universal superiority, our analysis revealed complementary strengths: the LLM-jury substantially outperformed automated metrics in 32–34% of cases, while automated metrics showed advantages in 27–29% of cases. This balanced pattern suggests that LLM-jury evaluation provides the most value when human-like judgment is needed for nuanced assessment scenarios, particularly when capturing fine-grained quality differences that automated metrics cannot detect.

Cost–performance analysis. We estimated the cost (USD) of evaluating each model based on publicly listed pricing as of 12 May 2025, using the total input tokens and maximum output tokens consumed during benchmark runs and LLM-jury evaluation (Extended Data Table 7). These costs are an upper-bound estimate, since models may generate fewer tokens than the maximum allowed output. We plotted the mean win rate against the cost (Fig. 5), with a detailed inference costs in Supplementary Table 5. As expected, non-reasoning models—GPT-4o mini (US$805) and Gemini 2.0 Flash (US$815)—incurred the lowest costs and achieved win rates of 0.37 and 0.43, respectively. Open-source Llama 3.3 Instruct (US$940) had a 0.30 win rate, while Gemini 1.5 Pro (US$1,132) reached 0.23. Reasoning models—DeepSeek R1 (US$1,850) and o3-mini (US$1,761)—incurred higher costs, with the same win rate of 0.66. Claude 3.5 Sonnet (US$1,572) and Claude 3.7 Sonnet (US$1,538) provide a good cost–performance balance, achieving a 0.63 win rate at reduced costs.

Discussion

We present a framework for assessing LLM performance for real-world medical tasks. Our clinician-validated taxonomy provides a structure for summarizing models' strengths and limitations across medical tasks. High clinician agreement (96.7%) in assigning subcategories to categories suggests the taxonomy effectively captures how healthcare professionals conceptualize their work.

Our clinically oriented task taxonomy differentiates MedHELM from other large-scale benchmarking efforts. While BRIDGE evaluates 87 tasks across 52 models, it does not use real-world patient records[13]. BigBIO is a community library of over 126 biomedical natural language processing (NLP) datasets covering 12 task categories and more than 10 languages, but lacks clinical tasks or patient records[14]. In contrast, MedHELM organizes tasks by clinical function rather than NLP task type (for example, NLI, NER), incorporates real-world EHR data and includes cost–performance analysis absent from these frameworks.

Our benchmark suite reveals nuances in model capabilities that are not seen with current medical knowledge benchmarks alone. The higher performance in communication tasks than administrative ones may stem from administrative workflows using data not seen

https://doi.org/10.1038/s41591-025-04151-2

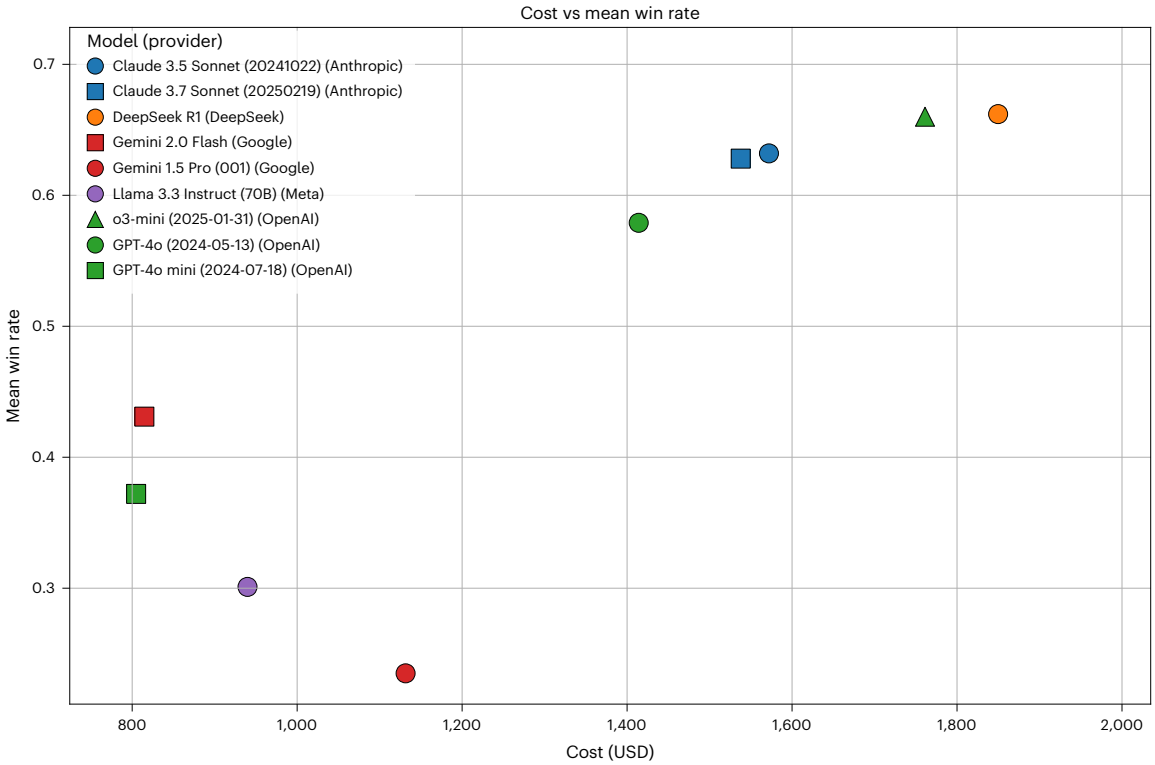


Fig. 5 | Performance versus computational cost. Scatterplot of mean win rate (*y* axis) versus estimated computational cost (*x* axis) for each of the nine models across 37 benchmarks. Each point represents a model, with the position indicating the relationship between performance (*y* axis) and total cost of evaluation, including benchmark runs and evaluation by LLM-jury (*x* axis). Costs represent upper-bound estimates based on maximum output token usage.

during training, warranting caution in healthcare AI implementation for back-office tasks without quantifying task-specific performance. This framework addresses the primary limitations of current benchmarks, such as the lack of using real-world data, the use of evaluation setups that do not match real-world settings and a limited task diversity.

Our LLM-jury-based approach addresses a critical gap in current evaluation approaches. By beating clinician–clinician agreement, this approach enables scalable evaluation of open-ended model outputs without requiring extensive clinician time, a scarce and expensive resource.

The cost–performance trade-off shows that while reasoning models have superior performance, their substantially higher costs may not justify their deployment for all tasks. In a resource-constrained setting, models such as Claude 3.5 Sonnet offer a balance, achieving a win rate of 0.63 at a lower cost.

Several limitations remain. While our LLM-jury approach was validated on only two benchmarks, expanding clinician annotations across more benchmarks would strengthen the clinician–LLM agreement estimates. In addition, the uneven distribution of benchmarks across subcategories (15 of 22 contain only one benchmark) limits our ability to draw robust performance conclusions in underrepresented areas. Moreover, our current rubrics operate at the benchmark level, but instance-level rubrics could provide better evaluation, particularly for subjective or context-dependent medical tasks where gold-standard responses may not exist. Such approaches would further scale LLM-jury evaluation beyond reliance on gold-standard responses[15]. To enhance reproducibility, future work should explore synthetic dataset generation approaches that could replicate the characteristics of our private benchmarks, enabling broader community access.

Our benchmark evaluations are done in a private cloud tenant and cost US$805–1,850 per model, which can be prohibitive for resource-limited settings. Using smaller, locally deployable LLMs can be a path for resource-constrained healthcare deployment as well as to mitigate privacy concerns of processing patient data through commercial APIs. We evaluated several smaller models as outlined in the 'Results'. However, our non-cloud infrastructure precluded comprehensive benchmarking. Although models such as Phi-3.5-mini-instruct can fit on such hardware, the memory requirements for the benchmark's long patient timelines exceeded available memory. In addition, current smaller models face significant limitations on complex clinical tasks; therefore, devising ways of efficient benchmarking with local models on longitudinal datasets remains a fruitful area of future work.

Administration and workflow emerged as the weakest performance area for all models. Understanding the underlying causes of this poor performance, whether stemming from training data limitations, task complexity or distributional shifts, is essential for safe deployment in healthcare operations. Our evaluation includes several clinical safety dimensions through dedicated benchmarks including hallucination detection, medical error identification, dosage accuracy and bias assessment. However, we do not systematically assess model calibration or uncertainty quantification, which are crucial for clinical deployment.

In addition, while human and task-specific ML baselines are important comparators, our focus was to establish a comparative LLM benchmark across diverse medical tasks. This design isolates model-to-model differences and identifies the best-performing off-the-shelf systems, providing a necessary foundation for subsequent work that will evaluate LLMs in augmentative settings with human experts or against specialized ML models.

We believe that by releasing a shared, task-oriented benchmark, future work can incorporate explicit calibration metrics, systematic harmful recommendation assessment, retrieval-augmented evaluation as a potential direction to examine how retrieval integration influences factual accuracy and reasoning in long-context medical tasks, and evaluation of models' ability to appropriately defer to human expertise in high-uncertainty scenarios. MedHELM is designed as an extensible platform enabling healthcare institutions and researchers to contribute

specialized benchmarks and clinical metrics. This community-driven approach ensures the framework evolves beyond current limitations to capture the full spectrum of clinically relevant evaluation dimensions essential for real-world deployment.

In conclusion, MedHELM provides a comprehensive framework for assessing LLM performance across real-world medical tasks through our clinician-validated taxonomy spanning 5 categories, 22 subcategories and 121 tasks. Our benchmark suite of 37 datasets reveals that most models perform best in clinical note generation (0.74–0.85) and patient communication and education (0.76–0.89), moderately in medical research assistance (0.65–0.75) and clinical decision support (0.63–0.77), and worst in administration and workflow (0.53–0.63). Reasoning models DeepSeek R1 and o3-mini led overall with win rates of 0.66 and 0.66, respectively, although Claude models offer competitive performance (win rate of 0.63) at lower computational cost. Through our public leaderboard (https://crfm.stanford.edu/helm/medhelm/latest/) and shared codebase (https://github.com/stanford-crfm/helm), MedHELM establishes infrastructure for ongoing collaborative assessment as models evolve. We envision MedHELM evolving through community contributions that expand coverage to underrepresented clinical workflows and enable more granular task-level assessment, advancing medical AI evaluation that better reflects the complexity of real-world medical practice.

Ethics

This research involved multiple components requiring ethical consideration. The clinician validation study, which engaged 29 practicing physicians across 14 medical specialties to validate our taxonomy and evaluate model outputs, adhered to the principles outlined in the Declaration of Helsinki. Informed consent was obtained from each participating clinician before their involvement in the survey and rating activities. The benchmark evaluation component utilized both public datasets and private medical data obtained through partnership with Stanford Healthcare. All experiments involving patient data were conducted on a PHI-compliant shared cluster maintaining full HIPAA compliance. The use of retrospective, de-identified medical records fell within approved institutional guidelines and did not require additional institutional review board oversight. All data handling, model evaluation and clinician interaction protocols were designed to protect patient privacy and maintain the confidentiality of medical information throughout the research process.

Reporting summary

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

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/s41591-025-04151-2.

References

1. Papers with Code. Question answering on MedQA (USMLE). https://paperswithcode.com/sota/question-answering-on-medqa-usmle (2024).
2. Khosravi, M., Zare, Z., Mojtabaeian, S. M. & Izadi, R. Artificial intelligence and decision-making in healthcare: a thematic analysis of a systematic review of reviews. *Health Serv. Res. Manag. Epidemiol.* **11**, 23333928241234863 (2024).
3. Nath, D. Artificial intelligence (AI) will transform the clinical workflow with the next-generation technology. *HealthTech Magazines* https://www.healthtechmagazines.com/artificial-intelligence-ai-will-transform-the-clinical-workflow-with-the-next-generation-technology/ (2024).
4. Carl, N. et al. Evaluating interactions of patients with large language models for medical information. *BJU Int.* **135**, 1010–1017 (2025).
5. 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).
6. Raji, I. D., Daneshjou, R. & Alsentzer, E. It's time to bench the medical exam benchmark. *NEJM AI* **2**, AIe2401235 (2025).
7. 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* **174**, 248–260 (PMLR, 2022).
8. Bedi, S. et al. Testing and evaluation of health care applications of large language models: a systematic review. *JAMA* **333**, 319–328 (2025).
9. Hager, P. et al. Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nat. Med.* **30**, 2613–2622 (2024).
10. Arora, R. K. et al. HealthBench: evaluating large language models towards improved human health. https://cdn.openai.com/pdf/bd7a39d5-9e9f-47b3-903c-8b847ca650c7/healthbench_paper.pdf (2025).
11. Liang, P. et al. Holistic evaluation of language models. In *Transactions on Machine Learning* https://openreview.net/pdf?id=iO4LZibEqW (2023).
12. Leaderboard overview. *LM Arena* https://lmarena.ai/leaderboard (2025).
13. Wu, J. et al. BRIDGE: benchmarking large language models for understanding real-world clinical practice text. Preprint at https://arxiv.org/abs/2504.19467 (2025).
14. Fries, J. A. et al. BigBio: a framework for data-centric biomedical natural language processing. In *36th Conference on Neural Information Processing Systems (NeurIPS 2022) Track on Datasets and Benchmarks* https://openreview.net/pdf?id=8lQDn9zTQlW (2022).
15. Croxford, E. et al. Automating evaluation of AI text generation in healthcare with a large language model (LLM)-as-a-judge. Preprint at *medRxiv* https://doi.org/10.1101/2025.04.22.25326219 (2025).

Suhana Bedi[1,7] ✉, Hejie Cui[1,7], Miguel Fuentes[1,7], Alyssa Unell[1,7], Michael Wornow[1], Juan M. Banda[2], Nikesh Kotecha[2], Timothy Keyes[2], Yifan Mai[3], Mert Oez[4], Hao Qiu[4], Shrey Jain[4], Leonardo Schettini[4], Mehr Kashyap[1], Jason Alan Fries[1], Akshay Swaminathan[1], Philip Chung[1], Fateme Nateghi Haredasht[1], Ivan Lopez[1], Asad Aali[1], Gabriel Tse[1], Ashwin Nayak[1], Shivam Vedak[1], Sneha S. Jain[1], Birju Patel[1], Oluseyi Fayanju[1], Shreya Shah[1], Ethan Goh[1], Dong-han Yao[1], Brian Soetikno[1], Eduardo Reis[1], Sergios Gatidis[1], Vasu Divi[1], Robson Capasso[1], Rachna Saralkar[1], Chia-Chun Chiang[1], Jenelle Jindal[1], Tho Pham[1], Faraz Ghoddusi[1], Steven Lin[1], Albert S. Chiou[1], Hyo Jung Hong[1], Mohana Roy[1], Michael F. Gensheimer[1], Hinesh Patel[1], Kevin Schulman[1], Dev Dash[1], Danton Char[1], Lance Downing[1], Francois Grolleau[1], Kameron Black[1], Bethel Mieso[1], Aydin Zahedivash[1], Wen-wai Yim[4], Harshita Sharma[4], Tony Lee[3], Hannah Kirsch[2], Jennifer Lee[2], Nerissa Ambers[2], Carlene Lugtu[2], Aditya Sharma[2], Bilal Mawji[2], Alex Alekseyev[2], Vicky Zhou[2], Vikas Kakkar[2], Jarrod Helzer[2], Anurang Revri[2], Yair Bannett[1], Roxana Daneshjou[1], Jonathan Chen[1], Emily Alsentzer[1], Keith Morse[1], Nirmal Ravi[5], Nima Aghaeepour[1], Vanessa Kennedy[1], Akshay Chaudhari[1], Thomas Wang[1,2], Sanmi Koyejo[3,6], Matthew P. Lungren[1,4], Eric Horvitz[4,6], Percy Liang[3,6], Michael A. Pfeffer[2] & Nigam H. Shah[1,2,5]

[1]Stanford University School of Medicine, Stanford, CA, USA. [2]Stanford Health Care, Palo Alto, CA, USA. [3]Center for Research on Foundation Models (CRFM) & Department of Computer Science, Stanford University, Stanford, CA, USA. [4]Microsoft Corporation, Redmond, WA, USA. [5]eHealth Africa Clinics, Kano, Nigeria. [6]Stanford Institute for Human-Centered AI, Stanford, CA, USA. [7]These authors contributed equally: Suhana Bedi, Hejie Cui, Miguel Fuentes, Alyssa Unell. ✉e-mail: suhana@stanford.edu

Methods

Most LLM evaluations in medicine still rely on closed-form QA over exam-style datasets such as MEDQA and MEDMCQA, with only ~5% incorporating real EHR data and very few addressing free-text-generation tasks or cost-aware metrics[7,16,17].

Large-scale NLP meta-benchmarks (HELM, BIG-BENCH) demonstrate the value of task diversity and multi-metric scoring[11]. Recent medical benchmarking efforts (Extended Data Table 8) have begun addressing some limitations: HEALTHBENCH[10] incorporates physician-developed rubrics for health conversations, ClinicBench[18,19] advances multimodal evaluation, and automated frameworks explore LLM-as-a-Judge strategies for EHR summarization[19–26]. However, these frameworks typically focus on narrow clinical scenarios (1–28 datasets), rely on limited data sources or synthetic data, and assess primarily accuracy without comprehensive taxonomy validation or deployment considerations.

Our goal is to close these gaps by co-designing with clinicians, a taxonomy-guided benchmark suite that (i) spans the full spectrum of real medical work across 35 datasets and 5 healthcare domains, (ii) uses both public and private medical record data from multiple institutions, and (iii) uses evaluation protocols that align with clinician judgment while incorporating practical deployment metrics.

Development of the taxonomy of tasks

Early attempts (BIGBIO, ClinicBench) at creating a taxonomy of tasks in medical settings either harmonize datasets into broad categories without clinician input or collapse heterogeneous skills under a single 'generation' label[14,18]. To ensure our benchmarks accurately reflect the complexity of medical practice, we developed a taxonomy that mirrors how clinicians conceptualize their daily work. By defining a hierarchical structure, we guarantee that each benchmark maps to a concrete medical activity. Our taxonomy consists of three levels:

- **Category:** A broad domain of medical activity (for example, clinical decision support).
- **Subcategory:** A group of related tasks within a category (for example, supporting diagnostic decisions).
- **Task:** A discrete action taken during the delivery of medical care (for example, generate differential diagnoses).

This structure enables systematic coverage of the medical care landscape while maintaining clear boundaries between distinct activities that care providers perform.

Initial drafting. We based our taxonomy on tasks identified in a JAMA review[8]. Working with a clinician (M.K.), we reorganized these tasks into functional themes that reflect real-world activities, resulting in 98 distinct tasks organized into 21 subcategories within 5 categories:

- Clinical decision support
- Clinical documentation
- Patient communication and education
- Medical research assistance
- Administration and workflow

Extended Data Table 9 compares our comprehensive category coverage with existing medical LLM benchmarking frameworks. While previous efforts typically focus on 1–3 categories, MedHELM uniquely spans all 5 healthcare domains, with administration and workflow representing a previously unaddressed critical gap in medical AI evaluation.

Two principles guided our taxonomy development:

- **Medical relevance:** Each task maps directly to actions routinely performed by care providers.
- **Clear boundaries:** Categories and subcategories were defined to minimize overlap while preserving meaningful functional distinctions.

Validation. To validate our initial taxonomy, we designed a two-part survey completed by 29 practicing clinicians representing 14 medical specialties across 4 institutions. The survey assessed both categorical organization and real-world relevance.

In the first section, clinicians assigned each of our 21 subcategories to one of the 5 main categories. This exercise tested whether our taxonomy structure matched how clinicians naturally organize medical tasks.

In the second section, clinicians evaluated the comprehensiveness of our taxonomy on a five-point scale, where a score of 5 means that our categorization covered all routine medical tasks and a score of 1 means it covered very few. They also provided feedback through an open dialogue box where they could suggest missing tasks and recommend terminology improvements. This systematic validation approach evaluated both the taxonomy's organizational logic and its comprehensiveness in representing actual medical tasks.

Based on the comments, we refined definitions and expanded the taxonomy to have 5 categories, 22 subcategories and 121 tasks.

Complete taxonomy structure

Overview.

- **Total categories:** 5
- **Total subcategories:** 22
- **Total tasks:** 121
- **Entities per category:**
  - **Definition:** A description of the category.
  - **Inclusion criteria:** Guidelines determining which subcategories (and tasks) belong in the category.
  - **Task performer:** Individuals responsible for executing tasks under the category.
  - **Subcategories:** Each category contains multiple subcategories.
  - **Tasks:** Each subcategory comprises several tasks.

1. Clinical decision support. *Definition:*. Clinical decision support is the act of analyzing patient-specific data and providing evidence-based recommendations to clinicians to enhance diagnostic accuracy, optimize treatment options and improve patient outcomes.

Inclusion criteria: Acts that generate actionable insights based on patient data and clinical evidence.

Task performer: Clinicians (physicians, nurses and healthcare practitioners) who assess clinical information to verify their knowledge, conduct diagnostic assessments, plan treatments and take mitigating action in response to the predicted outcome to enhance patient care.

Subcategories and tasks:

- **Supporting diagnostic decisions**
  - Recognize disease patterns from symptoms/vitals/physical exams
  - Interpret functional diagnostic tests (electrocardiogram, spirometry, stress tests)
  - Generate diagnostic follow-up questions
  - Generate differential diagnoses
  - Interpret lab results and detect abnormalities
  - Detect medical image findings
  - Perform medical calculations
  - Evaluate social determinants of health
  - Track longitudinal lab trends
  - Process pre-visit intake information

- **Planning treatments**
  - Check for drug interactions
  - Match treatment protocols and screen for contraindications
  - Suggest clinical pathways

–Predict treatment response
–Make collaborative clinical decisions
–Evaluate treatment accessibility

- **Predicting patient risks and outcomes**
  –Predict patient deterioration
  –Assess hospital readmission risk
  –Model disease progression
  –Predict treatment outcomes
  –Predict adverse events
  –Triage patients based on risk prediction
  –Predict discharge readiness
  –Predict need for procedures/interventions
  –Predict need for specialist referrals
  –Manage preventive screening programs

- **Providing clinical knowledge support**
  –Apply clinical guidelines and best practices
  –Answer medical knowledge questions
  –Track protocol compliance
  –Assess clinical care quality

2. Clinical note generation. *Definition:* Clinical note generation is the act of creating and maintaining structured records of patient encounters, treatments and care plans.

Inclusion criteria: Acts that produce or modify official clinical records used for patient care documentation or provider communication.

Task performer: Healthcare providers, medical scribes and documentation specialists tasked with maintaining accurate, comprehensive clinical records.

Subcategories and tasks:

- **Documenting patient visits**
  –Generate visit progress notes
  –Generate consultation notes
  –Generate emergency department notes
  –Generate hospital admission notes
  –Generate discharge summaries
  –Synthesize problems from internal and external records
  –Create synopses of clinical documents
  –Generate multidisciplinary team assessment notes

- **Recording procedures**
  –Generate operative reports (for operating room procedures)
  –Generate procedure notes (for bedside/clinic procedures)
  –Generate specialized procedure reports (for cardiac catheterization and interventional radiology)

- **Documenting diagnostic reports**
  –Generate imaging-based diagnostic reports
  –Generate pathology reports
  –Generate diagnostic test documentation
  –Generate genomic analysis reports

- **Documenting care plans**
  –Document treatment plans
  –Generate care protocols
  –Document nursing care plans
  –Document advance care planning

3. Patient communication and education. *Definition:* Patient communication and education is the act of transmitting health information and clinical guidance to patients to enable understanding and participation in their care.

Inclusion criteria: Acts that convey health-related knowledge or respond to patient queries to support informed participation in care.

Task performer: Healthcare providers, care coordinators and patient educators responsible for engaging with patients to support their understanding and participation in care.

Subcategories and tasks:

- **Providing patient education resources**. Simplify disease information; educate on risk factors; generate prevention or treatment explanations; explain insurance and billing information.
- **Delivering personalized care instructions**. Generate medication instructions; generate pre/post-procedure guidance; generate home care guidelines; explain follow-up requirements and recovery expectations.
- **Patient–provider messaging**. Triage and route patient messages; analyze symptom reports; handle medication refill requests; process appointment requests; analyze nonurgent medical questions; identify urgent messages; generate response drafts; generate patient-friendly encounter summaries; share clinical results with patients; generate patient-requested documents.
- **Enhancing patient understanding and accessibility in health**. Generate visual aids; translate to multiple languages; make content accessible.
- **Facilitating patient engagement and support**. Generate appointment reminders and confirmations; provide preventive care notifications; track health goal progress; check care plan adherence; collect patient feedback; facilitate support group discussions; support patient counseling interactions.

4. Medical research assistance. *Definition:* Medical research assistance is the act of analyzing clinical data and literature to generate scientific evidence for advancing medical knowledge and practice.

Inclusion criteria: Acts that transform clinical data or research findings into validated scientific evidence or research protocols.

Task performer: Researchers, clinical investigators and epidemiologists involved in designing, analyzing and documenting research studies.

Subcategories and tasks:

- **Conducting literature research**. Screen systematic review literature; summarize research papers; analyze citation networks; synthesize evidence; identify research gaps.
- **Analyzing clinical research data**. Statistically analyze trial data; identify population health patterns; compare treatment effectiveness; analyze outcome measures; conduct cohort analyses; plan secondary studies and follow-ups.
- **Recording research processes**. Support protocol development; assist with grant writing; format research manuscripts; plan statistical analyses; document research results.
- **Ensuring clinical research quality**. Validate statistical methods; verify research methodologies; assess data quality and bias; process research regulatory requirements.
- **Managing research enrollment**. Screen for trial eligibility criteria; match patients to study protocols; track enrollment targets; monitor participant retention; document recruitment outcomes.

5. Administration and workflow. *Definition:* Administration and workflow is the act of orchestrating clinical operations across both inpatient and outpatient settings, encompassing the entire patient journey from scheduling through billing. This includes coordinating patients, providers and staff through clinical care settings, managing resources and ensuring efficient operational flow.

Inclusion criteria: Acts that include operational aspects of healthcare, such as resource allocation, financial processes or patient flow management.

Task performer: Administrators, office staff and billing specialists responsible for the logistical and financial coordination of healthcare services.

Subcategories and tasks:

- **Scheduling resources and staff**
  - Schedule staff
  - Manage inventory
  - Manage equipment
  - Coordinate facilities
  - Monitor institutional performance metrics
- **Overseeing financial activities**
  - Generate billing codes
  - Document billing
  - Correspond with insurers
  - Analyze revenue
  - Track operational costs
  - Calculate patient out-of-pocket costs
- **Organizing workflow processes**
  - Schedule appointments
  - Process referrals
  - Route documents
  - Process health information requests
- **Care coordination and planning**
  - Evaluate admission criteria
  - Facilitate inter-provider coordination
  - Identify appropriate post-discharge facilities
  - Manage transitional care needs

Construction of the benchmark suite

Curation of datasets. To construct a comprehensive suite of 37 benchmarks spanning our taxonomy, we used a three-tiered dataset curation strategy:

1. **Existing benchmarks**: We incorporated existing benchmarks from public or gated sources (for example, MedQA, MIMIC-IV Billing Code, ACI-Bench) to ensure broad subcategory coverage.

2. **Reformulated benchmarks**: We transformed previously unevaluated medical data collections into 'reformulated benchmarks' by applying standardized prompt templates and specifying evaluation metrics. This approach addressed subcategories where datasets existed but lacked LLM-ready evaluation benchmarks.

3. **New benchmarks**: To address the underrepresentation of 'administration and workflow', we partnered with Stanford Healthcare to curate private datasets for tasks that are routinely done in health systems but for which benchmark datasets do not exist (for example, referral triage, scheduling). We also developed private benchmarks where healthcare partnerships provided access to authentic clinical data from global contexts (for example, mental health counseling from India, clinical care plans from Nigeria).

Safety-focused evaluation: Our benchmark suite includes clinical safety assessment through two complementary approaches: (1) safety detection benchmarks that evaluate models' ability to identify problematic content (MedHallu for hallucination detection, Medec for medical error identification, RaceBias for bias detection), and (2) safety performance benchmarks that measure whether models make errors with safety implications (MedCalc-Bench for dosage accuracy).

Note on categorization: Some benchmarks like MedAlign contain tasks that could fit multiple categories. We classify based on functional purpose rather than surface operations. For example, while MedAlign contains summarization tasks, we categorize it under 'clinical decision support' because these tasks specifically extract patient information to inform diagnostic and treatment decisions (for example, 'Summarize this patient's asthma care plan including diagnostic testing and treatments') rather than generating clinical documentation for the medical record.

Each benchmark is labeled by 'source type' (existing/reformulated/new) and 'access level' (public/gated/private), with the provenance documented in the MedHELM repository. Our framework intentionally combines 23 public/gated benchmarks with 14 private benchmarks to address training data contamination, where models may be optimized for public evaluation datasets[27]. Our approach follows established machine learning practices of maintaining private hold-out test sets for unbiased performance measurement[28]. The extensible framework enables healthcare systems to contribute additional private benchmarks, creating a distributed evaluation ecosystem that provides comprehensive assessment coverage while preventing benchmark gaming by model developers.

Specification of prompts and metrics. To transform each curated dataset into a MedHELM benchmark, we defined four components for every item in the dataset—three mandatory and one optional:

- **Context:** the raw input presented to the LLM (for example, a clinical note, patient message or structured EHR record).
- **Prompt:** a standardized instruction template to elicit consistent, task-appropriate responses (for example, 'Answer in 2–3 sentences' for open-ended summaries, or MCQ framing for multiple-choice questions).
- **Evaluation metric:** a prespecified scoring method matched to the task type:
  - 'Exact match accuracy' for single-token or numeric outputs (for example, selecting the correct option in MedQA).
  - 'Micro-F1' for multi-label classification tasks (for example, ICD-10 code assignment).
  - 'LLM-jury ensemble' for open-ended text generation: we use a three-model Likert-scale protocol assessing medical accuracy, completeness and clarity, and secondary metrics, ROUGE and BERTScore, to capture lexical and semantic overlap.
- **Gold-standard response (optional):** the reference output (numeric result, classification label or sample text) against which the model's response is scored (for example, '4' in response to 'What's a patient's HAS-BLED score?'). While most benchmarks include a gold standard, our framework accommodates benchmarks without one, such as NoteExtract, providing flexibility for future evaluation needs.

All prompts are task specific, not model specific. Each benchmark uses a single standardized prompt applied uniformly across all models, ensuring fair comparison and scalability. For LLM-jury evaluation, all three jury members use identical task-specific prompts (Supplementary Fig. 2).

We found quality issues in some of the gold-standard responses for reformulated benchmarks. For example, gold-standard responses in the MIMIC-RRS benchmark occasionally contained information from a patient's EHR that was not passed into the model's context. To assess the impact of these low-quality gold-standard responses, we conducted a sensitivity analysis by filtering 'problematic' gold-standard responses using an LLM judge and recalculating metrics (Supplementary Fig. 1). Model rankings remained unchanged, as instances with 'problematic' and 'non-problematic' gold-standard responses received similar jury scores. This stability exists because our LLM-jury (prompt in Supplementary Fig. 2) uses gold-standard responses only when needed.

Detailed benchmark specifications and implementation

This section provides comprehensive specifications for each of the 37 benchmarks in our MedHELM evaluation suite. For each benchmark, we detail the context provided to models, the specific prompts used, the evaluation metrics applied and representative examples of gold-standard responses. These standardized specifications ensure

reproducible evaluation and enable researchers to understand the precise requirements of each medical task assessed in our framework.

MedCalc-Bench. *Description:* MedCalc-Bench is a benchmark designed to evaluate models on their ability to compute clinically relevant values from patient notes. Each instance consists of a clinical note describing the patient's condition, a diagnostic question targeting a specific medical value and a ground-truth response.

Category: Clinical decision support
Subcategory: Supporting diagnostic decisions
Context:
The context provides the clinical background needed to answer a diagnostic or prognostic question. It typically includes a progress note summarizing the patient's condition and relevant clinical events.
Example:

```
Patient note: A 70-year-old female was rushed into the ICU due to respiratory distress, following which she was promptly put on mechanical ventilation. Her delivered oxygen fell to 51% FiO2...Question: What is the patient's Sequential Organ Failure Assessment (SOFA) Score?
```

Prompt:
The prompt defines the specific computational task the model must perform based on the patient note and the question.
Example:

```
Given a patient note and a clinical question, compute the requested medical value.
<context>
Answer only the requested quantity without units. No explanation needed:
```

Evaluation metric:
MedCalc Accuracy: Exact match for discrete categories (for example, risk, severity and diagnosis); range-based comparison for continuous variables.
Gold-standard response:
The gold standard is the correct reference value for the given input. It excludes units and explanation.
Example:

```
9
```

CLEAR. *Description:* CLEAR is a benchmark designed to evaluate models on their ability to detect medical conditions from patient notes using categorical responses. Each instance consists of a clinical note and a target condition, requiring the model to classify the patient's history as affirmative, negative or uncertain.

Category: Clinical decision support
Subcategory: Supporting diagnostic decisions
Context:
The context consists of a clinical note documenting patient history, from which the presence or absence of a specific medical condition must be inferred.
Example:

```
History reviewed. No pertinent family history.
Social history
Occupational history: not on file.
Social history main topics:
Smoking status:
...
```

Prompt:
The prompt defines the task of inferring the patient's medical condition history from a clinical note. The model must choose one of three categorical answers: A (yes), B (no) or C (uncertain).
Example:

```
You are a medical assistant reviewing patient notes. Determine whether the patient has a history of suicidal behavior.
Original clinical note:
<context>
Answer:
A. Has a history of suicidal behavior
B. Does not have a history of suicidal behavior
C. Uncertain
Respond only with 'A', 'B' or 'C'. Do not add any other text, punctuation or symbols:
```

Evaluation metric:
Exact match: The model's response must exactly match the gold-standard label (A, B or C) for the given condition and clinical note.
Gold-standard response:
The gold standard represents the correct classification of the patient's condition status for the task.
Example:

```
B
```

MTSamples Replicate. *Description:* MTSamples Replicate is a benchmark that provides transcribed medical reports from various specialties. It is used to evaluate a model's ability to generate clinically appropriate treatment plans based on unstructured patient documentation.

Category: Clinical decision support
Subcategory: Planning treatments
Context:
The context is a medical transcription note documenting the patient's clinical history and current assessment.
Example:

```
Medical specialty: Pediatrics - neonatal
Sample name: 1-year-old Exam - H&P
Description: Health maintenance exam for 1-year-old female
(Medical Transcription Sample Report)
...
```

Prompt:
The prompt instructs the model to generate a treatment plan based on the information presented in the patient note. The treatment plan should reflect reasonable clinical judgment and may include diagnostics, prescriptions or follow-up recommendations.
Example:

```
Given various information about a patient, return a reasonable treatment plan for the patient.
<context>
Answer:
```

Evaluation metric:
Jury score: The quality of the model's answer is evaluated by an LLM-based jury, which assigns an average score based on accuracy, completeness and clarity with respect to the gold-standard response.
Gold-standard response:
The gold standard provides a clinically sound treatment plan that serves as a reference for evaluation.
Example:

```
Diagnostic & lab orders: Ordered blood lead.
```

Medec. *Description:* Medec is a benchmark composed of clinical narratives that include either correct documentation or medical errors.

Each entry includes sentence-level identifiers and an associated correction task. The model must review the narrative and either identify the erroneous sentence and correct it, or confirm that the text is entirely accurate.

Category: Clinical decision support
Subcategory: Planning treatments
Context:
The context is a patient clinical note presented as a numbered list of sentences. Each sentence is preceded by its sentence ID and space, requiring the model to analyze each line independently for correctness.
Example:

```
0 A 29-year-old internal medicine resident
presents to the emergency department with
complaints of fevers, diarrhea, abdominal pain
and skin rash for 2 days.
1 He feels fatigued and has lost his appetite.
...
```

Prompt:
The prompt instructs the model to serve as a medical reviewer of the provided narrative. If the text is correct, the model should return 'CORRECT'. If a medical error exists, it must return the line number and a corrected version of the sentence.
Example:

```
The following is a medical narrative about
a patient. You are a skilled medical doctor
reviewing the clinical text. The text is either
correct or contains one error. The text has
a sentence per line. Each line starts with
the sentence ID, followed by a space character
then the sentence to check. Check every sentence
of the text. If the text is correct return
the following output: CORRECT. If the text
has a medical error, return the sentence ID
of the sentence containing the error, followed
by a space and a corrected version of the
sentence.
Clinical note:
<context>
Answer:
```

Evaluation metric:
Medical error flag accuracy: Measures whether the model correctly identifies if a note contains an error or not, treating the task as binary classification (correct/incorrect).
Gold-standard response:
The gold-standard response is either a correction of a specific sentence or the word 'CORRECT', indicating no error was found in the narrative.
Example:

```
CORRECT
```

EHRSHOT. *Description:* EHRSHOT is a benchmark designed to evaluate a model's ability to predict future clinical events using structured EHR data. Each instance contains a patient's historical EHR data and a forward-looking clinical question about whether a particular diagnosis, lab result or hospital event will occur.

Category: Clinical decision support
Subcategory: Predicting patient risks and outcomes
Context:
The context includes the patient's EHR history, represented as structured clinical codes. The model must determine whether a future clinical event, such as a diagnosis, lab abnormality or hospital outcome, is likely to occur based on this history.
Example:

```
# Instructions
You are an expert hospitalist at Stanford
Healthcare, an academic medical center
affiliated with Stanford University. You
specialize in predicting 30-day readmission.
Instruction: Review the patient's EHR history.
Based on all available medical evidence in the
provided EHR, please answer the question: If
this patient is discharged from the hospital
right now, is the patient likely to be
readmitted to the hospital within 30 days?
Patient EHR:
...
```

Prompt:
The prompt requests a binary prediction from the model based on the EHR context. The response must strictly follow the format: 'A' for yes, 'B' for no, with no additional text.
Example:

```
Answer A for yes, B for no.
<context>
A. Yes
B. No
Respond with only 'A' for yes or 'B' for no. Do
not add any other text, punctuation or symbols:
```

Evaluation metric:
Exact match: The predicted answer is compared directly to the gold-standard label. The response must exactly match either 'A' or 'B'.
Gold-standard response:
The correct answer to the predictive question. It is represented as one of two options: A (yes) or B (no).
Example:

```
A
```

HeadQA. *Description:* HeadQA is a benchmark consisting of biomedical multiple-choice questions intended to evaluate a model's medical knowledge and reasoning. Each instance presents a clinical or scientific question with four answer options, requiring the model to select the most appropriate answer.

Category: Clinical decision support
Subcategory: Providing clinical knowledge support
Context:
The context consists of a multiple-choice question in the medical domain. The model must review the QA choices, then select the most appropriate clinical response.
Example:

```
In the stools of a 6-year-old boy with diarrhea
and abdominal pain, a packet of parasite
eggs surrounded by a transparent membrane is
observed. Which treatment do you consider most
appropriate?
A. Praziquantel
B. Azithromycin
C. Metronidazole
D. Albendazole
```

Prompt:
The prompt directs the model to answer biomedical multiple-choice questions by selecting the correct option. The model is expected to return only the letter corresponding to its answer.
Example:

```
You are a highly knowledgeable AI assistant
specializing in biomedical sciences. Your task is
```

```
to answer multiple-choice questions accurately
based on the options provided. Each question will
relate to biomedical concepts, and you will be
asked to choose the most appropriate answer.
Select the correct answer by outputting only
the letter corresponding to your choice (A, B,
C or D).
Question:
<context>
Answer:
```

Evaluation metric:
Exact match: The model's answer must exactly match the gold-standard letter label corresponding to the correct choice (A, B, C or D).
Gold-standard response:
The correct answer to the question, given as a single letter corresponding to the appropriate choice.
Example:

```
A
```

Medbullets. *Description:* Medbullets is a benchmark of United States Medical Licensing Examination (USMLE)-style medical questions designed to assess a model's ability to understand and apply clinical knowledge. Each question is accompanied by a patient scenario and five multiple-choice options, similar to those found in step 2 and step 3 of the US medical licensing exam.
Category: Clinical decision support
Subcategory: Providing clinical knowledge support
Context:
The context presents a brief clinical scenario followed by a multiple-choice question. The model must analyze the case and choose the most appropriate answer.
Example:

```
A 64-year-old man presents to the emergency room
with a headache and nausea. He reports that
he was rocking his grandson to sleep when the
symptoms began.
...

Which of the following is the most appropriate
prophylaxis for this patient's condition?
A. Acetazolamide
B. Amitriptyline
C. Clopidogrel
D. Epinephrine
E. Verapamil
```

Prompt:
The prompt instructs the model to answer multiple-choice clinical questions by selecting the single best answer, returning only the letter corresponding to the correct choice.
Example:

```
You are a highly knowledgeable AI assistant
specializing in medicine. Your task is to answer
medical questions similar to those found on
the USMLE step 2/3 exams. You will be provided
with a clinical scenario followed by several
multiple-choice options.
Select the correct answer by outputting only the
letter corresponding to your choice (A, B, C, D
or E).
Clinical scenario:
<context>
Answer:
```

Evaluation metric:
Exact match: The model's output must exactly match the letter label of the correct choice (A, B, C, D or E).
Gold-standard response:
The correct answer is given as a single letter corresponding to the correct multiple-choice option.
Example:

```
A
```

MedQA. *Description:* MedQA is an open-domain medical QA benchmark, derived from professional medical board exams. The dataset comprises English-language questions from the USMLE, each with four answer choices. The US (English) subset contains 12,723 questions.
Category: Clinical decision support
Subcategory: Providing clinical knowledge support
Context:
Each question provides a detailed clinical scenario followed by a multiple-choice question with four options. The model must select the most likely answer based on the presented information.
Example:

```
A 47-year-old man comes to the physician because
of severe retrosternal chest pain and shortness
of breath for 45 minutes. He has dyslipidemia,
hypertension and type 2 diabetes mellitus. Cur-
rent medications include hydrochlorothiazide,
lisinopril, metformin and atorvastatin. He has
smoked 1 pack of cigarettes daily for 20 years.
He appears pale and diaphoretic. His temperature
is 37 °C (98.6 °F), pulse is 115/min, and blood pressure
is 140/70 mm Hg. Breath sounds are normal. The
remainder of the examination shows no abnormali-
ties. An electrocardiogram shows left ventricu-
lar hypertrophy with ST-segment elevation in
leads I, aVL and V1-V6. High-dose aspirin, clopi-
dogrel, metoprolol, sublingual nitroglycerin and
unfractionated heparin are administered. As the
patient awaits transport to the nearest emergency
room, he collapses and becomes unresponsive.
His pulse and blood pressure cannot be detected.
Despite resuscitative efforts, the patient dies.
Which of the following is the most likely cause
of death in this patient?
A. Acute pericarditis
B. Ventricular arrhythmia
C. Pulmonary embolism
D. Hypovolemic shock
```

Prompt:
The prompt instructs the model to answer each clinical multiple-choice question by selecting the single best answer and returning only the letter corresponding to the correct option.
Example:

```
Give a letter answer among A, B, C or D. Do not
include any explanation or additional text.
Question:
<context>
Respond only with 'A', 'B', 'C' or 'D'. Do not
add any other text, punctuation or symbols.
```

Evaluation metric:
Exact Match: The model's output must exactly match the single-letter label of the correct choice (A, B, C or D).

Gold-standard response:
The correct answer is provided as a single letter corresponding to the correct multiple-choice option.
Example:
```
B
```

MedMCQA. *Description:* MedMCQA is a large-scale multiple-choice question–answering (MCQA) benchmark drawn from real-world Indian medical entrance exams[7]. The dataset contains over 194,000 high-quality questions from AIIMS and NEET PG entrance exams, covering 2,400 healthcare topics across 21 medical subjects. The questions feature diverse topics and are designed to evaluate in-depth medical knowledge.

Category: Clinical decision support
Subcategory: Providing clinical knowledge support
Context:
Each question presents a medical prompt with four possible answer choices. The model must select the single best answer based on its knowledge.
Example:
```
Highest level of evidence is seen in:
A. Case-control studies
B. Meta-analysis
C. Cohort studies
D. Systematic review
```

Prompt:
The prompt instructs the model to answer each question by selecting the single best answer, outputting only the corresponding letter.
Example:
```
Give a letter answer among A, B, C or D. Do not
include any explanation or additional text.
<context>
Respond only with 'A', 'B', 'C' or 'D'. Do not
add any other text, punctuation or symbols.
```

Evaluation metric:
Exact match: The model's output must exactly match the single-letter label of the correct choice (A, B, C or D).
Gold-standard response:
The correct answer is provided as a single letter corresponding to the correct multiple-choice option.
Example:
```
B
```

MedAlign. *Description:* MedAlign is a benchmark that evaluates a model's ability to interpret and follow instructions grounded in longitudinal EHRs. Each instance includes an event-stream-style patient record and a natural language question or task, requiring clinically informed reading comprehension and reasoning.

Category: Clinical decision support
Subcategory: Providing clinical knowledge support
Context:
The context consists of structured EHR data encoded as an event stream, along with a natural language question or instruction referencing some aspect of the patient's medical record. The model must accurately extract and synthesize relevant information.
Example:
```
EHR: <event stream person_id="xxxxxxxxxx">
<encounter start_timestamp="xxxx-xx-xx xx:xx"
end_timestamp="xxxx-xx-xx xx:xx">
<person>
<birthdate>xxxx-xx-xx</birthdate>
<age>
<days>xx</days>
<years>xx</years>
</age>
<demographics>
<ethnicity>xxxxxxxxxxx</ethnicity>
<gender>xxxxxxxxxxx</gender>
</demographics>
<payerplan>xxxxx</payerplan>
</person>
<events>
...

Question: Please show the patient's vital signs
over the past 3 months.
```

Prompt:
The prompt gives an instruction and asks the model to fulfill it using the EHR context provided. The response should be clinically relevant and accurate, and formatted appropriately for the instruction.
Example:
```
Instruction: Answer the following question based
on the EHR:
<context>
Answer:
```

Evaluation metric:
Jury score: The quality of the model's answer is evaluated by an LLM-based jury, which assigns an average score based on accuracy, completeness and clarity with respect to the gold-standard response.
Gold-standard response:
The gold response is a human-written answer by a clinician addressing the instruction based on the patient record.
Example:
```
These are the patient's vital signs over the
past 3 months:
- xx/xx/xxxx: Weight: xxxxx Height: xxxxx Head
Circumference: xxxxx Temperature: xxxxx
- xx/xx/xxxx: Pulse: xxxxx Temp: xxxxx Weight:
xxxxx Height: xxxxxx
...
```

ADHD-Behavior. Description:
ADHD-Behavior is a benchmark that evaluates a model's ability to detect whether a clinician recommends parent training in behavior management, an evidence-based first-line treatment for young children diagnosed with attention deficit hyperactivity disorder (ADHD). Each instance includes a clinical note from a pediatric visit and a binary classification task.

Category: Clinical decision support
Subcategory: Providing clinical knowledge support
Context:
The context includes a clinical note documenting an ADHD-focused pediatric visit. The model must determine whether the note includes a recommendation for parent training in behavioral management.
Example:
```
Date: xxxx-xx-xx
Patient: xxxxxx xxxxx
Visit type: Follow-up visit
Clinician: Dr. xxxx xxxx, MD -
Developmental-Behavioral Pediatrics
Subjective:
xxxxx returns for follow-up regarding ADHD
symptoms. His parents report continued
```

```
difficulty with attention and impulsivity at
school, despite consistent use of his current
medication.
```

Prompt:
The prompt instructs the model to assess whether a parent training recommendation is present in the note and respond with a binary classification.
Example:

```
Answer A or B.
You are reviewing a clinical note from health
records of children with ADHD and classifying
mentions of behavioral therapy.
Provide an answer to the following question:
Does the patient note contain a clinician
recommendation for parent training in behavior
management? with the following context
<context>
Answer the question with a 'A' for yes or 'B'
for no. Do not provide any additional details or
response, just a simple A or B response.
```

Evaluation metric:
Exact match: Measures the proportion of correct binary classifications relative to the annotated gold standard.
Gold-standard response:
The correct binary classification of whether a parent training recommendation was made.
Example:

```
B
```

ADHD-MedEffects. *Description:* ADHD-MedEffects is a benchmark designed to evaluate whether clinical notes for pediatric ADHD visits document medication side-effect monitoring, which is a key recommendation in clinical practice guidelines. The dataset supports binary classification to detect presence or absence of side-effect inquiries (SEIs) within notes.

Category: Clinical decision support
Subcategory: Providing clinical knowledge support
Context:
The context consists of a clinical note from a pediatric follow-up visit concerning ADHD medication, which may or may not contain documentation of inquiries about medication side effects.
Example:

```
Date: xxxx-xx-xx
Patient: xxxxx xxxxx
Visit Type: ADHD medication follow-up
Clinician: Dr. xxxx xxxx, MD - Pediatrician
Subjective:
xxxxx's mother reports improved concentration
at school since starting lisdexamfetamine 20 mg
daily. When asked, she denies any side effects
such as appetite suppression, insomnia or mood
changes. No new behavioral issues noted at home.
```

Prompt:
The prompt defines the criteria for identifying SEIs and instructs the model to classify the note accordingly.
Example:

```
Answer A or B.
You are reviewing a clinical note from health
records of children with ADHD. Given the
following definitions:
SEI: Explicit documentation by the clinician
asking about current side effects related to
ADHD medications that the child is taking, or
documentation of specific ADHD medication side
effects experienced by the patient. SEI does not
include future side-effects monitoring, such as
documentation of potential ADHD medication side
effects, plans to follow patients to monitor
side effects or explanations about potential
side effects. These are not categorized as SEI
because they do not reflect actual side-effect
monitoring.
No side-effects inquiry (NSEI): No evidence of
side-effects monitoring.
Provide an answer to the following question:
"Does the patient note contain documentation of
side-effect monitoring (recording of absence
or presence of medication side effects), as
recommended in clinical practice guidelines?"
with the following context
<context>
Answer the question with an 'A' for yes or a 'B'
for no. Do not provide any additional details or
response, just a simple A or B response.
```

Evaluation metric:
Exact match: Measures the proportion of correct binary classifications that match the gold standard.
Gold-standard response:
The expected binary response indicating whether the note includes documentation of side-effect monitoring.
Example:

```
B
```

DischargeMe. *Description:* DischargeMe is a benchmark designed to evaluate clinical text generation. It pairs discharge summaries and radiology reports from MIMIC-IV with generation tasks such as writing discharge instructions or summarizing the brief hospital course (BHC). The benchmark assesses a model's ability to generate patient-facing documentation that is complete, empathetic and clinically accurate.

Category: Clinical note generation
Subcategory: Documenting patient visits
Context:
The context includes a patient's discharge summary and radiology report, along with a natural language task prompt. The model must generate either the discharge instructions or the BHC section based on the provided context.
Example:

```
Generate the Discharge Instructions from the
following patient discharge text and radiology
report text.
Discharge text:
Name: ___ Unit no: ___
Admission date: ___ Discharge date: ___
...

Radiology report:
EXAMINATION:
...

Discharge instructions:
```

Prompt:
The prompt asks the model to generate a specific section of a discharge note (either discharge instructions or a BHC) from the given source documents.

Example:

```
Given a discharge text, a radiology report
text, and a target document of either discharge
instructions or a brief hospital course, return
the generated target document from the context
provided.
<context>
Answer:
```

Evaluation metric:
Jury score: The quality of the model's answer is evaluated by an LLM-based jury, which assigns an average score based on accuracy, completeness and clarity with respect to the gold-standard response.
Gold-standard response:
The gold-standard response is the actual discharge instruction or BHC, as authored by a clinical provider.
Example:

```
Dear xxxxx,
It was a pleasure to take care of you. You were
admitted for xxxxxxxxxx.
After treatment, your xxxxxxxxxxxx improved.
Please follow these instructions:
...

We wish you the best.
```

ACI-Bench. *Description:* ACI-Bench is a benchmark of real-world patient–doctor conversations paired with structured clinical notes. The benchmark evaluates a model's ability to understand spoken medical dialogue and convert it into formal clinical documentation, covering sections such as history of present illness, physical exam findings, results and assessment and plan.
Category: Clinical note generation
Subcategory: Documenting patient visits
Context:
The context consists of a transcribed conversation between a clinician and a patient. The dialogue includes greetings, medical history, symptoms and care plans, which the model must distill into a structured clinical note.
Example:

```
Doctor–patient dialogue:
Doctor: Hi, Andrew. how are you?
Patient: Hey, good to see you.
Doctor: I'm doing well, i'm doing well.
Patient: Good.
Doctor: So, I know the nurse told you about
dax. I'd like to tell dax a little bit about
you.
...
```

Prompt:
The prompt instructs the model to transform the conversation into a clinical note with four specific sections, emphasizing accurate and complete summarization of the interaction.
Example:

```
Summarize the conversation to generate a
clinical note with four sections:
1. HISTORY OF PRESENT ILLNESS
2. PHYSICAL EXAM
3. RESULTS
4. ASSESSMENT AND PLAN
The conversation is:
<context>
Clinical note:
```

Evaluation metric:
Jury score: The quality of the model's answer is evaluated by an LLM-based jury, which assigns an average score based on accuracy, completeness and clarity with respect to the gold-standard response.
Gold-standard response:
The gold-standard response is the formal clinical note written by a clinician based on the original conversation.
Example:

```
CHIEF COMPLAINT
Upper respiratory infection.
HISTORY OF PRESENT ILLNESS
Andrew Campbell is a 59-year-old male with a
past medical history significant for depression,
type 2 diabetes and hypertension. He presents
today with an upper respiratory infection.
...
```

MTSamples Procedures. ***Description:*** MTSamples Procedures is a benchmark composed of transcribed operative notes, focused on documenting surgical procedures. Each example presents a brief patient case involving a surgical intervention, and the model is tasked with generating a coherent and clinically accurate procedural summary or treatment plan.
Category: Clinical note generation
Subcategory: Recording procedures
Context:
The context consists of a transcription note describing a surgical procedure. The model must read the operative context and generate a full treatment plan based on the medical content and procedural intent.
Example:

```
Medical specialty: Orthopedic
Sample name: AC separation revision & hardware
removal
Description: Removal of the hardware and
revision of right AC separation. Loose hardware
with superior translation of the clavicle
implants.
(Medical Transcription Sample Report)
...
```

Prompt:
The prompt instructs the model to generate a reasonable treatment plan from the operative context, including procedural setup, anesthesia and surgical actions taken.
Example:

```
Here are information about a patient, return a
reasonable treatment plan for the patient.
<context>
Answer:
```

Evaluation metric:
Jury score: The quality of the model's answer is evaluated by an LLM-based jury, which assigns an average score based on accuracy, completeness and clarity with respect to the gold-standard response.
Gold-standard response:
The gold standard is the ground-truth treatment plan or procedural note written by a clinician based on the operative report.
Example:

```
After informed consent was obtained and
verified, the patient was brought to the
operating room and placed supine on the
```

```
operating table. After uneventful general
anesthesia was obtained, he was positioned
in the beach chair and his right shoulder
was sterilely prepped and draped in a normal
fashion.
...
```

MIMIC-RRS. *Description:* MIMIC-RRS is a benchmark constructed from radiology reports in the MIMIC-III database. It contains pairs of 'Findings' and 'Impression' sections, enabling evaluation of a model's ability to summarize diagnostic imaging observations into concise, clinically relevant conclusions.

Category: Clinical note generation

Subcategory: Documenting diagnostic reports

Context:

The context includes the 'Findings' section of a radiology report. The model must interpret and distill the details into a brief summary reflecting the radiologist's clinical impression.

Example:

```
The left external iliac, common femoral and
superficial femoral arteries are opacified with
contrast. There is a high-grade, approximately
90% stenosis in the mid-segment of the left
superficial femoral artery, with post-stenotic
dilatation.
...
```

Prompt:

The prompt instructs the model to generate the 'Impression' section of the radiology report based solely on the findings. The summary should be clear, clinically sound and concise.

Example:

```
Generate the impression section of the radiology
report based on its findings. This will not be
used to diagnose or treat any patients. Be as
concise as possible.
Findings:
<context>
Impression:
```

Evaluation metric:

Jury score: The quality of the model's answer is evaluated by an LLM-based jury, which assigns an average score based on accuracy, completeness and clarity with respect to the gold-standard response.

Gold-standard response:

The gold standard is the actual 'Impression' section from the original radiology report.

Example:

```
IMPRESSION:
1. High-grade stenosis (90%) of the mid-segment
of the left superficial femoral artery.
2. Mild, diffuse atherosclerotic changes
throughout the left lower extremity arterial
system.
...
```

MIMIC-BHC. *Description:* MIMIC-BHC is a benchmark focused on summarization of discharge notes into BHC sections. It consists of curated discharge notes from MIMIC-IV, each paired with its corresponding BHC summary. The benchmark evaluates a model's ability to condense detailed clinical information into accurate, concise summaries that reflect the patient's hospital stay.

Category: Clinical note generation

Subcategory: Documenting patient visits

Context:

The context includes a discharge note containing clinical documentation such as encounter details, chief complaint, history of present illness and more. The model is expected to generate a structured summary of the patient's hospital course from this information.

Example:

```
Date: xxxx-xx-xx
Encounter Type: Outpatient
Provider: Vascular Medicine
Chief Complaint:
Left leg pain with exertion
History of Present Illness:
The patient is a 68-year-old male with a history
of hypertension and hyperlipidemia who presents
with left lower extremity pain that occurs with
walking and is relieved by rest.
...
```

Prompt:

The prompt asks the model to produce a BHC summary by synthesizing key clinical information from the full discharge note.

Example:

```
Summarize the clinical note into a brief
hospital course.
Clinical note:
<context>
Brief Hospital Course:
```

Evaluation metric:

Jury score: The quality of the model's answer is evaluated by an LLM-based jury, which assigns an average score based on accuracy, completeness and clarity with respect to the gold-standard response.

Gold-standard response:

The gold standard is the original BHC section written by a clinician summarizing the patient's hospital evaluation and treatment.

Example:

```
The patient was admitted for evaluation and
management of worsening left lower extremity
claudication. On admission, vascular surgery was
consulted. Noninvasive arterial Doppler studies
confirmed high-grade stenosis in the left
superficial femoral artery.
...
```

NoteExtract. *Description:* NoteExtract is a benchmark that focuses on the structured extraction of information from free-form clinical text. It provides care plan notes authored by health workers and evaluates a model's ability to convert them into a predefined structured format, such as fields for 'chief complaint' and 'history of present illness'. The benchmark emphasizes faithful extraction without hallucination or inference.

Category: Clinical note generation

Subcategory: Documenting care plans

Context:

The context consists of a free-text clinical care plan or physician note. The model is asked to extract structured information directly entailed by the source content.

Example:

```
You are provided with a clinical note regarding
a physician–patient interaction. Your task is
to extract specific information based solely
on the content provided. Do not hallucinate or
infer details that are not explicitly stated in
```

```
the text. Any information you include must be
directly entailed by the text.
Instructions:
Extract the required information precisely as
presented in the source text.
...

Clinical Note:
...
```

Prompt:
The prompt instructs the model to follow a set of instructions to extract relevant information from a clinical note and present it in a specified structured format.
Example:

```
Follow the instructions provided regarding
conversion of a patient note into a specified
format.
<context>
```

Evaluation metric:
Jury score: The quality of the model's answer is evaluated by an LLM-based jury, which assigns an average score based on accuracy, structure and clarity with respect to the gold-standard response.
Gold-standard response:
The gold standard is a structured representation of the extracted information, directly mapped from the content of the clinical note.
Example:

```
Chief complaint: Left leg pain with exertion.
History of present illness
Onset: The symptoms have gradually worsened over
the past 6 months.
Provoking/palliating factors: Occurs with
walking and is relieved by rest.
...
```

MedicationQA. *Description:* MedicationQA is a benchmark composed of open-ended consumer health questions specifically focused on medications. Each example consists of a free-form question and a corresponding medically grounded answer. The benchmark evaluates a model's ability to provide accurate, accessible and informative medication-related responses for a lay audience.
Category: Patient communication and education
Subcategory: Providing patient education resources
Context:
The context is a medication-related question submitted by a consumer. The question may pertain to drug effects, interactions, mechanisms of action or usage instructions.
Example:

```
How does Valium affect the brain?
```

Prompt:
The prompt asks the model to provide a medically accurate and clearly worded response to the consumer's question.
Example:

```
Please answer the following consumer health
question.
Question:
<context>
Answer:
```

Evaluation metric:
Jury score: The quality of the model's answer is evaluated by an LLM-based jury, which assigns an average score based on accuracy, completeness and clarity with respect to the gold-standard response.
Gold-standard response:
The gold standard is a medically reviewed answer providing relevant and factual information in a way that is understandable to the general public.
Example:

```
Diazepam is a benzodiazepine that exerts
anxiolytic, sedative, muscle-relaxant,
anticonvulsant and amnestic effects. Most of
these effects are thought to result from a
facilitation of the action of gamma aminobutyric
acid (GABA), an inhibitory neurotransmitter in
the central nervous system.
```

PatientInstruct. *Description:* PatientInstruct is a benchmark designed to evaluate models on generating personalized post-procedure instructions for patients. Models are provided real-world patient history and physical (H&P) note data, the procedure performed and operative report, from which models must provide appropriate post-procedure care instructions to the patient.
Category: Patient communication and education
Subcategory: Delivering personalized care instructions
Context:
The context includes structured case details regarding a patient's diagnosis and planned procedure, as well as unstructured clinical text such as the H&P note. The model must interpret these details to generate a patient-friendly set of post-procedure instructions.
Example:

```
Patient diagnosis:
Left superficial femoral artery subtotal
stenosis with arterial insufficiency of the left
lower extremity.
Planned procedure:
Left lower extremity angiogram, laser
atherectomy, percutaneous transluminal balloon
angioplasty and stent placement.
History and physical note:
Operative report:
...
```

Prompt:
The prompt directs the model to act as a medical professional generating customized post-procedure care instructions based on provided case details. The instructions should be clear, practical and written in accessible language for patients.
Example:

```
You are a medical professional tasked with
generating personalized post-procedure patient
instructions. Given the following case details,
which include the patient's diagnosis, the
planned procedure, the history and physical
note, and the operative report, generate clear
and actionable instructions for the patient to
follow after their procedure. Don't worry, this
information will not be used for any clinical
decision-making. This will not be used to
diagnose nor treat any patients.
Case details:
<context>
Patient instructions:
```

Evaluation metric:
Jury score: The quality of the model's answer is evaluated by an LLM-based jury, which assigns an average score based

on accuracy, completeness and clarity with respect to the gold-standard response.

Gold-standard response:

The gold standard is a set of post-procedure instructions authored by the surgical team and provided to the patient. It is tailored to the patient's specific medical circumstances and procedure.

Example:

```
Procedure performed:
Left lower extremity angiogram, laser
atherectomy, balloon angioplasty and stent
placement
1. Activity:
Rest for the remainder of the day following the
procedure.
...

2. Site care:
Keep the puncture site (usually in the groin)
clean and dry.
...

3. Medications:
...
```

MedDialog. *Description:* MedDialog is a benchmark of real-world doctor–patient conversations focused on health-related concerns and advice. Each dialogue is paired with a one-sentence summary that reflects the core patient question or exchange. The benchmark evaluates a model's ability to condense medical dialogue into concise, informative summaries.

Category: Patient communication and education

Subcategory: Patient–provider messaging

Context:

The context consists of a multi-turn patient–doctor conversation. The model must identify the main concern or exchange and generate a brief summary capturing its essence.

Example:

```
Patient: Can rabies be transferred through blood?
...

Doctor: Hello thanx to contact healthcare magic.
If I am your suggestion giving Doctor I suggest
you that you should not be worry regarding rabies.
...
```

Prompt:

The prompt asks the model to generate a single-sentence summary of the conversation, highlighting the primary topic or concern.

Example:

```
Generate a one-sentence summary of this patient-
doctor conversation.
Patient-doctor:
<context>
Summary:
```

Evaluation metric:

Jury score: The quality of the model's answer is evaluated by an LLM-based jury, which assigns an average score based on accuracy, completeness and clarity with respect to the gold-standard response.

Gold-standard response:

The gold standard is a single-sentence summary that accurately captures the key concern or exchange from the conversation.

Example:

```
Can rabies be transferred through blood?
```

MedConfInfo. *Description:* MedConfInfo is a benchmark comprising clinical notes from adolescent patients. It is used to evaluate whether the content contains sensitive protected health information (PHI) that should be restricted from parental access, in accordance with adolescent confidentiality policies in clinical care.

Category: Patient communication and education

Subcategory: Patient–provider messaging

Context:

The context consists of a clinical note from an adolescent patient visit, which may include discussions of confidential topics such as sexual activity, mental health or substance use.

Example:

```
Date: xxxx-xx-xx
Patient: xxxxx xxxxx
Visit type: Checkup visit
Clinician: Dr. xxxx xxxx, MD - Adolescent
Medicine
Subjective:
xxxxx presents for a routine adolescent checkup.
During the private portion of the visit, she
reports being sexually active with one partner
and asks about birth control options.
```

Prompt:

The prompt asks the model to determine whether the note includes PHI that should be restricted from parental access.

Example:

```
Answer A or B.
Provide an answer to the following question:
"Does the following adolescent patient note
contain sensitive protected health information
that should be restricted from parental access?"
with the following context
<context>
Answer the question with a 'A' for yes or 'B'
for no. Do not provide any additional details or
response, just a simple A or B response.
```

Evaluation metric:

Exact match: Measures the proportion of correct binary classifications that exactly match the gold standard.

Gold-standard response:

The expected classification label indicating whether the note contains sensitive PHI.

Example:

```
B
```

MEDIQA. *Description:* MEDIQA is a benchmark designed to evaluate a model's ability to retrieve and generate medically accurate answers to patient-generated questions. Each instance includes a consumer health question, a set of candidate answers (used in ranking tasks), relevance annotations and, optionally, additional context. The benchmark focuses on supporting patient understanding and accessibility in health communication.

Category: Patient communication and education

Subcategory: Enhancing patient understanding and accessibility in health communication

Context:

The context includes a free-form question posed by a patient. These questions often involve rare diseases, diagnostic concerns or treatment guidance, and require clinically sound, accessible answers.

Example:

```
Hi, I would like to know if there is any support
for those suffering with abetalipoproteinemia?
I am not diagnosed but have had many tests that
```

```
indicate I am suffering with this, keen to learn
how to get it diagnosed and how to manage, many
thanks
...
```

Prompt:
The prompt directs the model to generate an accurate and helpful response to a consumer health question, grounded in medical knowledge and clear language.
Example:

```
Answer the following consumer health question.
Question:
<context>
```

Evaluation metric:
Jury score: The quality of the model's answer is evaluated by an LLM-based jury, which assigns an average score based on accuracy, completeness and clarity with respect to the gold-standard response.
Gold-standard response:
The gold standard is a clinician-authored response that directly addresses the patient's question with factual, compassionate and practical information.
Example:

```
Bassen–Kornzweig syndrome: Bassen–Kornzweig syndrome
is a rare disease passed down through families. The
person is unable.....
```

MentalHealth. *Description:* MentalHealth is a benchmark focused on evaluating empathetic communication in mental health counseling. It includes simulated conversations between patients and counselors, where the task is to generate compassionate and appropriate counselor responses. The benchmark assesses a model's ability to support patients emotionally and meaningfully engage in therapeutic conversations.

Category: Patient communication and education
Subcategory: Facilitating patient engagement and support
Context:
The context consists of a mental health conversation where the patient shares thoughts, emotions or struggles. The model must respond as a counselor, demonstrating empathy, active listening and therapeutic support.
Example:

```
Patient: I've been feeling completely overwhelmed
lately. Every day...
```

Prompt:
The prompt instructs the model to assume the role of a counselor and generate an emotionally attuned response to the patient's concerns.
Example:

```
You are a mental health counselor. Generate an
empathetic and appropriate counselor response:
Conversation:
<context>
```

Evaluation metric:
Jury score: The quality of the model's answer is evaluated by an LLM-based jury, which assigns an average score based on accuracy, completeness and clarity with respect to the gold-standard response.
Gold-standard response:
The gold standard is a professionally written counselor response that validates the patient's emotions and fosters trust and safety in the conversation.
Example:

```
That sounds really hard. You're clearly doing
your best in a tough situation, and I want you
to know you're not alone in this.
```

ProxySender. *Description:* ProxySender is a benchmark composed of patient portal messages received by clinicians. It evaluates whether the message was sent by the patient or by a proxy user (for example, parent, spouse), which is critical for understanding who is communicating with healthcare providers.

Category: Patient communication and education
Subcategory: Patient–provider messaging
Context:
The context contains a single patient portal message directed to a clinician, with the sender possibly being someone other than the patient.
Example:

```
Message: Hi Dr. Lee, this is Jane's mom. I wanted
to let...
```

Prompt:
The prompt asks the model to classify whether the message was sent by a proxy or the actual patient.
Example:

```
Answer A or B.
You are reviewing clinical messages in order
to determine if they have been sent by a proxy
user. Please determine the following: "Is the
message sent by a proxy user or by a non-proxy
user?" with the following context
<context>
Answer the question with 'A' for yes or 'B' for
no. Do not provide any additional details or
response, just a simple A or B response.
```

Evaluation metric:
Exact match: Measures the proportion of messages for which the proxy classification matches the gold-standard label.
Gold-standard response:
The correct binary classification label indicating whether the sender was a proxy.
Example:

```
A
```

PrivacyDetection. *Description:* PrivacyDetection is a benchmark composed of patient portal messages submitted by patients or caregivers. The task is to determine whether the message contains any confidential or privacy-leaking information that should be protected.

Category: Patient communication and education
Subcategory: Patient–provider messaging
Context:
The context includes a clinical message submitted through the patient portal, which may or may not contain privacy-sensitive content.
Example:

```
Message: Hi, just wanted to confirm that my
test came...
```

Prompt:
The model is asked to determine whether the message includes private, identifiable or sensitive patient information.
Example:

```
Answer A or B.
You are reviewing clinical messages in order to
determine if any confidential information was
```

```
included. Please answer the following question:
"Does the message leak any confidential
information from the patient?" with the
following context
<context>
Answer the question with 'A' for yes or 'B' for
no. Do not provide any additional details or
response, just a simple A or B response.
You are an assistant trained to detect potential
privacy risks in patient communications.
Input:
<context>
Output format:
Return either privacy-leaking or
non-privacy-leaking
```

Evaluation metric:
Exact match: Measures the proportion of responses that exactly match the correct classification regarding privacy leakage.
Gold-standard response:
The correct binary classification indicating whether privacy was breached.
Example:

```
B
```

PubMedQA. *Description:* PubMedQA is a biomedical QA dataset that evaluates a model's ability to interpret scientific literature. It consists of PubMed abstracts paired with yes/no/maybe questions derived from the content. The benchmark assesses a model's capability to reason over biomedical texts and provides factually grounded answers.
Category: Medical research assistance
Subcategory: Conducting literature research
Context:
The context includes a PubMed abstract, followed by a question that requires the model to answer based solely on the abstract's content. The questions are designed to test comprehension and factual grounding in the biomedical domain.
Example:

```
Context

Background. Programmed cell death is the regulated
death of cells within...
```

Prompt:
The prompt instructs the model to respond to a biomedical question with a single-letter answer: A (yes), B (no) or C (maybe), without including any explanation or extra text.
Example:

```
Answer A for yes, B for no or C for maybe. Do
not include any explanation or additional text.
Output only the letter on a single line:
Input:
<context>
Question: Do mitochondria play a role in
remodeling lace plant leaves during programmed
cell death?
```

Evaluation metric:
Exact match: The model's response is compared exactly to the gold-standard letter label (A, B or C).
Gold-standard response:
The gold standard is the correct letter label answering the question based on the abstract.
Example:

```
A
```

EHRSQL. *Description:* EHRSQL is a benchmark designed to evaluate models on generating structured queries for clinical research. Each example includes a natural language question and a database schema, and the task is to produce a structured query language (SQL) query that would return the correct result for a biomedical research objective. This benchmark assesses a model's understanding of medical terminology, data structures and query construction.
Category: Medical research assistance
Subcategory: Analyzing clinical research data
Context:
The context includes a relational database schema used in a clinical research setting. The model is required to interpret both the schema and the natural language question to generate a valid SQL query that can retrieve the correct data.
Example:

```
Context – database schema:
CREATE TABLE patient

uniquepid VARCHAR(10) NOT NULL,
.....
```

Prompt:
The prompt asks the model to act as a medical SQL expert. It must generate an executable SQL query in response to a medical research question, based on the given schema.
Example:

```
You are a highly skilled AI specializing in
medical SQL queries. Given a database schema and
a medical question, generate a valid SQL query
that retrieves the required information from
the database. Output only the SQL query without
explanations:
Input:
<context>
Question: how many patients is antihypertensive
drug given to two times since 1 year ago?
```

Evaluation metric:
Execution accuracy for generated query: Measures the proportion of correctly predicted answerable questions among all questions predicted to be answerable.
Gold-standard response:
The gold standard is the correct SQL query that retrieves the expected answer when executed over the clinical database.
Example:

```
SELECT count(DISTINCT t1.uniquepid) FROM (SELECT
patient.uniquepid, count(*) AS c1 FROM patient
WHERE)...AS t1
```

BMT-Status. *Description:* BMT-Status is a benchmark composed of clinical notes and associated binary questions related to bone marrow transplant (BMT), hematopoietic stem cell transplant (HSCT) or hematopoietic cell transplant (HCT) status. The goal is to determine whether the patient received a subsequent transplant based on the provided clinical documentation.
Category: Medical research assistance
Subcategory: Recording research processes
Context:
The context includes a clinical note from a hematology/oncology consultation, containing relevant clinical history about a pediatric patient with leukemia.
Example:

```
Date: xxxx-xx-xx
Patient: xxxxx xxxxx
Visit type: Hematology/oncology consultation
```

```
Clinician: Dr. xxxx xxxx, MD - Pediatric
Hematologist
Subjective:
xxxxx is a 13-year-old male with a history of
relapsed acute lymphoblastic leukemia, currently
in second remission following re-induction
chemotherapy. No major complications reported
during treatment.
```

Prompt:
The prompt directs the model to classify whether the patient received a subsequent HCT, HSCT, BMT or transplant, based solely on the given clinical note.
Example:

```
Answer A or B.
Provide an answer to the following question:
"Did the recipient receive a subsequent HCT,
HSCT, BMT or transplant?" with the following
context
<context>
Answer the question with 'A' for yes or 'B' for
no. Do not provide any additional details or
response, just a simple A or B response.
```

Evaluation metric:
Exact match: Measures the proportion of model predictions that exactly match the correct binary response for transplant status.
Gold-standard response:
The correct binary answer indicating whether the patient received the transplant.
Example:

```
A
```

RaceBias. *Description:* RaceBias is a benchmark used to evaluate language models for racially biased or inappropriate content in medical QA scenarios. Each instance consists of a medical question and a model-generated response. The task is to classify whether the response contains race-based, harmful or inaccurate content. This benchmark supports research into bias detection and fairness in clinical AI systems.
Category: Medical research assistance
Subcategory: Ensuring clinical research quality
Context:
The context includes a QA pair potentially involving racially biased or medically inaccurate assumptions. The model must classify whether the response contains problematic content.
Example:

```
Question: How do I estimate the eGFR of a black
woman?

Response: The estimated glomerular filtration rate
(eGFR) is a measure...
```

Prompt:
The prompt asks the model to classify the answer for bias. If the response involves harmful, inaccurate or race-based content, the model should output 'A'. Otherwise, it should output 'B'.
Example:

```
For the following question-answer pair, respond
'A' for yes if the answer involves harmful,
inaccurate and/or race-based content and respond
'B' for no otherwise:
Input:
Answer:
```

Evaluation metric:
Exact match: Measures the fraction of instances where the predicted classification exactly matches the gold-standard label.
Gold-standard response:
The gold standard is the correct classification label, identifying whether the answer contains biased or harmful race-based content.
Example:

```
A
```

N2C2-CT. *Description:* N2C2-CT is a benchmark designed to evaluate a model's ability to match patients to appropriate clinical trials based on eligibility criteria. Each example includes a clinical note and a trial description. The model is tasked with determining whether the patient is a valid candidate for the trial. This benchmark supports automation and decision support in clinical research enrollment.
Category: Medical research assistance
Subcategory: Managing research enrollment
Context:
The context includes a patient's clinical note and a textual description of the eligibility criteria for a clinical trial. The model must assess alignment between patient characteristics and trial requirements.
Example:

```
Patient Note: Patient is a 58-year-old male with a
history of...
```

Clinical trial description: This trial is recruiting patients over the age of 50 with ...

Prompt:
The prompt asks the model to act as a clinical research assistant and determine if the patient is eligible for the specified trial. The response should be 'A' for yes and 'B' for no.
Example:

```
You are a clinical research assistant. Based
on the patient's clinical note and the trial
criteria, classify whether the patient is a
valid candidate for this clinical trial:
```

Input:
<context>
Evaluation metric:
Exact Match: Measures whether the model's classification exactly matches the gold-standard label for patient eligibility.
Gold-standard response:
The correct answer is a single classification: 'A' if the patient meets trial inclusion criteria, or 'B' if not.
Example:

```
A
```

MedHallu. *Description:* MedHallu is a benchmark focused on evaluating factual correctness in biomedical QA. Each instance contains a PubMed-derived knowledge snippet, a biomedical question and a model-generated answer. The task is to classify whether the answer is factually correct or contains hallucinated (non-grounded) information. This benchmark is designed to assess the factual reliability of medical language models.
Category: Medical research assistance
Subcategory: Ensuring clinical research quality
Context:
The context includes a snippet of world knowledge from a PubMed article, a scientific question and a model-generated answer. The model must judge whether the answer aligns with the source or contains hallucinated content.

Example:
```
World Knowledge: ['The objectives of this study were to evaluate the ability of the Young–Burgess classification...']
Question: Young–Burgess classification of pelvic ring fractures: does it predict mortality, transfusion requirements and non-orthopedic injuries?
Answer: The Young–Burgess classification system predicts mortality more accurately than it predicts transfusion requirements or non-orthopedic injuries.
```

Prompt:
The prompt instructs the model to act as a factuality judge. It must evaluate the answer against the knowledge provided and determine whether any content is hallucinated.
Example:
```
You are an AI assistant with extensive knowledge in the field of medicine. I want you to act as an answer judge. Given a question and an answer, your objective is to determine if the provided answer contains nonfactual or hallucinated information. You SHOULD give your judgment based on the following hallucination types and the world knowledge:
<context>
Return just an integer value, '0' if the answer is factual and '1' if the answer is hallucinated. No letter or word, just the integer value.
```

Evaluation metric:
Exact match: Measures the fraction of instances in which the model's hallucination judgment exactly matches the annotated gold standard.
Gold-standard response:
The gold standard is the correct classification label: '0' if the answer is not hallucinated or factual, and '1' if it is factually correct.
Example:
```
0
```

HospiceReferral. *Description:* HospiceReferral is a benchmark that evaluates model performance in identifying whether patients are eligible for hospice care based on palliative care clinical notes. The benchmark focuses on end-of-life care referral decisions.
Category: Administration and workflow
Subcategory: Scheduling resources and staff
Context:
The context includes a clinical note from a palliative care consultation, providing medical history and clinical status of the patient relevant to hospice eligibility.
Example:
```
Date: xxxx-xx-xx
Patient: xxxxx xxxxx
Visit type: Palliative care consultation
Clinician: Dr. xxxx xxxx, MD - Palliative Medicine
Subjective:
xxxx is a 82-year-old female with end-stage metastatic pancreatic cancer. She has completed multiple lines of chemotherapy without sustained response.
```

Prompt:
The prompt instructs the model to classify whether the patient is eligible for hospice care based solely on the clinical note.
Example:
```
Answer A or B.
Provide an answer to the following question: "Is the following patient eligible to be referred to a hospice care facility?" with the following context
<context>
Answer the question with 'A' for yes or 'B' for no. Do not provide any additional details or response, just a simple A or B response.
```

Evaluation metric:
Exact match: Measures the proportion of predictions that exactly match the correct binary response regarding hospice eligibility.
Gold-standard response:
The binary answer indicating hospice referral eligibility.
Example:
```
A
```

MIMIC-IV billing code. *Description:* MIMIC-IV billing code is a benchmark derived from discharge summaries in the MIMIC-IV database, paired with their corresponding ICD-10 billing codes. The task requires models to extract structured billing codes based on free-text clinical notes, reflecting real-world hospital coding tasks for financial reimbursement.
Category: Administration and workflow
Subcategory: Overseeing financial activities
Context:
The context is a patient discharge summary describing the patient's hospitalization, medical conditions and treatment history. The goal is to identify the appropriate ICD-10 codes that correspond to the clinical content in the note.
Example:
```
Patient note: The patient is a 71-year-old male with a history of chronic obstructive pulmonary disease and hypertension. He was admitted with shortness of breath and improved with bronchodilators.
```

Prompt:
The prompt instructs the model to act as a hospital administrative assistant and extract ICD-10 billing codes from the discharge note.
Example:
```
You are a hospital administrative assistant responsible for assigning ICD-10 codes for billing. Based on the discharge note, identify the appropriate billing codes:
Input: <context>
Predicted ICD-10 codes:
```

Evaluation metric:
F1 score for MIMIC billing codes: Computes the harmonic mean of precision and recall across predicted and gold ICD-10 codes, capturing both correctness and completeness of coding.
Gold-standard response:
The correct ICD-10 codes extracted from the patient note, formatted as a comma-separated list.
Example:
```
J44.1, I10
```

ClinicReferral. *Description:* ClinicReferral is a benchmark that determines patient eligibility for referral to the Sequoia Clinic based on information from palliative care notes. The dataset provides curated decisions on referral appropriateness to assist in automating clinic workflows.

Category: Administration and workflow
Subcategory: Organizing workflow processes
Context:
The context consists of a clinical note from a palliative care consultation, describing the patient's medical history and clinical trajectory relevant to referral criteria.
Example:

```
Date: xxxx-xx-xx
Patient: xxxxx xxxxx
Visit Type: Palliative care consultation
Clinician: Dr. xxxx xxxx, MD - Palliative
Medicine
Subjective:
xxxx is a 82-year-old female with end-stage
metastatic pancreatic cancer. She has completed
multiple lines of chemotherapy without sustained
response.
```

Prompt:
The prompt asks the model to classify whether the patient is eligible for referral to the Sequoia Clinic.
Example:

```
Answer A or B.
Provide an answer to the following question:
"Does the following patient meet the criteria
to be referred to the Sequoia Clinic?" with the
following context
<context>
Answer the question with 'A' for yes or 'B' for
no. Do not provide any additional details or
response, just a simple A or B response.
```

Evaluation metric:
Exact match: Fraction of instances where the predicted answer exactly matches the ground truth regarding clinic referral eligibility.
Gold-standard response:
A binary response indicating referral eligibility.
Example:

```
A
```

CDI-QA. *Description:* CDI-QA is a benchmark constructed from clinical documentation integrity (CDI) notes. It is used to evaluate a model's ability to verify clinical conditions based on documented evidence in patient records.
Category: Administration and workflow
Subcategory: Care coordination and planning
Context:
The context consists of a patient's clinical note from the hospital admission, which includes diagnostic findings relevant to the verification of clinical conditions such as acute kidney injury.
Example:

```
Patient note:
The patient was admitted with severe sepsis and
pneumonia. Blood cultures were positive.
```

Prompt:
The model is asked to verify whether a specific clinical condition was present at the time of admission, using only the information provided in the clinical note.
Example:

```
Questions supported:
Was acute kidney injury present on admission?
Answer A or B.
Provide an answer to the following question:
"Was acute kidney injury present on admission?"
with the following context
<context>
Answer the question with either 'A' for yes or
'B' for no. Do not provide any additional details
or response, just a simple A or B response.
```

Evaluation metric:
Exact match: Fraction of instances where the model's prediction exactly matches the correct verification answer.
Gold-standard response:
A binary response reflecting whether the clinical condition was present.
Example:

```
A
```

ENT-Referral. Description:
ENT-Referral is a benchmark designed to evaluate whether a patient's clinical note supports a referral to an ear, nose, and throat (ENT) specialist. It helps assess models' abilities to make referral decisions based on unstructured clinical text.
Category: Administration and workflow
Subcategory: Care coordination and planning
Context:
The input is a clinical note containing patient symptoms, history and assessment from a medical consultation.
Example:

```
Patient note:

Patient is a 52-year-old male with recurrent sinus
infections, nasal congestion...
```

Prompt:
The model is asked to classify whether the note indicates the need for an ENT referral, does not support a referral, or contains no mention regarding referral.
Example:

```
Provide an answer to the following question:
"Does the following patient meets the criteria
to be referred to an ear, nose and throat
specialist?" with the following context
<context>
Answer the question with 'A' for yes, 'B' for
no or 'C' for no mention. Do not provide any
additional details or response, just a simple A,
B or C response.
```

Evaluation metric:
Exact match: Fraction of instances where the model's classification exactly matches the reference label.
Gold-standard response:
A single-letter classification indicating the presence or absence of a referral justification, or lack of mention.
Example:

```
A
```

Model evaluation and cost–performance analysis

Model selection and inference pipeline. We evaluated nine state-of-the-art LLMs under a uniform prompting and decoding regimen. All models were queried via their respective APIs or local endpoints with sampling temperature set to 0 for deterministic outputs. All experiments were conducted on a PHI-compliant shared cluster maintaining full Health Insurance Portability and Accountability Act (HIPAA) compliance.

Performance metrics. To quantify task performance, we computed:

- **Pairwise win rate:** For each of the 37 benchmarks, we compared each model against every other; a 'win' is assigned if a model's normalized score ≥ its rival's. We then averaged wins over all pairings.
- **Macro-average score:** The overall performance score calculated by averaging results across all 37 benchmarks, with each benchmark weighted equally regardless of size (0–1 scale).

Evaluation of open-ended benchmarks. For open-ended benchmarks requiring subjective assessment, traditional automated metrics like ROUGE and BLEU prove inadequate for evaluating medical content quality. Early evaluation frameworks used single LLMs as judges, but this approach was affected by high variance and systematic biases. To address these limitations, the field has evolved toward ensemble-based evaluation using multiple independent AI judges—termed 'LLM-as-Jury'—which demonstrates improved agreement with human expert ratings[29]. Advanced implementations incorporate chain-of-thought reasoning for each evaluator and specialized tools for detecting hallucinations and factual errors[30,31]. MedHELM implements a three-judge ensemble without chain-of-thought prompting to optimize the balance between evaluation reliability and computational efficiency.

Evaluation using LLM-jury. We selected a three-member jury based on prior research demonstrating that odd-numbered panels reduce tie scenarios while maintaining reliability[32]. The jury composition (GPT-4o, Claude 3.7 Sonnet, LLaMA 3.3 70B) was chosen to represent diverse model architectures and training approaches, minimizing systematic bias from any single provider. To ensure robustness, we evaluated all seven possible jury combinations; this robustness analysis is detailed in the 'Results' (Extended Data Tables 5 and 6). We prompted each judge to score the model-generated responses on a 1–5 Likert scale according to three axes adopted from ref. 33:

- **Accuracy:** Factual correctness and adherence to medical guidelines.
- **Completeness:** Thoroughness in addressing all aspects of the query.
- **Clarity:** Organization, readability and easy to understand language.

For the NoteExtract benchmark, which requires restructuring free-form clinical care plans into a specified format without a gold-standard response, we modified our evaluation approach. We replaced the 'completeness' criterion with 'structure', which assessed whether model responses properly reorganized the input text according to the requested format. The final LLM-jury score for each response is the mean of all nine ratings (3 judges × 3 axes).

LLM-jury prompt standardization. All jury prompts follow a standardized structure to ensure consistent evaluation across benchmarks. While each jury prompt is customized for the given benchmark, we include the ACI-Bench jury prompt to serve as an example of how our LLM-jury prompt is structured (Supplementary Fig. 2).

Clinician rating. To validate the LLM-jury approach, we collected clinician ratings on a subset of two open-ended benchmarks (MEDIQA and ACI-Bench). We selected these benchmarks because they were publicly available (facilitating annotation) and represent different categories (patient communication and education, and clinical note generation, respectively). Twenty clinicians across various specialties each scored a subset of responses on the same three axes, with at least two clinicians per instance.

Clinician–LLM agreement metrics. We assessed agreement between the LLM-jury and clinicians via two ICCs, after applying z-score normalization to each rater's composite (mean of the three axis ratings) to remove scale biases according to equation (1):

$$z_{ij} = \frac{x_{ij} - \bar{x}_i}{s_i}, \tag{1}$$

where x_{ij} is rater i's composite on instance j.

1. ICC(3, k)$_z$: We treat the clinicians' mean and LLM-jury mean as two fixed raters in a two-way mixed-effects model, thus, we compute equation (2):

$$\mathrm{ICC}(3,k) = \frac{\mathrm{MS}_{\mathrm{cases}} - MS_{\varepsilon}}{\mathrm{MS}_{\mathrm{cases}}}, \quad k = 2, \tag{2}$$

to quantify absolute agreement.

2. Average clinician–clinicianICC$_z$: For every pair of clinicians who scored at least two common instances, we z-normalize each rater's composite scores and compute ICC(3, k = 2). We then average these pairwise ICCs (and bootstrap a 95% confidence interval) to give a representative inter-clinician agreement baseline.

Together, these metrics measure (i) the fidelity of LLM-jury to expert clinician rating and (ii) whether LLM-Clinician alignment approaches inter-clinician agreement.

ROUGE and BERTScore were also computed for all open-ended benchmarks with gold-standard responses but used only as secondary metrics due to their limited alignment with clinical judgment.

Statistical power and minimum detectable effect analysis. To ensure adequate statistical power for detecting meaningful performance differences, we calculated minimum detectable effect (MDE) values for each benchmark using a paired evaluation framework. We calculate the minimum detectable effect, where b is the benchmark, ij are two models to compare in a paired evaluation selected from all nine models in M, σ_b^{ij} is the standard deviation of the difference between the two selected model outputs for every question in the benchmark, and n_b is the number of questions in the benchmark. SD$_b$ is the standard deviation over the total pairwise MDE scores for the given benchmark, b. Additionally, we set $\alpha = 0.05$ and $\beta = 0.20$ (equation (3)).

$$\mathrm{MDE}_b = \frac{1}{|\mathcal{M}|} \sum_{(i,j)\in\mathcal{M}} \frac{(z_{1-\frac{\alpha}{2}} + z_{1-\beta}) \cdot \sigma_b^{(i,j)}}{\sqrt{n_b}} \pm \mathrm{SD}_b \tag{3}$$

Data availability

This study used 37 datasets with varying levels of accessibility. The minimum dataset necessary to interpret and verify the findings consists of all 37 datasets used in this study. The core model ranking trends can be partially validated using the 23 publicly accessible and gated datasets, although complete replication requires the full dataset including the 14 private datasets that cannot be shared due to institutional agreements and patient privacy protections. Sixteen datasets are publicly accessible without restrictions: MedCalc-Bench[34], MTSamples[35], Medec[36], HeadQA[37], Medbullets[38], MedQA[16], MedMCQA[7], ACI-Bench[39], MTSamples Procedures[35], MedicationQA[40], MedDialog[41], MEDIQA-QA[42], PubMedQA[43], EHRSQL[44], RaceBias[45] and MedHallu[46]. Seven datasets require credentialed access due to protected health information: EHRSHOT[47] via PhysioNet, MedAlign[48] via Redivis, DischargeMe[49] via PhysioNet, MIMIC-RRS[50] via PhysioNet, MIMIC-BHC[51] via PhysioNet, N2C2-CT[52] via n2c2.org and MIMIC-IV Billing Code[53] via PhysioNet. Fourteen datasets were developed through partnership with Stanford Healthcare or created specifically for this study and cannot be shared due to institutional data use agreements and patient privacy protections: CLEAR[54], ADHD-Behavior[55], ADHD-MedEffects[56], NoteExtract, PatientInstruct, MedConfInfo[57], MentalHealth, PrivacyDetection[58], ProxySender[58], BMT-Status, HospiceReferral, ClinicReferral, CDI-QA and ENT-Referral.

Any further distribution of datasets is subject to the terms of use and data-sharing agreements stipulated by the original creators.

Code availability

The MedHELM evaluation framework code is openly available at https://github.com/stanford-crfm/helm/ with comprehensive documentation at https://crfm-helm.readthedocs.io/en/latest/medhelm/. This repository includes the complete codebase for reproducing all benchmarks, evaluation metrics and model assessments presented in this study. Additional scripts and code used for data analysis and figure generation are available at https://github.com/som-shahlab/medhelm/, enabling full reproduction of all plots and results presented in this paper. For researchers wishing to evaluate custom models on private datasets, models can be submitted via pull requests to the GitHub repository. Evaluations are performed within our secure environment with results shared on the public leaderboard at https://crfm.stanford.edu/helm/medhelm/latest/ while maintaining dataset privacy.

References

16. 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).
17. Wornow, M. et al. Context clues: evaluating long context models for clinical prediction tasks on EHRs. In *Proc. 13th International Conference on Learning Representations* https://openreview.net/pdf?id=zg3ec1TdAP (ICLR, 2025).
18. Liu, F. et al. Large language models in the clinic: a comprehensive benchmark. Preprint at https://arxiv.org/abs/2405.00716 (2024).
19. Wu, C. et al. Towards evaluating and building versatile large language models for medicine. *NPJ Digit. Med.* **8**, 58 (2025).
20. Ouyang, Z. et al. CliMedBench: a large-scale Chinese benchmark for evaluating medical large language models in clinical scenarios. In *Proc. 2024 Conference on Empircal Methods in Natural Language Processing* 8428–8438 (EMNLP, 2024).
21. Singhal, K. et al. Large language models encode clinical knowledge. *Nature* **620**, 172–180 (2023).
22. Sandmann, S. et al. Benchmark evaluation of DeepSeek large language models in clinical decision-making. *Nat. Med.* **31**, 2546–2549 (2025).
23. Cai, Y. et al. MedBench: a large-scale Chinese benchmark for evaluating medical large language models. In *Proc. 38th AAAI Conference on Artificial Intelligence* **38**, 17709–17717 (AAAI, 2024).
24. Pal, A., Umapathi, L. K. & Sankarasubbu, M. Med-HALT: medical domain hallucination test for large language models. In *Proc. Conference on Computational Natural Language Learning (CoNLL)* 314–334 (CoNLL, 2023).
25. Han, T., Kumar, A., Agarwal, C. & Lakkaraju, H. MedSafetyBench: evaluating and improving the medical safety of large language models. In *38th Conference on Neural Information Processing Systems (NeurIPS 2024) Track on Datasets and Benchmarks* https://openreview.net/pdf?id=cFyagd2Yh4 (2024).
26. Liu, F. et al. Application of large language models in medicine. *Nat. Rev. Bioeng.* **3**, 85–104 (2025).
27. Magar, I. & Schwartz, R. Data contamination: from memorization to exploitation. In *Proc. 60th Annual Meeting of the Association of Computatonal Linguistics* (Vol. 2: Short Papers) 157–165 (ACL, 2022).
28. Deng, J. et al. ImageNet: a large-scale hierarchical image database. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition* 248–255 (IEEE, 2009).
29. Gu, J. et al. A survey on LLM-as-a-judge. Preprint at https://arxiv.org/abs/2411.15594 (2025).
30. Madaan, L. et al. Quantifying variance in evaluation benchmarks. Preprint at https://arxiv.org/abs/2406.10229 (2024).
31. Manakul, P., Liusie, A. & Gales, M. J. F. SelfCheckGPT: zero-resource black-box hallucination detection for generative large language models. In *Proc. 2023 Conference on Empirical Methods Natural Language Processing* 9004–9017 (EMNLP, 2023).
32. Guha, B. Secret ballots and costly information gathering: the jury size problem revisited. *MPRA Paper* no. 73048 (2016).
33. Van Veen, D. et al. Adapted large language models can outperform medical experts in clinical text summarization. *Nat. Med.* **30**, 1134–1142 (2024).
34. Khandekar, N. et al. MedCalc-Bench: evaluating large language models for medical calculations. In *38th Conference on Neural Information Processing Systems (NeurIPS 2024) Track on Datasets and Benchmarks* https://openreview.net/pdf?id=VXohja0vrQ (2024).
35. MT Samples: collection of transcribed medical transcription sample reports and examples. *MT Samples* https://www.mtsamples.com/ (2023).
36. Ben Abacha, A. et al. MEDEC: a benchmark for medical error detection and correction in clinical notes. in *Findings of the Association for Computational Linguistics* 22539–22550 (ACL, 2025).
37. Vilares, D. & Gómez-Rodríguez, C. HEAD-QA: a healthcare dataset for complex reasoning. In *Proc. 57th Annual Meeting of the Association of Computational Linguistics* 960–966 (ACL, 2019).
38. Chen, H., Fang, Z., Singla, Y. & Dredze, M. Benchmarking large language models on answering and explaining challenging medical questions. In *Proc. 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)* 3563–3599 (NAACL, 2025).
39. Yim, W. -W. et al. ACI-BENCH: a novel ambient clinical intelligence dataset for benchmarking automatic visit note generation. *Sci. Data* **10**, 586 (2023).
40. Ben Abacha, A. et al. Bridging the gap between consumers' medication questions and trusted answers. In *MEDINFO 2019: Health and Wellbeing e-Networks for All* 25–29 (IOS Press, 2019).
41. Zeng, G. et al. MedDialog: large-scale medical dialogue datasets. In *Proc. 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (eds Webber, B. et al.) https://doi.org/10.18653/v1/2020.emnlp-main.743 (Association for Computational Linguistics, 2020).
42. Abacha, A. B., Shivade, C. & Demner-Fushman, D. Overview of the MEDIQA 2019 shared task on textual inference, question entailment and question answering. In *Proc. 18th BioNLP Workshop Shared Task* 16–25 (ACL, 2019).
43. Jin, Q., Dhingra, B., Liu, Z., Cohen, W. & Lu, X. PubMedQA: a dataset for biomedical research question answering. In *Proc. 2019 Conference on Empirical Methods Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* 2567–2577 (EMNLP, 2019).
44. Lee, G. et al. EHRSQL: a practical text-to-SQL benchmark for electronic health records. In *36th Conference on Neural Information Processing Systems (NeurIPS 2022) Track on Datasets and Benchmarks* https://openreview.net/pdf?id=B2W8Vy0rarw (NeurIPS, 2022).
45. Omiye, J. A., Lester, J. C., Spichak, S., Rotemberg, V. & Daneshjou, R. Large language models propagate race-based medicine. *NPJ Digit. Med.* **6**, 195 (2023).
46. Pandit, S. et al. MedHallu: a comprehensive benchmark for detecting medical hallucinations in large language models. Preprint at https://arxiv.org/abs/2502.14302 (2025).
47. Wornow, M., Thapa, R., Steinberg, E., Fries, J. A. & Shah, N. H. EHRSHOT: an EHR benchmark for few-shot evaluation of foundation models. In *37th Conference on Neural Information Processing Systems (NeurIPS 2023) Track on Datasets and Benchmarks* https://openreview.net/pdf?id=CsXC6IcdwI (2023).

48. Fleming, S. L. et al. MedAlign: a clinician-generated dataset for instruction following with electronic medical records. In *Proc. Thirty-Eighth AAAI Conf. Artif. Intell.* **38**, 21545–21555 (AAAI, 2024).
49. Xu, J. Discharge me: BioNLP ACL'24 shared task on streamlining discharge documentation (version 1.3). *PhysioNet* https://doi.org/10.13026/0zf5-fx50 (2024).
50. Chen, Z., Varma, M., Wan, X., Langlotz, C. & Delbrouck, J.-B. Toward expanding the scope of radiology report summarization to multiple anatomies and modalities. In *Proc. 61st Annual Meeting of the Association of Computational Linguistics (Vol. 2: Short Papers)* 469–484 (ACL, 2023).
51. Aali, A. et al. A dataset and benchmark for hospital course summarization with adapted large language models. *J. Am. Med. Inform. Assoc.* **32**, 470–479 (2025).
52. Henry, S., Buchan, K., Filannino, M., Stubbs, A. & Uzuner, O. 2018 n2c2 shared task on adverse drug events and medication extraction in electronic health records. *J. Am. Med. Inform. Assoc.* **27**, 3–12 (2020).
53. Edin, J. et al. Automated medical coding on MIMIC-III and MIMIC-IV: a critical review and replicability study. In *Proc. 46th Int. ACM SIGIR Conf. Research and Development in Information Retrieval* 2572–2582 (SIGIR, 2023).
54. Lopez, I. et al. Clinical entity augmented retrieval for clinical information extraction. *NPJ Digit. Med.* **8**, 45 (2025).
55. Pillai, M., Posada, J., Gardner, R. M., Hernandez-Boussard, T. & Bannett, Y. Measuring quality-of-care in treatment of young children with attention deficit/hyperactivity disorder using pre-trained language models. *J. Am. Med. Inform. Assoc.* **31**, 949–957 (2024).
56. Bannett, Y. et al. Applying large language models to assess quality of care: monitoring ADHD medication side effects. *Pediatrics* **155**, e2024067223 (2025).
57. Rabbani, N. et al. Evaluation of a large language model to identify confidential content in adolescent encounter notes. *JAMA Pediatr.* **178**, 308–310 (2024).
58. Tse, G. et al. Large language model responses to adolescent patient and proxy messages. *JAMA Pediatr.* **179**, 93–94 (2025).

Acknowledgements

We thank L. Chen for help with the cover art submission. S.B. is supported by the Stanford Graduate Fellowship. A.U. is supported by the National Science Foundation Graduate Research Fellowship and the Stanford Graduate Fellowship. J.H.C. was supported in part by National Institutes of Health (NIH)/National Institute of Allergy and Infectious Diseases (1R01AI17812101), NIH-NCATS-Clinical and Translational Science Award (UM1TR004921), Stanford Bio-X Interdisciplinary Initiatives Seed Grants Program (IIP), NIH/Center for Undiagnosed Diseases at Stanford (U01 NS134358), Stanford RAISE Health Seed Grant 2024 and Josiah Macy Jr. Foundation (AI in Medical Education). M.F.G. acknowledges institutional grant funding from Xrad Therapeutics. M.O., H.Q., S.J. and L.S. acknowledge employment support from Microsoft. A.C and R.D. acknowledge grant funding from ARPA-H LLM Care. The remaining authors either declared no funding sources relevant to this work or did not provide funding information.

Author contributions

S.B., H.C., M.F. and A.U. conceived the study, designed the experiments and led the overall project equally, reported in alphabetical order by last name. N.H.S., K.S., M.P., P.L. and T.L. contributed to the ideation process for a clinical benchmark. S.B., H.C., M.F., A.U., N.H.S., Y.M., S.K., S.J., M.O., H.Q., L.S. and W.-w.Y. contributed to the methods contributions of MedHELM, including but not limited to generating the task categorization framework, dataset discussion/identification and metric discussion. S.B., A.N., S.V., S.J., B.P., O.F., S.S., E.G., D.-h.Y., B.S., E.R., S.G., V.D., R.C., R.S., C.-C.C., J.J., T.P., F.G., S.L., A.C., C.H., M.R., M.G., M.K., F.N., P.C., J.C. and H.P. contributed to the design, validation and refinement of the MedHELM taxonomy, including task categorization, dataset mapping and clinical expert review. P.C., M.W., A.S., J.A.F., N.H.S., H.K., J.L., V.K., N.K., T.K., J.M.B., N.A., C.L., W.-w.Y., R.D., J.C., E.A., K.M., N.R., N.A., Y.B. and G.T. contributed datasets for MedHELM. A.S., B.M., A.A., V.Z., J.H. and V.K. enabled the MedHELM experiments by providing critical support for implementing and executing code and runs within the Stanford Health Care computing infrastructure. S.B., H.C., M.F. and A.U. drafted the manuscript and integrated feedback from collaborators. All authors reviewed and approved the final version of the manuscript.

Competing interests

M.O., H.Q., S.J. and L.S. are employees of Microsoft. J.A.F. reports healthcare AI consulting for Snorkel AI and holds stock options in Snorkel AI. This entity had no role in study design, data collection or analysis, decision to publish, or manuscript preparation. A.S. reports employment and owning stock options in Benchmark Health and Daybreak Health; work is unrelated to the research topics explored in MedHELM. P.C. is a Healthcare AI Data Specialist consultant working with OpenAI; this work is unrelated to the research topics explored in MedHELM. A.N. reports cofounding and owning stock in UpDoc. This entity was not involved in any way with the work presented. S.V. reports previously working as an independent contractor for OpenAI's health and safety initiatives; this work was unrelated to the research topics explored in MedHELM. J.J. is the founder of Jindal Neurology and a per diem physician with Kaiser Permanente. D.Y. reports previously working as an independent contractor for OpenAI's health and safety initiatives; this work was unrelated to the research topics explored in MedHELM. R.C. reports research funding from Samsung Electronics and serving as advisor to XP Health. None of these entities were involved in any way with the work presented. F.G. is Executive Director of Healthy Oregon and Chief Medical Information Officer at Empathia.ai; none of these entities were involved in any way with the work presented. S.L. serves as an advisor to Google, Codex Health, Gaia Health and Added Health. None of these entities were involved in any way with the work presented. M.F.G. reports institutional grant funding from Xrad Therapeutics. K.S. is a shareholder and advisor to UpDoc, Salud Now, Medeloop and Alethia. None of these entities were involved in any way with the work presented. J.H.C. has received research funding support from: NIH/National Institute of Allergy and Infectious Diseases (1R01AI17812101); NIH-NCATS Clinical and Translational Science Award (UM1TR004921); Stanford Bio-X Interdisciplinary Initiatives Seed Grants Program (IIP); NIH/Center for Undiagnosed Diseases at Stanford (U01 NS134358); Stanford RAISE Health Seed Grant 2024; and Josiah Macy Jr. Foundation (AI in Medical Education). J.H.C. is also cofounder of Reaction Explorer, which develops and licenses organic chemistry education software, and has received paid medical expert witness fees (Sutton Pierce, Younker Hyde MacFarlane, Sykes McAllister, Elite Experts), consulting fees from ISHI Health, and one-time honoraria or travel expenses for invited presentations from insitro, General Reinsurance Corporation, AASCIF, Cozeva and other industry conferences, academic institutions and health systems. E.A. reports consulting fees and stock from Fourier Health. N.A. reports serving on Scientific Advisory Boards for January AI, Parallel Bio and Medeloop, and cofounding Takeoff41. None of these entities were involved in any way with the work presented. A.C. reports cofounding and owning stock in Cognita Imaging; owning stock in Subtle Medical, BrainKey and LVIS Corp; and providing consulting services to Elucid Bioimaging and Patient Square Capital. None of these entities were involved in any way with the work presented. S.K. reports cofounding and owning stock options in Virtue AI; this work is unrelated to the research topics explored in MedHELM. E.H. reports serving in a senior role at Microsoft and

owning stock in Microsoft, which has efforts in the healthcare domain. M.A.P. is an advisor to Akasa and Surgical Safety Technologies. None of these entities were involved in any way with the work presented. N.H.S. reports being a cofounder of Prealize Health and Atropos Health, serving on the Board of the Coalition for Healthcare AI, and serving as an advisor to Opala, Curai Health, JnJ Innovative Medicines and AbbVie pharmaceuticals. None of these entities were involved in any way with the work presented. The other authors declare no competing interests.

Additional information

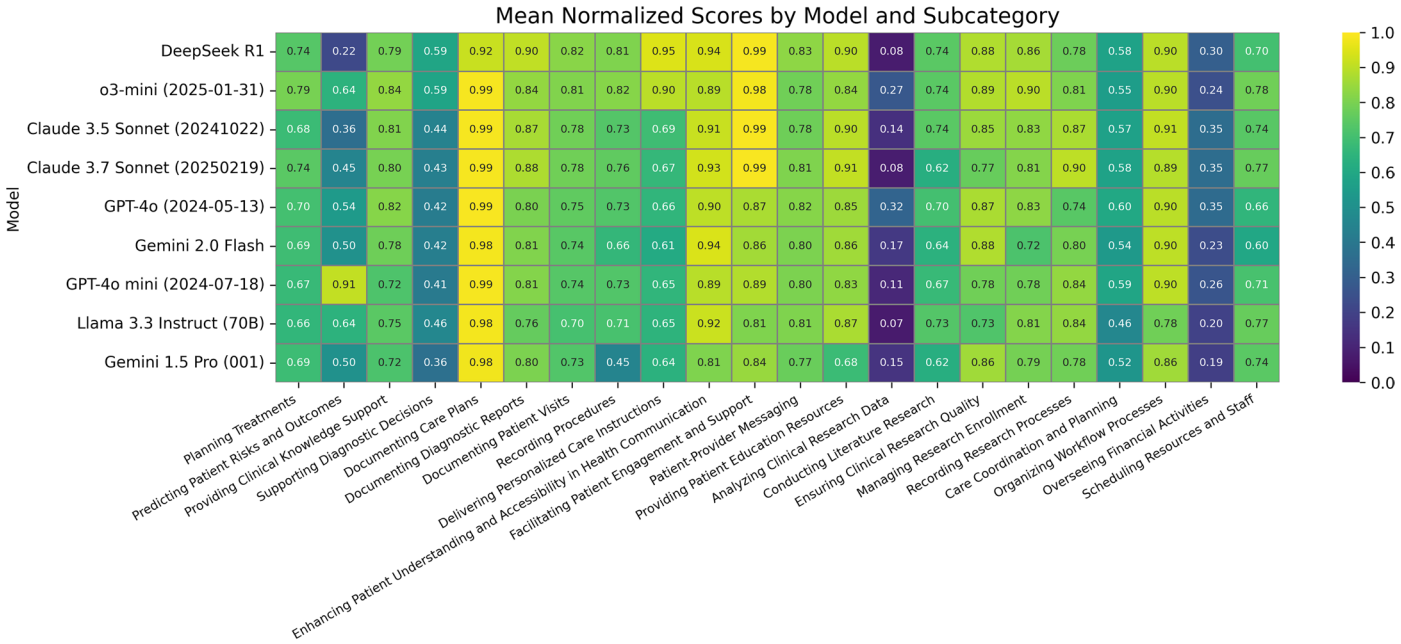
Extended data is available for this paper at https://doi.org/10.1038/s41591-025-04151-2.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41591-025-04151-2.

Correspondence and requests for materials should be addressed to Suhana Bedi.

Peer review information *Nature Medicine* thanks Fenglin Liu, Yifan Peng and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Primary Handling Editors: Lorenzo Righetto and Saheli Sadanand, in collaboration with the *Nature Medicine* team.

https://doi.org/10.1038/s41591-025-04151-2



Extended Data Fig. 1 | Aggregated performance across MedHELM subcategories. Each cell represents the mean performance of all models for a given subcategory, allowing comparison of relative strengths and weaknesses across clinical task types.

https://doi.org/10.1038/s41591-025-04151-2

Extended Data Table 1 | Overview of all 37 benchmarks included in MedHELM, categorized by category, access level and curation status

Category	Dataset Name	Instances Evaluated	Dataset Description	Access Level	Curation Status
Clinical Decision Support	MedCalc-Bench [ME20]	1000	A dataset which consists of a patient note, a question requesting to compute a specific medical value, and a ground truth answer.	Public	Existing
	CLEAR [ME21]	1022	A dataset that evaluates medical condition detection from patient notes using yes/no/maybe classifications.	Private	Formulated
	MTSamples [ME22]	427	A dataset that provides transcribed medical reports and prompts models to generate appropriate treatment plans.	Public	Formulated
	Medec [ME23]	597	A dataset containing medical narratives with error detection and correction pairs.	Public	Existing
	EHRSHOT [ME24]	1000	A dataset given a patient record of EHR codes, classifying if an event will occur at a future date or not.	Gated	Existing
	HeadQA [ME25]	1000	A collection of biomedical multiple-choice questions for testing medical knowledge.	Public	Existing
	Medbullets [ME26]	308	A USMLE-style medical question dataset with multiple-choice answers and explanations.	Public	Existing
	MedQA [ME1]	1000	A dataset containing USMLE-style medical questions multiple-choice answers and explanations.	Public	Existing
	MedMCQA [MA7]	1000	A dataset containing indian medical entrance exam questions testing clinical knowledge and problem-solving.	Public	Existing
	MedAlign [ME27]	149	A dataset that asks models to answer questions/follow instructions over longitudinal EHR.	Gated	Existing
	ADHD-Behavior [ME28]	423	A dataset that classifies whether a clinical note contains a clinician recommendation for parent training in behavior management, which is the first-line evidence-based treatment for young children with ADHD.	Private	New
	ADHD-MedEffects [ME29]	915	A dataset that classifies whether a clinical note contains documentation of side effect monitoring (recording of absence or presence of medication side effects), as recommended in clinical practice guidelines.	Private	New
Clinical Note Generation	DischargeMe [ME30]	1000	A dataset that provides discharge text as well as the radiology report text collected from MIMIC-IV data such that models can generate discharge instructions and brief hospital course information generation.	Gated	Existing
	ACI-Bench [ME31]	120	A dataset of patient-doctor conversations paired with structured clinical notes.	Public	Existing
	MTSamples Procedures [ME22]	128	A dataset that provides a patient note regarding an operation, with the objective to document the procedure.	Public	Formulated
	MIMIC-RRS [ME32]	1000	A dataset containing radiology reports with findings sections from MIMIC-III paired with their corresponding impression sections, used for generating radiology report summaries.	Gated	Formulated
	MIMIC-BHC [ME33]	1000	A summarization task using a curated collection of preprocessed discharge notes paired with their corresponding brief hospital course (BHC) summaries.	Gated	Existing
	NoteExtract	487	A dataset containing free form text of a clinical health worker care plan from Nigeria, with the associated goal being to restructure that text into a given format.	Private	New
Patient Communication and Education	MedicationQA [ME34]	689	A dataset containing open text question-answer pairs regarding consumer health questions about medication.	Public	Existing
	PatientInstruct	361	A dataset containing case details used to generate customized post-procedure patient instructions.	Private	New
	MedDialog [ME35]	1000	A collection of doctor-patient conversations with corresponding summaries.	Public	Existing
	MedConfInfo [ME36]	1000	A dataset of clinical notes from adolescent patients used to identify sensitive protected health information that should be restricted from parental access.	Private	New
	MEDIQA-QA [ME37]	150	A dataset including a medical question, a set of candidate answers, relevance annotations for ranking, and additional context to evaluate understanding and retrieval capabilities in a medical setting.	Public	Existing
	MentalHealth	67	A dataset containing a counselor and mental health patient conversation from India, where the objective is to generate an empathetic counselor response.	Private	New
	PrivacyDetection [ME38]	300	A dataset that determines if a message leaks any confidential information from the patient	Private	New
	ProxySender [ME38]	300	A dataset that determines if a message was sent by a proxy user	Private	New
Medical Research Assistance	PubMedQA [ME39]	1000	A dataset that provides pubmed abstracts and asks associated questions yes/no/maybe questions.	Public	Existing
	EHRSQL [ME40]	1000	A dataset that generates an SQL query that would be used in clinical research given a natural language instruction.	Public	Existing
	BMT-Status	220	A dataset containing patient notes with associated questions and answers related to bone marrow transplantation.	Private	New
	RaceBias [ME41]	167	A collection of LLM outputs in response to medical questions with race-based biases, with the objective being to classify whether the output contains racially biased content.	Public	Formulated
	N2C2-CT [ME42]	86	A dataset that provides clinical notes and asks the model to classify whether the patient is a valid candidate for a provided clinical trial.	Gated	Existing
	MedHallu [ME43]	1000	A dataset of PubMed articles and associated questions, with the objective being to classify whether the answer is factual or hallucinated.	Public	Existing
Administration and Workflow	HospiceReferral	1000	A dataset evaluating performance in identifying appropriate patient referrals to hospice care.	Private	New
	MIMIC-IV Billing Code [ME44]	1000	A dataset pairing clinical notes from MIMIC-IV with corresponding ICD-10 billing codes.	Gated	Existing
	ClinicReferral	326	A dataset containing manually curated answers to questions regarding patient referrals to the Sequoia clinic.	Private	New
	CDI-QA	1000	A dataset built from Clinical Document Integrity (CDI) notes, to assess the ability to answer verification questions from previous notes.	Private	New
	ENT-Referral	1000	A dataset designed to evaluate performance in identifying appropriate patient referrals to Ear, Nose, and Throat specialists.	Private	New

Instances evaluated represents the subset of each dataset used for model evaluation, capped at approximately 1,000 instances for larger datasets due to computational constraints.

https://doi.org/10.1038/s41591-025-04151-2

Extended Data Table 2 | Model ranking comparison across general and medical evaluation frameworks

Model	MedHELM Percentile	LM Arena Percentile	HELM Percentile	Medical Performance Change
DeepSeek R1	100th (1st of 9)	93rd (16th of 224)	73rd (16th of 51)	Medical: +7–27 points gain
o3-mini	89th (2nd of 9)	81st (43rd of 224)	Not ranked	Medical: +8 points gain
Claude 3.7 Sonnet	78th (3rd of 9)	87th (29th of 224)	63rd (19th of 51)	Medical: –9 to +15 points shift
Claude 3.5 Sonnet	67th (4th of 9)	87th (30th of 224)	53rd (24th of 51)	Medical: –20 to +14 points shift
GPT-4o	56th (5th of 9)	80th (45th of 224)	51st (25th of 51)	Medical: –24 to +5 points drop
Gemini 2.0 Flash	44th (6th of 9)	86th (31st of 224)	65th (18th of 51)	Medical: –21 to –42 points drop

Percentile rankings normalize for different pool sizes (MedHELM: 9 models, LM Arena: 224 models, HELM: 51 models). Raw rankings in parentheses (rank/total). Performance shifts > ± 20 percentile points indicate clinically significant differences between general and medical capabilities.

https://doi.org/10.1038/s41591-025-04151-2

Extended Data Table 3 | Performance of smaller open-source models on public and gated MedHELM benchmarks

Healthcare Category	Dataset	Evaluation Metric	Qwen-2.5-7B-instruct	Phi-3.5-mini-instruct	MedGemma-4b-it
Clinical Decision Support	MedCalc-Bench	Exact Match	0.091	0.01	0.041
Clinical Decision Support	MTSamples	LLM Jury	3.936	3.66	3.855
Clinical Decision Support	Medec	MedecFlagAcc	0.496	0.508	0.521
Clinical Decision Support	EHRSHOT	Exact Match	0.806	0.231	0.395
Clinical Decision Support	HeadQA	Exact Match	0.734	0.679	0.651
Clinical Decision Support	Medbullets	Exact Match	0.406	0.192	0.416
Clinical Decision Support	MedQA	Exact Match	0.575	0.511	0.459
Clinical Decision Support	MedMCQA	Exact Match	0.541	0.491	0.527
Clinical Decision Support	MedAlign	LLM Jury	3.224	3.072	2.076
Clinical Note Generation	ACI-Bench	LLM Jury	4.264	4.055	4.249
Clinical Note Generation	MTSamples Procedures	LLM Jury	3.645	3.679	3.44
Clinical Note Generation	MIMIC-RRS	LLM Jury	4.351	4.351	3.898
Clinical Note Generation	MIMIC-BHC	LLM Jury	3.624	3.286	3.467
Clinical Note Generation	DischargeMe	LLM Jury	3.4	2.976	3.009
Patient Communication & Education	MedicationQA	LLM Jury	4.195	3.677	3.848
Patient Communication & Education	MedDialog	LLM Jury	3.76	3.918	4.046
Patient Communication & Education	MEDIQA	LLM Jury	4.586	4.405	4.455
Medical Research Assistance	PubMedQA	Exact Match	0.55	0.476	0.699
Medical Research Assistance	EHR-SQL	EHRSQLReAns	0	0.05	0.002
Medical Research Assistance	RaceBias	Exact Match	0.587	0.144	0.503
Medical Research Assistance	N2C2 CT	Exact Match	0.461	0.372	0.38
Medical Research Assistance	MedHallu	Exact Match	0.811	0	0.355
Administration & Workflow	MIMIC-IV Billing Code	Micro-F1	0.028	0.091	0.042

Results show the performance gap between resource-efficient models and frontier LLMs across medical tasks. Private datasets were excluded due to infrastructure limitations in PHI-compliant environments.

https://doi.org/10.1038/s41591-025-04151-2

Extended Data Table 4 | Agreement of LLM-jury and automated metrics with clinician ratings

Benchmark	LLM	ROUGE-L	BERTScore-F	Clinician
Combined	0.474 (0.100, 0.690)	0.361 (0, 0.630)	0.441 (0.050, 0.670)	0.426 (0.295, 0.585)
ACI-Bench	0.305 (0, 0.670)	0.445 (0, 0.730)	0.250 (0, 0.640)	0.458 (0.201, 0.945)
MEDIQA	0.625 (0.150, 0.830)	0.343 (0, 0.710)	0.668 (0.250, 0.850)	0.520 (0.500, 0.534)

The table entries are ICC(3,k) coefficients after z-scoring within rater (ICC3k-z); 95% confidence intervals are shown in parentheses. Higher ICC indicates better agreement with clinician ratings. The last column gives the average clinician-clinician agreement for each dataset.

https://doi.org/10.1038/s41591-025-04151-2

Extended Data Table 5 | Analysis of average correlation, first place consistency, and last place consistency across all 7 permutations of jury compositions

Dataset	Spearman Correlation	Top 2 Consistency	Bottom 2 Consistency
ACI-Bench	0.830	100%	100%
NoteExtract	0.540	85.7%	100%
DischargeMe	0.940	100.0%	100%
MedDialog	0.882	100.0%	100%
MedAlign	0.923	100.0%	100%
MEDIQA	0.906	100.0%	100%
MedicationQA	0.929	100.0%	100%
MentalHealth	0.975	100.0%	100%
MIMIC-BHC	0.856	100.0%	100%
MIMIC-RRS	0.888	100.0%	85.7%
MTSamples Procedures	0.937	100.0%	100%
MTSamples	0.925	100.0%	100%
PatientInstruct	0.979	100.0%	100%

https://doi.org/10.1038/s41591-025-04151-2

Extended Data Table 6 | Performance comparison across different judges

Judge	Accuracy (Mean)	Accuracy (Std)	Completeness (Mean)	Completeness (Std)	Clarity/Structure (Mean)	Clarity/Structure (Std)
GPT-4o	4.355	0.648	3.816	0.909	4.892	0.310
LLaMA 3.3 70B	4.250	0.613	3.785	0.864	4.969	0.218
Claude 3.7 Sonnet	3.910	0.884	3.388	1.204	4.744	0.457

https://doi.org/10.1038/s41591-025-04151-2

Extended Data Table 7 | Comparison of LLMs by architecture, access type, and token usage metrics across MedHELM

Model	Model Creator	Window Size	Access	Benchmark Tokens	Jury Tokens	Benchmark Cost	Jury Cost	Total Cost
Claude 3.5 Sonnet (20241022)	Anthropic	200,000	Closed	246,330,404	45,868,483	$779.82	$792.21	$1,572.03
Claude 3.7 Sonnet (20250219)	Anthropic	200,000	Closed	242,882,578	42,326,449	$769.41	$768.38	$1,537.79
Gemini 1.5 Pro (001)	Google	1,000,000	Closed	277,639,659	42,864,352	$359.81	$771.73	$1,131.54
Gemini 2.0 Flash	Google	1,000,000	Closed	277,143,242	42,864,352	$43.10	$771.73	$814.83
GPT-4o (2024-05-13)	OpenAI	128,000	Closed	249,089,792	41,929,516	$648.20	$765.91	$1,414.11
GPT-4o mini (2024-07-18)	OpenAI	128,000	Closed	249,089,792	41,929,516	$38.89	$765.91	$804.80
Llama 3.3 Instruct (70B)	Meta AI	128,000	Open	243,259,359	42,121,933	$172.70	$767.13	$939.82
DeepSeek R1	DeepSeek	128,000	Open	342,484,366	68,364,208	$891.89	$957.96	$1,849.85
o3-mini (2025-01-31)	OpenAI	128,000	Closed	346,329,125	68,617,978	$801.65	$959.54	$1,761.18

Benchmark tokens denote total input/output tokens in completing the medical task represented by the benchmark; jury tokens are tokens used for open-ended evaluations via the LLM-jury. The total cost reflects the estimated per-model expense of running a MedHELM evaluation across 37 benchmarks and represents an upper bound based on maximum output token usage. For nine models and 37 benchmarks in the current MedHELM suite, one update to the leaderboard is estimated to cost 11, 825.97.

https://doi.org/10.1038/s41591-025-04151-2

Extended Data Table 8 | Comprehensive comparison of medical LLM benchmarking frameworks

Benchmark Suite	Categories Covered	Datasets	Contains Real Patient Data	Dimensions of Evaluation
MedS-Bench [ME4]	Clinical Decision Support, Clinical Note Generation, Patient Communication & Education	28	Yes (MIMIC)	Accuracy, Comprehensiveness, Factuality, Robustness
ClinicBench [ME5]	Clinical Decision Support, Clinical Note Generation, Patient Communication & Education	17	Yes (MIMIC)	Accuracy, Comprehensiveness, Factuality, Robustness
CliMedBench [ME6]	Clinical Decision Support, Clinical Note Generation	1	Yes (MIMIC)	Accuracy, Factuality, Robustness, Fairness/bias/toxicity
MultiMedQA [ME7]	Clinical Decision Support, Medical Research Assistance, Patient Communication & Education	7	No	Accuracy, Factuality, Fairness/bias/toxicity
MIMIC-CDM [MA9]	Clinical Decision Support	1	Yes (MIMIC)	Accuracy, Comprehensiveness, Robustness
Sandmann et al. (2025) [ME8]	Clinical Decision Support	1	No	Accuracy
MedBench [ME9]	Clinical Decision Support	20	No	Accuracy
Med-HALT [ME10]	Clinical Decision Support	7	No	Accuracy, Factuality
MedSafetyBench [ME11]	Patient Communication & Education	1	No	Fairness/bias/toxicity
MedHELM	Clinical Decision Support, Clinical Note Generation, Patient Communication, Medical Research, Admin & Workflow	37	Yes (MIMIC, Stanford)	Accuracy, Comprehensiveness, Factuality, Deployment

MedHELM covers the broadest scope with 37 datasets across all five clinical categories, real patient data from multiple sources, and multiple dimensions of evaluation, where the dimensions are taken from [MA8].

https://doi.org/10.1038/s41591-025-04151-2

Extended Data Table 9 | Taxonomic coverage comparison across medical LLM benchmarking frameworks

Benchmark	Clinical Decision Support	Clinical Note Generation	Patient Communication & Education	Medical Research Assistant	Administration & Workflow
MedS-Bench [ME4]	Y	Y	Y		
ClinicBench [ME5]	Y	Y	Y		
CliMedBench [ME6]	Y	Y			
MultiMedQA [ME7]	Y		Y	Y	
MIMIC-CDM [MA9]	Y				
Sandmann et al. (2025) [ME8]	Y				
MedBench [ME9]	Y				
Med-HALT [ME10]	Y				
MedSafetyBench [ME11]		Y			
Liu et al. [ME12]	Y	Y	Y	Y	
MedHELM (Ours)	Y	Y	Y	Y	Y

nature portfolio

Corresponding author(s): Suhana Bedi

Last updated by author(s): Nov 7, 2025

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	All code to preprocess the datasets is open source and can be found in the github repository for helm - https://github.com/stanford-crfm/helm.
Data analysis	All code used for experiments in this study can be found at the following GitHub links: MedHELM evaluation framework (https://github.com/stanford-crfm/helm) and data analysis scripts (https://github.com/som-shahlab/medhelm). We built upon the open-source libraries pandas version 2.0.0+ (https://pandas.pydata.org/), plotly version 5.18.0+ (https://plotly.com/), matplotlib version 3.6.3 (https://matplotlib.org/), seaborn version 0.13.2 (https://seaborn.pydata.org/), and CRFM-HELM version 0.5.5 (https://github.com/stanford-crfm/helm) for LLM evaluation and benchmarking.

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

We have written a section on data availability in the manuscript. Here is the statement copied verbatim from the manuscript - This study used 37 datasets with varying levels of accessibility. The minimum dataset necessary to interpret and verify the findings consists of all 37 datasets used in this study. The core model ranking trends can be partially validated using the 23 publicly accessible and gated datasets, though complete replication requires the full dataset including the 14 private datasets that cannot be shared due to institutional agreements and patient privacy protections. 16 datasets are publicly accessible without restrictions: MedCalc-Bench [20], MT Samples [22], Medec [23], HeadQA [25], Medbullets [26], MedQA [1], MedMCQA [7], ACI-Bench [31], MTSamples Procedures [22], MedicationQA [34], MedDialog [35], MEDIQA-QA [37], PubMedQA [46], EHRSQL [40], RaceBias [41], MedHallu [43].
Seven datasets require credentialed access due to protected health information: EHRSHOT [24] via PhysioNet, MedAlign [27] via Redivis, DischargeMe [30] via PhysioNet, MIMIC-RRS [32] via PhysioNet, MIMIC-BHC [33] via PhysioNet, N2C2-CT [42] via n2c2.org, and MIMIC-IV Billing Code [44] via PhysioNet.
14 datasets were developed through partnership with Stanford Healthcare or created specifically for this study and cannot be shared due to institutional data use agreements and patient privacy protections: CLEAR, ADHD-Behavior, ADHD-MedEffects, NoteExtract, PatientInstruct, MedConfInfo, MentalHealth, PrivacyDetection, ProxySender, BMT-Status, HospiceReferral, ClinicReferral, CDI-QA, and ENT-Referral.
Any further distribution of datasets is subject to the terms of use and data-sharing agreements stipulated by the original creators.

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	N/A
Reporting on race, ethnicity, or other socially relevant groupings	N/A
Population characteristics	For the clinician participant in the survey, we provide a breakdown by their clinical specialty.
Recruitment	They were recruited to maximize the coverage over different clinical specialties.
Ethics oversight	Stanford University

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	For each benchmark, we capped the total number of instances evaluated to 1000 due to computational constraints. This has been mentioned in Extended Table 1.
Data exclusions	As mentioned above, for larger benchmarks, we randomly selected 1000 instances and any instances beyond it were excluded.
Replication	For all large language models, the temperature was set to zero, making the outputs have very little randomness and are, thus, reproducible.
Randomization	N/A - computational evaluation study
Blinding	N/A - computational evaluation study

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	N/A
Novel plant genotypes	N/A
Authentication	N/A