HEALTHCOREBENCH: MEASURING EVALUATION ARTIFACTS FOR RELIABLE MEDICAL LLM EVALUATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Medical evaluation of large language models (LLMs) increasingly relies on heterogeneous benchmarks, yet benchmark failure is not always a reliable indicator of limited medical intelligence: unsolved samples may reflect genuine capability gaps or evaluation artifacts such as ambiguity, incorrect references, or outdated knowledge. We formulate medical benchmark construction as a **failure attribution** problem: identifying samples where failures can be meaningfully attributed to capability limitations. We introduce **HealthCoreBench**, a unified medical evaluation benchmark distilled from **47** language benchmark sources spanning **165,120** samples, of which only **13,987** (**8.47%**) are retained. HealthCoreBench uses progressive model-based filtering to identify frontier-challenging samples, followed by clinician validation to flag and correct unreliable failures. Auditing large-scale medical benchmark collections reveals that: 1) unreliable failures are prevalent, and difficulty filtering alone can even enrich evaluation artifacts, since selecting samples that strong models fail on preferentially retains defective rather than genuinely hard questions; 2) these artifacts create false successes and shift model scores unevenly, with physician validation alone changing overall scores by up to **5.18** points across models and distorting apparent medical capability; and 3) retaining only clinician-validated failures preserves model rankings while cutting average token consumption by **77.17%** (from **82.82M** to **18.91M** tokens). Overall, unreliable failures undermine evaluation validity, and failure attribution offers a path forward. We release HealthCoreBench to shift medical evaluation from "unsolved as failure" toward "failure attributable to capability limitations."

1 INTRODUCTION

Large language models (LLMs) are increasingly evaluated on medical benchmarks that measure knowledge (Jin et al., 2021; Pal et al., 2022; Hendrycks et al., 2021), clinical reasoning (Zuo et al., 2025; Chen et al., 2024), safety (Pal et al., 2023; Pandit et al., 2025), and multimodal understanding (Lau et al., 2018; Liu et al., 2021; Hu et al., 2024). Existing benchmarks have enabled systematic comparison of medical AI systems across diverse tasks, but their interpretation relies on a fundamental assumption: *a model failure reflects a limitation of model capability*. As models approach the frontier of medical performance, this assumption becomes increasingly important because evaluation failures are often used to identify remaining capability gaps.

However, a benchmark failure is not necessarily a valid measure of medical intelligence. An unsolved sample may represent a genuine limitation in medical knowledge or reasoning, but it may also arise from evaluation artifacts, including ambiguous questions, incorrect reference answers, outdated information, unreliable scoring criteria, or insufficient multimodal evidence (Gema et al., 2024; Asadi et al., 2026). Consequently, selecting samples solely based on model difficulty may produce benchmarks that are challenging but unreliable, where poor performance cannot be confidently interpreted as evidence of missing capabilities.

This issue is particularly relevant in the current evaluation landscape. The rapid development of medical benchmarks has produced a large and heterogeneous collection of datasets covering different specialties, languages, modalities, and task formats. While this diversity improves coverage, evaluating all existing samples is increasingly expensive, and the information value of individual samples varies substantially (Polo et al., 2024; Zeng & Papailiopoulos, 2026). Many samples are al-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

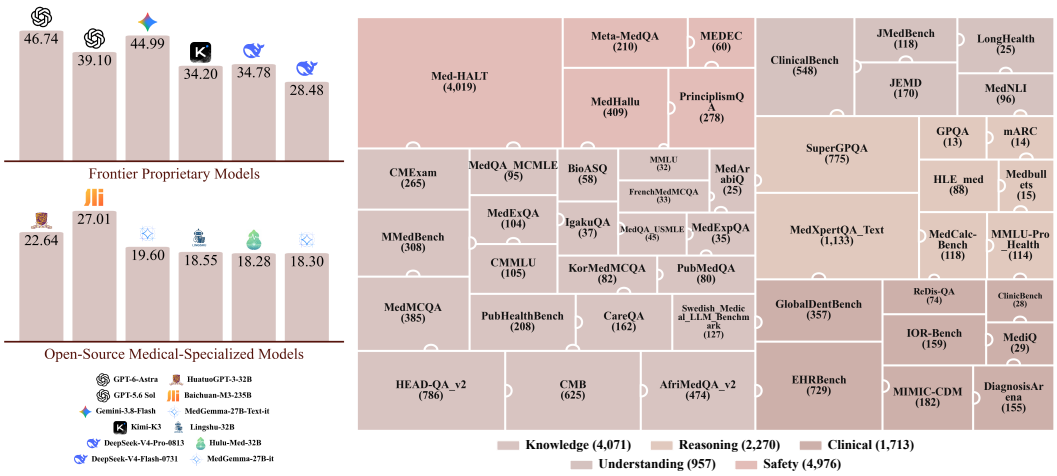

Figure 1: **Overview of HealthCoreBench.** *Left*: Overall performance of frontier proprietary models and open-source medical-specialized models. *Right*: Sample composition across source benchmarks and five capability categories.

ready solved reliably by strong models and provide limited discrimination, whereas some remaining difficult samples may not provide meaningful evidence about model capabilities. Therefore, effective evaluation requires not only identifying where models fail, but also determining whether those failures are valid.

We formulate medical benchmark construction as a **failure attribution** problem:

$$\underbrace{\text{Observed Failure}}_{\text{benchmark outcome}} = \underbrace{\text{Capability Failure}}_{\text{desired signal}} + \underbrace{\text{Evaluation Artifact}}_{\text{noise}}$$

The goal is to identify a subset of evaluation samples where model failures can be reasonably attributed to limitations in medical intelligence rather than defects in the evaluation itself. This formulation shifts benchmark construction from collecting difficult questions to validating informative failures.

To address this problem, we introduce **HealthCoreBench**, a medical language benchmark constructed through a multi-stage failure attribution pipeline. Starting from 47 existing medical benchmark entries, HealthCoreBench first applies progressively stronger and heterogeneous model panels to identify samples that remain challenging near the capability frontier. Since model difficulty alone does not guarantee validity, the remaining samples are further examined through automated auditing and physician validation to identify and remove unreliable failures caused by question defects, incorrect references, or insufficient evidence.

HealthCoreBench contains **13,987** validated samples spanning medical knowledge, clinical reasoning, safety, and understanding (Figure 1). Retaining only **8.47%** of the original **165,120** samples, it preserves the ranking of evaluated models while cutting average token consumption by **77.17%** (from **82.82M** to **18.91M** tokens). Beyond efficiency, our analysis exposes how unreliable failures distort evaluation: difficulty filtering alone enriches evaluation artifacts, raising the estimated defect rate from **1.36%** in the initial pool to **4.20%** among samples surviving the strongest model tier (a **3.09**$\times$ enrichment), and physician validation alone shifts overall scores by up to **5.18** points across models. These findings demonstrate that a non-trivial fraction of samples that appear difficult for frontier models are not valid indicators of medical capability limitations, underscoring the importance of validating failures before drawing conclusions from benchmark results.

Contributions. Our contributions are threefold: (i) we formulate medical benchmark construction as a failure attribution problem, distinguishing capability limitations from evaluation artifacts; (ii) we introduce HealthCoreBench, comprising 13,987 physician-validated samples drawn from 47 source benchmarks; and (iii) we show that difficulty filtering can enrich evaluation artifacts, while

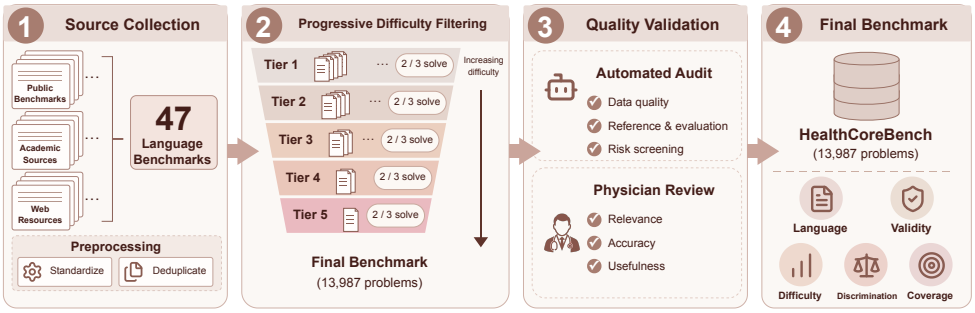

Figure 2: **Overview of HealthCoreBench Construction.** Samples from 47 benchmarks are standardized and deduplicated, then progressively filtered through five model tiers to retain challenging questions. Automated audits and physician review further assess data quality and correct identified issues, producing a final set of 13,987 samples.

validation changes model scores by up to 5.18 points and preserves model rankings with 77.17% lower token consumption.

2 HealthCoreBench: Benchmark Construction via Failure Attribution

Reliable medical evaluation requires distinguishing genuine capability limitations from failures caused by defects in the evaluation itself. We therefore formulate medical benchmark construction as a **failure attribution** problem. In §2.1, we formalize the distinction between capability failures and evaluation artifacts. We then describe how this formulation is operationalized in the construction of HealthCoreBench in §2.2. Finally, §2.3 presents the statistics of the resulting benchmark.

2.1 Problem Formulation

Given a collection of medical evaluation samples

$$\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N},$$

where x_i denotes the input (including optional multimodal evidence) and y_i represents the reference answer or evaluation target, a model failure is observed when a model prediction $\hat{y}_i$ does not match the expected outcome, i.e., $\hat{y}_i \neq y_i$.

However, an observed failure does not directly indicate a capability limitation. We introduce a latent validity variable:

$$z_i = \begin{cases} 1, & \text{if the failure reflects a genuine capability limitation,} \\ 0, & \text{if the failure results from an evaluation artifact.} \end{cases}$$

The goal of failure attribution is therefore to identify the subset:

$$\mathcal{D}^* = \{(x_i, y_i) \in \mathcal{D} : z_i = 1\},$$

where model failures can be meaningfully interpreted as evidence of missing medical capabilities.

This formulation introduces two requirements for reliable medical evaluation. First, evaluation items should remain challenging enough to expose limitations of current models. Second, the resulting failures should provide valid measurements of the intended capability rather than arise from ambiguous questions, incorrect references, outdated information, or insufficient evidence.

Table 1: **Language-model panels used for progressive difficulty filtering.** Each tier contains models from different families within a broadly comparable capability range. Only samples not reliably solved at one tier advance to the next. **# Filtered** reports the number of samples reliably solved and removed at each tier, and **# To Next Tier** reports the number of remaining samples passed to the next tier.

Tier	Models	Model Scale	# Filtered	# To Next Tier
1	Qwen3-32B gpt-oss-20b Phi-4	∼30B	117953	47167
2	LongCat-Flash-Lite Qwen2.5-72B-Instruct Llama-3.3-70B-Instruct	∼70B	14493	32674
3	Qwen3.5-122B-A10B gpt-oss-120b Mistral-Medium-3.5-128B	∼120B	8374	24300
4	Qwen3-VL-235B-A22B-Instruct MiniMax-M2.7 InternVL3.5-241B-A28B	∼200B	2944	21356
5	Gemini-3.6-Flash GPT-5.6-Terra Claude-Sonnet-5	Closed-source frontier	6893	14463

2.2 Benchmark Collection and Construction

Guided by this formulation, and as illustrated in Figure 2, we build HealthCoreBench through a three-stage pipeline: source benchmark collection (§2.2.1), progressive difficulty filtering (§2.2.2), and physician validation with automated pre-screening (§2.2.3). The first stage targets challenging samples near the capability frontier, while the latter stages ensure the retained failures reflect genuine capability limitations rather than evaluation artifacts.

2.2.1 Source Benchmark Collection

Selection Principles. We select source benchmarks according to three principles: **(1) adoption**, prioritizing evaluations widely used in research and industry; **(2) reliability**, favoring benchmarks supported by peer-reviewed research and reproducible evaluation protocols; **(3) coverage**, retaining sources that capture underrepresented capabilities, languages, specialties, or modalities. Together, these principles balance established evaluation practice with broad medical coverage.

Benchmark Sources. Following these principles, we collect **47 medical benchmark** entries. The language benchmarks span medical knowledge, clinical reasoning, numerical calculation, long-context understanding, and safety. Table 5 summarizes the collection, with the complete inventory provided in Appendix B.

Task Standardization. Benchmarks differ substantially in input structure, answer format, prompting, and evaluation criteria. We map each sample to a unified representation while preserving the original task and scoring protocol whenever possible. To ensure reliable evaluation, we exclude open-ended tasks and retain only tasks with objectively verifiable answers. We standardize prompts and output parsing, trace composite benchmarks to their sources, and deduplicate samples, yielding **165,120 samples** for difficulty filtering.

2.2.2 Progressive Difficulty Filtering

Model Panels. Question difficulty should not be determined by the idiosyncratic strengths or weaknesses of a single model. We therefore construct five progressively stronger panels for evaluation, with three models in each panel. Models within each tier represent a broadly comparable capability range but differ in architecture, training data, and model family. Panel-level agreement therefore provides a more robust difficulty signal than the outcome of any individual model. Table 1 summarizes the model panels and the number of samples filtered at each tier.

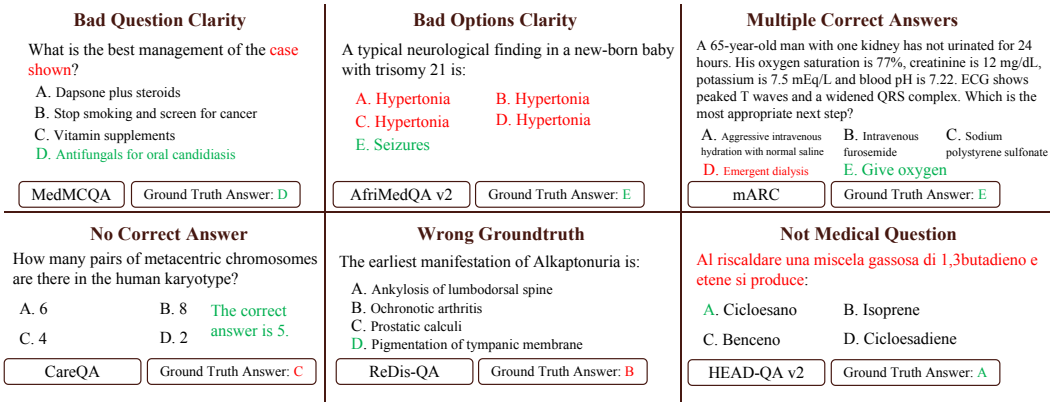

Figure 3: **Representative error cases from different medical benchmarks.** Examples of difficult samples whose observed model failures cannot be directly attributed to capability limitations. The cases illustrate common artifacts identified during quality screening, including *Bad Question Clarity*, *Bad Options Clarity*, *Multiple Correct Answers*, *No Correct Answer*, *Wrong Ground Truth*, and *Not Medical Question*.

Filtering Protocol. Filtering proceeds sequentially from Tier 1 to Tier 5. At each tier, every sample is evaluated by three models, with each model independently answering the sample three times. For a sample q, model m, and tier t, we first define model-level stability as

$$s_{m,t}(q) = \mathbb{I}\left[\frac{1}{3}\sum_{r=1}^{3}\mathbb{I}(m \text{ answers } q \text{ correctly in run } r) \geq \frac{2}{3}\right].$$

A model therefore solves a sample reliably only if it answers correctly in at least two of the three runs. We then define the tier-level pass rate over the three-model panel $\mathcal{M}_t$ as

$$p_t(q) = \frac{1}{|\mathcal{M}_t|}\sum_{m\in\mathcal{M}_t} s_{m,t}(q).$$

A sample is considered reliably solved at tier t when at least two of the three models solve it reliably, i.e., $p_t(q) \geq 2/3$. Otherwise, it advances to tier $t+1$. This two-level majority criterion reduces sensitivity to both within-model stochasticity and isolated model failures, while progressive filtering concentrates evaluation on samples that remain challenging across increasingly capable and heterogeneous model families.

2.2.3 Physician Validation with Automated Pre-Screening

Final quality decisions require medical expertise, but expert review is costly at scale. To reduce physician workload, we first apply an automated screening step that pre-annotates every sample remaining after progressive filtering, flagging likely evaluation defects so that physicians can focus their effort on the samples most in need of expert judgment.

Automated Pre-Screening. We develop a hierarchical annotation protocol to distinguish genuinely difficult samples from those affected by evaluation defects (Gema et al., 2024); Figure 3 shows representative error cases identified across source benchmarks. The protocol covers three dimensions:

- **Medical Relevance.** Whether the sample evaluates medically relevant knowledge.
- **Question Assessment.** Whether the question is unclear, incomplete, or malformed (*Bad Question Clarity*), or its options are ambiguous, duplicated, irrelevant, or incorrectly parsed (*Bad Options Clarity*).

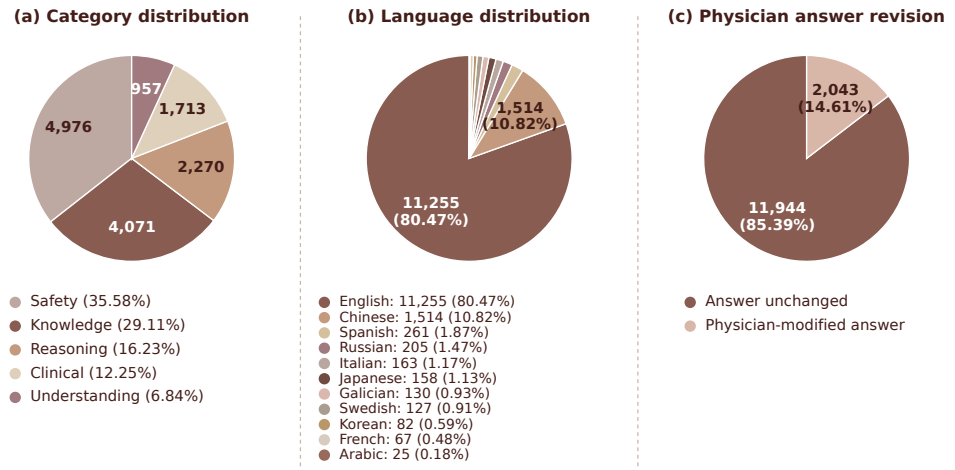

Figure 4: **Composition of the HealthCoreBench.** Distribution of the final 13,987 samples across (a) capability categories, (b) languages, and (c) physician answer revision status.

- **Ground-Truth Verification.** Whether no option is correct (*No Correct Answer*), multiple options are medically defensible (*Multiple Correct Answers*), or the provided label is incorrect (*Wrong Ground Truth*).

A sample is marked *YES* only when none of these issues is identified. Multiple frontier auditors independently apply the protocol to every sample, providing a binary judgment and an evidence-grounded rationale for each error type. These annotations serve as screening signals rather than final decisions: physicians receive the complete original sample together with all structured annotations and evidence, and use them to prioritize review. The complete audit prompt used for automated pre-screening is provided in Box D.

Physician Review. Each sample is independently reviewed by at least **three physicians**, matched to the relevant specialty whenever possible. Reviewers prioritize samples flagged by the automated auditors while also examining all remaining samples for potential false negatives. They verify whether the question is clear and answerable, the options are well formed, and the reference answer is medically correct, current, and uniquely supported. Each sample is classified as **retain**, **repair**, or **exclude**, with a rationale recorded for the decision.

Adjudication and Re-evaluation. Samples with uncertain or conflicting judgments are adjudicated by **two senior physicians**. Valid samples are retained, correctable defects are repaired, and irreparable samples are excluded. Because a correction may alter the empirical difficulty of a sample, every repaired sample is re-evaluated through the full progressive filtering pipeline before inclusion. We report physician agreement and the proportions of retained, repaired, and excluded samples in Table 6.

2.3 Benchmark Statistics

Overall Composition. HealthCoreBench contains **13,987 samples** after progressive difficulty filtering and physician validation. Figure 4 summarizes the resulting benchmark across capability categories, languages, and physician answer revision status. Across capability categories, Safety constitutes the largest subset with 4,976 samples (35.58%), followed by Knowledge with 4,071 (29.11%), Reasoning with 2,270 (16.23%), Clinical with 1,713 (12.25%), and Understanding with 957 (6.84%). The resulting benchmark therefore spans all five capability dimensions, with Safety and Knowledge together accounting for the majority of the final samples.

Source Composition. The final samples remain distributed across the 47 source benchmarks, with substantial variation in their contributions after progressive filtering and validation. Med-HALT contributes the largest subset with 4,019 samples, followed by MedXpertQA (Text) with 1,133,

Table 2: **Model performance on the HealthCoreBench.** Scores are reported for frontier and medically specialized language models across medical capability dimensions (Knowledge, Reasoning, Clinical, Understanding, and Safety) and overall.

Model	Knowledge	Reasoning	Clinical	Understanding	Safety	Overall
Frontier Proprietary Models						
GPT-6-Astra	**50.97**	**38.90**	**35.73**	**46.50**	50.70	**46.74**
GPT-5.6 Sol	46.18	27.93	35.38	30.83	41.28	39.10
Gemini-3.8-Flash	45.10	35.24	29.95	46.08	**54.32**	44.99
Kimi-K3	42.08	22.95	22.42	16.41	40.35	34.20
DeepSeek-V4-Pro-0813	39.89	21.67	22.88	18.70	43.75	34.78
DeepSeek-V4-Flash-0731	31.79	17.31	21.13	11.70	36.62	28.48
Qwen3.8-27B	30.34	15.77	19.09	13.69	32.17	26.11
Open-Source Medical-Specialized Models						
HuatuoGPT-3-32B	25.92	9.25	17.98	14.52	29.22	22.64
Baichuan-M3-235B	30.14	14.27	20.08	15.57	34.85	27.01
MedGemma-27B-Text-it	22.62	9.56	22.18	10.55	22.57	19.60
Lingshu-32B	23.93	8.06	15.18	11.91	21.36	18.55
Hulu-Med-32B	23.41	8.77	18.62	11.60	19.59	18.28
MedGemma-27B-it	22.11	6.78	26.56	11.70	18.87	18.30

HEAD-QA v2 with 786, SuperGPQA (Medicine & Biology) with 775, EHRBench with 729, and CMB with 625. Other source benchmarks contribute smaller but complementary subsets. Complete per-benchmark retention statistics are reported in Table 7, while Figure 6 visualizes the proportions of removed, retained, and physician-modified samples for each source benchmark.

Language Distribution. HealthCoreBench covers **11 languages**. English accounts for 11,255 samples (80.47%), followed by Chinese with 1,514 (10.82%). Although the final benchmark remains predominantly English, the retained multilingual subsets enable evaluation across a broader range of linguistic settings.

Physician Answer Revision. Among the final 13,987 samples, 11,944 (85.39%) retain their original reference answers, while 2,043 (14.61%) are included after physician modification.

3 Experiments

3.1 Experimental Setup

Evaluation Models. We evaluate a diverse set of frontier and medically specialized models across language and multimodal settings: *Frontier models* include GPT-6-Astra, GPT-5.6 Sol, Gemini-3.8-Flash, Kimi-K3, DeepSeek-V4-Pro-0813, DeepSeek-V4-Flash-0731, and Qwen3.8-27B. *Medical language models* include HuatuoGPT-3-32B (Chen et al., 2026), Baichuan-M3-235B (Dou et al., 2026), MedGemma-27B-Text-it (Sellergren et al., 2025), Lingshu-32B (Xu et al., 2025), Hulu-Med-32B (Jiang et al., 2025a), and MedGemma-27B-it (Sellergren et al., 2025).

None of the evaluation models participate in benchmark filtering, quality screening, or sample selection, ensuring that reported performance is not biased by the construction process.

Evaluation Protocol. All models are evaluated in a *zero-shot* setting without demonstrations, tools, or external retrieval. Each model receives identical task content through its native chat template. We set the temperature to 0 and enable the highest reasoning-effort setting available for each model.

3.2 Overall Benchmark Results

Overall Results. Table 2 reports performance on the 13,987 samples retained through difficulty filtering and physician validation. Since the evaluated models did not participate in benchmark

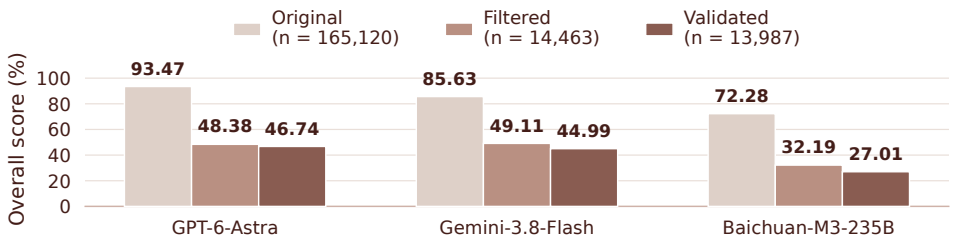

Figure 5: **Overall scores across benchmark construction stages.** Model scores decrease markedly after filtering, reducing score saturation caused by overly simple questions.

construction, these results examine whether the validated challenges extend beyond the construction models. We highlight three findings:

- **Difficulty persists after validation and extends to models outside the construction panels.** GPT-6-Astra achieves the highest overall score of *46.74*, followed by Gemini-3.8-Flash at *44.99*. Substantial difficulty therefore remains after physician review, correction, and re-evaluation. This supports the motivation for failure attribution: identifying challenging questions whose validity has been examined before interpreting model errors as evidence of capability limitations.
- **Parameter count alone does not explain performance on validated challenges.** Baichuan-M3-235B scores *27.01* overall, only *0.90* points above Qwen3.8-27B (*26.11*), despite its substantially larger nominal parameter count. Qwen3.8-27B also scores higher on Knowledge (*30.34* versus *30.14*) and Reasoning (*15.77* versus *14.27*). Larger parameter counts therefore do not consistently correspond to stronger performance across the evaluated capabilities.
- **Some specialized models remain competitive on validated clinical tasks.** MedGemma-27B-it achieves a Clinical score of *26.56*, exceeding Kimi-K3 (*22.42*) and DeepSeek-V4-Pro-0813 (*22.88*), despite a substantially lower overall score (*18.30* versus *34.20* and *34.78*). Its clinical performance also contrasts with a Reasoning score of *6.78*. These results reveal a localized strength on the validated clinical subset alongside substantial weaknesses in other capabilities, showing that lower overall performance need not imply uniformly weaker medical capability.

3.3 In-Depth Analysis

Evaluation Artifacts Create False Successes. We evaluate GPT-6-Astra, Gemini-3.8-Flash, and Baichuan-M3-235B on the original pool, the difficulty-filtered set, and the physician-validated set. As shown in Figure 5, difficulty filtering substantially lowers overall scores. Physician validation further reduces them by 1.64, 4.12, and 5.18 points, respectively. The unequal changes show that evaluation defects can affect models differently and alter their apparent performance.

Improved Evaluation Efficiency with Preserved Rankings. We compare token consumption, overall scores, and rankings on the original pool and HealthCoreBench. Table 3 shows that average token consumption falls from *82.82M* to *18.91M*, a reduction of **77.17%**, while the ranking of the three models remains unchanged. HealthCoreBench thus preserves their observed ordering at substantially lower inference cost.

Difficulty Filtering Can Enrich Evaluation Artifacts. We test whether selecting samples that stronger models fail to solve also concentrates evaluation defects. We group samples by the tier at which they are first reliably solved and separately consider those remaining unsolved after Tier 5. We randomly audit 3,000 samples per resolved group, review all 2,944 samples resolved at Tier 4, and use the physician reviews of all 14,463 unresolved samples. Physicians assess the original questions without access to model responses or filtering outcomes. Samples requiring correction or exclusion are counted as defective. As shown in Table 4, the estimated defect rate rises from 1.36%

Table 3: **Token consumption, overall scores, and rankings on the original pool and HealthCoreBench.**

	Original: 165,120 samples			Final: 13,987 samples		
Model	**Token Consumption (M)**	**Overall**	**Rank**	**Token Consumption (M)**	**Overall**	**Rank**
GPT-6-Astra	83.47	93.47	1	19.82	46.74	1
Gemini-3.8-Flash	91.26	85.63	2	23.41	44.99	2
Baichuan-M3-235B	73.73	59.28	3	13.49	27.01	3

Table 4: **Evaluation defects across difficulty-filtering outcomes.** Groups are defined by the first tier at which samples are reliably solved. The Tier 4 group and the group unsolved after Tier 5 are reviewed in full. The initial-pool rate is estimated by weighting group rates by their pool sizes.

Sample Group	**Pool Size**	**Audited**	**Defect Rate (%)**
Initial pool	165,120	29,407	1.36
Solved at Tier 1	117,953	3,000	0.80
Solved at Tier 2	14,493	3,000	1.40
Solved at Tier 3	8,374	3,000	2.00
Solved at Tier 4	2,944	2,944	2.79
Solved at Tier 5	6,893	3,000	3.40
Unsolved after Tier 5	14,463	14,463	4.20

in the initial pool to 4.20% among samples surviving Tier 5, a 3.09-fold enrichment. This pattern motivates a separate validity check before retained failures are interpreted as capability limitations.

4 Related Work

Medical Benchmarks for LLMs. Medical LLMs are evaluated through examination-style QA (Jin et al., 2021; Pal et al., 2022; Jin et al., 2019), multilingual extensions (Wang et al., 2024; Liu et al., 2023; Labrak et al., 2022), and increasingly on capabilities beyond knowledge recall, including medical calculation (Khandekar et al., 2024), diagnostic reasoning (Chen et al., 2024; Zuo et al., 2025), long-context understanding (Adams et al., 2025), factual reliability (Pal et al., 2023; Pandit et al., 2025), open-ended interactions (OpenAI, 2025; Hicks et al., 2026), and tool-using settings (Schmidgall et al., 2024). Collectively they offer broad coverage but differ substantially in format, difficulty, language, and evaluation criteria, making consistent comparison difficult.

Benchmark Reliability and Saturation. Valid comparison requires evaluation items that remain both correct and discriminative. Many medical benchmarks are increasingly saturated (Warner et al., 2026), yet low accuracy need not imply meaningful difficulty, since questions may contain incorrect references, multiple defensible answers, insufficient evidence, or outdated guidelines (Gema et al., 2024), and multimodal items add failure modes such as degraded images and weak visual dependence (Asadi et al., 2026). Difficult questions must therefore be distinguished from broken or underspecified ones before supporting credible conclusions about model capability. **Appendix A provides a fuller discussion.**

5 Conclusion

We study medical benchmark construction as a failure attribution problem: a question is useful for measuring capability only when its evaluation target is valid. HealthCoreBench combines progressive difficulty filtering with physician validation, retaining 13,987 samples from an initial pool of 165,120 across 47 benchmarks. Physician review changes measured model scores, showing that difficult samples can also contain evaluation defects. The resulting benchmark remains challenging for frontier models: the highest overall score is 46.74. HealthCoreBench preserves model rankings while reducing average token consumption by 77.17%. These findings underscore the need to validate the causes of model failures before interpreting them as medical capability gaps.

AI USE STATEMENT

Generative AI systems were used in two roles in this work. First, as part of benchmark construction, language models were used for progressive difficulty filtering and automated quality screening. Their judgments served as screening signals; physicians reviewed the candidate samples and made the final validity decisions. Second, generative AI tools assisted with manuscript organization, translation, and language editing. The authors reviewed and validated all AI-assisted outputs, verified the reported results, and take responsibility for the final text, claims, data, code, figures, and benchmark artifacts.

REFERENCES

Lisa Adams, Felix Busch, Tianyu Han, Jean-Baptiste Excoffier, Matthieu Ortala, Alexander Löser, Hugo J. W. L. Aerts, Jakob Nikolas Kather, Daniel Truhn, and Keno Bressem. LongHealth: A question answering benchmark with long clinical documents. *Journal of Healthcare Informatics Research*, 2025. doi: 10.1007/s41666-025-00179-2. URL https://arxiv.org/abs/2401.14490.

Mohammad Asadi, Jack W O'Sullivan, Fang Cao, Tahoura Nedaee, Kamyar Rajabalifardi, Fei-Fei Li, Ehsan Adeli, and Euan Ashley. Mirage: The illusion of visual understanding. *arXiv preprint arXiv:2603.21687*, 2026.

BIG-Bench Collaboration, Aarohi Srivastava, et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*, 2023. URL https://openreview.net/forum?id=uyTL5Bvosj.

Min Cao et al. OpenCompass: A universal evaluation platform for large language models. *arXiv preprint arXiv:2605.19276*, 2026. doi: 10.48550/arXiv.2605.19276. URL https://arxiv.org/abs/2605.19276.

Junying Chen, Xinyuan Xie, Ziniu Li, and Benyou Wang. Onepo: Direct one-stage policy optimization for sft-free domain adaptation. In *Forty-third International Conference on Machine Learning*, 2026.

Xi Chen et al. RareBench: Can large language models serve as rare diseases specialists? *arXiv preprint arXiv:2402.06341*, 2024. doi: 10.48550/arXiv.2402.06341. URL https://arxiv.org/abs/2402.06341.

Chengfeng Dou, Fan Yang, Fei Li, Jiyuan Jia, Qiang Ju, Shuai Wang, Tianpeng Li, Xiangrong Zeng, Yijie Zhou, Hongda Zhang, et al. Baichuan-m3: Modeling clinical inquiry for reliable medical decision-making. *arXiv preprint arXiv:2602.06570*, 2026.

Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Jiaqi Wang, et al. Vlmevalkit: An open-source toolkit for evaluating large multi-modality models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 11198–11201, 2024.

FHIR-AgentBench Authors. FHIR-AgentBench: Benchmarking large language model agents for realistic interoperable electronic health record question answering. *arXiv preprint arXiv:2509.19319*, 2025. doi: 10.48550/arXiv.2509.19319. URL https://arxiv.org/abs/2509.19319.

Leo Gao, Jonathan Tow, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Kyle McDonell, Niklas Muennighoff, et al. A framework for few-shot language model evaluation. Zenodo, 2021. URL https://doi.org/10.5281/zenodo.5371628.

Aryo Pradipta Gema, Joshua Ong Jun Leang, Giwon Hong, Alessio Devoto, Alberto Carlo Maria Mancino, Rohit Saxena, Xuanli He, Yu Zhao, Xiaotang Du, Mohammad Reza Ghasemi Madani, Claire Barale, Robert McHardy, Joshua Harris, Jean Kaddour, Emile van Krieken, and Pasquale Minervini. Are we done with mmlu? *arXiv preprint arXiv:2406.04127*, 2024. URL https://arxiv.org/abs/2406.04127.

Harbor Framework. Introducing Harbor-Index: A compact, diverse, challenging, and high-quality benchmark for agentic evaluation. Technical report and benchmark release, July 2026. URL https://harbor-index.org/. Harbor-Index 1.0.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=d7KBjmI3GmQ.

Rebecca S. Hicks et al. HealthBench Professional: Evaluating large language models on real clinician tasks. *arXiv preprint arXiv:2604.27470*, 2026. URL https://arxiv.org/abs/2604.27470.

Yutao Hu, Tao Li, Qu Lu, et al. OmniMedVQA: A new large-scale comprehensive evaluation benchmark for medical large vision-language models. *arXiv preprint arXiv:2402.09181*, 2024. doi: 10.48550/arXiv.2402.09181. URL https://arxiv.org/abs/2402.09181.

Songtao Jiang, Yuan Wang, Sibo Song, Tianxiang Hu, Chenyi Zhou, Bin Pu, Yan Zhang, Zhibo Yang, Yang Feng, Joey Tianyi Zhou, et al. Hulu-med: A transparent generalist model towards holistic medical vision-language understanding. *arXiv preprint arXiv:2510.08668*, 2025a.

Yixing Jiang et al. MedAgentBench: A realistic virtual electronic health record environment to benchmark medical large language model agents. *NEJM AI*, 2025b. doi: 10.1056/AIdbp2500144. URL https://ai.nejm.org/doi/full/10.1056/AIdbp2500144.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421, 2021. doi: 10.3390/app11146421.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W. Cohen, and Xinghua Lu. PubMedQA: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pp. 2567–2577. Association for Computational Linguistics, 2019. doi: 10.18653/v1/D19-1259. URL https://aclanthology.org/D19-1259/.

Nikhil Khandekar, Qiao Jin, Seong-Eun Kim, Jiayue Zhang, Qingyu Chen, and Zhiyong Lu. Medcalc-bench: Evaluating large language models for medical calculations. In *Advances in Neural Information Processing Systems*, 2024. URL https://github.com/ncbi-nlp/MedCalc-Bench. Datasets and Benchmarks Track.

Yanis Labrak, Adrien Bazoge, Richard Dufour, Mickael Rouvier, Emmanuel Morin, Béatrice Daille, and Pierre-Antoine Gourraud. FrenchMedMCQA: A french multiple-choice question answering dataset for the medical domain. In *Proceedings of the 13th International Workshop on Health Text Mining and Information Analysis*, pp. 41–46. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.louhi-1.5. URL https://aclanthology.org/2022.louhi-1.5/.

John P. Lalor, Hao Wu, and Hong Yu. Building an evaluation scale using item response theory. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 648–657. Association for Computational Linguistics, 2016. doi: 10.18653/v1/D16-1062. URL https://aclanthology.org/D16-1062/.

Jason J. Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. A dataset of clinically generated visual questions and answers about radiology images. *Scientific Data*, 5:180251, 2018. doi: 10.1038/sdata.2018.251. URL https://www.nature.com/articles/sdata2018251.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. Holistic evaluation of language models. *Transactions on Machine Learning Research*, 2023. URL https://openreview.net/forum?id=iO4LZibEqW.

Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. SLAKE: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In *2021 IEEE 18th International Symposium on Biomedical Imaging*, pp. 1650–1654. IEEE, 2021. doi: 10.1109/ISBI48211.2021.9434010.

Junling Liu, Peilin Zhou, Yantao Hua, Dading Chong, Zhiyang Tian, Andrew Liu, Huayu Wang, Chenyu You, Zibin Guo, Lei Zhu, et al. Benchmarking large language models on CMExam: A comprehensive chinese medical exam dataset. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/a48ad12d588c597f4725a8b84af647b5-Abstract-Datasets_and_Benchmarks.html.

Collins Nimo, Tobi Olatunji, et al. AfriMed-QA: A pan-african, multi-specialty, medical question-answering benchmark dataset. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2025. URL https://aclanthology.org/2025.acl-long.96/.

OpenAI. HealthBench: Evaluating large language models towards improved human health. *Technical Report*, 2025. URL https://openai.com/index/healthbench/.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. MedMCQA: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Proceedings of the Conference on Health, Inference, and Learning*, volume 174 of *Proceedings of Machine Learning Research*, pp. 248–260. PMLR, 2022. URL https://proceedings.mlr.press/v174/pal22a.html.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. Med-HALT: Medical domain hallucination test for large language models. In *Proceedings of the 27th Conference on Computational Natural Language Learning*, pp. 314–334. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.conll-1.21. URL https://aclanthology.org/2023.conll-1.21/.

Shrey Pandit, Jiawei Xu, Junyuan Hong, Zhangyang Wang, Tianlong Chen, Kaidi Xu, and Ying Ding. MedHallu: A comprehensive benchmark for detecting medical hallucinations in large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2025. URL https://aclanthology.org/2025.emnlp-main.143/.

Long Phan, Alice Gatti, Ziwen Han, Nathaniel Li, Josephina Hu, Hugh Zhang, Sean Shi, Michael Choi, Anish An, et al. Humanity's last exam. *arXiv preprint arXiv:2501.14249*, 2025. doi: 10.48550/arXiv.2501.14249. URL https://arxiv.org/abs/2501.14249.

Felipe Maia Polo, Lucas Weber, Leshem Choshen, Yuekai Sun, Gongjun Xu, and Mikhail Yurochkin. tinyBenchmarks: Evaluating large language models with fewer examples. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 40706–40727. PMLR, 2024. URL https://proceedings.mlr.press/v235/maia-polo24a.html.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. GPQA: A graduate-level google-proof question answering benchmark. *arXiv preprint arXiv:2311.12022*, 2024. doi: 10.48550/arXiv.2311.12022. URL https://arxiv.org/abs/2311.12022.

Samuel Schmidgall, Rojin Ziaei, Carl Harris, Eduardo Reis, Jeffrey Jopling, and Michael Moor. AgentClinic: A multimodal agent benchmark to evaluate ai in simulated clinical environments. *arXiv preprint arXiv:2405.07960*, 2024. doi: 10.48550/arXiv.2405.07960. URL https://arxiv.org/abs/2405.07960.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, et al. Medgemma technical report. *arXiv preprint arXiv:2507.05201*, 2025.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, 2023. doi: 10.1038/s41586-023-06291-2. URL https://www.nature.com/articles/s41586-023-06291-2.

Yiyou Sun, Xinyang Han, Weichen Zhang, Yuanbo Pang, Tianyu Wang, Yuhan Cao, Yixiao Huang, Chris Duroiu, Haoyun Zhang, Jeffrey Lin, et al. Agents' last exam. *arXiv preprint arXiv:2606.05405*, 2026.

David Vilares and Carlos Gómez-Rodríguez. HEAD-QA: A healthcare dataset for complex reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 960–966. Association for Computational Linguistics, 2019. doi: 10.18653/v1/P19-1092. URL https://aclanthology.org/P19-1092/.

Xidong Wang, Guiming Hardy Chen, Dingjie Song, Zhiyi Zhang, Zhihong Chen, Qingying Xiao, Feng Jiang, Jianquan Li, Xiang Wan, Benyou Wang, et al. CMB: A comprehensive medical benchmark in chinese. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 6184–6205. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.naacl-long.343. URL https://aclanthology.org/2024.naacl-long.343/.

Benjamin Warner, Ratna Sagari Grandhi, Max Kieffer, Aymane Ouraq, Saurav Panigrahi, Geetu Ambwani, Kunal Bagga, Nikhil Khandekar, Arya Hariharan, Nishant Mishra, Manish Ram, Shamus Sim Zi Yang, Ahmed Essouaied, Adepoju Jeremiah Moyondafoluwa, Robert Scholz, Bofeng Huang, Molly Beavers, Srishti Gureja, Anish Mahishi, Sameed Khan, Maxime Griot, Hunar Batra, Jean-Benoit Delbrouck, Siddhant Bharadwaj, Ronald Clark, Ashish Vashist, Anas Zafar, Leema Krishna Murali, Harsh Deshpande, Ameen Patel, William Brown, Johannes Hagemann, Connor Lane, Paul Steven Scotti, and Tanishq Mathew Abraham. Medmarks: A comprehensive open-source llm benchmark suite for medical tasks, 2026. URL https://arxiv.org/abs/2605.01417.

Weiyun Xu et al. Lingshu: A generalist foundation model for unified multimodal medical understanding and reasoning. *arXiv preprint arXiv:2506.07044*, 2025. doi: 10.48550/arXiv.2506.07044. URL https://arxiv.org/abs/2506.07044.

Yuchen Zeng and Dimitris Papailiopoulos. You don't need to run every eval. *arXiv preprint arXiv:2606.24020*, 2026.

Kaiwen Zuo and Yirui Jiang. MedHallBench: A new benchmark for assessing hallucination in medical large language models. *arXiv preprint arXiv:2412.18947*, 2024. doi: 10.48550/arXiv.2412.18947. URL https://arxiv.org/abs/2412.18947.

Yuxin Zuo, Shang Qu, Yifei Li, Zhangren Chen, Xuekai Zhu, Ermo Hua, Kaiyan Zhang, Ning Ding, and Bowen Zhou. MedXpertQA: Benchmarking expert-level medical reasoning and understanding. In *Proceedings of the 42nd International Conference on Machine Learning*, 2025. URL https://arxiv.org/abs/2501.18362.

A Related Work

Medical Benchmarks for LLMs. Language models have traditionally been evaluated using examination-style question answering. MedQA-USMLE (Jin et al., 2021), MedMCQA (Pal et al., 2022), PubMedQA (Jin et al., 2019), and the medical subsets of MMLU (Hendrycks et al., 2021) assess factual knowledge and short-form reasoning, while CMB (Wang et al., 2024), CMExam (Liu et al., 2023), FrenchMedMCQA (Labrak et al., 2022), HEAD-QA (Vilares & Gómez-Rodríguez, 2019), and AfriMedQA (Nimo et al., 2025) extend this paradigm across languages and healthcare systems. More recent benchmarks target capabilities beyond knowledge recall. MedCalc-Bench (Khandekar et al., 2024) evaluates medical calculations, RareBench (Chen et al., 2024) and MedXpertQA (Zuo et al., 2025) emphasize complex diagnostic reasoning, LongHealth (Adams et al., 2025) examines long-context understanding, and Med-HALT (Pal et al., 2023), MedHallu (Pandit et al., 2025), and MedHallBench (Zuo & Jiang, 2024) assess hallucination and factual reliability.

HealthBench (OpenAI, 2025) and HealthBench Professional (Hicks et al., 2026) evaluate open-ended healthcare interactions, while AgentClinic (Schmidgall et al., 2024), MedAgentBench (Jiang et al., 2025b), and FHIR-AgentBench (FHIR-AgentBench Authors, 2025) extend evaluation to interactive and tool-using settings. Collectively, these benchmarks provide broad coverage of medical language capabilities, but differ substantially in task format, difficulty, language, prompting protocol, and evaluation criteria, making comprehensive and consistent model comparison difficult.

Benchmark Reliability and Saturation. Benchmark performance is informative only when evaluation items remain valid and discriminative. Many established medical benchmarks are increasingly saturated by recent models, limiting their ability to distinguish frontier systems (Warner et al., 2026). At the other extreme, low accuracy does not necessarily imply meaningful difficulty. Medical questions may contain incorrect references, multiple defensible answers, insufficient evidence, outdated guidelines, or evaluation rules that reject semantically correct responses (Gema et al., 2024). Multimodal items introduce additional failure modes, including degraded images, missing views, annotation leakage, and weak dependence on visual evidence (Asadi et al., 2026). Consequently, selecting questions solely because strong models fail on them can preferentially retain unreliable items. Difficult questions must therefore be distinguished from broken or underspecified questions before they can support credible conclusions about model capability.

Benchmark Curation and Efficient Evaluation. General evaluation suites such as BIG-bench (BIG-Bench Collaboration et al., 2023), HELM (Liang et al., 2023), OpenCompass (Cao et al., 2026), VLMEvalKit (Duan et al., 2024), and the LM Evaluation Harness (Gao et al., 2021) standardize evaluation across heterogeneous tasks, while MultiMedQA (Singhal et al., 2023) and MedEvalKit (Xu et al., 2025) consolidate multiple medical datasets. These frameworks improve coverage and reproducibility but largely retain the scale, redundancy, and varying quality of their constituent benchmarks. Complementary efforts such as GPQA Diamond (Rein et al., 2024), Humanity's Last Exam (Phan et al., 2025), HealthBench Hard (OpenAI, 2025), Agents' Last Exam (Sun et al., 2026), and Harbor-Index (Harbor Framework, 2026) construct challenging subsets for evaluating frontier models. In particular, Harbor-Index (Harbor Framework, 2026) demonstrates that model-based difficulty filtering, automated auditing, and human review can produce a compact evaluation set from a substantially larger task pool. Related approaches based on item response theory and informative subset selection (Lalor et al., 2016; Polo et al., 2024), as well as benchmark-level score prediction methods (Zeng & Papailiopoulos, 2026), further seek to preserve model comparisons while reducing evaluation cost.

B Source Benchmarks

We collect a broad set of medical evaluation benchmarks spanning knowledge assessment, complex reasoning, multimodal understanding, and safety. From these source benchmarks, we curate a compact, diverse, challenging, and high-quality benchmark through a multi-stage selection pipeline.

B.1 Medical Benchmarks for LLM Evaluation

Table 5: Overview of selected medical benchmarks for LLM evaluation. The **Samples** column reports the number of samples in medical-related tasks. ⚠ Some benchmarks have inaccessible source data, causing sample-count discrepancies with the original papers.

#	Benchmark	Publication	Download Link	Year	Samples
1	MMLU (Medical)	ICLR 2021	Hugging Face	2021	1,089
2	PubMedQA	EMNLP-IJCNLP 2019	GitHub	2019	500
3	MedMCQA	CHIL 2022 (PMLR 174)	Hugging Face	2022	4,183
4	Medbullets	NAACL 2025	GitHub	2025	616
5	MedXpertQA (Text)	ICML 2025 (PMLR 267)	Hugging Face	2025	2,450
6	SuperGPQA (Medicine & Biology)	NeurIPS 2025	Hugging Face	2025	3,875
7	AfriMedQA v2	ACL 2025	Hugging Face	2025	3,957
8	CMB	NAACL 2024	Hugging Face	2024	11,408

(Continued on next page)

(Continued from previous page)

#	Benchmark	Publication	Download Link	Year	Samples
9	CareQA	NAACL 2025	Hugging Face	2025	5,621 (open 2,769)
10	HLE (Medicine, Text)	Nature	Hugging Face	2026	222
11	MedNLI	EMNLP 2018	PhysioNet	2018	1,422
12	MediQ	NeurIPS 2024	GitHub	2024	1,412
13	ReDis-QA	arXiv preprint	Hugging Face	2024	1,360
14	MMedBench	Nature Communications	Hugging Face	2024	8,518
15	GPQA (Biology)	COLM 2024	Hugging Face	2024	105
16	FrenchMedMCQA	LOUHI 2022	GitHub	2022	622
17	HEAD-QA v2	LREC 2026	Hugging Face	2026	2,675
18	CMMLU (Medicine)	Findings of ACL 2024	Hugging Face	2024	2,110
19	LongHealth	JHIR	GitHub	2025	400
20	MedHallu	EMNLP 2025	Hugging Face	2025	10,000
21	Meta-MedQA	Nature Communications	Hugging Face	2025	1,373
22	MedExQA	BioNLP 2024	Hugging Face	2024	940
23	MEDEC	Findings of ACL 2025	GitHub	2025	597 (correction 311)
24	KorMedMCQA	arXiv preprint	Hugging Face	2024	3,009
25	IgakuQA	arXiv preprint	GitHub	2023	1,987
26	JMedBench	COLING 2025	Hugging Face	2025	4,183
27	MedExpQA	AIM	Hugging Face	2024	500
28	BioASQ	BioASQ Challenge	GitHub	2013	5,483
29	ClinicBench	EMNLP 2024	GitHub	2024	5,212
30	MIMIC-CDM	Nature Medicine	GitHub	2024	2,400
31	EHRBench	arXiv preprint	Hugging Face	2025	21,221
32	MedArabiQ	MLHC 2025 (PMLR 298)	GitHub	2025	700
33	Swedish Medical LLM Benchmark	Frontiers in AI	GitHub	2025	3,178
34	DiagnosisArena	Findings of ACL 2026	Hugging Face	2026	915
35	CMExam	NeurIPS 2023	GitHub	2023	6,810
36	MedCalc-Bench	NeurIPS 2024	GitHub	2024	1,100
37	Med-HALT	CoNLL 2023	Hugging Face	2023	11,076
38	MMLU-Pro (Health & Biology)	NeurIPS 2024	Hugging Face	2024	1,535
39	PubHealthBench	arXiv preprint	Hugging Face	2025	7,929
40	mARC	Scientific Reports	GitHub	2025	100
41	JEMD	arXiv preprint	Hugging Face	2025	1,000
42	ClinicalBench	NeurIPS 2024 Workshop	GitHub	2024	1,498
43	IOR-Bench	arXiv preprint	Hugging Face	2025	1,476
44	PrinciplismQA	Findings of ACL 2026	Github	2026	4,325
45	MedQA-USMLE	Applied Sciences	Hugging Face	2021	1,273
46	MedQA-MCMLE	Applied Sciences	Hugging Face	2021	3,426
47	GlobalDentBench (OA)	arXiv preprint	Hugging Face	2026	4,055

C Sample Filtering

C.1 Language Benchmarks

Table 6 reports the number of language samples entering, removed, and retained at each progressive filtering stage, and Table 7 provides the corresponding per-benchmark retention counts. Figure 6 visualizes the filtering outcomes across all 47 source benchmarks.

Table 7: **Language benchmark sample retention across filtering stages.** Per-benchmark language sample counts retained after each filtering stage.

Benchmark	Initial	Tier 1	Tier 2	Tier 3	Tier 4	Tier 5	Post-QA	Final
MMLU (Medical)	1089	127	85	60	50	32	32	32
PubMedQA	500	193	148	116	108	80	80	80
MedMCQA	4183	1398	1005	731	623	396	385	385
Medbullets	298	129	98	67	57	26	15	15

(Continued on next page)

(Continued from previous page)

Benchmark	Initial	Tier 1	Tier 2	Tier 3	Tier 4	Tier 5	Post-QA	Final
MedXpertQA (Text)	2450	2070	1920	1684	1597	1134	1133	1133
SuperGPQA (Medicine & Biology)	3875	2545	2016	1557	1327	900	775	775
AfriMedQA v2	3904	1099	844	676	600	476	474	474
CMB	11200	4012	1680	1323	865	665	625	625
CareQA	5621	894	628	374	323	172	162	162
HLE (Medicine, Text)	147	139	129	125	122	100	88	88
MedNLI	1422	225	175	142	128	97	96	96
MediQ	1412	268	178	109	94	32	29	29
ReDis-QA	1360	284	177	135	111	74	74	74
MMedBench	8518	1721	903	603	495	323	308	308
GPQA (Biology)	105	42	30	27	24	14	13	13
FrenchMedMCQA	622	221	174	108	90	33	33	33
HEAD-QA v2	13375	2391	1693	1278	1148	798	786	786
CMMLU (Medicine)	2110	581	256	194	144	109	105	105
LongHealth	400	58	46	34	29	25	25	25
MedHallu	20000	2399	1686	1255	1079	409	409	409
Meta-MedQA	1373	516	420	318	292	229	210	210
MedExQA	940	187	151	124	117	105	104	104
MEDEC	597	218	171	127	109	67	60	60
KorMedMCQA	3009	858	546	351	278	83	82	82
IgakuQA	1987	535	337	197	153	61	37	37
JMedBench	1536	369	268	199	176	124	118	118
MedExpQA	500	121	88	50	44	35	35	35
BioASQ	1478	221	147	119	103	59	58	58
ClinicBench	413	169	138	115	109	28	28	28
MIMIC-CDM	2400	445	353	307	284	199	182	182
EHRBench	7721	2187	1633	1152	1025	729	729	729
MedArabiQ	300	141	93	57	49	25	25	25
Swedish Medical LLM Benchmark	5396	824	528	319	254	127	127	127
DiagnosisArena	915	488	392	282	263	160	155	155
CMExam	6810	2221	770	586	351	270	265	265
MedCalc-Bench	1100	572	446	263	239	118	118	118
Med-HALT	22152	10330	8065	5875	5584	4152	4019	4019
MMLU-Pro (Health & Biology)	1535	339	271	199	176	116	114	114
PubHealthBench	7929	1065	770	480	406	208	208	208
mARC	100	77	57	37	32	14	14	14
JEMD	1000	440	310	260	233	173	170	170
ClinicalBench	1498	947	906	828	822	548	548	548
IOR-Bench	1476	344	236	208	187	159	159	159
PrinciplismQA	2182	660	501	395	346	278	278	278
MedQA-USMLE	1273	289	204	97	82	47	45	45
MedQA-MCMLE	3426	696	205	164	116	96	95	95
GlobalDentBench (OA)	3483	1112	797	593	512	358	357	357
Total Language	**165120**	**47167**	**32674**	**24300**	**21356**	**14463**	**13987**	**13987**

Note: Post-QA includes both automated screening and physician review stages.

Table 6: **Language sample filtering statistics across progressive stages.** Each stage shows the number of language samples entering, removed, and retained for the next stage.

Stage	Input Samples	Removed	Retained	Removal Rate	Cumulative Retention
Initial Pool	165120	-	165120	-	100%
Tier 1 Filtering	165120	117953	47167	71.43%	28.57%
Tier 2 Filtering	47167	14493	32674	30.73%	19.79%
Tier 3 Filtering	32674	8374	24300	25.63%	14.72%
Tier 4 Filtering	24300	2944	21356	12.12%	12.93%
Tier 5 Filtering	21356	6893	14463	32.28%	8.76%
Automated Screening	14463	9697	4766	67.05%	2.89%
Physician Review	14463	476	13987	3.29%	8.47%
Final Benchmark	-	-	**13987**	-	**8.47%**

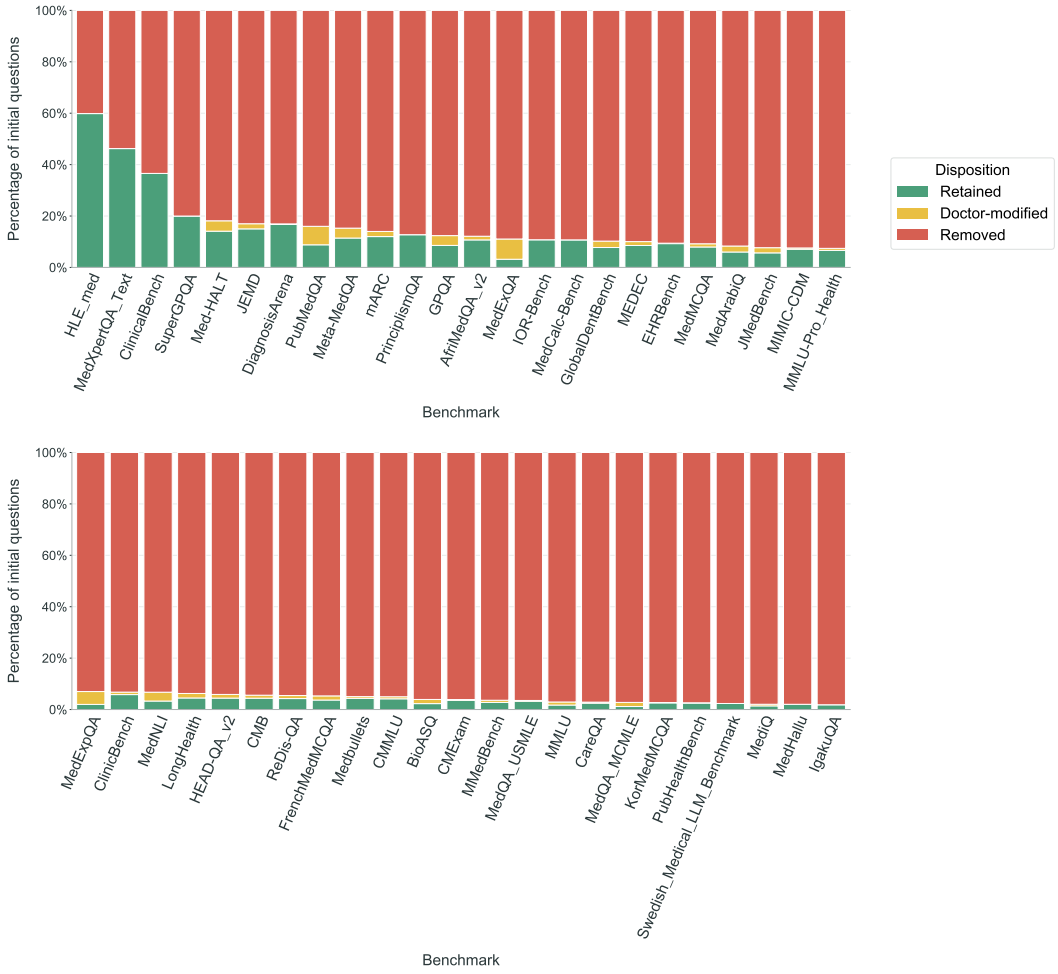

Figure 6: **Language benchmark filtering outcomes.** Percentages of initial questions removed, retained, or retained after physician answer correction across 47 benchmarks.

D Automated Pre-Screening Prompt

The following prompt is used for automated pre-screening of samples that remain after progressive difficulty filtering. Each sample is independently assessed across seven categories covering medical relevance, question and option clarity, answer validity, and ground-truth correctness. The auditor returns a structured judgment for every category together with a concise rationale. These outputs are

used only as screening signals to prioritize physician review and do not constitute the final validity decision.

Automated Pre-Screening Audit Prompt

```
You are performing a meticulous expert second-pass audit of questions
     from
medical and health benchmarks. Review every supplied item
    independently using
established medical, biomedical, public-health and quantitative
    knowledge and
the exact wording/options. Do not trust either the existing dataset
    or its
ground truth by default.

Assess all seven categories for every item:

\begin{enumerate}
  \item \textbf{Bad Question Clarity:} ambiguous, internally
      contradictory,
  truncated, malformed, missing indispensable context/image/table/
      statements,
  or not independently answerable as presented.

  \item \textbf{Bad Options Clarity:} incomplete, duplicated,
      malformed,
  materially overlapping or unclear options. Empty options are normal
       for
  genuinely free-response numeric tasks.

  \item \textbf{No Correct Answer:} none of the listed answers is
      correct; for
  free-response tasks, the requested result cannot be validly derived
      .

  \item \textbf{Multiple Correct Answers:} multiple choices are
      defensibly
  correct where uniqueness is required, or the correct multi-select
      set is not
  determinate. Do not flag this merely because a task intentionally
      permits
  multiple selections.

  \item \textbf{Wrong Groundtruth:} a unique correct answer or
      determinate
  answer set/value exists but differs from \texttt{ground\_truth}.

  \item \textbf{OK:} true only when the question/options are
      sufficiently clear
  for its task, the answer is determinate as required, and
  \texttt{ground\_truth} is correct; otherwise false.

  \item \textbf{Not Medical Question:} true only if the item is
      outside
  medicine, dentistry, biomedicine, healthcare, bioethics, public
      health,
  clinical operations, or medically relevant basic science.
\end{enumerate}

Interpret \texttt{answer\_format} correctly: \texttt{single\_choice}
    requires
```

```
one answer; \texttt{multi\_choice} may require a set; \texttt{yes\_no} and
\texttt{yes\_no\_maybe} are classification; \texttt{numeric} requires
calculation; label tasks may encode observed outcomes or risk classes. For
empirical patient-outcome labels, do not call the ground truth wrong solely
because the outcome cannot be predicted with certainty; only flag clarity when
indispensable information required by the stated task is actually absent or
malformed. Questions in any language must be audited in that language. If a
referenced image/table is indispensable but unavailable in the supplied item,
flag Bad Question Clarity.

Return exactly one review for every input \texttt{sample\_id}, preserving IDs
and order exactly. Every one of the seven reasons must be concise, specific
Chinese text, including categories marked false. Enforce logic: OK is true if
and only if all other six categories are false. If No Correct Answer is true,
normally Wrong Groundtruth should be false because no unique correct answer
exists. Do not skip or alter any \texttt{sample\_id}.
```