Original Paper

Medical Reasoning in Large Language Models: Development and Evaluation of the MedEvalArena Benchmark

Preethi Prem[1], BS; Vikasini Kuppa[2]; Kie Shidara[3], BS; Esmé Wheeler[3], BA; Gene Chang[4]; Feng Liu[5], PhD; Ahmed Alaa[4], PhD; Danilo Bernardo[3], MD

[1]Carle Illinois College of Medicine, University of Illinois Urbana-Champaign, Champaign, IL, United States
[2]College of Natural and Agricultural Sciences, University of California, Riverside, Riverside, CA, United States
[3]Weill Institute of Neurology and Neurosciences, University of California, San Francisco, San Francisco, CA, United States
[4]Department of Neurosciences, University of California, Berkeley, Berkeley, CA, United States
[5]Department of Industrial & Systems Engineering, Rutgers University, Piscataway, NJ, United States

Corresponding Author:

Danilo Bernardo, MD
Weill Institute of Neurology and Neurosciences, University of California, San Francisco
1825 Fourth Street, Suite 5A
San Francisco, CA 94143
United States
Phone: 1 415-353-2437
Email: dbernardoj@gmail.com

Abstract

Background: Large language models (LLMs) perform strongly on medical specialty board multiple-choice questions (MCQs) but underperform in complex medical reasoning. This gap indicates a need to improve LLM medical reasoning and evaluation paradigms.

Objective: This study aimed to develop and evaluate MedEvalArena, a dynamically regenerable benchmark for assessing medical reasoning in LLMs.

Methods: In a symmetric round-robin format, 6 frontier LLMs from the OpenAI, Grok, Gemini, Claude, Kimi, and DeepSeek families generated challenging board-style medical MCQs, adjudicated question validity on an ensemble LLM-as-judge panel, and answered the validated examination. Candidate items were generated until each model contributed 50 validated questions according to Association of American Medical Colleges (AAMC)/American Board of Medical Specialties (ABMS) physician specialty distributions, yielding a 300-question examination. Judges adjudicated candidate questions' logical validity and medical accuracy. We compared question-generation validity rates, assessed leave-one-out judge agreement, and tested for self-judging preference. Questions were audited for overlap with public MedQA (medical question-answering) benchmarks. We tested question-answering (QA) accuracy across LLMs using a binomial-logit generalized linear mixed model likelihood-ratio test. A generator-stratified 90-question subset was completed by 5 board-certified physicians for paired human-LLM analysis. We evaluated inference cost-accuracy tradeoffs. Multiple comparisons were adjusted by the Benjamini-Hochberg (BH) procedure.

Results: Judge-panel validity rates were significantly higher for questions generated by OpenAI (94.8%, Wilson 95% CI 85.9%-98.2%), Gemini (92.3%, 95% CI 83.2%-96.7%), and Claude (83.3%, 95% CI 72.6%-90.4%) models than those generated by Kimi (63.8%, 95% CI 54.3%-72.4%), Grok (56%, 95% CI 45.8%-65.8%), and DeepSeek (46.0%, 95% CI 37.4%-54.7%). Kimi also demonstrated significantly higher validity than DeepSeek (2-sample tests of proportions; BH-adjusted P=.01). The public-source similarity audit found no exact or near-duplicate matches. Leave-one-out judge agreement was moderate. Generator-specific same-family judging coefficients were nonnegative (0 to 11.0 percentage points); however, none remained significant after BH adjustment. There was no significant difference between the 6 frontier LLM QA accuracies, which ranged from 85.7% (257/300) to 91.7% (275/300), on the 300-question examination (generalized linear mixed model likelihood-ratio test P=.07). The study had 80% power to detect LLM accuracy differences of 9.2 percentage points; nonsignificance did not establish equivalence. On the 90-question subset, physician accuracy was 85.6% (95% CI 79.4%-91.7%; stratified normal with finite-population correction), compared with aggregate LLM performance of 94.4% (85/90; Wilson 95% CI 87.6%-97.6%). This difference was not significant (McNemar test, P=.08); however, the study was powered to detect

a human vs LLM accuracy difference of 14% to 15%. Answering-only costs varied by nearly 2 orders of magnitude, with multiple models on the cost-accuracy Pareto frontier.

Conclusions: MedEvalArena enables scalable, regenerable, peer-judged benchmarking of LLM medical reasoning. Question-generation validity differentiated model families more clearly than answering accuracy, suggesting that valid MCQ generation may complement QA benchmarks for assessing LLM medical reasoning.

J Med Internet Res 2026;28:e96885; doi: 10.2196/96885

Keywords: medical reasoning; large language model; LLM; AI; reasoning; benchmark

Introduction

The performance of large language models (LLMs) in medical reasoning tasks has quickly progressed [1], now matching or exceeding human physician performance on standardized medical examinations, such as the United States Medical Licensing Examination (USMLE) step examinations [2-4]. However, as LLMs continue to improve, their performance remains limited on medical benchmarks requiring open-ended reasoning and future planning [5-9], demonstrating the need for more robust LLM medical reasoning and evaluation frameworks.

Conventional medical reasoning evaluation relies on benchmarks that require academic peer review and validation [10]. However, LLM reasoning capabilities are improving rapidly, driving a continuous need for novel and more challenging benchmarks of reasoning capabilities [11]. This introduces a significant lag between the abilities of frontier LLMs and our understanding of their medical reasoning competence. Addressing these limitations motivates dynamic evaluation frameworks that are able to adapt alongside LLMs, specifically evaluation pipelines that enable rapid benchmark reinstantiation to facilitate tighter feedback-loops for LLM improvement.

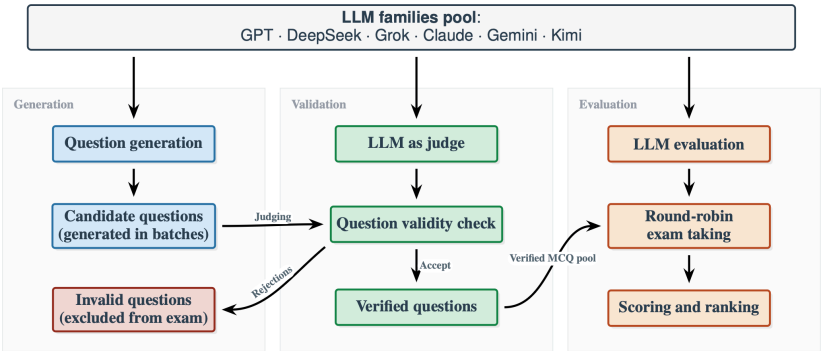
To mitigate benchmark saturation and dataset contamination while keeping pace with rapid model development, recent work has introduced LLM benchmarks that can be updated with human guidance. Approaches such as DynaBench [12] and LiveBench [13] are capable of intermittently updating assessments, with human oversight of task curation. BenchBuilder operationalizes updating at scale via automatic generation of a sizable, diverse benchmark of tasks at low cost [14]. More recently, Autobench incorporates the LLM-as-judge paradigm into dynamic benchmark generation, using a round-robin framework in which models generate questions, complete questions, and evaluate outputs via peer assessment [15].

Motivated by the potential of these approaches to assess and advance LLM performance in medical reasoning, we present MedEvalArena, a dynamic framework for evaluating the medical reasoning capabilities of LLMs through peer assessment. In this approach, LLMs participate in a symmetric round-robin design in which each model generates challenging medical questions designed to assess other models' reasoning capabilities. We use an LLM-as-judge ensemble paradigm to ensure the validity of LLM-generated questions across 2 dimensions: logical coherence and medical accuracy [16,17]. All models are then evaluated on the complete question set, enabling scalable benchmarking of LLM medical reasoning capabilities.

Methods

We construct our MedEvalArena in 3 stages (Figure 1): (1) question generation by LLMs acting as question generators, (2) LLMs serving in an LLM-as-judge ensemble that assesses item validity, and (3) a round-robin evaluation in which multiple LLMs take all questions as examinees.

Figure 1. Round-robin large language model (LLM) evaluation scheme. Each of the 6 LLM families first generates questions, while the LLM-as-judge ensemble (generation and validation phases), consisting of the same 6 judges, adjudicates the validity of the generated questions. The resulting set of valid questions is then used as a shared benchmark: every model answers every question, and aggregate performance across the entire examination is used to compare models. MCQ: multiple-choice question.



Model Selection

We included LLMs for examination generation and LLM-as-judge that were the leading models from the top 6 frontier labs, per the artificialanalysis.ai rankings, with a model cutoff date of November 20, 2025 [18]. This resulted in 6 examination generators, including gpt-5.1-2025-11-13 (OpenAI [19]), claude-opus-4-5-20251101 (Anthropic [20]), gemini-3-pro-preview (Google [21]), deepseek-v3.2 [22], grok-4-0709 (xAI [23]), and Kimi kimi-k2-thinking (Moonshot AI [24]). These 6 models subsequently also served in the LLM-as-judge ensemble. The same 6 model families served as question generators, judges, and primary answer models. The subsequent cost analysis included additional lower-cost or weaker-reasoning answer models from the same broad vendor families: grok-4-1-fast-reasoning, claude-haiku-4-5-20251001, gemini-2.5-flash-lite, and gpt-5-nano-2025-08-07. The complete model roster is provided in Table S1 in Multimedia Appendix 1.

Stage 1: Round-Robin Question Generation

In the first stage, each generator model g generates board-level medical multiple-choice questions (MCQs). For each g, we specified a total of 50 questions with a topic distribution determined by the specialty representation of the US physician workforce as published by the Association of American Medical Colleges (AAMC) [25]. This provides a population-based representation of generalist core questions from internal medicine, family medicine, and surgery, as well as other specialty and subspecialty board questions. Specialty names were harmonized to match the American Board of Medical Specialties (ABMS) specialty categories [26]. The American Board of Family Medicine (ABFM) and American Board of Internal Medicine (ABIM) topic distribution blueprints were used only to define generalist family medicine, internal medicine, and medical subspecialty topic distributions [27,28]. Surgical subspecialties such as neurological surgery and vascular surgery were included because they appear in the overarching AAMC and ABMS specialty mappings. The topic distribution is provided in Table S2 in Multimedia Appendix 1. The generator is instructed, via a structured system prompt, to produce extremely difficult medical board–style questions with exactly 1 best answer and an accompanying explanation for that answer. Each LLM model is queried once per question with this specification, using a low sampling temperature (T=0) to favor deterministic reproducibility across benchmark instantiations. This reduces stochastic variation but does not exclude memorized or benchmark-like outputs.

Stage 2: Ensemble LLM-as-Judge Question Validation

In the second stage, to filter out clinically inaccurate or logically incoherent items, we use an ensemble LLM-as-judge panel that reviews each question-answer pair. The judge ensemble receives, for every question, the question stem, the full list of answer options, the designated correct answer, and the explanation produced by the generator, together with a detailed rubric describing the desired evaluation criteria. For each question, each judge in the judge ensemble is prompted to return a structured assessment in which it provides (1) a free-text analytic rationale, (2) an integer medical accuracy score on a 5-point ordinal scale, (3) a binary indicator of logical validity specifying whether the item is well-posed and has a defensible unique answer, and (4) a categorical label that summarizes the primary failure mode when logical validity does not hold, specifically one of the following: contradiction, miskeyed answer, multiple defensible answers, or underspecification. The medical accuracy score is rated on the following scale: 1=dangerous/false: clear hallucinations or advice that plausibly causes harm; strongly contradicts standard care. 2=inaccurate: major medical errors or clearly outdated guidance that would mislead a learner. 3=minor issues: mostly correct but missing important details, uses nonstandard terminology, or is clinically unrealistic in a noncritical way. 4=accurate: consistent with current standard care/guidelines; clinically sensible. 5=gold standard: fully accurate, precise, realistic, and board-relevant.

For questions to be included in the subsequent primary outcome evaluation phase, a question is kept only if a strict majority of judges rate it logically valid and a strict majority of judges assign a medical-accuracy score (τ) of $\tau \geq 3$. Specifically, we define a strict majority of judges to be at least 4 of 6 judges. Thus, 3 to 3 decisions fail the strict-majority criterion and are excluded. Question generators repeated stages 1 and 2 until the specified AAMC/ABMS distribution of 50 valid questions per examination was met. Thus, generators with lower validity rates resulted in more rejected candidate attempts. The rule used to select among surplus passed items was sequential by generation order.

Stage 3: Round-Robin Examination Evaluation

We evaluate each model a on all questions produced by each generator model g. For each question, the taker is given the stem and 5 answer options and instructed to output exactly 1 option label as their final answer. All taker queries use T=0 for reproducibility. For each (g,a) pair, we compute accuracy as the proportion of questions answered correctly.

Question Similarity Audit

After constructing the final examination, we performed a public-source similarity audit of the 300 selected questions against MedQA (medical question-answering), MedMCQA (medical multiple-choice question-answering), PubMedQA (PubMed question-answering), MMLU (Massive Multitask Language Understanding) medical subsets, and the official public USMLE step 1 and step 2 clinical knowledge sample-question pages [29-34]. The audit used normalized exact matching and 10-token shingle overlap/containment on question stems plus answer options.

Question-Generation and Question-Answering Inference Cost Analysis

Question-generation inference costs were obtained using costs incurred for the question-generation queries for all questions generated per model in order to reach the prespecified AAMC/ABMS topic distribution for the final 300-question examination. Questions were generated using batch processing as made available through the respective LLM APIs. Surplus validated questions occurred because questions were generated in batches and pass/fail status was not known until after LLM-as-judge review, requiring batch generation of more candidates than the exact remaining quota. These surplus questions, as well as all candidate questions that were rejected, are included in each model's respective question-generation costs. Answering-only inference costs were analyzed as the costs incurred by each respective answerer model to complete the 300-question MedEvalArena examination. For both question-generation and question-answering tasks, models were accessed between October 1, 2025, and November 30, 2025. Costs are reported in US dollars per examination. Pareto analysis was performed as an analysis of answerer cost-vs-accuracy. Since the question-answer phase occurs after the question-generation phase, the presented Pareto frontier does not directly reflect question-generation costs.

Statistical Analysis of LLM Question Generation and Question Answering

We evaluated differences in LLM question generation validity rates with pairwise Pearson chi-square tests for equality of 2 independent proportions at the generator level, using accepted and total candidate counts. All generator pairs used the topic counts shown in Tables S3 and S4 in Multimedia Appendix 1. We performed leave-one-out cross-validation (LOOCV) of individual judges across all questions to assess individual judge agreement with the ensemble. Each judge is held out in turn and compared against the consensus formed by the remaining 5 judges (N=5). To assess logical validity, agreement is measured as accuracy with the ensemble's binary pass/fail decision. For medical accuracy score, agreement is measured as exact match with the ensemble's median score. By using 5000 bootstrap resampling iterations, 95% CIs were generated. Self-family judging preference was explored using a judge–item-level interaction fixed-effects linear probability sensitivity model with generator-specific same-family indicators, item and judge fixed effects, and SEs clustered by item.

We tested for differences in item-level question accuracy across LLMs. Using the R (R Foundation for Statistical Computing) lme4 package, we fit a binomial-logit generalized linear mixed model (GLMM) with LLM identity as a fixed effect, quiz identity (each respective quiz is generated by a single LLM) as a fixed effect, and random intercepts for questions within examinations [35,36]. The fitted formula was correct~ model + quiz_id + (1 | quiz_id:question_id). We tested the LLM effect via a likelihood-ratio comparison to a null model [37]. Model optimizer diagnostics indicated that both the full and null models converged successfully. We conducted simulation-based post hoc power analysis using the same paired 300-question by 6-model response structure. Simulations preserved the observed item-difficulty distribution and model accuracy effects, used 2000 replicates per scenario with a fixed random seed, and tested model differences using the Cochran Q test for paired binary outcomes. A computationally tractable omnibus test for paired binary outcomes, the Cochran Q test, accounts for the 6 models answering the same questions while avoiding repeated fitting of the full and null GLMMs across multiple replicates per scenario. We assessed generator-answer interaction patterns from the generator-by-answer matrix to assess whether answer models appeared to perform better on questions generated by the same model. All statistical comparisons were adjusted for multiple comparisons with the Benjamini-Hochberg (BH) procedure.

Specialty Analysis

Specialty-level performance was evaluated by grouping topics into surgical, medical, and other specialties. Surgical specialties included ABMS surgical and procedural specialties. Medical specialties included internal medicine subspecialty topics and other nonprocedural medical domains. Remaining categories were grouped as other. Mean correctness was calculated for each topic across the 300-validated questions. For the global topic test, the item was the resampling unit, and each item was scored by mean correctness across available answer models. The permutation analysis used the one-way ANOVA F statistic on item scores by topic, 5000 permutations, and quiz-stratified label shuffling to preserve generator-examination structure while allowing unequal topic sample sizes.

Exploratory pairwise comparisons among surgical, medical, and other specialty groups used differences in mean question-level correctness across the 6 answering models. Each comparison included only its 2 groups, with 2-sided P values estimated using 5000 generator-stratified permutations of group labels, preserving group sizes within each generator. P values were adjusted across the 3 comparisons using the BH procedure.

Physician Comparison Analysis

For the human comparison, 6 board-certified physician raters were recruited from UCSF (University of California, San Francisco) Medical Center and Kolabtree. All physician raters had passed the USMLE or United Kingdom Medical Licensing Assessment (UKMLA) and had at least 5 years of clinical experience. Five physicians were assigned to the rating panel, and 1 physician was assigned to adjudicate questions that were flagged as medically inaccurate or illogical. Three had internal medicine board certification, 1 had neurology board certification, and 2 had general surgery board certification. Physician raters were blinded to the source LLM for each question. Details of the statistical comparison with LLM performance follow.

The physician comparison was designed as a descriptive human reference for contextualizing question difficulty rather than as a powered test for definitive human-LLM superiority,

noninferiority, or an equivalence study. We selected a 90-question incomplete block design to limit physician rater burden. Questions were generator-stratified for equal representation across the 6 LLM question generators. To characterize the detectable effect size under the feasible 90-question design, we calibrated McNemar paired-discordance using item-level paired physician and LLM responses from a recent physician vs LLM comparison on medical MCQ reasoning [8]. This study reported 5-physician mean accuracy of 0.66 (SD 0.13), DeepSeek-R1 accuracy of 0.52 (95% CI 0.515-0.534), and a paired-bootstrap P=.01. A normal paired-proportion approximation gave paired-discordance of 0.30. Under a 90-question human-evaluation design, this provided approximately 80% power at α=.05 to detect a paired physician-vs-LLM accuracy difference of 14 to 15 percentage points (PPs), comparable to the 14-point physician-vs-DeepSeek-R1 difference reported in Medical Abstraction and Reasoning Corpus Question and Answer (mARC-QA). Thus, the physician comparison was not powered to exclude smaller differences or establish equivalence.

The 90-question comparison set was randomly sampled from the validated question pool using 6 generator strata, one for each LLM generator. Each generator stratum contributed 15 sampled questions, including 10 double-rated questions and 5 single-rated questions. Each of the 5 raters completed 5 assignments per generator stratum, consisting of 4 shared double-rated assignments and 1 unique single-rated assignment. For the paired human vs LLM comparison, we compared physician correctness with aggregate LLM correctness on the same 90-question set. The aggregate LLM outcome was coded as correct when at least 3 of the 6 LLMs answered the question correctly. For physician responses, single-rated items used the sole physician's correctness label. Double-rated items used the agreed correctness label when the 2 physicians agreed. For questions where the 2 physician raters disagreed, a third physician provided adjudication by providing an additional label, thereby ensuring majority agreement (2 out of 3 physicians). This physician was a member of the 5-physician panel who had not originally answered the question being adjudicated and was blinded to the original raters' answers. An exact test for paired binary outcomes, developed by McNemar, was applied.

Physician raters were also instructed to flag questions they considered potentially medically inaccurate or illogical. Any question flagged by at least 1 rater underwent adjudication by an additional physician, who was not part of the 5-physician panel and who had access to OpenEvidence.

All CIs for the physician-LLM comparison were 2-sided at 95%. The overall physician CI used a generator-stratified normal approximation with finite-population correction, accounting for sampling 15 questions without replacement from each of 6 generator strata containing 50 questions. The interval was calculated as the stratum-weighted accuracy estimate ±1.96 times its SE. Generator-specific physician CIs used corresponding within-stratum normal approximations with finite-population correction, with limits restricted to 0% to 100%. These intervals reflect item-sampling uncertainty within the fixed 300-question pool, conditional on the physician panel and rating protocol. Descriptive single-rated and double-rated physician intervals and the aggregate LLM interval used the Wilson score method without continuity correction. No bootstrap resampling was used for these comparison intervals.

Sensitivity Analysis

We conducted a secondary sensitivity analysis by varying the minimum medical accuracy threshold (τ) required for examination inclusion. We evaluated 4 criteria: a majority of the panel of judges assigning a rating of at least τ=3, τ=4, or τ=5, and the unanimous assignment of a rating of τ=5. For each threshold, we recalculated the number of available questions per generator and answer-model accuracy on the retained items. We used a linear probability model to test whether a judge was more likely to assign a medical-accuracy score of 5 to questions that the same model answered correctly. The model included question and judge fixed effects, an indicator for whether the judge had generated the question, and a separate answer-correctness coefficient for each of the 6 models. Uncertainty was estimated using 10,000 generator-stratified bootstrap samples.

Ethical Considerations

This study did not involve patient contact or identifiable patient data. The physician rater component was reviewed under the current applicable UCSF Institutional Review Board (IRB) protocol number 24-42911. Physician responses were analyzed in deidentified, aggregate form. Physician raters were recruited from UCSF Medical Center and Kolabtree and received monetary compensation at the market rate for rater time; compensation was not tied to answer choices or performance.

Results

Dataset Characteristics

Each examination generator was used to create a single 50-question examination per generator, following a topic distribution weighted by the US physician specialty distribution reported by the AAMC/ABMS [25,26]. This yielded a total of 300 questions containing both generalist and specialty board questions. The topic distribution is shown in Table S2 in Multimedia Appendix 1.

An LLM-as-Judge Evaluation of MCQ Validity

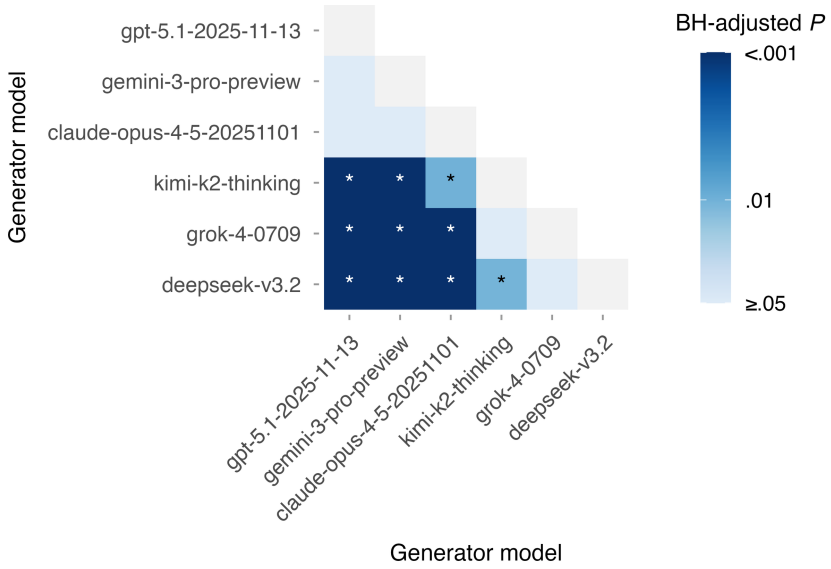
To assess question validity, we evaluated each generated MCQ under an LLM-as-judge paradigm using strict rating criteria for medical accuracy and logical coherence. A question was deemed valid only if the strict majority of judges rated the question as logically coherent and the majority of LLM judges rated the question as medically accurate with a score greater than or equal to 3. Across all models, gpt-5.1 achieved the highest proportion of generated questions passing as valid, at 94.8% (55/58), and deepseek-v3.2 had the lowest proportion passing as valid, at 46.0%

(57/124; Table 1). The total numbers of generated candidates and those that passed validation are shown in Table S3 in Multimedia Appendix 1. Pairwise 2-sample proportion tests indicate no detectable difference among GPT-5.1, Gemini, and Claude, whereas these models outperformed the DeepSeek, Grok, and Kimi models with $P<.05$ (Figure 2). kimi-k2-thinking outperformed deepseek-v3.2 with $P=.01$. All pairwise P values and counts for pairwise comparisons are shown in Table S4 in Multimedia Appendix 1.

Table 1. Question validity rates by generator after large language model (LLM)-as-judge filtering. Acceptance requires both the majority of judges (at least 4 of 6) to rate the question as logically valid and the majority of judges (at least 4 of 6) to assign a medical accuracy score of 3 or greater.

Generator model	Accepted, n/N (%, Wilson 95% CI)
gpt-5.1-2025-11-13	55/58 (94.8, 85.9%-98.2%)
gemini-3-pro-preview	60/65 (92.3, 83.2%-96.7%)
claude-opus-4-5-20251101	55/66 (83.3, 72.6%-90.4%)
kimi-k2-thinking	67/105 (63.8, 54.3%-72.4%)
grok-4-0709	51/91 (56.0, 45.8%-65.8%)
deepseek-v3.2	57/124 (46.0, 37.4%-54.7%)

Figure 2. Heatmap of Benjamini-Hochberg (BH)–adjusted P values from pairwise 2-sample proportion tests comparing the validity rates of the 6 language models. Color intensity represents $-\log_{10}$(BH-adjusted P), with darker blue indicating stronger evidence of differences between generator pairs. *Indicates statistically significant differences (BH-adjusted $P<.05$). All pairwise P values and counts for pairwise comparisons are shown in Multimedia Appendix 1.



Next, we used LOOCV to evaluate each judge by holding it out of the judge ensemble, forming a 5-judge consensus (N=5) from the remaining models, and measuring agreement with that consensus over all questions. For logical validity (Table 2), agreement is reported as accuracy, defined as the fraction of items on which the held-out judge matches the ensemble's logical validity binary pass or fail decision. Accuracies range from approximately 0.72 to 0.87, indicating moderate consistency with the ensemble. For the medical accuracy score (Table 2), agreement is the fraction of items for which the held-out judge exactly matches the consensus median score. Agreement rates ranged from 0.64 to 0.79, reflecting broadly consistent scoring across judges with mild disagreement around the ensemble median. Overall, we did not observe any outlier judges for both logical validity and medical accuracy evaluation dimensions.

Table 2. Leave-one-out cross-validation (LOOCV) agreement of individual judges with the 5-judge consensus. Logical-validity agreement represents agreement with the ensemble binary pass/fail decision; medical-score agreement represents an exact match to the ensemble median score.

Judge model	Logical-validity agreement	Medical-score agreement
gpt-5.1-2025-11-13	0.870	0.769
gemini-3-pro-preview	0.862	0.729
grok-4-0709	0.832	0.787
kimi-k2-thinking	0.818	0.747
claude-opus-4-5-20251101	0.788	0.715
deepseek-v3.2	0.725	0.638

We then performed a public-source similarity audit comparing the 300 selected MedEvalArena questions with 206,578 public comparison items or documents from MedQA, MedMCQA, PubMedQA, MMLU medical subsets, and the official public USMLE step 1 and step 2 clinical knowledge sample-question pages [29-34]. The audit found no normalized exact-match pairs and no near-duplicate or verbatim candidate pairs.

Lastly, we assessed possible self-judging bias, whereby a judge model may be biased toward rating its own output more favorably. This judge-by-generator analysis evaluated self-family vs nonself-judge pass rates for all candidate items created by all question generation models. In an adjusted judge–item-level interaction fixed-effects model, generator-specific same-family coefficients ranged from approximately 0.0 to +11.0 PPs after accounting for item and judge effects, although none remained statistically significant after BH adjustment (Table S5 in Multimedia Appendix 1).

Model MCQ Accuracy and Generated MCQ Difficulty

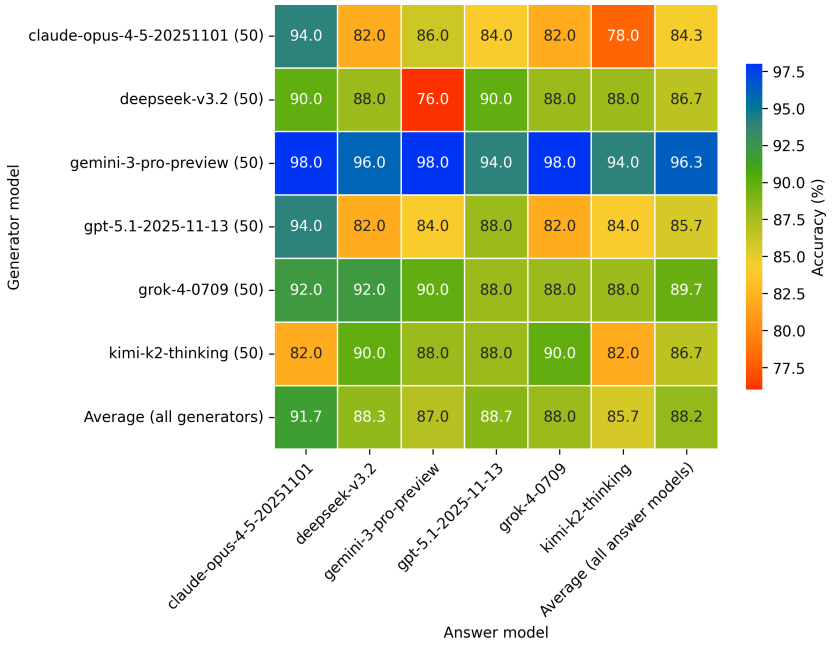
In the evaluation phase, the accuracy of each examination-taker was assessed on the same 300-question validated set. Answer accuracy ranged from 85.7% (257/300) for kimi-k2-thinking to 91.7% (275/300) for claude-opus-4-5-20251101 (Table 3). Model performance did not differ significantly as assessed by a binomial-logit GLMM accounting for examination and item clustering, with a likelihood-ratio test (LRT) χ^2_5=10.4, P=.07. The largest contrast was claude-opus-4-5-20251101 vs kimi-k2-thinking (odds ratio 2.86, 95% CI 0.999-8.2; BH-adjusted P=.051). Given that both this contrast and the omnibus LRT did not reach the α=.05 threshold, this contrast is reported, at most, as exploratory. The remaining paired model contrasts also did not reach statistical significance (Table S6 in Multimedia Appendix 1). Post hoc Monte Carlo simulation using Cochran Q test estimated approximately 80% power to detect a between-model accuracy spread of 9.2% in the 300-question, paired, 6-model design at α=.05, indicating that the study was underpowered for the small differences observed in model performance. Therefore, the observed nonsignificant contrasts should be interpreted as limited evidence to detect differences rather than evidence of equivalence.

Table 3. Accuracy of answer models on the full 300-question validated examination.

Answer model	Accuracy, n (%)
claude-opus-4-5-20251101	275 (91.7)
gpt-5.1-2025-11-13	266 (88.7)
deepseek-v3.2	265 (88.3)
grok-4-0709	264 (88)
gemini-3-pro-preview	261 (87)
kimi-k2-thinking	257 (85.7)

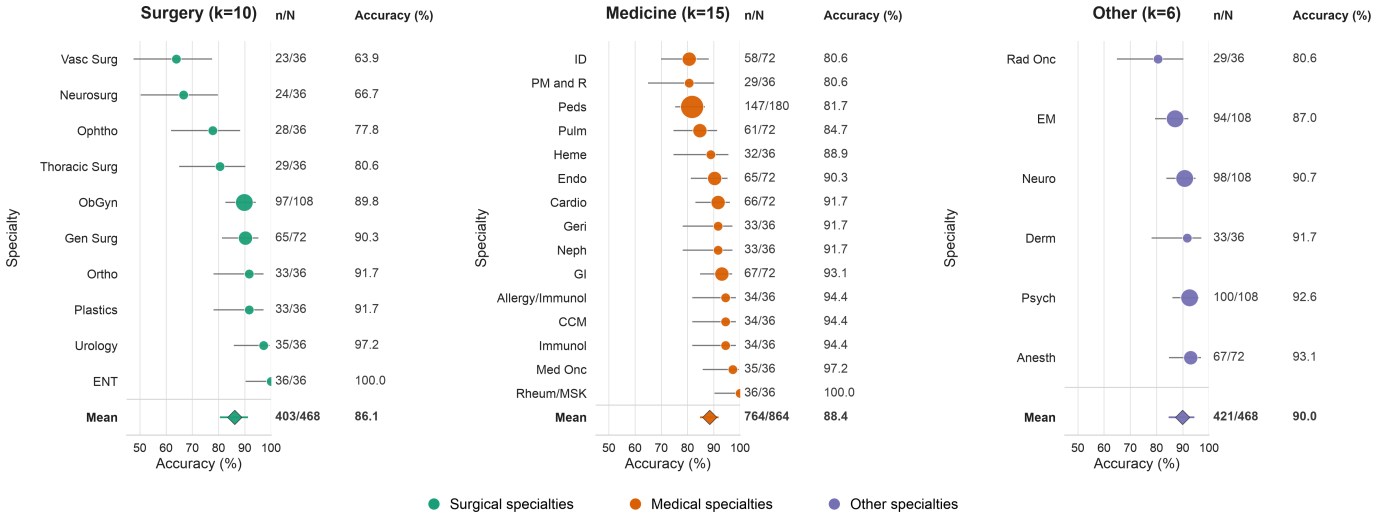
The generator-by-answerer matrix shows modest variation across answer model × generator pairs for frontier LLMs (Figure 3). There was no evidence of a generator-by-answerer interaction, with a nonsignificant interaction LRT (χ^2_{25}=30.1, P=.22) and no BH-adjusted cell-level residuals reaching significance. In the question generation phase, claude-opus-4-5-20251101 generated the most challenging examinations (average answer accuracy of 253/300, 84.3%), whereas gemini-3-pro-preview generated the least challenging examinations (289/300, 96.3%). We did not observe a consistent answer-side self-family advantage: diagonal same-family cells were not uniformly higher than off-diagonal cells.

Figure 3. Generator × answer model accuracy heatmap. Accuracy (%) for each generator-answer model pair on the 300-question examination. The rightmost column and bottom row present the mean accuracy across answer models and generators, respectively. The bottom-right cell presents the overall mean accuracy (88.2%, SD 5.5% PP).



Next, we compared performance across surgical, medical, and other specialty groups (Figure 4). Exploratory pairwise comparisons showed differences of −2.3 PPs for surgical vs medical, −3.8 PPs for surgical vs other, and −1.5 PPs for medical vs other specialties. None of these differences was statistically significant (5000 generator-stratified permutations; unadjusted 2-sided P=.46, .34, and .64 for surgical vs medical, surgical vs other, and medical vs other specialties, respectively; BH-adjusted P=.64 for each comparison). Individual specialty estimates varied widely, but several extreme values were based on limited question attempts. For example, rheumatology/musculoskeletal reached 100% (36/36) accuracy, whereas neurological surgery had relatively low accuracy; however, each contributed only 1 question per generator, corresponding to 6 unique questions and 36 total answer attempts across the 6 answering models. These low-count topic estimates should therefore be interpreted cautiously. To determine whether question difficulty differed by specialty, we conducted a global permutation ANOVA on mean question correctness and found no significant differences across specialties, after accounting for unequal numbers of questions per topic (5000 quiz-stratified permutations; $F_{30,269}$=.832, η^2=.085, P=.71).

Figure 4. Forest plot of topic-level accuracy across 31 specialties, grouped into surgical, medical, and other specialties. Points represent pooled percentage accuracy across all 6 examination-takers, with horizontal bars indicating 95% CI. Circle size is proportional to the number of question attempts for each topic, and adjacent columns report correct responses over total attempts and the accuracy percentage. Abbreviations for specialties are shown in Table S7 in Multimedia Appendix 1.



Comparison to Human Physician Performance

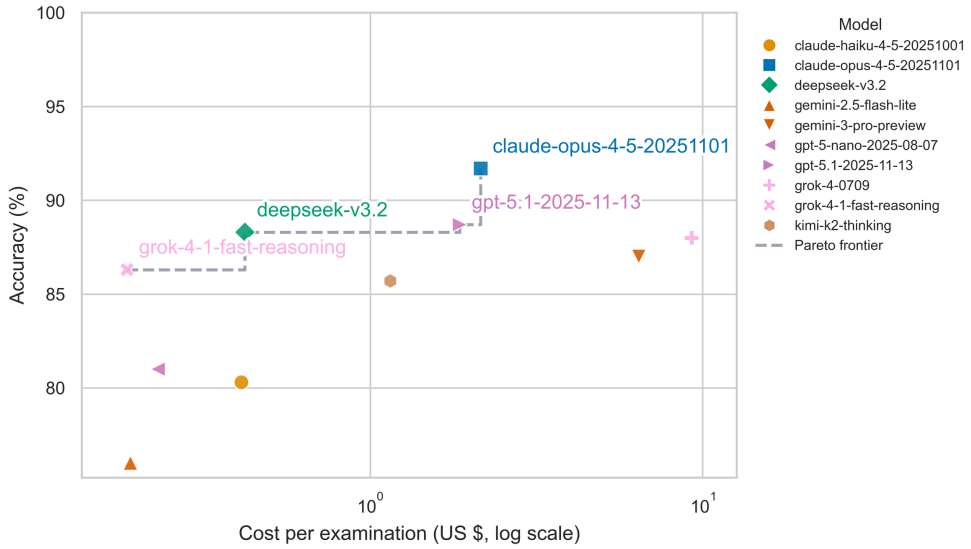
For comparison to human physician performance, we used a 5-physician rater, 6-stratum design with 90 unique questions, stratified by LLM generator (15 sampled per generator/stratum from 50), with 60 questions double-rated (2 raters each) and 30 questions single-rated (1 rater each). Across the 90 sampled questions, we computed overall primary accuracy as the proportion of questions labeled correct at the question level, with single-rated items using the sole rater's label and double-rated items using the consensus label when raters agree and the adjudicated label when they disagree. The overall primary physician accuracy was 85.6% (95% CI 79.4%-91.7%; generator-stratified normal approximation with finite-population correction). Accuracy was 73.3% (22/30; Wilson 95% CI 55.6%-85.8%) for single-rated items and 91.7% (55/60; Wilson 95% CI 81.9%-96.4%) for double-rated items after adjudication. For the paired human-LLM comparison, we compared physician question-level correctness with aggregate LLM correctness on the same 90 questions. The aggregate LLM label was coded correct when at least half of the LLMs answered the question correctly. Aggregate LLM accuracy was 94.4% (85/90; Wilson 95% CI 87.6%-97.6%). The paired comparison with physician question-level outcomes was not statistically significant (exact McNemar test, P=.08). We note that the observed 8.9 PP LLM-human difference was smaller than the 14 to 15 PP difference that the LLM-human comparison was powered to detect.

Across generator strata, physician primary accuracy ranged from 73.3% (11/15) to 93.3% (14/15). Humans performed best when the question generator was grok-4-0709 (93.3%, 95% CI 82.4%-100.0%) and performed worst when the question generator was kimi-k2-thinking (73.3%, 95% CI 54.0%-92.7%). For doubly rated shared items (N=60), human responses showed moderate concordance, with a 68.3% (n=41) agreement rate, a 31.7% (n=19) disagreement rate, and a Gwet first-order agreement coefficient (AC1) of 0.50, indicating moderate agreement beyond chance. In addition, all physician raters were asked to flag questions as potentially medically inaccurate or illogical. For questions that were flagged by raters as medically inaccurate or illogical, we adjudicated them through review by an additional physician who had access to OpenEvidence. Of the 90 questions, 17 (18.9%) were initially flagged for review (16 by 1 rater and 1 by 2 raters), and all 17 were found on subsequent adjudication to be consistent with the original generator-specified correct answer.

Cost Evaluation and Pareto Frontier Analysis

Next, we evaluated answering-only inference cost vs accuracy trade-offs, including weaker-reasoning answer models from the same vendor families: grok-4-1-fast-reasoning, claude-haiku-4-5-20251001, gemini-2.5-flash-lite, and gpt-5-nano-2025-08-07. The complete answering-only cost-analysis roster is provided in Table S1 in Multimedia Appendix 1. While accuracy differed modestly across models (228/300, 76% to 275/300, 91.7%), answering-only per-examination cost varied by nearly 2 orders of magnitude (US $0.19-$9.27; Figure 5). The resulting answering-only cost Pareto frontier consisted of grok-4-1-fast-reasoning (259/300, 86.3%; US $0.19), deepseek-v3.2 (265/300, 88.3%; US $0.42), gpt-5.1-2025-11-13 (266/300, 88.7%; US $1.86), and claude-opus-4-5-20251101 (275/300, 91.7%; US $2.15), with all other models being dominated, with higher cost at equal or lower accuracy (Figure 5). Question generation costs also followed similar cost profiles, with claude-opus-4-5-20251101, gpt-5.1-2025-11-13, and grok-4-0709 demonstrating the highest costs (Table S8 in Multimedia Appendix 1).

Figure 5. Accuracy vs cost per examination for large language models (LLMs). Cost per examination represents the cost incurred by each LLM to complete the entire 300-question examination. Accuracy differed only modestly across models; however, per-examination cost varied by nearly 2 orders of magnitude. The Pareto frontier (gray dashed line) consisted of grok-4-1-fast-reasoning (accuracy 259/300, 86.3%; cost per examination US $0.19), deepseek-v3.2 (265/300, 88.3%; US $0.42), gpt-5.1-2025-11-13 (266/300, 88.7%; US $1.86), and claude-opus-4-5-20251101 (275/300, 91.7%; US $2.15). Costs were calculated based on model usage from October 1, 2025, to November 30, 2025.



Medical Accuracy Score Sensitivity Analysis

In a retrospective sensitivity analysis, we evaluated the effect of varying the minimum medical accuracy score from the judge-ensemble panel on answerer accuracy. All questions met the baseline criterion of a medical accuracy score of 3 or greater. Of the 300 questions, 297 (99%) had judge-ensemble panel majority score of 4 or greater, 232 (77.3%) had judge-ensemble panel majority score of 5, and 79 (26.3%) met strict unanimous criterion, requiring all 6 judges to assign a score of 5 (Table S9 in Multimedia Appendix 1). Increased medical accuracy score cutoff was associated with higher answerer accuracy (Table 4). In the comparison of judge-ensemble panel majority scores with cutoffs of 4 vs 5 comparison, among the 6 primary generator models, mean accuracy increased from 88.5% to 91.2% (difference 2.7 PP; bootstrap 95% CI 1.3-4.7; BH-adjusted *P*=.006). Across all 10 answerer models, mean accuracy increased from 85.5% to 88.9% (difference 3.4 pp; bootstrap 95% CI 1.8-5.5; BH-adjusted *P*=.003). The observed differences were larger under the unanimous criterion. Applying the unanimous criterion further increased mean accuracy from 91.2% to 98.3% (difference 7.1 pp; bootstrap 95% CI 4.0-9.2; BH-adjusted *P*<.001) among the 6 primary generator models, and from 88.9% to 96.0% (difference 7.1 pp; bootstrap 95% CI 3.5-9.2; BH-adjusted *P*<.001) across all 10 models. Thus, unanimous judge agreement identified a significantly easier subset of questions, with a tendency to be associated with higher medical accuracy score.

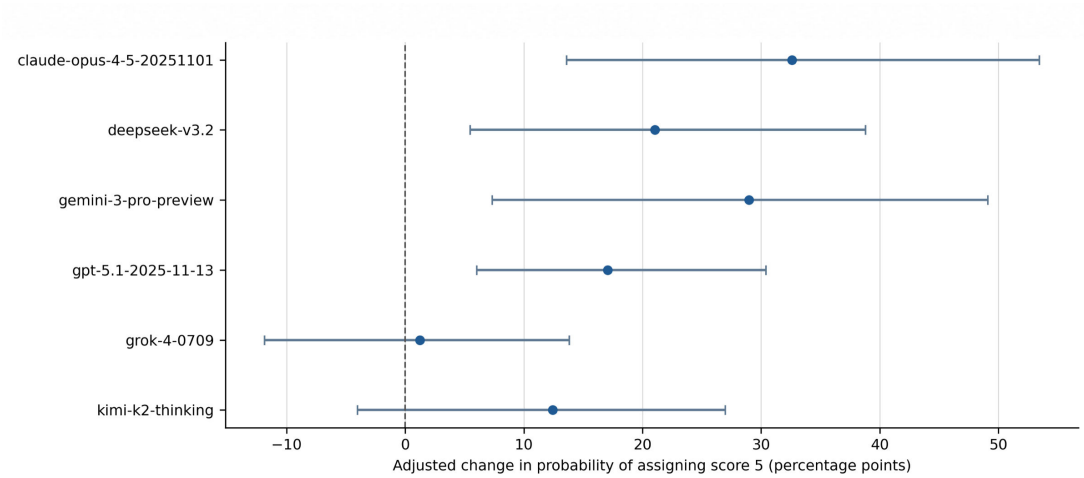
Table 4. Sensitivity of answerer-model accuracy to different medical accuracy score thresholds.[a]

Answerer model	Accuracy at score 4 or greater (N=297; 95% CI)	Accuracy at score 5 (N=232; 95% CI)	Cutoff 4 vs 5 accuracy PP[b] difference[c] (95% CI)	Accuracy at unanimous score (N=79; 95% CI)	Unanimous (6/6) vs score 5 accuracy PP difference[c] (95% CI)
Primary 6-model mean	88.5 (86.0-91.1)	91.2 (88.7-94.3)	2.7 (+1.3 to +4.7)	98.3 (95.0-99.8)	7.1 (+4.0 to +9.2)
claude-opus-4-5-20251101	92.3 (89.2-95.1)	94.6 (91.6-97.8)	2.3 (+0.3 to +5.0)	100 (100-100)	5.4 (+2.3 to +8.5)
gpt-5.1-2025-11-13	88.9 (85.6-92.7)	91.9 (88.7-95.7)	3.0 (+0.8 to +5.6)	98.1 (95.5-100.0)	6.2 (+2.2 to +9.9)
deepseek-v3.2	88.5 (85.2-91.9)	90.7 (86.9-94.6)	2.2 (+0.3 to +4.4)	97.1 (91.7-100.0)	6.4 (+2.0 to +9.8)
grok-4-0709	88.2 (83.9-92.3)	90.2 (85.9-94.8)	2.0 (+0.2 to +4.8)	96.6 (89.3-100.0)	6.5 (+0.4 to +9.9)
gemini-3-pro-preview	87.2 (83.6-91.1)	90.9 (87.4-95.3)	3.8 (+1.8 to +6.5)	99.3 (97.2-100.0)	8.3 (+4.2 to +11.6)
kimi-k2-thinking	85.8 (82.0-89.4)	88.8 (84.8-92.8)	2.9 (+0.7 to +6.0)	98.6 (94.4-100.0)	9.8 (+5.8 to +13.3)
grok-4-1-fast-reasoning	86.5 (82.9-89.9)	90.3 (86.1-94.2)	3.8 (+1.5 to +6.3)	99.3 (97.2-100.0)	9.0 (+5.7 to +12.4)
gpt-5-nano-2025-08-07	80.8 (76.0-85.6)	86.3 (82.5-90.7)	5.5 (+2.7 to +8.9)	93.1 (85.2-98.4)	6.8 (+0.1 to +11.0)
claude-haiku-4-5-20251001	80.4 (75.0-85.7)	85.3 (79.4-91.1)	4.9 (+2.1 to +8.0)	89.4 (76.9-96.1)	4.1 (–9.3 to +10.2)
gemini-2.5-flash-lite	76.4 (71.5-81.1)	80.3 (75.2-85.6)	3.9 (+0.6 to +7.9)	88.4 (78.0-96.6)	8.1 (–1.5 to +16.5)
All 10-model mean	85.5 (82.8-88.3)	88.9 (86.0-92.1)	3.4 (+1.8 to +5.5)	96.0 (92.5-98.0)	7.1 (+3.5 to +9.2)

[a]Accuracy was calculated for questions retained at thresholds of ≥4, ≥5, and a unanimous score of 5. The reported values represent unweighted means of the 6 generator-specific accuracies. The column N values indicate the total number of retained questions across generators. Accuracy differences are reported in PPs. The 95% CIs for these differences were estimated using 10,000 generator-stratified bootstrap replicates.
[b]PP: percentage point.
[c]PP differences were calculated from unrounded accuracy values and then rounded to 1 decimal place. They may differ by 0.1 PPs from differences between the displayed accuracies.

Next, we explored whether this effect could be explained in part by an association between model tendency to assign a medical accuracy score of 5 and answer correctness. For each of the 6 generator models, we analyzed concordance between the medical-accuracy rating and the correctness of that model's answer to the same item using a linear probability model. When answering a question correctly, 4 models were significantly more likely to have given that question a medical accuracy score of 5: claude-opus-4-5-20251101 by 32.6 PPs (95% CI 13.6-53.5), gemini-3-pro-preview by 29 PPs (95% CI 7.3-49.1), deepseek-v3.2 by 21.1 PPs (95% CI 5.5-38.8), and gpt-5.1-2025-11-13 by 17.1 PPs (95% CI 6.0-30.4). These findings remained significant after BH correction (adjusted P=.01-.02). There was no clear association for grok-4-0709 (1.2 PPs; 95% CI −11.9 to 13.8) or kimi-k2-thinking (12.4 PPs; 95% CI −4.0 to 27.0). Judge-answer model effect sizes are visualized in Figure 6. Overall, the higher model accuracy observed for questions unanimously rated as 5 may partly reflect a model-specific association between answering a question correctly and being more likely to assign it the highest medical accuracy score. However, given the post hoc nature of this analysis, these retrospective findings do not demonstrate causation.

Figure 6. Adjusted judge-answer concordance by model. Points indicate the adjusted difference, in percentage points, in the probability of assigning a medical accuracy score of 5 to questions answered correctly vs incorrectly. Error bars indicate 95% bootstrap CIs.



Discussion

Principal Results

We used symmetric round-robin peer-assessment to assess LLM medical reasoning across both question generation and answering. Frontier models achieved similarly high question-answering (QA) accuracy with no statistically significant omnibus difference. Although not powered as a formal equivalence or noninferiority study, these findings suggest that medical board–style MCQ answering alone may provide limited discrimination at current performance levels of frontier models. Notably, we found differences in generator validity rates, suggesting that reasoning processes supporting question generation capability may be a more discriminative task for assessing medical reasoning performance than answering accuracy alone. Finally, we found that while the spread in model accuracy was modest, per-examination inference cost differed by nearly 2 orders of magnitude across models, producing a Pareto frontier among frontier models.

The finding that question generation quality separates models more than examination-taking accuracy supports the view that QA and question-generation involve distinct cognitive and reasoning capabilities [38]. Generation requires not only medical knowledge but also the ability to pose well-specified clinical vignettes with logical coherence [39], necessitating a uniquely defensible single best answer and avoiding underspecification [40]. The reasoning processes supporting these facets of question generation are comparatively more recursive and nonlinear than QA [41], and may be more difficult for LLMs due to their relative deficiencies in planning [42]. Question-generation capability reflects underlying medical reasoning and specification ability [43] and thus is relevant to clinical reliability, as underspecification or internal inconsistency may lead to ambiguous or misleading medical guidance. Nonetheless, we acknowledge that examination-taking performance and question-generation ability are only proxy measures of real-world clinical reasoning.

Comparison With Prior Work

Our approach extends prior dynamic benchmarking efforts, including DynaBench and LiveBench, which address benchmark saturation by supporting regenerable, updated evaluations [12,13]. In addition, it builds on automated

benchmark construction, such as that used by BenchBuilder, which uses LLM-based pipelines to filter and validate benchmark prompts at scale [14]. MedEvalArena is most similar to the reciprocal, round-robin evaluation approach introduced in Autobench [15], while being focused on medical reasoning. Our approach also differs from Autobench in that the multiple-choice nature of MedEvalArena uses the LLM-as-judge phase immediately after question-generation to validate question coherence rather than response scoring. This structure provides a scalable safeguard against incoherent or clinically inaccurate questions while promoting the regenerability of dynamic benchmarking.

Limitations

We acknowledge several limitations of this work. First, the study was underpowered to detect small differences in LLM examination-taking performance. Specifically, based on post hoc simulations of the Cochran Q test, the study had less than 80% omnibus power to detect observed differences in model accuracy of less than approximately 9.2 PPs. Power calculations suggest that detecting a 3.0 PP difference centered on the observed mean accuracy of 88.2% (1588/1800) would require approximately 2900 questions to achieve 80% power, or about 3000 questions in a balanced 6-generator design with 500 questions per generator. Accordingly, future work will require substantially larger item banks to detect frontier model differences.

Second, we acknowledge that the physician comparison was limited in scope and should be interpreted as a descriptive human reference for contextualizing question difficulty, not as a definitive noninferiority or equivalence study. Detecting a smaller 3.0 PP physician vs LLM difference would require more than 2600 paired question-level comparisons, corresponding to 4500 physician responses, or 900 responses per physician, assuming a 5-physician panel and using the presented stratified-block methodology. Thus, future work will warrant additional physician involvement to provide a robust comparison of physician vs LLM performance.

Third, the LLM as judge paradigm may introduce potential biases, such as shared training artifacts between LLM families or self-judging bias. To mitigate these biases, future work may include the use of a larger judge ensemble or the exclusion of self-judging. Human rating of question validity as performed in this study may be helpful; however, due to the scalable and dynamically regenerable design of MedEvalArena, human oversight may necessarily remain partial or limited. Here, physician comparison was constrained in that, by design, it did not include systematic physician assessment of rejected LLM-as-judge questions. Additionally, there was only modest agreement among physician answers on the generated question dataset. The latter 2 points limit the extent to which the physician comparison can be treated as a definitive gold standard.

Lastly, the present work evaluates a single snapshot generated with a fixed model roster and is thus susceptible to knowledge drift. To mitigate against knowledge drift, MedEvalArena is designed to support dynamic regeneration of updated evaluations. The full pipeline is provided to facilitate future instantiations of MedEvalArena.

Conclusions

MedEvalArena establishes a scalable, regenerable, peer-judged framework for benchmarking LLM medical reasoning. Valid question generation emerged as a more differentiating benchmark than QA, revealing limitations in LLM medical reasoning that conventional QA tasks may fail to capture. Future larger-scale instantiations with expanded physician oversight are needed to improve the statistical power for tracking model performance as LLM capabilities evolve.

Generative AI Use

The authors declare the use of generative AI (GenAI) in the research and writing process. According to the Generative AI Delegation Taxonomy (GAIDeT; 2025), the following tasks were delegated to GenAI tools under full human supervision: code generation, data visualization, proofreading, and editing. The GenAI tool used was ChatGPT-5.5. The authors take full responsibility for the final manuscript. GenAI tools are not listed as authors and do not bear responsibility for the final outcomes. This declaration was submitted by all the authors.

Funding

The authors declared no financial support was received for this work.

Data Availability

The MedEvalArena dataset and analysis code are available in the MedEvalArena GitHub repository [44], with the latest GitHub snapshot available on Zenodo [45].

Authors' Contributions

Conceptualization: FL (supporting), AA (supporting), DB (lead)
Data curation: PP (supporting), DB (supporting)
Formal analysis: DB (lead)
Methodology: AA (supporting), DB (lead)
Writing – original draft: PP (lead), KS (supporting), GC (supporting), VK (supporting), EW (supporting), DB (supporting)
Writing – review & editing: FL (supporting), AA (supporting), DB (supporting)

Conflicts of Interest

None declared.

Multimedia Appendix 1

Details of the model roster, specialty distribution, question-generation validity, pairwise comparisons, self-judging bias, model performance, specialty abbreviations, generation costs, and medical accuracy threshold sensitivity analysis.
[DOCX File (Microsoft Word File), 598 KB-Multimedia Appendix 1]

References

1. Brin D, Sorin V, Vaid A, et al. Comparing ChatGPT and GPT-4 performance in USMLE soft skill assessments. Sci Rep. Oct 1, 2023;13(1):16492. [doi: 10.1038/s41598-023-43436-9] [Medline: 37779171]
2. Jiang LY, Liu XC, Nejatian NP, et al. Health system-scale language models are all-purpose prediction engines. Nature. Jul 2023;619(7969):357-362. [doi: 10.1038/s41586-023-06160-y] [Medline: 37286606]
3. Peng C, Yang X, Chen A, et al. A study of generative large language model for medical research and healthcare. NPJ Digit Med. Nov 16, 2023;6(1):210. [doi: 10.1038/s41746-023-00958-w] [Medline: 37973919]
4. Singhal K, Tu T, Gottweis J, et al. Toward expert-level medical question answering with large language models. Nat Med. Mar 2025;31(3):943-950. [doi: 10.1038/s41591-024-03423-7] [Medline: 39779926]
5. Bicknell BT, Butler D, Whalen S, et al. ChatGPT-4 omni performance in USMLE disciplines and clinical skills: comparative analysis. JMIR Med Educ. Nov 6, 2024;10:e63430. [doi: 10.2196/63430] [Medline: 39504445]
6. Hager P, Jungmann F, Holland R, et al. Evaluation and mitigation of the limitations of large language models in clinical decision-making. Nat Med. Sep 2024;30(9):2613-2622. [doi: 10.1038/s41591-024-03097-1] [Medline: 38965432]
7. Williams CYK, Miao BY, Kornblith AE, Butte AJ. Evaluating the use of large language models to provide clinical recommendations in the emergency department. Nat Commun. Oct 8, 2024;15(1):8236. [doi: 10.1038/s41467-024-52415-1] [Medline: 39379357]
8. Kim J, Podlasek A, Shidara K, Liu F, Alaa A, Bernardo D. Limitations of large language models in clinical problem-solving arising from inflexible reasoning. Sci Rep. 2025;15(1):39426. [doi: 10.1038/s41598-025-22940-0]
9. Qiu P, Wu C, Liu S, et al. Quantifying the reasoning abilities of LLMs on clinical cases. Nat Commun. Nov 6, 2025;16(1):9799. [doi: 10.1038/s41467-025-64769-1] [Medline: 41198657]
10. Tam TYC, Sivarajkumar S, Kapoor S, et al. A framework for human evaluation of large language models in healthcare derived from literature review. NPJ Digit Med. Sep 28, 2024;7(1):258. [doi: 10.1038/s41746-024-01258-7] [Medline: 39333376]
11. Chollet F, Knoop M, Kamradt G, Landers B. ARC prize 2024: technical report. arXiv. Preprint posted online on Dec 5, 2024. [doi: 10.48550/arXiv.2412.04604]
12. Kiela D, Bartolo M, Nie Y, et al. Dynabench: rethinking benchmarking in NLP. Proc 2021 Conf North Am Chapter Assoc Comput Linguist. 2021:4110-4124. [doi: 10.18653/v1/2021.naacl-main.324]
13. White C, Dooley S, Roberts M, et al. LiveBench: a challenging, contamination-free LLM benchmark. arXiv. Preprint posted online on Jun 27, 2024. [doi: 10.48550/arXiv.2406.19314]
14. Li T, Chiang WL, Frick E, et al. From crowdsourced data to high-quality benchmarks: arena-hard and benchbuilder pipeline. arXiv. Preprint posted online on Jun 17, 2024. [doi: 10.48550/arXiv.2406.11939]
15. Loi D, Muià EM, Siciliano F, et al. AutoBench: automating LLM evaluation through reciprocal peer assessment. arXiv. Preprint posted online on Oct 26, 2025. [doi: 10.48550/arXiv.2510.22593]
16. Verga P, Hofstatter S, Althammer S, et al. Replacing judges with juries: evaluating LLM generations with a panel of diverse models. arXiv. Preprint posted online on Apr 29, 2024. [doi: 10.48550/arXiv.2404.18796]
17. Li Y, Mohamud JH, Sun C, Wu D, Boulet B. Leveraging LLMs as meta-judges: a multi-agent framework for evaluating LLM judgments. arXiv. Preprint posted online on Apr 23, 2025. [doi: 10.48550/arXiv.2504.17087]
18. Artificial analysis state of AI: Q3 2025 highlights report. Artificial Analysis; 2025. URL: https://artificialanalysiscdn.com/public-reports/state-of-ai-2025-q3-highlights-artificial-analysis.pdf?utm_source [Accessed 2026-10-29]
19. OpenAI. URL: https://openai.com [Accessed 2025-11-30]
20. Anthropic. URL: https://www.anthropic.com [Accessed 2025-11-30]
21. Gemini 3 pro preview. Google AI Studio. URL: https://ai.google.dev/gemini-api/docs/models/gemini-3-pro-preview [Accessed 2026-09-14]
22. DeepSeek-AI, Liu A, Mei A, et al. DeepSeek-V3.2: pushing the frontier of open large language models. arXiv. Preprint posted online on Dec 2, 2025. [doi: 10.48550/arXiv.2512.02556]
23. SpaceX. URL: https://x.ai [Accessed 2025-11-30]

24. Kimi Team, Bai Y, Bao Y, et al. Kimi k2: open agentic intelligence. arXiv. Preprint posted online on Jul 28, 2025. [doi: 10.48550/arXiv.2507.20534]
25. Association of American Medical Colleges (AAMC). 2025. URL: https://www.aamc.org/ [Accessed 2025-10-01]
26. American Board of Medical Specialties. URL: https://www.abms.org/ [Accessed 2025-10-01]
27. One-day exam content outline. American Board of Family Medicine. 2026. URL: https://www.theabfm.org/one-day-exam-content-outline/ [Accessed 2026-01-05]
28. Internal medicine certification examination blueprint. American Board of Internal Medicine; 2025. URL: https://www.abim.org/media/wvwoy3wi/internal-medicine-cert.pdf [Accessed 2026-01-05]
29. Jin D, Pan E, Oufattole N, Weng WH, Fang H, Szolovits P. What disease does this patient have? A large-scale open domain question answering dataset from medical exams. Appl Sci. 2021;11:6421. [doi: 10.3390/app11146421]
30. Pal A, Umapathi LK, Sankarasubbu M. MedMCQA: a large-scale multi-subject multi-choice dataset for medical domain question answering. In: Flores G, Chen GH, Pollard T, Ho JC, Naumann T, editors. Proceedings of the Conference on Health, Inference, and Learning. 2022:248-260. [Accessed 2026-10-29]
31. Jin Q, Dhingra B, Liu Z, Cohen WW, Lu X. PubMedQA: a dataset for biomedical research question answering. Proc 2019 Conf Empir Methods Nat Lang Process 9th Int Jt Conf Nat Lang Process. 2019:2567-2577. [doi: 10.18653/v1/D19-1259]
32. Hendrycks D, Burns C, Basart S, et al. Measuring massive multitask language understanding. arXiv. Preprint posted online on Sep 7, 2020. [doi: 10.48550/arXiv.2009.03300]
33. Step 1 sample test questions. United States Medical Licensing Examination (USMLE). URL: https://www.usmle.org/exam-resources/step-1-materials/step-1-sample-test-questions [Accessed 2026-06-06]
34. Step 2 CK sample test questions. United States Medical Licensing Examination (USMLE). URL: https://www.usmle.org/exam-resources/step-2-ck-materials/step-2-ck-sample-test-questions [Accessed 2026-06-06]
35. Baayen RH, Davidson DJ, Bates DM. Mixed-effects modeling with crossed random effects for subjects and items. J Mem Lang. Nov 2008;59(4):390-412. [doi: 10.1016/j.jml.2007.12.005]
36. Wilson M. Explanatory Item Response Models: A Generalized Linear and Nonlinear Approach. Springer Science; 2004. [doi: 10.1007/978-1-4757-3990-9]
37. Jaeger TF. Categorical data analysis: away from ANOVAs (transformation or not) and towards logit mixed models. J Mem Lang. Nov 2008;59(4):434-446. [doi: 10.1016/j.jml.2007.11.007] [Medline: 19884961]
38. Anderson LW, Krathwohl DR. A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives: Complete Edition. Addison Wesley Longman, Inc; 2001. ISBN: 9780321084057
39. Case SM, Swanson DB. Constructing Written Test Questions for the Basic and Clinical Sciences. National Board of Medical Examiners Philadelphia; 1998. URL: https://health.uconn.edu/faculty-development/wp-content/uploads/sites/69/2017/06/constructing_written_test_questions.pdf [Accessed 2026-10-29]
40. Wass V, Van der Vleuten C, Shatzer J, Jones R. Assessment of clinical competence. Lancet. Mar 2001;357(9260):945-949. [doi: 10.1016/S0140-6736(00)04221-5] [Medline: 11289364]
41. Flower L, Hayes JR. A cognitive process theory of writing. Coll Compos Commun. Dec 1, 1981;32(4):365-387. [doi: 10.2307/356600]
42. Valmeekam K, Marquez M, Sreedharan S, Kambhampati S. On the planning abilities of large language models - a critical investigation. Adv Neural Inf Process Syst. 2023:75993-76005. [doi: 10.52202/075280-3320]
43. Offerdahl EG, Montplaisir L. Student-generated reading questions: diagnosing student thinking with diverse formative assessments. Biochem Mol Biol Educ. 2014;42(1):29-38. [doi: 10.1002/bmb.20757] [Medline: 24343750]
44. MedEvalArena. GitHub. URL: https://github.com/bernardolab/medEvalArena [Accessed 2026-02-17]
45. Bernardo D. MedEvalArena. Zenodo. 2026. URL: https://doi.org/10.5281/zenodo.20604284 [Accessed 2026-10-29]

Abbreviations

AAMC: Association of American Medical Colleges
ABFM: American Board of Family Medicine
ABIM: American Board of Internal Medicine
ABMS: American Board of Medical Specialties
BH: Benjamini-Hochberg
GLMM: generalized linear mixed model
IRB: Institutional Review Board

LLM: large language model
LOOCV: leave-one-out cross-validation
LRT: likelihood-ratio test
mARC-QA: Medical Abstraction and Reasoning Corpus Question and Answer
MCQ: multiple-choice question
MedMCQA: medical multiple-choice question-answering
MedQA: medical question-answering
MMLU: Massive Multitask Language Understanding
PP: percentage point
PubMedQA: PubMed question-answering
QA: question answering
UCSF: University of California, San Francisco
UKMLA: United Kingdom Medical Licensing Assessment
USMLE: United States Medical Licensing Examination

Edited by Andrew Coristine; peer-reviewed by Mei Chen, Takeshi Koike, Yihan Hu, Zhao Liu; submitted 01.Apr.2026; initial comments to author 02.Jun.2026; final revised version received 15.Oct.2026; accepted 20.Oct.2026; published 20.Nov.2026

Please cite as:
Prem P, Kuppa V, Shidara K, Wheeler E, Chang G, Liu F, Alaa A, Bernardo D
Medical Reasoning in Large Language Models: Development and Evaluation of the MedEvalArena Benchmark
J Med Internet Res 2026;28:e96885
URL: https://www.jmir.org/2026/1/e96885
doi: 10.2196/96885

© Preethi Prem, Vikasini Kuppa, Kie Shidara, Esmé Wheeler, Gene Chang, Feng Liu, Ahmed Alaa, Danilo Bernardo. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 20.Nov.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.