Beyond Overall Accuracy: A Psychometric Deep Dive into the Topic-Specific Medical Capabilities of 80 Large Language Models

Zhimeng Luo* **Lixin Wu†** **Adam Frisch‡** **Daqing He***
*School of Computing and Information, University of Pittsburgh
†Department of Educational Psychology, University of Illinois Urbana–Champaign
‡Department of Emergency Medicine, University of Pittsburgh
{zhl123,dah44}@pitt.edu lixinwu2@illinois.edu frischan@upmc.edu

Abstract

As Large Language Models (LLMs) are increasingly proposed for high-stakes medical applications, there has emerged a critical need for reliable and accurate evaluation methodologies. Traditional accuracy metrics fail inadequately as they neither capture question characteristics nor offer topic-specific insights. To address this gap, we introduce MEDIRT, a rigorous evaluation framework grounded in Item Response Theory (IRT), the gold standard in high-stakes educational testing. Unlike previous research relying on archival data, we prospectively gathered fresh responses from 80 diverse LLMs on a balanced, 1,100-question USMLE-aligned benchmark. Using one unidimensional two-parameter logistic IRT model per topic, we estimate LLM's latent model ability jointly with question difficulty and discrimination, yielding more stable and nuanced performance rankings than accuracy alone. Notably, we identify distinctive "spiky" ability profiles, where overall rankings can be misleading due to highly specialized model abilities. While `GPT-5` was the top performer in a majority of domains (8 of 11), it was outperformed in Social Science and Communication by `Claude-3-opus`, demonstrating that even an overall 23rd-ranked model can hold the top spot for specific competencies. Furthermore, we demonstrate IRT's utility in auditing benchmarks by identifying flawed questions. We synthesize these findings into a practical decision-support framework that integrates our multi-factor competency profiles with operational metrics. This work establishes a robust, psychometrically grounded methodology essential for the safe, effective, and trustworthy deployment of LLMs in healthcare.

1 Introduction

Large Language Models (LLMs) have demonstrated remarkable capabilities in high-stakes domains like medicine, achieving expert-level performance on standardized medical licensing exams [Singhal et al., 2023, Nori et al., 2023]. This progress has been tracked by a growing ecosystem of benchmarks for different levels of expertise, such as MedQA [Jin et al., 2021] and MedMCQA [Pal et al., 2022], and MedXpertQA [Zuo et al., 2025]. However, while existing benchmarks [Shah et al., 2025] are valuable, the prevailing paradigm for evaluating these models is built on a foundation of aggregate accuracy, a metric that masks critical nuances and faces two fundamental challenges threatening its long-term viability.

First, the reliance on overall accuracy assumes all questions are equally informative, yet test items inherently vary in quality, difficulty, and diagnostic power. This approach fails to distinguish between

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: The Second Workshop on GenAI for Health: Potential, Trust, and Policy Compliance.

models that master difficult concepts and those that merely solve easier problems. Second, this item-level oversight creates a system-level crisis: without a unified scale, scores are incomparable across different benchmarks. Consequently, researchers are forced into an unsustainable practice of reporting an ever-expanding list of benchmark scores, making comprehensive evaluation increasingly cumbersome and difficult to synthesize. These limitations highlight a pressing need for a more principled evaluation framework.

To overcome these blind spots, recent work has increasingly turned to Item Response Theory (IRT), the gold standard for high-stakes educational and psychological testing [Lord and Novick, 1968]. Unlike simple accuracy, IRT provides a sophisticated statistical framework to model the interaction between a subject's latent ability and the properties of individual test items (i.e., questions). By doing so, it can co-estimate a model's underlying capability and characterize each question's difficulty and discriminability on a unified, continuous scale. However, existing applications of IRT in machine learning often diverge from its core purpose or contain critical methodological flaws. For instance, early work utilized IRT for dataset curation rather than to measure model capabilities [Lalor et al., 2016], while other influential studies built IRT models on archival, outdated data [Rodriguez et al., 2021, Polo et al., 2024]. More recent applications have misapplied psychometric models to improve efficiency, such as wrongly using unidimensional IRT to assess abilities across multiple distinct domains [Zhuang et al., 2023, Zhou et al., 2025]. Even sophisticated approaches, such as using Multidimensional IRT (MIRT) for LLM routing [Song et al., 2025], can suffer from limited sample sizes and models with weakly interpretable, data-driven latent dimensions. This leaves a critical gap for healthcare GenAI (Generative AI): the lack of prospective, domain-specific evaluation built from the ground up to provide deep, diagnostic measurement of clinical knowledge in LLMs.

In this work, we apply IRT to conduct a rigorous, topic-level evaluation of diverse LLMs, moving beyond simple accuracy to a psychometrically sound assessment of their capabilities. We introduce MEDIRT, a framework to prospectively apply IRT for measuring the latent medical knowledge of LLMs. Our approach involves creating a balanced 1,100-item benchmark aligned with the USMLE content specifications [Federation of State Medical Boards (FSMB) and National Board of Medical Examiners (NBME), 2025] and systematically evaluating 80 different LLMs. Unlike previous research that relies on archival data, we collected fresh response data using a standardized API protocol via OpenRouter [OpenRouter, 2025]. This methodology ensures true comparability across both proprietary and open-source models while simultaneously capturing operational data on cost and latency. The resulting framework yields not only robust estimates of overall ability but also fine-grained, topic-specific specialty profiles across 11 medical topics, making the findings directly interpretable for applications in medical education, licensing, and clinical decision support. Our work makes four principal contributions:

- We establish a psychometrically rigorous evaluation methodology grounded in IRT that jointly estimates model abilities and question characteristics, separating the model's true ability from item difficulty and discrimination properties. This approach provides more stable and interpretable performance rankings than traditional accuracy-based measures.
- We developed a pipeline to gather fresh response data from 80 diverse models on a balanced 1,100-question medical benchmark, avoiding the limitations of outdated archival datasets.
- Our large-scale analysis reveals a previously unrecognized complex and heterogeneous landscape of abilities from fine-grained, multi-factor competency profiles, demonstrating that even top-performing models exhibit domain-specific weaknesses.
- We synthesize our analytical findings into a practical decision support framework, delivered as an interactive platform[1], that integrates our competency profiles with operational metrics. This provides practitioners with a clear, evidence-based pathway to ensure safe, effective, and trustworthy deployment for specific medical applications.

2 MEDIRT Framework

As shown in Figure 1, MEDIRT is a methodological framework designed to move beyond conventional accuracy and provide topic-wise, item-adjusted evaluations of LLMs in medical domains. The framework consists of three integrated stages: first, we construct a balanced benchmark dataset

[1]An interactive leaderboard with the benchmark is available at `https://huggingface.co/spaces/Pitt-iRiS-Lab/MedIRT`.

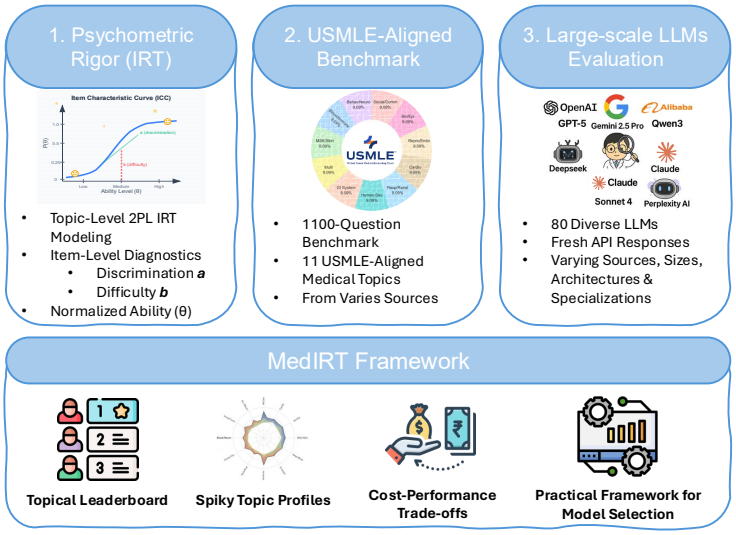


Figure 1: **An Overview of MEDIRT Framework**, illustrating the three critical phases including (1) Topic-level 2PL IRT modeling, (2) USMLE topic-aligned benchmark, (3) Large-scale LLMs cohort.

aligned to the USMLE content outline to ensure content validity and mitigate measurement bias; second, we execute a standardized LLM inference protocol that systematically collects both response data and operational metrics; and third, we perform psychometric evaluation by fitting 11 independent unidimensional two-parameter logistic (2PL) IRT models—one per medical topic—to obtain item parameters and model ability estimates that are inherently adjusted for item difficulty and discrimination; Together, these components yield robust, topic-specific ability estimates that support principled, cost-aware model selection in medical contexts.

Table 1: **Distribution of Medical Topics**, according to the *USMLE Step 1 Test Content Specifications* in Original Dataset, labeled by `GPT-oss-120b` model. Abbr. refers to abbreviations of each topic used in later sections.

Medical Topic	**Abbr.**	**#Samples (%)**	**#Final Sample**
Musculoskeletal, Skin & Subcutaneous Tissue	MSK/Skin	1,693 (21.4%)	100
Multisystem Processes & Disorders	Multi	1,012 (12.8%)	100
Reproductive & Endocrine Systems	Repro/Endo	926 (11.7%)	100
Behavioral Health & Nervous Systems/Special Senses	Behav/Neuro	849 (10.7%)	100
Blood & Lymphoreticular/Immune Systems	Blood/Immune	743 (9.4%)	100
Human Development	Dev	658 (8.3%)	100
Cardiovascular System	Cardio	606 (7.7%)	100
Respiratory & Renal/Urinary Systems	Resp/Renal	545 (6.9%)	100
Gastrointestinal System	GI	471 (6.0%)	100
Biostatistics & Epidemiology/Population Health	Bio/Epi	294 (3.7%)	100
Social Sciences: Communication and Interpersonal Skills	Comm	109 (1.4%)	100
Total		**7,906**	**1,100**

2.1 The Medical Knowledge Benchmark with USMLE-Aligned Topics

We constructed a topic-balanced evaluation dataset by integrating questions from three established medical benchmarks: MedQA Test (1,273 questions), MedMCQA Dev (4,183 questions), and MedXpertQA Test (2,450 questions), totaling 7,906 questions. To ensure topic-specific evaluation aligned with medical education standards, we employed `GPT-oss-120b` [Agarwal et al., 2025] to automatically classify all questions according to USMLE Step 1 Test Content Specifications. We used `GPT-oss-120b` for labeling given its strong MedQA performance [Vals AI, 2025] at lower cost than proprietary models, outperforming `MedGemma-27b` [Sellergren et al., 2025]. Verified by one domain expert, the topic labels achieve a very low error rate of 1%, indicating the quality of the automatic labeling for our dataset. Table 1 presents the topic distribution in our original dataset before balanced sampling. The classification yielded 11 primary medical topics with substantial

representation variations, ranging from 1,693 questions in Musculoskeletal, Skin & Subcutaneous Tissue (21.4%) to 109 questions in Social Sciences: Communication and Interpersonal Skills (1.4%). To create a balanced evaluation set, we implemented stratified sampling to select 100 questions per topic, resulting in a final dataset of 1,100 questions. The composition of 544 questions from MedMCQA, 382 questions from MedXpertQA, and 174 questions from MedQA, shows the final dataset composition across source benchmarks. This balanced approach ensures equal representation across medical specialties while maintaining sufficient sample sizes for reliable IRT parameter estimation.

2.2 Large Language Model Cohort

The study evaluated a cohort of Large Language Models (LLMs), selected to represent the breadth and diversity of the current GenAI landscape. 80 LLMs were randomly sampled from the OpenRouter API [OpenRouter, 2025] to represent the current landscape of proprietary and open-source systems. This cohort includes models that vary significantly across several key dimensions: **Size and Architecture** ranges from efficient 3B-parameter models (`Llama-3.2-3B` [Grattafiori et al., 2024]) to frontier systems exceeding 400B parameters (`Hermes-3-405B` [Teknium et al., 2024]), including both dense architectures and mixture-of-experts (MoE) [Lo et al., 2025] designs; **Access and Origin** encompass leading proprietary APIs (`GPT-5`, `Claude-Sonnet-4` [Anthropic, 2025], `Gemini-2.5-Pro` [Gemini Team, Google, 2025]) and prominent open-source alternatives (`Llama-4-Maverick` [Meta AI / Hugging Face, 2025], `Qwen-3-30B` [Yang et al., 2025], `DeepSeek-V3` [Liu et al., 2024]); **Specializations** include base models, instruction-tuned variants, and reasoning-optimized systems, with most being general-purpose but some specialized for domains like code generation (`Codex-Mini` [OpenAI, 2025]) to assess skill transferability. Detailed model statistics for all 80 models are presented in the Appendix Table S3.

To ensure reliable evaluation, we applied strict selection criteria: (1) API accessibility and standardized interfaces, (2) operational reliability (error rates $< 5\%$), (3) reasonable inference times (< 120 seconds per query), and (4) consistent availability during the evaluation period. After applying these criteria, 80 models successfully completed the full evaluation protocol and were included in our final analysis. All models were evaluated under identical inference parameters to ensure fair comparison: **Temperature** 0 for deterministic sampling; **Maximum tokens** 3000 to accommodate single-letter responses with a safety margin for reasoning; **Reasoning mode** was set to low to minimize computational overhead; **Retry** maximum three attempts per question to handle transient API failures. This standardized configuration ensures that observed performance differences reflect genuine model capabilities rather than parameter optimization artifacts. Simultaneously, our operational telemetry captures authentic computational costs associated with each model's inference process, providing decision-makers with complete information for deployment planning.

2.3 Topic-Level 2PL IRT Modeling

We treat each of the 11 USMLE topics as separate areas of knowledge and build one model per topic using the two-parameter logistic (2PL) framework. Instead of fitting a single, complex multidimensional model, this design provides stable estimates and clear, topic-level profiles of performance.

2.3.1 Why Topic-Level Models?

Medical training evaluates knowledge domain by domain (e.g., Cardiology, Pharmacology), and our framework mirrors this structure [Downing, 2003, Holmboe et al., 2018]. To operationalize this, we modeled each of the 11 USMLE-aligned topics independently using unidimensional 2PL models, producing ability scores $\theta_{m,t}$ that are directly interpretable for diagnosing model strengths and weaknesses in specific subject areas.

Our choice of topic-level unidimensional models provided the optimal balance of statistical rigor and practical utility for this study. While a single multidimensional model is theoretically appealing, attempts to fit one to our dataset (N=80 models, 1,100 items) resulted in estimation instability, a known challenge with highly correlated dimensions and potential local item dependencies [Reckase, 1997, Yen, 1984]. Conversely, forcing a simple unidimensional model across all topics risks creating biased, composite estimates of ability [Reckase, 1979]. By modeling each of the 11 medical domains

independently, we achieved stable, reliable results and generated the interpretable, topic-specific diagnostics essential for high-stakes evaluation.

2.3.2 IRT Model Specification

For each medical topic t, the probability that model m answers item i correctly is defined under the two-parameter logistic (2PL) framework as:

$$\Pr(X_{imt} = 1 \mid \theta_{m,t}) = \sigma\big(a_{i,t}(\theta_{m,t} - b_{i,t})\big), \qquad \sigma(x) = \tfrac{1}{1+e^{-x}}.$$

In this formulation, $\theta_{m,t}$ represents the latent ability of model m on topic t, $b_{i,t}$ captures the item's location on the ability scale (i.e., its difficulty), and $a_{i,t}$ indicates how sharply the item discriminates between stronger and weaker models. This framework, widely used in educational and psychological measurement [Birnbaum, 1968], provides a principled way to separate model proficiency from item characteristics. Abilities are estimated on a standardized scale (mean 0, SD 1) using marginal maximum likelihood procedures, ensuring comparability across topics and models.

To ensure measurement consistency, we evaluated the internal reliability of each of the 11 topic-specific scales. All showed high reliability (marginal reliability > 0.93; see Appendix Table S1), meeting accepted thresholds for high-stakes testing [Downing, 2003]. For interpretability, we also report a composite measure by averaging standardized topic scores, $\hat{\Theta}_m = \frac{1}{11}\sum_{t=1}^{11} Z_{m,t}$, where $Z_{m,t}$ is the standardized ability estimate for topic t. This unweighted aggregation reflects the balanced design of our benchmark, assigning equal weight to each medical domain in accordance with USMLE content specifications, and facilitates meaningful comparisons of overall performance across models.

3 Results

3.1 Experiment Settings

Our evaluation is grounded in a standardized protocol to ensure that all models are assessed under identical conditions, guaranteeing the reproducibility and fairness of our results. This comprehensive evaluation environment ensures direct comparability across all evaluated models and enables statistically valid performance assessments.

For prompt design, a zero-shot, multiple-choice question (MCQ) format was utilized in the study. The prompt frames the task as a "closed-book multiple-choice medical exam" and strictly instructs the model to return **only one letter** corresponding to the single best option, explicitly forbidding any supplementary words or explanations. The exact prompt templates used in the study are shown in Section 2 of the Appendix. We extracted final answers from the raw text using an automated parsing script that requires strict adherence to the specified output format, treating any deviation as an instruction-following failure.

3.1.1 Evaluation Metrics

Our framework employs three complementary metric classes that provide comprehensive model assessment:

Accuracy-based metrics provide baseline comparisons with existing benchmarks, calculated as the percentage of correct answers.

Psychometric metrics, our primary evaluation metric, use *IRT* to estimate each model's latent ability (θ) across 11 medical specialties. Unlike overall accuracy, IRT produces difficulty-adjusted scores on a standardized scale (mean $= 0$, SD $= 1$), where positive values indicate above-average proficiency. This yields interpretable "ability profiles" that reveal topic-specific strengths and weaknesses, offering a robust foundation for high-stakes evaluation.

Operational metrics capture deployment-relevant telemetry such as API costs and inference latency, supporting multi-objective optimization that balances capability, efficiency, and economic constraints.

Table 2: **Dual-Ranking Model Performance Leaderboard.** Top-15 models ranked by their mean IRT ability (θ), the mean of the 11 topic-specific ability scores. Acc refer to the mean accuracy for all topics. Time and Cost refer to the mean response time per question (seconds) and total evaluation cost (USD), respectively. Rows where the two ranks differ are bolded in the rank cells to indicate a "rank-flip."

Model	Rank (θ)	Rank (Acc)	Mean Ability ↑	Mean Acc(%) ↑	Mean Time (s) ↓	Total Cost ($) ↓
openai/gpt-5	1	1	2.394	74.4	26.68	2.88
google/gemini-2.5-pro	2	2	1.925	68.8	49.26	5.21
openai/codex-mini	3	3	1.873	66.9	24.89	2.58
openai/gpt-oss-120b	4	4	1.826	63.4	2.80	0.10
openai/gpt-5-nano	5	5	1.356	61.2	31.20	0.11
openai/gpt-4o	**6**	**7**	1.263	58.6	4.70	0.67
x-ai/grok-3-mini	**7**	**6**	1.258	60.7	3.22	0.42
anthropic/claude-sonnet-4	**8**	**9**	1.241	58.0	6.24	1.62
meta-llama/llama-4-maverick	**9**	**8**	1.155	58.3	2.06	0.04
anthropic/claude-3.7-sonnet	**10**	**11**	1.117	57.6	10.21	1.01
google/gemini-2.5-flash	**11**	**12**	1.086	57.4	6.97	0.08
moonshotai/kimi-k2	**12**	**10**	1.086	57.6	4.40	0.05
openai/gpt-4.1-mini	13	13	1.050	56.9	25.87	0.11
qwen/qwen3-30b-a3b	**14**	**15**	0.916	56.4	1.06	0.10
openai/gpt-oss-20b	**15**	**14**	0.914	56.7	1.98	0.08

3.2 Overall Performance

Eighty LLMs were evaluated on the benchmark, producing two distinct performance rankings: one based on conventional accuracy and the other on psychometrically grounded IRT ability estimates. Complete results for all 80 models are reported in Appendix Table S4. Table 2 presents the top 15 performing models, primarily ranked by their mean composite IRT ability, with their corresponding accuracy rank shown for direct comparison. The results reveal a consistent pattern: among the highest-performing models, accuracy and IRT-based ability estimates were in close agreement. The top five models, led by `openai/gpt-5`, maintain identical rank orderings across both metrics, establishing a clear top tier of performance. Beyond this tier, however, rank shifts begin to appear, reflecting meaningful differences between accuracy and ability. For example, `openai/gpt-4o` improves from seventh place in accuracy to sixth in ability, suggesting better performance on more difficult questions. In contrast, `moonshotai/kimi-k2` declines from tenth in accuracy to twelfth in ability, indicating that its observed accuracy may be disproportionately driven by success on easier questions.

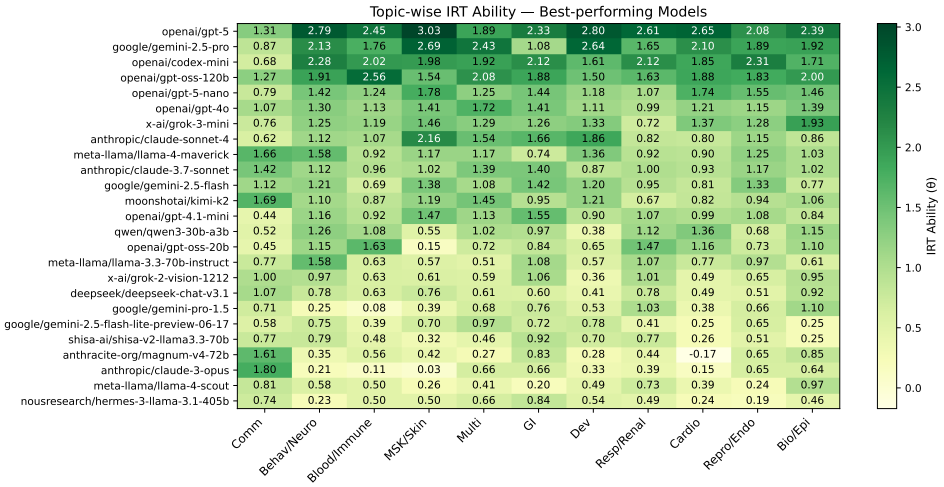
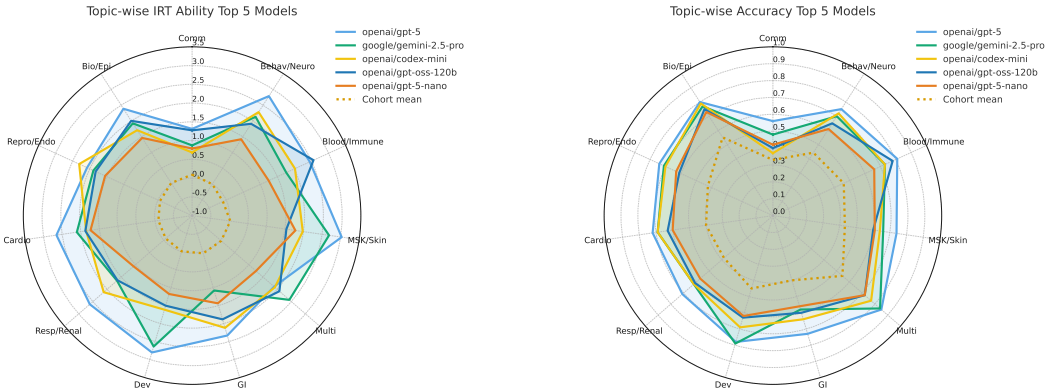


Figure 2: **Heatmap of topic-wise IRT Ability (θ) for the top 25 Models.** Rows list models with the highest *mean* ability across topics (sorted descending); columns are topic abbreviations, as shown in Table 1.

4 Analysis

4.1 Topic-Specific Competency Profiles

Our topic-level decomposition reveals that even top-performing models display highly uneven specialty profiles, as shown in Figure 2. No single model demonstrates mastery across all medical specialties. While `GPT-5` demonstrates exceptional ability in most areas (e.g., Respiratory & Renal, Musculoskeletal, with ability $\theta > 2.3$), other models lead in specific domains. `Gemini-2.5-pro`, for instance, is the strongest in Multisystem Processes & Disorders ($\theta = 2.43$), and even the overall 23rd-ranked `Claude-3-opus` outperformed all others in Social Science and Communication ($\theta = 1.80$). These distinct ability "fingerprints" highlight critical trade-offs for deployment: while `GPT-5` presents as a strong generalist, other leading models exhibit pronounced variations in proficiency across specialties. Importantly, this granular view also uncovers systemic weaknesses across the entire 80-model cohort. Even after adjusting for item difficulty, domains such as Multisystem Processes and Communication/Interpersonal Skills remain universally challenging, marking key priorities for future development in clinical LLMs. The complete topic-wise estimations are provided in Appendix Table S5.



(a) Radar chart of topic-wise IRT Ability (θ).

(b) Radar chart of topic-wise accuracy.

Figure 3: **Radar charts of topic-wise IRT Ability (θ) and accuracy for the top five Models.** Topic abbreviations are the same as in Table 1.

The radar visualizations in Figure 3 highlight why our psychometric metric (θ) provides a more veridical account of a model's capabilities than overall accuracy. Specifically, accuracy is an inherently confounded metric: a high score on one topic may be an artifact of easy questions rather than evidence of superior ability, as shown in Panel b). In contrast, our IRT-derived ability scores are difficulty-adjusted, yielding a much cleaner signal of a model's latent capacity, as shown in Panel a). Furthermore, this perspective enables a more strategic, task-oriented model selection.

4.2 Item Parameter Landscape

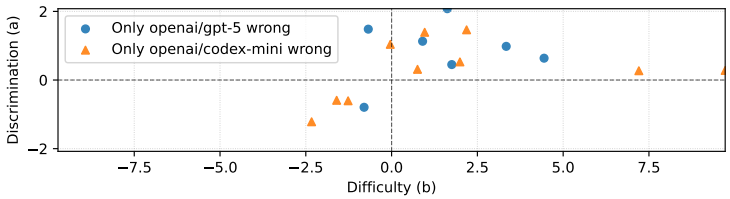


Figure 4: **Item missed by model for the for the Reproductive/Endocrine topic in IRT space** (difficulty b vs. discrimination a). Each point is a missed item. Circles = items only `GPT-5` missed; triangles = items only `Codex-mini` missed. Dashed lines mark $b = 0$ (vertical) and $a = 0$ (horizontal).

To illustrate the added value of IRT-derived ability estimates over raw accuracy, Figure 4 contrasts `GPT-5` and `Codex-mini` on Reproductive & Endocrine Systems. Although their raw accuracies are

similar (GPT-5 even a little bit higher), IRT analysis reveals distinct error profiles: GPT-5 misses higher-discrimination items, which are especially informative for distinguishing between high- and low-ability systems. In contrast, Codex-mini struggles with lower-discrimination ones, resulting in higher ability compared to GPT-5. This case study underscores how IRT ability estimates provide a more diagnostic and interpretable view of model competency than aggregate accuracy alone.

Shifting the focus from model performance to the properties of the test questions enables an audit of the benchmark itself. The relatively high average discrimination (mean $a = 1.96$) indicates that the benchmark is well suited for making fine-grained distinctions among top-performing systems. Detailed statistics of item parameters are provided in Appendix Table S2.

However, our IRT analysis also identified a distinct subset of questions ($N = 15$) with negative discrimination ($a < 0$) and positive difficulty ($b > 0$), indicating a paradoxical situation in which higher-ability models were more likely to respond incorrectly to these items compared to lower-ability models. A clear example is one question concerning "rapid prototyping," which yielded a discrimination of $a = -0.342$ and difficulty of $b = 0.77$, was consistently missed by top models such as GPT-5. A closer inspection revealed a design pathology: its "All of the above" answer format was compromised by imprecise distractors, rendering the intended answer unsound. This case illustrates how IRT can serve as a powerful tool for quality control, flagging poorly constructed or miscategorized items that may distort the evaluation.

Table 3: **Top Cost- and Time-Normalized Performance (Ability).** Models are ranked by θ/\$ (mean IRT ability per total cost). C and T refer to total cost (USD) and mean time (seconds), respectively. The other metric is Ability/s (mean θ per second), computed using mean response time (s/question). Ranks are reported for θ and θ/s among all 80 LLMs (larger is better).

Model	$\theta \uparrow$	**C (\$)** $\downarrow$	**T (s)** $\downarrow$	$\theta/\$ \uparrow$	**θ/s** $\uparrow$	**Rank θ**	**Rank θ/s**
meta-llama/llama-3.3-70b-instruct	0.829	0.01	1.99	82.905	0.417	16	5
meta-llama/llama-4-maverick	1.155	0.04	2.05	28.871	0.563	9	3
moonshotai/kimi-k2	1.086	0.05	4.40	21.728	0.247	12	6
openai/gpt-oss-120b	1.826	0.10	2.81	18.261	0.650	4	2
google/gemini-2.5-flash	1.086	0.08	6.97	13.580	0.156	11	7
openai/gpt-5-nano	1.356	0.11	31.20	12.329	0.043	5	9
deepseek/deepseek-chat-v3.1	0.688	0.05	6.63	13.760	0.104	18	8
openai/gpt-oss-20b	0.914	0.08	1.98	11.430	0.462	15	4
openai/gpt-4.1-mini	1.050	0.11	25.87	9.550	0.041	13	10
qwen/qwen3-30b-a3b	0.916	0.10	1.06	9.165	0.865	14	1

4.3 Cost-Performance Trade-offs

Economic considerations play crucial roles in practical LLM deployment decisions. We analyze this by examining the relationship between model capability (mean IRT ability) and the associated financial cost and inference latency of the evaluation. Top cost- and time-normalized performance (based on mean IRT ability) is shown in Table 3. Practical deployment considerations reveal a clear Pareto frontier [Jahan et al., 2016] in the cost-performance landscape. Considering the joint objectives of (i) mean IRT ability (θ), (ii) cost-normalized performance ($\theta/\$$), and (iii) time-normalized performance (θ/s), the non-dominated set comprises four models: GPT-oss-120b, Llama-4-maverick, Llama-3.3-70b-instruct, and Qwen3-30b-a3b. Each occupies a distinct operating point on the efficiency–performance landscape: GPT-oss-120b attains the highest absolute ability (strongest θ); Llama-3.3-70b-instruct is the cost leader (maximizing $\theta/\$$); Qwen3-30b-a3b is the latency/throughput leader (maximizing θ/s); and Llama-4-maverick provides a balanced trade-off—high θ with favorable $\theta/\$$ and θ/s.

5 Discussion

5.1 A Diagnostic Framework for Evaluating Model and Benchmark Integrity

This work demonstrates that moving beyond simple accuracy to a psychometric approach like IRT provides a "magnifying glass" for LLM evaluation. As shown in Figure 4, IRT reveals that GPT-5 and Codex-mini fail on different item types despite similar accuracies, highlighting contrasts in

their error profiles. Crucially, an IRT-based evaluation also provides a unified scale for all questions, making it possible to merge multiple benchmarks onto a single scoring scale in the future. IRT offers a more stable and interpretable measure of a model's latent knowledge by disentangling its capability from the characteristics of the test questions. This transforms evaluation from a simple leaderboard ranking into a rich diagnostic exercise, which is critical in high-stakes domains where understanding the reason for a model's failure is paramount.

A key application of this framework, discovered in Section 4.2, is a dual-probe diagnostic methodology, which leverages items with strong negative discrimination (where high-ability models fail more often than weaker ones). This method mandates a two-step workflow. First, the item is treated as a probe for benchmark integrity, prompting domain experts to validate it for flaws like ambiguity or factual errors. If the item is validated as sound, its function shifts to a probe for model integrity. In this role, it can expose subtle, systematic reasoning fallacies in high-ability models, such as a tendency to overthink simple questions. This process enhances evaluation rigor by first validating the measurement instrument itself before using it to identify otherwise undetectable model failures.

5.2 Implications for Deploying Trustworthy AI in Healthcare

Our findings demonstrate that trustworthiness in a medical LLM cannot be a monolithic concept. The highly variable, "spiky" performance profiles (Figure 3) mean that a model's reliability is context- and domain-dependent. A model trustworthy for cardiology might be dangerously unreliable for patient communication. For AI to be deployed safely, its performance envelope must be precisely defined at a granular level, moving beyond aggregate scores that can mask critical weaknesses.

Based on this, we synthesize our findings into a practical, multi-criteria framework for selecting an LLM for a medical application. Practitioners should: (a) **Assess global performance** by examining the ability–accuracy scatter plot in Table 2; (b) **Ensure domain alignment** by inspecting the model's topic-specific radar chart in Section 4.1; (c) **Evaluate cost-performance trade-offs** to identify models that provide "good enough" capability at a sustainable cost follow Section 4.3; (d) **Confirm competence** by validating the chosen model on a small panel of high-discrimination questions relevant to the specific use case. This structured approach allows healthcare professionals to select models based on demonstrated competencies in relevant domains, mitigating the risk of deploying a model that fails unexpectedly in a high-stakes clinical situation.

5.3 Limitations

This study has several limitations. First, its reliance on static multiple-choice medical questions does not fully capture the dynamic nature of clinical reasoning. Second, our evaluation was shaped by practical constraints. Our reliance on the OpenRouter API excluded specialized medical LLMs from our analysis, and the token and time limits we imposed for cost-efficiency may have compromised the full potential of high-capacity reasoning models. Third, the domain-specific analysis may not reflect the interdisciplinary integration required in actual medical practice.

6 Conclusion

In this paper, we present a large-scale, fine-grained evaluation of 80 LLMs across 11 distinct factors of medical competence. Our analysis moves beyond monolithic performance metrics, uncovering a complex and heterogeneous landscape of model abilities. We find that even top-performing models exhibit notable, domain-specific weaknesses, and that essential skills related to communication and quantitative reasoning are systematically underdeveloped across the current generation of LLMs. These results deliver a clear and urgent message: safe, effective, and trustworthy deployment of generative AI in health requires moving beyond the illusion of competence created by single-score benchmarks. Adopting detailed, multi-factor competency profiles is a critical step toward accurately characterizing the true capabilities of these LLMs. Also, a dual-probe diagnostic workflow that uses negative-discrimination questions has been discovered to enhance benchmark integrity while exposing subtle model reasoning failures. Moreover, we synthesize our findings into a practical, multi-criteria framework for practitioners, integrating our competency profiles with cost-performance analysis to guide evidence-based model selection. Together, these contributions establish a foundation for more principled evaluation and informed deployment of medical LLMs.

References

Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K Arora, Yu Bai, Bowen Baker, Haiming Bao, et al. gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925*, 2025.

Anthropic. Introducing claude opus 4 and claude sonnet 4. Website blog post, May 2025. URL `https://www.anthropic.com/news/claude-4`. Claude Sonnet 4: a high-performance AI model for coding and reasoning. Retrieved from Anthropic's announcement.

Allan Birnbaum. Some latent trait models and their use in inferring an examinee's ability. *Statistical theories of mental test scores*, 1968.

Steven M Downing. Validity: on the meaningful interpretation of assessment data. *Medical Education*, 37(9):830–837, 2003. doi: 10.1046/j.1365-2923.2003.01594.x.

Federation of State Medical Boards (FSMB) and National Board of Medical Examiners (NBME). USMLE content outline. `https://www.usmle.org/sites/default/files/2022-01/USMLE_Content_Outline_0.pdf`, 2025. Accessed: 2025-08-26.

Gemini Team, Google. Gemini 2.x: Pushing the frontier with advanced reasoning, multimodality, long context, and next-generation agentic capabilities. Technical report, Google DeepMind, June 2025. Technical report introducing Gemini 2.5 Pro and Gemini 2.5 Flash; includes benchmark results and architectural details.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. URL `https://arxiv.org/abs/2407.21783`.

Eric S. Holmboe, Jose Biller, Elizabeth G. Donahue, Paul George, Francisco J. Longo, Sarah Scheiber, Elizabeth Sinz, and NEJM Knowledge+ Team. Redesigning continuing medical education: The case for competency-based education. *Academic Medicine*, 93(10):1461–1464, 2018. doi: 10.1097/ACM.0000000000002324.

Ali Jahan, Kevin L Edwards, and Marjan Bahraminasab. *Multi-criteria decision analysis for supporting the selection of engineering materials in product design*. Butterworth-Heinemann, 2016.

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.

John P Lalor, Hao Wu, and Hong Yu. Building an evaluation scale using item response theory. In *Proceedings of the conference on empirical methods in natural language processing. Conference on empirical methods in natural language processing*, volume 2016, page 648, 2016.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.

Ka Man Lo, Zeyu Huang, Zihan Qiu, Zili Wang, and Jie Fu. A closer look into mixture-of-experts in large language models. In Luis Chiruzzo, Alan Ritter, and Lu Wang, editors, *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 4427–4447, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-195-7. doi: 10.18653/v1/2025.findings-naacl.251. URL `https://aclanthology.org/2025.findings-naacl.251/`.

Frederic M Lord and Melvin R Novick. *Statistical theories of mental test scores*. Addison-Wesley, Reading, MA, 1968.

Meta AI / Hugging Face. Llama 4 maverick (17b-active, 128-expert mixture-of-experts, multimodal). Model card on Hugging Face, April 2025. URL `https://huggingface.co/meta-llama/Llama-4-Maverick-17B-128E-Original`. Multimodal Mixture-of-Experts model: 17 B active parameters out of 400 B total, released April 5, 2025 under the Llama 4 Community License :contentReference[oaicite:1]index=1.

Harsha Nori, Yin Tat Lee, Sheng Zhang, Dean Carignan, Richard Edgar, Nicolo Fusi, Nicholas King, Jonathan Larson, Yuanzhi Li, Weishung Liu, Renqian Luo, Scott Mayer McKinney, Robert Osazuwa Ness, Hoifung Poon, Tao Qin, Naoto Usuyama, Chris White, and Eric Horvitz. Can generalist foundation models outcompete special-purpose tuning? case study in medicine. *arXiv preprint arXiv:2311.16452*, 2023.

OpenAI. Codex-mini (codex-mini-latest): a fine-tuned version of o4-mini for cli use. Model release announcement & documentation, May 2025. URL `https://openai.com/index/introducing-codex/`. Codex-Mini is a compact, low-latency coding model (fine-tuned from o4-mini) released May 16, 2025 via Codex CLI and Responses API; supports long contexts (200k tokens), multilingual and vision inputs, but does not support tool calling.

OpenRouter. OpenRouter: Universal API for large language models. `https://openrouter.ai/`, 2025. Accessed: 2025-08-27.

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

Felipe Maia Polo, Lucas Weber, Leshem Choshen, Yuekai Sun, Gongjun Xu, and Mikhail Yurochkin. tinybenchmarks: evaluating llms with fewer examples. *arXiv preprint arXiv:2402.14992*, 2024.

Mark D Reckase. Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of educational statistics*, 4(3):207–230, 1979.

Mark D Reckase. The past and future of multidimensional item response theory. volume 21, pages 25–36. SAGE PUBLICATIONS, INC. 2455 Teller Road, Thousand Oaks, CA 91320, 1997.

Pedro Rodriguez, Joe Barrow, Alexander Miserlis Hoyle, John P Lalor, Robin Jia, and Jordan Boyd-Graber. Evaluation examples are not equally informative: How should that change nlp leaderboards? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4486–4503, 2021.

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.

Nigam Shah, Mike Pfeffer, and Percy Liang. Holistic evaluation of large language models for medical applications, 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.

Wei Song, Zhenya Huang, Cheng Cheng, Weibo Gao, Bihan Xu, GuanHao Zhao, Fei Wang, and Runze Wu. Irt-router: Effective and interpretable multi-llm routing via item response theory. *arXiv preprint arXiv:2506.01048*, 2025.

Ryan Teknium, Jeffrey Quesnelle, and Chen Guang. Hermes 3 technical report. *arXiv preprint arXiv:2408.11857*, 2024.

Vals AI. Medqa benchmark leaderboard (august 26, 2025). `https://www.vals.ai/benchmarks/medqa-08-26-2025`, 2025. Accessed: September 4, 2025.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Wendy M Yen. Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement*, 8(2):125–145, 1984.

Hongli Zhou, Hui Huang, Ziqing Zhao, Lvyuan Han, Huicheng Wang, Kehai Chen, Muyun Yang, Wei Bao, Jian Dong, Bing Xu, et al. Lost in benchmarks? rethinking large language model benchmarking with item response theory. *arXiv preprint arXiv:2505.15055*, 2025.

Yan Zhuang, Qi Liu, Yuting Ning, Weizhe Huang, Rui Lv, Zhenya Huang, Guanhao Zhao, Zheng Zhang, Qingyang Mao, Shijin Wang, et al. Efficiently measuring the cognitive ability of llms: An adaptive testing perspective. 2023.

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. *arXiv preprint arXiv:2501.18362*, 2025.