

ReMedQA: Are We Done With Medical Multiple-Choice Benchmarks?

Alessio Cocchieri* Luca Ragazzi* Giuseppe Tagliavini Gianluca Moro*

{a.cocchieri, l.ragazzi, giuseppe.tagliavini, gianluca.moro}@unibo.it

Department of Computer Science and Engineering,
University of Bologna, Italy

Abstract

Medical multiple-choice question answering (MCQA) benchmarks show that models achieve near-human accuracy, with some benchmarks approaching saturation—leading to claims of clinical readiness. Yet a single accuracy score is a poor proxy for competence: models that change answers under minor input perturbations cannot be considered reliable. We argue that reliability underpins accuracy—only consistent predictions make correctness meaningful. We release **ReMedQA**, a new benchmark that augments three standard medical MCQA datasets with open-ended answers and systematically perturbed options. Building on this, we introduce **ReAcc** and **ReCon**, two reliability metrics: ReAcc measures the proportion of questions answered correctly across all variations, while ReCon measures the proportion answered consistently regardless of correctness. Our evaluation shows that high MCQA accuracy masks low reliability: models remain sensitive to format and perturbation changes, and domain specialization offers no robustness gain. MCQA underestimates smaller models while inflating large ones that exploit structural cues—with some exceeding 50% accuracy even when the original questions are hidden. This shows that, despite near-saturated accuracy, we are not yet done with medical MCQA benchmarks.¹

1 Introduction

Multiple-choice question answering (MCQA) is the dominant paradigm for assessing medical knowledge in large language models (LLMs) (Singhal et al., 2022; Nori et al., 2023a; Christophe et al., 2024b; Labrak et al., 2024; Sellergren et al., 2025), and is widely used to evaluate retrieval-augmented generation systems in healthcare (Frisoni et al., 2024; Xiong et al., 2024). While MCQA perfor-

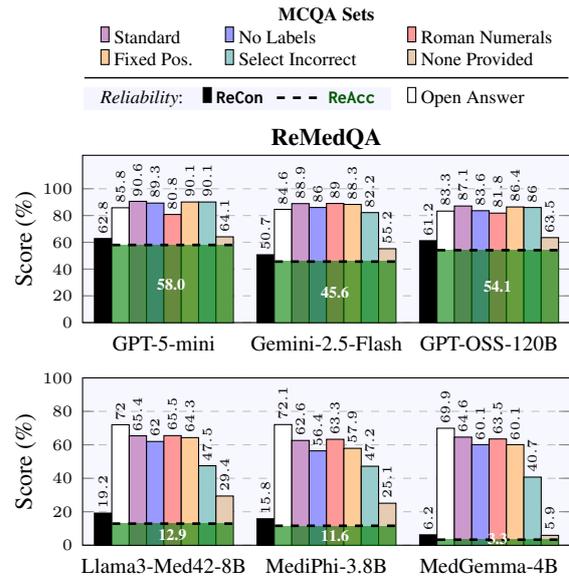


Figure 1: **LLM evaluation on ReMedQA**. Colored: MCQA accuracy; white: open-answer accuracy. **ReAcc** (green): % of questions answered *correctly* across all variations; **ReCon** (black): % of questions answered *consistently* (same prediction) across all variations.

mance does not reflect clinical competence, it remains a practical and scalable screening tool for model comparison, and a necessary—though not sufficient—basis for assessing medical knowledge.

Several commonly used MCQA benchmarks show models achieving human-level or even super-human performance, with accuracy approaching saturation (Nori et al., 2023b; Saab et al., 2024; Etzine et al., 2025). This may suggest that clinical reasoning is nearly solved, motivating the release of increasingly difficult benchmarks (Griot et al., 2025a; Zuo et al., 2025) and analyses (Hosseini et al., 2024; Gu et al., 2025; Lamparth et al., 2025). However, a single MCQA accuracy score remains an imperfect proxy for medical competence. First, it can reward superficial test-taking strategies such as distractor elimination or the exploitation of statistical patterns (Bedi et al., 2025; Griot et al., 2025b).

*Equal contribution (co-first authors).

¹<https://disi-unibo-nlp.github.io/remedqa>

Second, the paradigm is misaligned with real-world practice: physicians must reason openly, integrate knowledge, and generate conclusions without being guided by predefined options.

Importantly, accuracy alone, without reliability, offers a misleading picture of competence. Following standard medical terminology, *reliability* refers to the consistency or reproducibility of an assessment result across equivalent cases or formats (Fenton et al., 2017; Ahmed and Ishtiaq, 2021). A reliable physician reaches the same conclusion even when information is presented in slightly different ways. By contrast, a model that produces inconsistent outputs cannot be trusted. Prior work has shown that even minor input perturbations in MCQA (e.g., reordering answer options) can alter model predictions (Zheng et al., 2023; Wang et al., 2024), casting doubt on the significance of accuracy scores reported in existing benchmarks.

To remedy these limitations, we introduce a new benchmark, **ReMedQA** (Reliable Medical QA), which aims to reconceptualize how medical knowledge is evaluated. At its core, ReMedQA treats open-ended answers as the primary signal of competence: what matters is whether a model can articulate correct knowledge without being guided by options, reflecting how clinical practice typically operates. Closed-form MCQA, along with meaning-preserving perturbations, is repurposed as a scalable mechanism to evaluate model reliability in terms of output stability. Building on this design, we introduce **ReAcc** and **ReCon**, two reliability-oriented metrics. ReAcc captures the proportion of questions answered correctly across all variations (i.e., open and closed settings), whereas ReCon measures answer consistency across variations regardless of correctness. By jointly tracking accuracy and consistency, our metrics elevate reliability to a core dimension of evaluation—broadly applicable, yet critical in the medical domain.

We construct ReMedQA by augmenting three clinical benchmarks—MedQA (Jin et al., 2020), MedMCQA (Pal et al., 2022), and the medical subsets of MMLU (Hendrycks et al., 2021)—with open answers and systematically perturbed options, enabling ReAcc and ReCon to be measured at scale.

Our findings reveal that high MCQA accuracy often fails to reflect reliability (see Figure 1). Across all datasets and model families, performance remains highly sensitive to format and perturbation changes. Reliability scales with model size but not with domain specialization: large general-purpose

systems achieve the best balance between ReAcc and ReCon, whereas medical-tuned models offer no robustness gain despite their domain exposure. Interestingly, smaller models are often underestimated by MCQA, performing better in open-ended answering than in multiple-choice formats, showing how traditional benchmarks can misrepresent model capabilities. Finally, additional diagnostics show that frontier models can answer correctly without seeing the question, exploiting statistical patterns in the answer options. Overall, these results challenge the notion that accuracy alone signals understanding, highlighting the need for evaluation frameworks where reliability stands alongside accuracy as a primary metric in medical QA.

To sum up, our contributions are as follows:

- We introduce **ReAcc** and **ReCon**, two complementary reliability metrics capturing accuracy consistency and prediction consistency across input perturbations and formats.
- We release **ReMedQA**, a benchmark that augments standard MCQA medical datasets with open answers and systematic perturbations.
- We present the first large-scale assessment of LLMs’ medical reliability, revealing behaviors overlooked by standard evaluations.

2 Related Work

In MCQA evaluation, numerous works have shown that models are highly sensitive to superficial input changes, often leveraging spurious patterns rather than robust reasoning (Ribeiro et al., 2020; Gardner et al., 2021). Within general-purpose MCQA, recent studies have explored model consistency from different angles. Some works applied perturbations to question structure and distractors (Wang et al., 2025a), while others tested stability under option order (Pezeshkpour and Hruschka, 2023; Li et al., 2024) and symbol variation (Yang et al., 2025). In the medical domain, adversarial stress tests reveal that LLMs can be misled by minor, clinically irrelevant changes (Ness et al., 2024; Vishwanath et al., 2025). Other findings highlight vulnerabilities to format-specific phenomena, including elimination-based reasoning (Balepur et al., 2023), sensitivity to “None of the above” options (Pal et al., 2023; Tam et al., 2025), accuracy fluctuations driven by prompt design or answer extraction strategy (Molfese et al., 2025), and performance degradation in open-ended setups (Singh et al., 2025).

Our Position Together, prior studies expose the fragility of MCQA accuracy: models often collapse under stress tests, but this analysis typically ends there. No work provides a framework for quantifying *reliability*—namely, whether model predictions remain stable across input perturbations and formats. Our work fills this gap in three ways. First, we introduce the notion of reliability as a core dimension that must accompany accuracy, operationalized through two complementary metrics, ReAcc and ReCon. Second, rather than probing models with a few ad-hoc perturbations, we systematically incorporate the most rigorous and broadly studied stress tests from the literature. Third, we apply this framework to the most widely used medical MCQA benchmarks, yielding new insights into the current state of LLM competence and releasing ReMedQA, a novel benchmark designed for large-scale, reliability-oriented clinical evaluation.

3 Methodology

We now describe the design of ReMedQA and how we prioritize reliability in medical MCQA.

3.1 Task Setup and Perturbation Formats

Let $x = (q, o_1, \dots, o_n)$ denote an MCQA instance, consisting of a question q and a set of n candidate options $\{o_1, \dots, o_n\}$, with $n = 4$ in our experiments. To evaluate prediction stability, we assess each model across **7 semantically equivalent, literature-grounded input variants** that differ only in surface form or answer format: the standard MCQA format, 5 MCQA perturbations, and one open-ended reformulation. We define prediction stability *strictly*: a model is considered reliable only if it produces consistent predictions across *all* seven variants. Full prompts and implementation details are reported in Appendix B.

Standard MCQA In the standard setting—corresponding to the original dataset format—the model receives the full MCQA instance x and outputs a predicted option $\hat{o} = f(x) \in \{o_1, \dots, o_n\}$.

MCQA Perturbations Following prior work on robustness in MCQA (Pezeshkpour and Hruschka, 2023; Li et al., 2024; Tam et al., 2025; Wang et al., 2025a; Yang et al., 2025), we construct 5 MCQA variants that preserve the underlying clinical semantics while modifying surface features (Figure 2):

1. **Fixed Position.** We always place the correct option in position D, which is the least fre-

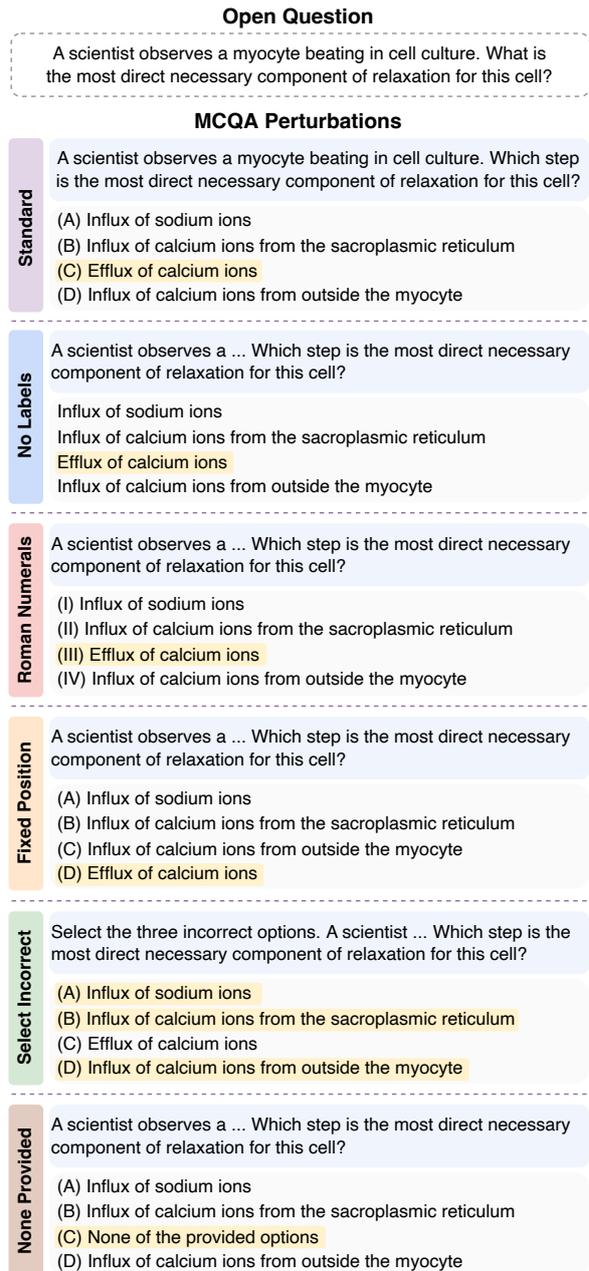


Figure 2: **Semantic-preserving input perturbations in ReMedQA.** Consistent predictions across sets indicate model reliability. The correct answer is highlighted.

quent position of gold options in the datasets (see Appendix C), minimizing positional priors while stressing robustness to option order.

2. **No Labels.** Standard answer labels (A/B/C/D) are removed, forcing models to output the content of the chosen option rather than a label.
3. **Select Incorrect.** Models must choose all incorrect options instead of the correct one, probing if they can identify distractors.
4. **Roman Numerals.** Standard labels are re-

placed with {I, II, III, IV}, testing robustness to an alternative labeling scheme.

5. **None Provided.** The gold option is replaced with the string “None of the provided options”, requiring models to recognize when the correct answer is absent from the listed options.

Open-Ended Reformulation Finally, we consider an open-ended variant in which the model receives only the question q and generates a free-form answer $\hat{a} = f(q)$. When possible, \hat{a} is mapped back to one of the original options using a two-step LLM-based procedure (see Section 3.3), yielding a standardized prediction \hat{o}_{open} .

3.2 Reliability Metrics

For each item x , let $\mathcal{P}(x) = \{\text{Pert}_1, \dots, \text{Pert}_7\}$ denote the set of semantic-preserving input perturbations. For each $p \in \mathcal{P}(x)$, the model produces a prediction \hat{o}_p aligned to one of the original options. We evaluate model behavior over this full set.

ReCon Measures whether predictions are consistent across all versions of the same item:

$$\text{Con}(x) = \mathbf{1}\{\hat{o}_p = \hat{o}_{p'} \ \forall p, p' \in \mathcal{P}(x)\}.$$

Dataset-level score:

$$\text{ReCon} = \frac{1}{|D|} \sum_{x \in D} \text{Con}(x) \quad (1)$$

ReAcc Measures whether all predictions are correct (i.e., match the gold option o^*):

$$\text{Acc}(x) = \mathbf{1}\{\hat{o}_p = o^* \ \forall p \in \mathcal{P}(x)\}.$$

Dataset-level score:

$$\text{ReAcc} = \frac{1}{|D|} \sum_{x \in D} \text{Acc}(x) \quad (2)$$

3.3 Open-Answer Mapping

To compute reliability across open and closed formats, each free-form open answer \hat{a} is mapped to one of the original MCQA options (standard version) using a two-step LLM-as-judge pipeline.

① Candidate Mapping Given \hat{a} and the MCQA item x , the first judge proposes an alignment to one or more candidate options, or to a special “no clear match” label when none of the options adequately corresponds to \hat{a} . To avoid bias, we intentionally select a judge that is external to our evaluated baselines; specifically, we use **Qwen3-32B** with thinking mode enabled (see Appendix B).

② Verification If the candidate alignment is labeled as “no clear match”, we exclude the answer from the set of predictions used to compute ReCon. However, we perform a second verification step to assess its correctness, which impacts ReAcc. This is necessary because, when responding to reformulated open-answer prompts, models may produce answers that are only partially misaligned with the options (e.g., due to differences in specificity), leading to answers that appear incorrect but are in fact valid. To mitigate this, we employ a second, stronger judge with proven medical expertise to prevent spurious rejections of otherwise correct answers. Given its consistently strong empirical performance across our datasets (Sellergren et al., 2025), we use **Gemini-2.5-Flash** as the verifier. When Gemini-2.5-Flash itself is the evaluated model, we instead adopt **GPT-5-mini** as the verifier to avoid bias in the decision process.

Decision Rules (i) If a single option is selected, it is assigned as \hat{o}_{open} . (ii) If multiple options are returned (e.g., “A, B”), the response is treated as both incorrect (ReAcc) and inconsistent (ReCon). (iii) If no clear mapping is produced, the response is marked as inconsistent. (iv) If the verification step confirms the validity of an initially missing match, the response is labeled as correct; otherwise, it is labeled as incorrect.

Human Validation To evaluate the reliability of the candidate mapping (①), we manually inspected 300 randomly sampled alignments (100 per dataset). Agreement between the pipeline and human annotators was 95.5%, with no systematic bias toward any particular option. For the verification step (②), we observed that the “no clear match” label accounts for at most 10% of answers per dataset, and in some model–dataset combinations this proportion falls below 5%. To further validate the reliability of Gemini-2.5-Flash as the verifier, we randomly sampled 200 “no clear match” cases across all models and datasets. These samples were reviewed by a medical expert with over 15 years of specialization, allowed to consult online medical documentation and textbooks (see Appendix B for details). The assessment established that approximately 81% of Gemini’s judgments were fully correct, 7% were partially acceptable, and the remaining cases were factually incorrect. Despite this small fraction of errors, the verification process proved to be highly reliable and introduced only negligible impact on the validity of our results.

	MedQA	MedMCQA	MMLU (<i>medical subsets</i>)					
			Clinical	Genetics	Anatomy	Pro Med.	Biology	College Med.
Answer Options	A/B/C/D	A/B/C/D	A/B/C/D	A/B/C/D	A/B/C/D	A/B/C/D	A/B/C/D	A/B/C/D
Avg. Question Words	118.2	14.1	11.1	12.3	13.7	105.5	22.4	48.8
Test Size (orig.)	1,273	4,183	265	100	135	272	144	173
After Filtering	1,259	1,000	193	82	125	254	109	132

Table 1: **Dataset statistics.** We report option format, average question length, and test set size. *After Filtering* (highlighted row) indicates items retained in ReMedQA after removing instances not reliably convertible to open-ended form. For MedMCQA, we sampled 1,000 items via stratified sampling by subject.

4 Experimental Setup

4.1 Datasets

To build ReMedQA, we draw from three prominent medical corpora, yielding eight English-language MCQA tasks spanning specialties such as genetics, anatomy, and clinical reasoning. These datasets reflect both real-world scenarios encountered by medical professionals and exam formats commonly used in licensing and entrance tests. We focus only on datasets with 4-option choice format, to ensure consistency across our findings and structural compatibility with our framework. Consequently, we excluded datasets such as PubMedQA (Jin et al., 2019), whose binary yes/no format is incompatible with open-ended reformulation.

- **MedQA** (Jin et al., 2020) contains professional questions from the U.S. Medical License Exam (USMLE), focusing on clinical reasoning over patient cases, such as disease presentation, diagnostic decision-making, and pharmacological treatment. We use the official test split, which contains 1,273 questions.
- **MedMCQA** (Pal et al., 2022) includes questions from Indian medical entrance exams (AIIMS/NEET), testing foundational and applied medical knowledge, across 2,400 healthcare topics grouped into 21 medical subjects. Following common practices, we use the validation set, which includes 4,183 questions.
- **MMLU** (Hendrycks et al., 2021) covers exam questions from 57 subjects. Following common approach in medical literature, we focus on its six medical subsets: clinical knowledge, medical genetics, anatomy, professional medicine, college biology, and college medicine (1,089 questions in total).

Data statistics are shown in Table 1. For each dataset, every item was expanded into semantic-

preserving perturbations (Section 3.1) and evaluated using standardized prompts (see Appendix B).

Open-Answer Conversion We reformulated each MCQ into an open-ended version by removing the answer options and prompting an LLM to generate a free-text question (see Appendix B). The clinical vignette, when present, was kept exactly as in the original item, and only the final question sentence was rewritten in open-ended form. Questions whose validity relied on the specific wording of options (e.g., exclusion terms, subtle lexical differences, or comparisons across multiple choices) could not be reliably reformulated. Convertibility was automatically assessed by the same LLM, and only convertible items were retained (representative examples are reported in Appendix B). This yielded reduced but consistent subsets across the three benchmarks (see Table 1). After reformulation, MedMCQA resulted in approximately 3,000 valid samples. Yet, to improve scalability in our experiments, we further sampled 1,000 questions using stratified sampling across subjects, ensuring that the final subset preserved the original domain distribution. For the reformulation, we used GPT-4.1, the most effective non-reasoning OpenAI model available at the time of writing. The model was instructed to directly output its response in a fixed schema, ensuring consistency while avoiding unnecessary token usage and reducing inference cost. To ensure reliability, we manually inspected a subset of reformulated instances for each dataset. GPT-4.1 consistently produced faithful and accurate transformations, confirming that this step—being simple and highly constrained—can be safely automated without human intervention.

4.2 Models

We evaluate a diverse set of 11 LLMs that differ along key dimensions: (1) backbone architecture, (2) training objective (general-purpose vs. medical-

specialized), (3) availability (open- vs. closed-source), (4) parameter scale, ranging from 3.8B to 120B, and (5) intended optimization (reasoning-intensive tasks vs. general conversational use). Our selection includes both widely deployed commercial models and open-source research releases, as well as domain-specialized medical variants (see Appendix A for details). This breadth allows us to analyze how reliability is affected not only by scale, but also by domain specialization and accessibility.

Criteria for Medical Model Selection For the inclusion of medical-specialized models, we applied a set of minimum eligibility criteria. Specifically, each model was required to: (i) provide a publicly accessible model card (e.g., via Hugging Face), (ii) include clear and verifiable attribution with the name and contact information of the responsible individual or institution, (iii) specify explicit licensing terms governing its use, (iv) have at least one associated technical report or publication describing the current or a prior version of the model, (v) disclose details on the model architecture and backbone used, and (vi) ensure compatibility with the vLLM inference library. These criteria were established to promote accountability, transparency, and adherence to sound machine learning practices, including the prevention of test set leakage.

4.3 Implementation Details

All models are evaluated in a zero-shot setting, without any few-shot demonstrations. To ensure consistency with the evaluation protocol proposed by Sellergren et al. (2025), we adopted the following setup: (i) all evaluations were performed with a single inference run per example; (ii) greedy decoding (temperature = 0) combined with Chain-of-Thought (CoT) prompting (Wei et al., 2022) was used for non-reasoner models; (iii) default temperature and top- k settings were used for reasoner models; and (iv) generalist non-reasoner models were provided with a short system persona, such as “You are a medical expert”, which empirically improves their medical reasoning performance.

For models we ran locally, we fixed the random seed to 42 to ensure reproducibility. Full prompt templates are reported in Appendix B.

Reasoning Effort To ensure cost efficiency and fair comparison across experiments, we standardized the thinking budget for all reasoning models. Closed-source models, namely GPT-5-mini and

Gemini-2.5-Flash, were evaluated using a *low* reasoning effort setting to account for their higher API costs. In particular, Gemini-2.5-Flash was assigned a thinking budget of 1024, which Gemini’s documentation indicates as equivalent to the *low* reasoning setting in the OpenAI API. Due to their substantially lower inference costs, open-source models from the GPT-OSS family were evaluated with a *medium* reasoning effort. This setup ensures that all reasoning models operate under comparable computational constraints while maintaining practical cost considerations.

Environment Experiments were conducted on a workstation equipped with a single NVIDIA RTX 3090 GPU (24 GB VRAM), used for models up to 8B parameters. For efficient local inference, we adopted the vLLM library, following each model’s default precision settings. Llama-3.3-70B and GPT-OSS family models, although open-source, were accessed through Together Batch AI due to hardware constraints. Closed-source models were evaluated via their respective APIs (OpenAI Batch API for GPT models and Gemini Batch API for Gemini-2.5-Flash). This setup ensured consistent and reproducible conditions across all models.

5 Results

We report model performance on ReMedQA using the proposed reliability metrics (ReAcc and ReCon). We first present results across datasets and evaluation formats, and then analyze performance across experimental conditions to characterize model behavior under a reliability-oriented evaluation.

5.1 Main Findings

Figure 1 shows model performance on ReMedQA, comparing open-ended responses with all MCQA perturbation variants and reporting the averaged reliability metrics across datasets. Table 2, in contrast, breaks down ReAcc and ReCon by dataset within ReMedQA. We further provide a fine-grained analysis of model accuracy across perturbation types in Appendix F. Together, these results highlight the following key observations.

❶ **High accuracy on perturbations does not imply reliability.** Models struggle to maintain stable predictions, as reflected by substantially lower ReAcc and ReCon values compared to accuracy on a single perturbation (see Figure 1), and even relative to accuracy averaged across perturbations (see the last column in Table 2). This is particularly true

ReMedQA	Pro Med.		College Med.		Anatomy		Clinical		Biology		Genetics		MMLU Avg		MedQA		MedMCQA		Avg		
	RA	RC	RA	RC	RA	RC	RA	RC	RA	RC	RA	RC	RA	RC	RA	RC	RA	RC	RA	RC	Acc
Large Models																					
GPT-5-mini 🧠	76.4	77.5	58.3	66.7	65.6	73.6	61.1	68.9	67.0	71.3	85.4	86.6	69.0	74.1	65.9	68.2	39.1	46.0	58.0	62.8	84.4
GPT-OSS-120B 🧠	72.0	76.8	59.8	65.9	54.4	65.6	57.0	65.3	70.6	74.3	79.3	84.1	65.5	72.0	62.4	67.6	34.4	43.9	<u>54.1</u>	<u>61.2</u>	81.7
Gemini-2.5-Flash 🧠	57.9	60.5	46.2	55.7	52.0	62.4	46.6	52.8	60.6	67.9	72.0	73.2	55.9	62.1	47.6	50.7	33.2	39.4	45.6	50.7	<u>82.0</u>
GPT-OSS-20B 🧠	62.6	66.7	47.7	55.7	52.0	59.2	48.7	54.9	60.6	65.1	64.6	75.3	56.0	62.8	49.4	54.0	24.9	30.4	43.4	49.1	76.0
Llama-3.3-70B	53.5	56.5	34.8	40.9	48.8	53.6	42.5	45.6	55.0	56.0	63.4	64.6	49.7	52.9	34.0	37.9	23.6	28.6	35.8	39.8	75.2
Small Models																					
Llama-3-8B	9.8	17.7	9.8	15.9	12.8	27.2	11.9	20.2	15.6	25.7	23.2	37.8	13.9	24.1	6.1	11.9	6.8	16.0	8.9	17.3	55.1
Llama3-Med42-8B 🩺	15.7	18.9	18.2	24.6	20.8	29.6	17.6	24.4	15.6	22.9	24.4	36.6	18.7	26.2	9.1	13.2	10.8	18.2	<u>12.9</u>	<u>19.2</u>	58.0
Phi-3.5-mini	23.6	25.6	9.8	24.2	26.4	35.2	18.1	26.9	29.4	32.1	19.5	40.2	21.1	30.7	12.3	15.4	11.0	13.8	14.8	20.0	<u>57.1</u>
MediPhi-3.8B 🩺	11.8	16.1	14.4	18.2	20.0	27.2	18.7	20.2	21.1	23.9	24.4	29.3	18.4	22.5	8.6	13.4	7.9	11.5	11.6	15.8	54.9
Gemma-3-4B	5.9	8.3	5.3	9.8	16.0	19.2	6.7	9.4	12.8	17.4	23.2	23.8	11.7	14.6	3.3	6.2	3.8	6.8	6.3	9.2	48.6
MedGemma-4B 🩺	2.8	4.3	2.3	5.3	4.8	10.4	3.6	5.7	6.4	9.2	8.5	11.0	4.7	7.6	2.5	5.8	2.7	5.3	3.3	6.2	52.1

Table 2: **Performance on ReMedQA using ReAcc (RA), ReCon (RC), and Accuracy (Acc).** Values in columns with bolded headers are used to compute RA and RC averages. Reported Acc is obtained by averaging accuracy across all perturbation variants (see Tables 5 and 6 in Appendix). Dashed lines group base models with their medical-specialization (marked with 🩺). 🧠 = Reasoning models. Best scores are in bold; second-best are underlined.

for smaller models, whose outputs fluctuate sharply under distractor changes.

② **Accuracy varies strongly between open-ended and MCQA formats.** Smaller models often achieve higher accuracy in the open-ended setting than under MCQA perturbations, indicating that distractor options can hinder rather than help their reasoning. In contrast, large models benefit from the structured MCQA format and typically outperform their open-ended counterparts. Across all model sizes, the largest accuracy drop occurs when the correct option is replaced with “None of the provided options”, while the “Select Incorrect” variant mainly affects smaller models, with larger systems handling this task more robustly (see Figure 1).

③ **Domain specialization does not guarantee reliability.** Medical-specialized models (e.g., MedGemma) show no robustness advantage over their general-purpose alternatives and often underperform in both ReAcc and ReCon despite medical tuning. Figure 3 shows this pattern by comparing correctness and consistency across families. While reliable models should cluster near the upper-right diagonal, all systems—particularly medical ones (green markers)—fall well below this region.

5.2 Dataset-Specific Reliability

We now analyze model performance in greater depth in terms of reliability metrics across datasets, as summarized in Table 2. Overall, MMLU medical subsets are the easiest, followed by MedQA, while MedMCQA is the most unstable bench-

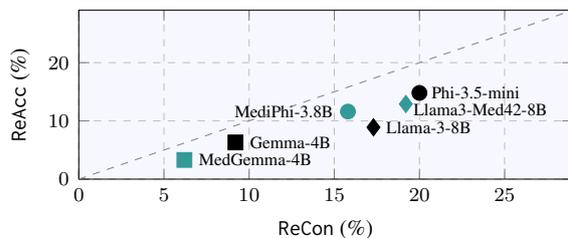


Figure 3: **ReAcc vs. ReCon across model families on ReMedQA.** Green markers denote medical-specialized variants. The dashed diagonal marks the ideal region, where accuracy retention and consistency are balanced.

mark: even the strongest models remain below 40% ReAcc. Among large models, GPT-5-mini consistently achieves the best performance, ranking highest in ReAcc, ReCon, and averaged accuracy across perturbations. The model reaches its highest reliability on MMLU (up to ~69% ReAcc and ~74% ReCon). However, it still loses ~30 points on MedMCQA on both metrics. Smaller models exhibit similar trends, though with reduced gaps across datasets. Notably, the most reliable small model is Phi-3.5-mini, a general-purpose system, whereas the most accurate small model is Llama3-Med42-8B, reinforcing that domain specialization does not imply higher reliability. In contrast, the Gemma-3 family shows the weakest robustness overall, with ReAcc and ReCon frequently dropping below 10% across most subsets, highlighting severe sensitivity to input perturbations.

These patterns reflect differences in dataset design. MMLU primarily tests factual recall through

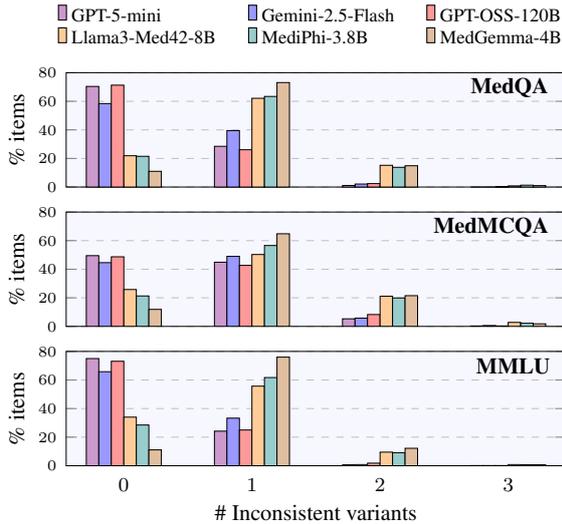


Figure 4: **Distribution of answer consistency across datasets.** Each bar shows the % of samples for which a model changes its prediction across input formats.

short, unambiguous questions, favoring surface-level lexical matching. MedQA contains longer, clinically grounded vignettes that require multi-step reasoning, resulting in greater separation between small and large models but improved internal stability. MedMCQA, composed of short exam-style questions, is particularly sensitive to distractor noise and format perturbations, exposing reliability weaknesses even in frontier systems.

5.3 Prediction Stability

Figure 4 shows how frequently models change their predictions across the seven input variants of each question. Each bar represents the percentage of questions with a given number of inconsistent predictions. For example, #0 denotes identical answers across all formats, regardless of correctness, whereas #1 indicates that the model changes its prediction exactly once across the seven variants, resulting in two distinct selected options. Most instability is concentrated at one or two variants. A finer-grained analysis of single-flip inconsistencies (#1), reported in Appendix E, shows that they are predominantly caused by the “None Provided”, “Select Incorrect”, and “Open” formats. Clear scale effects emerge: unlike large models, small ones exhibit flatter distributions with inconsistencies spread across multiple variant counts. Dataset typology also influences stability: MMLU shows the highest overall consistency, MedQA intermediate levels, and MedMCQA the lowest, mirroring the aggregate reliability patterns reported in Table 2.

5.4 Open-Ended Pair-Wise Consistency

Since ReMedQA treats open-ended answering as the primary signal of competence—reflecting how clinical reasoning operates in practice—we analyze pair-wise consistency, measuring whether each MCQA perturbation individually preserves a model’s original free-form decision, in contrast to ReCon in Table 2, which requires agreement across all perturbations. A detailed pair-wise analysis across all perturbations and datasets is provided in the Appendix F. Figure 5 reports model pair-wise consistency with respect to the original open-ended prediction. Overall, alignment with open-ended responses is sensitive to MCQA format changes, indicating that MCQA structure can partially mask underlying reasoning instability. Across models, the “No Labels” perturbation yields the highest consistency, suggesting that removing option identifiers best preserves the semantics of open-ended reasoning. In contrast, as noted for accuracy (see Section 5.1), “None Provided” is the least stable variation, frequently inducing divergences from the original answer. Task-inverting formats such as “Select Incorrect” disproportionately affect smaller and domain-specialized models, which struggle to reliably invert their original open-ended reasoning. Larger models maintain higher consistency overall, reinforcing that robustness is driven more by model capacity than by medical specialization.

5.5 Options Only Evaluation

Recent work (Balepur et al., 2024) has shown that LLMs can answer multiple-choice questions with surprisingly high accuracy even when the question text is removed and only the answer options are provided. This behavior arises from the exploitation of statistical artifacts and surface patterns learned during training, making LLMs exceptionally strong MCQA test-takers (Balepur et al., 2025). A reliable evaluation framework should expose, rather than reward, such shortcut behavior.

We introduce an *Options Only* analysis with two objectives: (i) to examine, for the first time, whether this phenomenon persists in a substantially more challenging domain—medical MCQA—and when using modern reasoning LLMs, and (ii) to demonstrate that high accuracy in the absence of the question is an unreliable indicator of true understanding. This differs from prior analyses (Balepur et al., 2024): medical benchmarks such as MedQA contain long clinical vignettes, and answer options

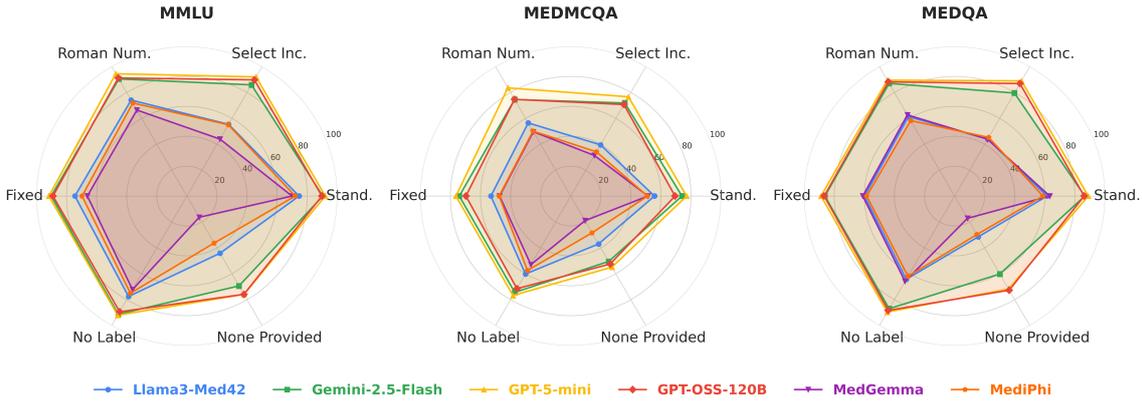


Figure 5: **Pair-wise consistency with open-ended answers across MCQA perturbations.** Each radar plot shows the percentage of samples for which a model’s prediction remains consistent with its original open-ended answer.

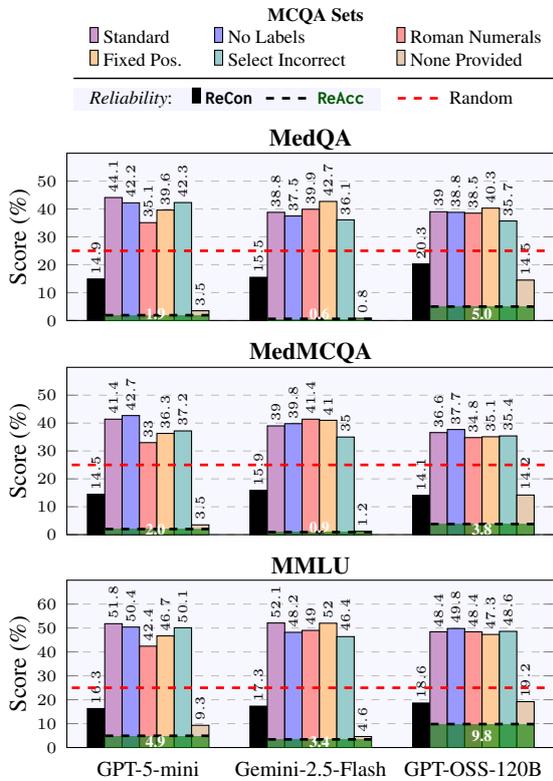


Figure 6: **Options Only performance.** Each model is evaluated by providing only the MCQA answer options, without the question prompt. The red dashed line indicates the random-guess baseline.

correspond to subtle differential diagnoses or treatment decisions, making question reconstruction from options alone considerably harder (see illustrative examples in Appendix D).

Figure 6 reports results for the three most reliable models—GPT-5-mini, Gemini-2.5-Flash, and GPT-OSS-120B—evaluated using only the answer options (see Appendix D for examples). Despite the increased difficulty, all models achieve non-trivial

accuracy (typically 40–50%) across datasets—well above the random baseline. This indicates that models can still partially exploit superficial cues within the options, such as lexical overlap, stylistic regularities, or positional biases. Performance is consistently highest on Standard splits, suggesting that existing benchmarks contain exploitable surface patterns that inflate apparent competence. In contrast, reliability metrics collapse under this study, revealing that such option-based predictions are highly unstable across perturbations. This confirms that the observed accuracy reflects fragile statistical shortcuts rather than robust medical reasoning, underscoring the need for reliability-centered evaluation of medical MCQA benchmarks.

6 Conclusion

We introduced ReMedQA, the first benchmark assessing the reliability of LLMs in medical MCQA. Unlike prior datasets, it systematically tests consistency across controlled perturbations and open-answer variants, with complementary metrics—ReAcc and ReCon—that separate accuracy from robustness. Experiments show that high MCQA accuracy overstates competence: models remain format-sensitive, and reliability grows with scale rather than specialization. We hope ReMedQA encourages evaluations that consider not only *what* models answer correctly, but also *how* consistently they do so—a prerequisite for trustworthy clinical and biomedical use (Wang et al., 2025b).

These results underscore a persistent gap between LLM and human reasoning, a limitation observed consistently across different domains and assessment frameworks (Cocchieri et al., 2025c,d).

Limitations

While ReMedQA introduces a new perspective on reliability, it has several limitations. First, our analysis primarily targets answer accuracy rather than the underlying reasoning process, offering only a partial view of model cognition. Second, for scalability and cost efficiency, (i) evaluations rely on one inference per problem without exploring self-consistency, and (ii) the reasoning models operate under a limited thinking budget, which may underestimate their capabilities. Third, while ReMedQA builds on prior work on robustness in MCQA, it currently covers only a subset of perturbations; other variants such as position swaps, semantic paraphrases, and counterfactual edits remain unexplored. Fourth, although ReMedQA measures reliability in open-domain QA, it does not evaluate medical competence in downstream tasks such as text summarization (Moro et al., 2022, 2023a,b; Moro and Ragazzi, 2023; Ragazzi et al., 2024, 2025), or entity recognition (Cocchieri et al., 2025a,b). Finally, our open-answer mapping procedure is precision-oriented and validates a subset of model-retrieved candidates. While supported by a strong external judge and human validation, this design may overlook systematic mapping errors or hidden biases. A more exhaustive evaluation—based on independently manually mapping a random subset and comparing multiple mapper models—could help uncover such edge cases.

Future work will expand our evaluation by enriching ReMedQA with more complex datasets and new perturbations (e.g., synonym substitution, paraphrasing) to better probe model robustness. Moreover, while our current reliability metrics adopt a deliberately strict formulation—motivated by the need for rigorous evaluation in high-stakes domains such as medicine—we will explore softer variants, including weighted reliability measures that account for the number and distribution of errors within a perturbation set. We also plan to incorporate multimodal inputs (e.g., radiology images) to assess diagnostic reasoning in more realistic settings. Finally, we will generalize our metrics to other medical tasks (Domeniconi et al., 2014; Lena et al., 2015; Frisoni and Moro, 2020) where reliability is a critical indicator of trustworthiness.

Ethical Considerations

This study does not involve human subjects, patient data, or personally identifiable information.

All benchmarks used are publicly available and contain de-identified text. Our goal is to promote safer and more reliable evaluation of medical language models, not to deploy them in clinical settings. Nevertheless, we caution that high benchmark performance does not imply readiness for real-world medical use; LLM outputs should never replace professional medical judgment.

Acknowledgements

Research partially supported by AI-PACT project (CUP B47H22004450008, B47H22004460001); National Plan PNC-I.1 DARE initiative (PNC0000002, CUP B53C22006450001); PNRR Extended Partnership FAIR (PE00000013, Spoke 8); 2024 Scientific Research and High Technology Program, project “AI analysis for risk assessment of empty lymph nodes in endometrial cancer surgery”, the Fondazione Cassa di Risparmio in Bologna; Chips JU TRISTAN project (G.A. 101095947).

References

- Marah I Abdin, Sam Ade Jacobs, and Ammar Ahmad Awan et al. 2024. [Phi-3 technical report: A highly capable language model locally on your phone](#). CoRR, abs/2404.14219.
- Ishtiaq Ahmed and Sundas Ishtiaq. 2021. Reliability and validity: importance in medical research. *Methods*, 12(1):2401–2406.
- Nishant Balepur, Shramay Palta, and Rachel Rudinger. 2023. [It’s not easy being wrong: Large language models struggle with process of elimination reasoning](#). In *Annual Meeting of the Association for Computational Linguistics*.
- Nishant Balepur, Abhilasha Ravichander, and Rachel Rudinger. 2024. [Artifacts or abduction: How do LLMs answer multiple-choice questions without the question?](#) In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10308–10330, Bangkok, Thailand. Association for Computational Linguistics.
- Nishant Balepur, Rachel Rudinger, and Jordan Lee Boyd-Graber. 2025. [Which of these best describes multiple choice evaluation with LLMs? a\) forced B\) flawed C\) fixable D\) all of the above](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3394–3418, Vienna, Austria. Association for Computational Linguistics.
- Suhana Bedi, Yixing Jiang, Philip Chung, Sanmi Koyejo, and Nigam Shah. 2025. Fidelity of medical

- reasoning in large language models. JAMA Network Open, 8(8):e2526021–e2526021.
- Clément Christophe, Praveen K. Kanithi, Prateek Munjal, Tathagata Raha, Nasir Hayat, Ronnie Rajan, Ahmed Al-Mahrooqi, Avani Gupta, Muhammad Umar Salman, Gurpreet Gosal, Bhargav Kanakiya, Charles Chen, Natalia Vassilieva, Boulbaba Ben Amor, Marco AF Pimentel, and Shadab Khan. 2024a. Med42 - evaluating fine-tuning strategies for medical llms: Full-parameter vs. parameter-efficient approaches. CoRR, abs/2404.14779.
- Clément Christophe, Praveen K. Kanithi, Tathagata Raha, Shadab Khan, and Marco AF Pimentel. 2024b. Med42-v2: A suite of clinical llms. CoRR, abs/2408.06142.
- Alessio Cocchieri, Giacomo Frisoni, Marcos Martínez Galindo, Gianluca Moro, Giuseppe Tagliavini, and Francesco Candoli. 2025a. OpenBioNER: Lightweight open-domain biomedical named entity recognition through entity type description. In Findings of the Association for Computational Linguistics: NAACL 2025, pages 818–837, Albuquerque, New Mexico. Association for Computational Linguistics.
- Alessio Cocchieri, Marcos Martínez Galindo, Giacomo Frisoni, Gianluca Moro, Claudio Sartori, and Giuseppe Tagliavini. 2025b. ZeroNER: Fueling zero-shot named entity recognition via entity type descriptions. In Findings of the Association for Computational Linguistics: ACL 2025, pages 15594–15616, Vienna, Austria. Association for Computational Linguistics.
- Alessio Cocchieri, Luca Ragazzi, Paolo Italiani, Giuseppe Tagliavini, and Gianluca Moro. 2025c. “what do you call a dog that is incontrovertibly true? dogma”: Testing LLM generalization through humor. In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 22922–22937, Vienna, Austria. Association for Computational Linguistics.
- Alessio Cocchieri, Luca Ragazzi, Giuseppe Tagliavini, Lorenzo Tordi, Antonella Carbonaro, and Gianluca Moro. 2025d. Can large language models win the international mathematical games? In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 9645–9671, Suzhou, China. Association for Computational Linguistics.
- Gheorghe Comanici, Eric Bieber, and Mike Schaeckermann et al. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. CoRR, abs/2507.06261.
- Jens Kleesiek, and Paul Vozila. 2025. A modular approach for clinical slms driven by synthetic data with pre-instruction tuning, model merging, and clinical-tasks alignment. In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025, pages 19352–19374. Association for Computational Linguistics.
- Giacomo Domeniconi, Marco Masseroli, Gianluca Moro, and Pietro Pinoli. 2014. Discovering new gene functionalities from random perturbations of known gene ontological annotations. In International Conference on Knowledge Discovery and Information Retrieval.
- Abhimanyu Dubey, Abhinav Jauhri, and Abhinav Pandey et al. 2024. The llama 3 herd of models. CoRR, abs/2407.21783.
- Bryan Etzine, Masoud Hashemi, Nishanth Madhusudhan, Sagar Davasam, Roshnee Sharma, Sathwik Tejaswi Madhusudhan, and Vikas Yadav. 2025. Revitalizing saturated benchmarks: A weighted metric approach for differentiating large language model performance. CoRR, abs/2503.05551.
- Joshua J Fenton, Anthony Jerant, Richard L Kravitz, Klea D Bertakis, Daniel J Tancredi, Elizabeth M Magann, and Peter Franks. 2017. Reliability of physician-level measures of patient experience in primary care. Journal of General Internal Medicine, 32(12):1323–1329.
- Giacomo Frisoni, Alessio Cocchieri, Alex Presepi, Gianluca Moro, and Zaiqiao Meng. 2024. To generate or to retrieve? on the effectiveness of artificial contexts for medical open-domain question answering. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 9878–9919, Bangkok, Thailand. Association for Computational Linguistics.
- Giacomo Frisoni and Gianluca Moro. 2020. Phenomena explanation from text: Unsupervised learning of interpretable and statistically significant knowledge. In International Conference on Data Technologies and Applications.
- Matt Gardner, William Merrill, Jesse Dodge, Matthew E. Peters, Alexis Ross, Sameer Singh, and Noah A. Smith. 2021. Competency problems: On finding and removing artifacts in language data. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021, pages 1801–1813. Association for Computational Linguistics.
- Maxime Griot, Coralie Hemptinne, Jean Vanderdonckt, and Demet Yuksel. 2025a. Large language models lack essential metacognition for reliable medical reasoning. Nature communications, 16(1):642.

- Maxime Griot, Jean Vanderdonck, Demet Yuksel, and Coralie Hemptinne. 2025b. [Pattern recognition or medical knowledge? the problem with multiple-choice questions in medicine](#). In [Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics \(Volume 1: Long Papers\)](#), pages 5321–5341, Vienna, Austria. Association for Computational Linguistics.
- Yu Gu, Jingjing Fu, Xiaodong Liu, Jeya Maria Jose Valanarasu, Noel Codella, Reuben Tan, Qianchu Liu, Ying Jin, Sheng Zhang, Jinyu Wang, et al. 2025. [The illusion of readiness: Stress testing large frontier models on multimodal medical benchmarks](#). [arXiv preprint arXiv:2509.18234](#).
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. [Measuring massive multitask language understanding](#). In [9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021](#). OpenReview.net.
- Pedram Hosseini, Jessica M. Sin, Bing Ren, Bryceton G. Thomas, Elnaz Nouri, Ali Farahanchi, and Saeed Hasanzpour. 2024. [A benchmark for long-form medical question answering](#). [CoRR](#), abs/2411.09834.
- Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2020. [What disease does this patient have? A large-scale open domain question answering dataset from medical exams](#). [CoRR](#), abs/2009.13081.
- Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W. Cohen, and Xinghua Lu. 2019. [Pubmedqa: A dataset for biomedical research question answering](#). In [Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019](#), pages 2567–2577. Association for Computational Linguistics.
- Aishwarya Kamath, Johan Ferret, and Shreya Pathak et al. 2025. [Gemma 3 technical report](#). [CoRR](#), abs/2503.19786.
- Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. 2024. [Biomistral: A collection of open-source pretrained large language models for medical domains](#). In [Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024](#), pages 5848–5864. Association for Computational Linguistics.
- Max Lamparth, Declan Grabb, Amy Franks, Scott Gershan, Kaitlyn N. Kunstman, Aaron Lulla, Monika Drummond Roots, Manu Sharma, Aryan Shrivastava, Nina Vasan, and Colleen Waickman. 2025. [Moving beyond medical exam questions: A clinician-annotated dataset of real-world tasks and ambiguity in mental healthcare](#). [CoRR](#), abs/2502.16051.
- Pietro Di Lena, Giacomo Domeniconi, Luciano Margara, and Gianluca Moro. 2015. [Gota: Go term annotation of biomedical literature](#). [BMC Bioinformatics](#), 16.
- Wangyue Li, Liangzhi Li, Tong Xiang, Xiao Liu, Wei Deng, and Noa Garcia. 2024. [Can multiple-choice questions really be useful in detecting the abilities of llms?](#) In [Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation \(LREC-COLING 2024\)](#), pages 2819–2834.
- Francesco Maria Molfese, Luca Moroni, Luca Gioffre, Alessandro Sciré, Simone Conia, and Roberto Navigli. 2025. [Right answer, wrong score: Uncovering the inconsistencies of llm evaluation in multiple-choice question answering](#). [ArXiv](#), abs/2503.14996.
- Gianluca Moro and Luca Ragazzi. 2023. [Align-then-abstract representation learning for low-resource summarization](#). [Neurocomputing](#), 548:126356.
- Gianluca Moro, Luca Ragazzi, and Lorenzo Valgimigli. 2023a. [Carburacy: Summarization models tuning and comparison in eco-sustainable regimes with a novel carbon-aware accuracy](#). In [Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023](#), pages 14417–14425. AAAI Press.
- Gianluca Moro, Luca Ragazzi, Lorenzo Valgimigli, and Davide Freddi. 2022. [Discriminative marginalized probabilistic neural method for multi-document summarization of medical literature](#). In [Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics \(Volume 1: Long Papers\)](#), pages 180–189, Dublin, Ireland. Association for Computational Linguistics.
- Gianluca Moro, Luca Ragazzi, Lorenzo Valgimigli, and Lorenzo Molfetta. 2023b. [Retrieve-and-rank end-to-end summarization of biomedical studies](#). In [Similarity Search and Applications - 16th International Conference, SISAP 2023, A Coruña, Spain, October 9-11, 2023, Proceedings, volume 14289 of Lecture Notes in Computer Science](#), pages 64–78. Springer.
- Robert Osazuwa Ness, Katie Matton, Hayden S. Helm, Sheng Zhang, Junaid Bajwa, Carey E. Priebe, and Eric Horvitz. 2024. [Medfuzz: Exploring the robustness of large language models in medical question answering](#). [ArXiv](#), abs/2406.06573.
- Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. 2023a. [Capabilities of gpt-4 on medical challenge problems](#). [Preprint, arXiv:2303.13375](#).
- 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, Hoi-fung Poon, Tao Qin, Naoto Usuyama, Chris White, and Eric Horvitz. 2023b. [Can generalist foundation models outcompete special-purpose tuning? case study in medicine](#). [Preprint](#), arXiv:2311.16452.
- OpenAI. 2025. [gpt-oss-120b & gpt-oss-20b model card](#). [Preprint](#), arXiv:2508.10925.
- Ankit Pal, Logesh Kumar Umapathi, and Malaikanan Sankarasubbu. 2022. [Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering](#). In [Conference on Health, Inference, and Learning, CHIL 2022, 7-8 April 2022, Virtual Event](#), volume 174 of [Proceedings of Machine Learning Research](#), pages 248–260. PMLR.
- Ankit Pal, Logesh Kumar Umapathi, and Malaikanan Sankarasubbu. 2023. [Med-HALT: Medical domain hallucination test for large language models](#). In [Proceedings of the 27th Conference on Computational Natural Language Learning \(CoNLL\)](#), pages 314–334, Singapore. Association for Computational Linguistics.
- Pouya Pezeshkpour and Estevam Hruschka. 2023. [Large language models sensitivity to the order of options in multiple-choice questions](#). In [NAACL-HLT](#).
- Luca Ragazzi, Paolo Italiani, Gianluca Moro, and Matia Panni. 2024. [What are you token about? differentiable perturbed top-k token selection for scientific document summarization](#). In [Findings of the Association for Computational Linguistics: ACL 2024](#), pages 9427–9440, Bangkok, Thailand. Association for Computational Linguistics.
- Luca Ragazzi, Gianluca Moro, Lorenzo Valgimigli, and Riccardo Fiorani. 2025. [Cross-document distillation via graph-based summarization of extracted essential knowledge](#). [IEEE Transactions on Audio, Speech and Language Processing](#), 33:518–527.
- Marco Túlio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. [Beyond accuracy: Behavioral testing of NLP models with checklist](#). In [Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020](#), pages 4902–4912. Association for Computational Linguistics.
- Khaled Saab, Tao Tu, and Wei-Hung et al. 2024. [Capabilities of gemini models in medicine](#). [CoRR](#), abs/2404.18416.
- Andrew Sellergren, Sahar Kazemzadeh, and Tiam Jaroensri et al. 2025. [Medgemma technical report](#). [CoRR](#), abs/2507.05201.
- Shrutika Singh, Anton Alyakin, Daniel Alexander Alber, Jaden Stryker, Ai Phuong S. Tong, Karl L. Sangwon, Nicolas K. Goff, Mathew de la Paz, Miguel Hernandez-Rovira, Ki Yun Park, Eric Claude Leuthardt, and Eric Karl Oermann. 2025. [It is too many options: Pitfalls of multiple-choice questions in generative AI and medical education](#). [CoRR](#), abs/2503.13508.
- Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Kumar Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Nathaneal Schärli, Aakanksha Chowdhery, Philip Andrew Mansfield, Blaise Agüera y Arcas, Dale R. Webster, Gregory S. Corrado, Yossi Matias, Katherine Chou, Juraj Gottweis, Nenad Tomasev, Yun Liu, Alvin Rajkumar, Joelle K. Barral, Christopher Semturs, Alan Karthikesalingam, and Vivek Natarajan. 2022. [Large language models encode clinical knowledge](#). [CoRR](#), abs/2212.13138.
- Zhi Rui Tam, Cheng-Kuang Wu, Chieh-Yen Lin, and Yun-Nung Chen. 2025. [None of the above, less of the right: Parallel patterns between humans and llms on multi-choice questions answering](#). [ArXiv](#), abs/2503.01550.
- Krithik Vishwanath, Anton Alyakin, Daniel Alexander Alber, Jin Vivian Lee, Douglas Kondziolka, and Eric Karl Oermann. 2025. [Medical large language models are easily distracted](#). [arXiv preprint arXiv:2504.01201](#).
- Haochun Wang, Sendong Zhao, Zewen Qiang, Bing Qin, and Ting Liu. 2024. [Beyond the answers: Reviewing the rationality of multiple choice question answering for the evaluation of large language models](#). [arXiv e-prints](#), pages arXiv–2402.
- Haochun Wang, Sendong Zhao, Zewen Qiang, Nuwa Xi, Bing Qin, and Ting Liu. 2025a. [LLMs may perform MCQA by selecting the least incorrect option](#). In [Proceedings of the 31st International Conference on Computational Linguistics](#), pages 5852–5862, Abu Dhabi, UAE. Association for Computational Linguistics.
- Yinuo Wang, Robert E. Mercer, Frank Rudzicz, Sudipta Singha Roy, Pengjie Ren, Zhumin Chen, and Xindi Wang. 2025b. [Trustworthy medical question answering: An evaluation-centric survey](#). [CoRR](#), abs/2506.03659.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, et al. 2022. [Chain-of-thought prompting elicits reasoning in large language models](#). In [NeurIPS 2022](#).
- Chaoyi Wu, Weixiong Lin, Xiaoman Zhang, Ya Zhang, Weidi Xie, and Yanfeng Wang. 2024. [Pmc-llama: toward building open-source language models for medicine](#). [J. Am. Medical Informatics Assoc.](#), 31(9):1833–1843.
- Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. 2024. [Benchmarking retrieval-augmented generation for medicine](#). In [Findings of the Association for Computational Linguistics: ACL 2024](#), pages 6233–6251, Bangkok, Thailand. Association for Computational Linguistics.

Zhen Yang, Ping Jian, and Chengzhi Li. 2025. [Option symbol matters: Investigating and mitigating multiple-choice option symbol bias of large language models](#). In [Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies \(Volume 1: Long Papers\)](#), pages 1902–1917, Albuquerque, New Mexico. Association for Computational Linguistics.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. 2023. [Large language models are not robust multiple choice selectors](#). [ArXiv](#), abs/2309.03882.

Yuxin Zuo, Shang Qu, Yifei Li, Zhangren Chen, Xuekai Zhu, Ermo Hua, Kaiyan Zhang, Ning Ding, and Bowen Zhou. 2025. [Medxpertqa: Benchmarking expert-level medical reasoning and understanding](#). [CoRR](#), abs/2501.18362.

A Models

We evaluated 11 distinct LLMs. Given the focus of our benchmark on medical MCQA, we prioritized models with strong reasoning abilities and, where available, models specifically adapted to the medical domain. Table 3 provides an overview of the models and their sources.

OpenAI We included the most recent OpenAI models available at the time of writing. We tested **GPT-5-mini**, released on August 7, 2025,² a faster and more cost-efficient variant of GPT-5 optimized for well-structured tasks. We also evaluated **GPT-OSS-20B** and **GPT-OSS-120B** (OpenAI, 2025), open-weight models introduced on August 5, 2025, which represent the most powerful openly released alternatives from OpenAI. They are designed as strong reasoners, producing intermediate reasoning steps before delivering their final answers.

Gemini-2.5 The Gemini-2.5 family (Comanici et al., 2025) represents Google’s most advanced AI models. As reasoner model, Gemini-2.5 systems are designed to reason through intermediate steps before producing a final response, which leads to improved accuracy and robustness. In our evaluation, we focus on **Gemini-2.5-Flash**, a cost-efficient, high-speed variant optimized for practical deployment. Notably, prior studies (Sellergren et al., 2025) demonstrate its strong performance in medical tasks, making it a particularly relevant model for our benchmark.

Gemma-3 The Gemma-3 family (Kamath et al., 2025) consists of lightweight, state-of-the-art open models released by Google, developed using the same research and technologies underlying the Gemini series. Gemma-3 models are multimodal, capable of processing both text and image inputs while generating text outputs, and are available as open-weight pre-trained and instruction-tuned variants. Within this family, **MedGemma** (Sellergren et al., 2025) refers to specialized variants adapted for medical text and image understanding. For our benchmark, we focus on **MedGemma-4B** and its base counterpart **Gemma-3-4B**.

Llama-3 The Llama-3 family (Dubey et al., 2024), released by Meta, comprises pretrained and instruction-tuned generative text models. It is available in multiple sizes, including 8B and 70B pa-

rameter variants. Building on this, **Med42-v2** is a suite of open-access clinical LLMs developed by M42 (Christophe et al., 2024b), instruction- and preference-tuned to broaden access to medical knowledge. In our evaluation, we used **Llama3-Med42-8B** alongside its base counterpart **Llama-3-8B**. We also included **Llama-3.3-70B-Instruct**, the strongest model in the series, which achieves performance comparable to the larger Llama-3.1-405B-Instruct while offering improved efficiency.

Phi-3 The **Phi-3.5-mini** model (Abdin et al., 2024) is a lightweight, state-of-the-art open model trained on synthetic data and carefully filtered public web sources, with a strong emphasis on high-quality, reasoning-dense content. It belongs to the broader Phi-3 family and supports a context length of up to 128K tokens. Building on this foundation, the **MediPhi** model collection (Corbeil et al., 2025) extends Phi-3.5-mini-instruct into seven specialized small language models (3.8B parameters each) tailored for medical and clinical applications, designed in a modular fashion. For our experiments, we focused on the **MediPhi-Instruct** variant.

B Prompts and Guidelines

Figure 11 shows the prompt for converting MedQA into open-ended format. Figure 12 illustrates the prompt used to map multiple-choice questions to open-ended answers, while Figure 13 presents the prompt for handling valid alternative answers. Table 14 reports examples of multiple-choice questions that cannot be reformulated into an open-ended format. Tables 9 and 10 report the prompts for non-reasoner and reasoner models, respectively. Table 11 shows the prompts adopted for the Options Only diagnostic modality. Finally, Figure 14 provides the guidelines given to the medical expert.

C Fixed Position Choice

In our experiments, we fix the position **D** as the shifted option. As shown in Table 4, this choice is motivated by its being the least frequent correct answer in both the training and test splits of MedQA and MedMCQA. While MMLU does not provide an official training set, both MedQA and MedMCQA are widely used for training LLMs (Christophe et al., 2024a; Wu et al., 2024; Sellergren et al., 2025). Consequently, many models are likely to have been exposed to their training data, potentially developing mild positional biases toward more frequent options. Selecting **D**, the

²<https://openai.com/it-IT/index/introducing-gpt-5/>

Model	Snapshot / Repository	URL
<i>API Models</i>		
Gemini-2.5-Flash	gemini-2.5-flash	https://ai.google.dev/gemini-api/docs/models
GPT-5-mini	gpt-5-mini-2025-08-07	https://platform.openai.com/docs/models/gpt-5-mini
GPT-OSS-120B	openai/gpt-oss-120b	https://www.together.ai/models/gpt-oss-120b
GPT-OSS-20B	openai/gpt-oss-20b	https://www.together.ai/models/gpt-oss-20b
Llama-3.3-70B	meta-llama/Llama-3.3-70B-Instruct-Turbo	https://www.together.ai/models/llama-3-3-70b
<i>Hugging Face Models</i>		
Llama-3-8B	meta-llama/Meta-Llama-3-8B-Instruct	https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct
Llama3-Med42-8B	m42-health/Llama3-Med42-8B	https://huggingface.co/m42-health/Llama3-Med42-8B
Gemma-3-4B	google/gemma-3-4b-it	https://huggingface.co/google/gemma-3-4b-it
MedGemma-4B	google/medgemma-4b-it	https://huggingface.co/google/medgemma-4b-it
Phi-3.5-mini	microsoft/Phi-3.5-mini-instruct	https://huggingface.co/microsoft/Phi-3.5-mini-instruct
MediPhi-3.8B	microsoft/MediPhi-Instruct	https://huggingface.co/microsoft/MediPhi-Instruct

Table 3: **Models source.** Repository identifiers and access URLs for all models used in our experiments, covering both API-based and open-source releases.

Dataset	Option A	Option B	Option C	Option D
<i>Training Sets</i>				
MedQA	2,584 (25.4%)	2,654 (26.1%)	2,557 (25.1%)	2,383 (23.4%)
MedMCQA	53,591 (29.3%)	47,826 (26.2%)	42,442 (23.2%)	38,963 (21.3%)
<i>Test Sets</i>				
MedQA	353 (27.7%)	309 (24.3%)	346 (27.2%)	265 (20.8%)
MedMCQA	1,348 (32.2%)	1,085 (25.9%)	925 (22.1%)	825 (19.7%)

Table 4: **Distribution of gold answer positions across datasets.** Option D consistently shows the lowest frequency as the correct answer in both MedQA and MedMCQA (training and test splits), motivating its selection as the fixed shifted position in our experiments.

least common correct answer, allows us to mitigate biases and better stress-test the robustness of models under fixed-position conditions.

D Inferential Difficulty in *Options Only*

To clarify why *Options Only* inference is substantially harder in medical MCQA than in general-domain benchmarks, Figure 15 contrasts representative examples from ARC-Challenge and MedQA. These examples illustrate why *Options Only* accuracy should be interpreted differently in the medical domain. Unlike prior analyses focused on general benchmarks (Balepur et al., 2024), medical MCQs pair long clinical vignettes with nuanced diagnostic or treatment options, sharply limiting the feasibility of question reconstruction from answer choices alone. This distinction motivates re-evaluating *Options Only* behavior when analyzing modern, reasoning LLMs in medical settings.

Tables 12 and 13 show cases where GPT-5-mini answers correctly even when the question text is

omitted, relying solely on the answer options. This behavior points to the exploitation of statistical regularities—due to shallow pattern recognition or mild data leakage—allowing models to appear correct without true reasoning. Table 7 illustrates accuracy scores across perturbations under *Options Only* modality, for each subset of ReMedQA.

E Analysis of Single-Flip Inconsistencies

We now analyze cases where models change their prediction exactly once across the seven input variants. Figure 10 reports the distribution of perturbation types responsible for these single-flip inconsistencies, attributing each question to the format variant that triggers the answer change. Across models and datasets, most single-flip inconsistencies are caused by the “None Provided”, “Select Incorrect”, and “Open” formats, indicating that deviations from the standard multiple-choice structure disproportionately affect prediction stability. To further characterize the nature of these failures, Figure 9 complements the perturbation-type analysis by quantifying the *severity* of single-flip inconsistencies. Severity is defined based on the degree of majority agreement across the seven perturbations: *minor* inconsistencies correspond to strong agreement ($\geq 80\%$, e.g., a 6:1 split), *moderate* inconsistencies reflect weaker majorities (60–80%, e.g., 5:2), and *major* inconsistencies indicate near-balanced predictions ($< 60\%$, e.g., 4:3). Figure 9 shows that the majority of single-flip inconsistencies fall into the minor category, where one prediction deviates while the remaining perturbations remain consistent. This trend holds across datasets and models, suggesting that most failures arise from isolated format sensitivities rather than sys-

tematic instability. However, smaller and medically specialized models exhibit a higher proportion of moderate and major inconsistencies—particularly on MedMCQA—indicating that when these models fail, disagreements are more evenly distributed across perturbations. In contrast, larger frontier models concentrate almost exclusively in the minor regime, suggesting that model scale mitigates not only the frequency but also the severity of format-induced prediction shifts.

F Further Results

Accuracy We provide a fine-grained analysis of model performance in terms of accuracy across MCQA perturbations. Tables 5 and 6 report accuracy scores for large and small models, respectively, while Table 7 presents results for the *Options Only* setting. Together, these tables detail model behavior under each perturbation and report average accuracy both across perturbations for each dataset and across datasets for each perturbation, offering a comprehensive view of model sensitivity.

Pair-Wise Consistency Figure 7 presents a comprehensive agreement grid for five models, illustrating answer consistency across pair-wise question perturbations over all ReMedQA datasets. Each cell reports the percentage of questions for which two perturbation formats produce identical predictions for a given model and dataset.

Figure 8 shows reliability across perturbation types, reporting the percentage of samples where a model preserves its original MCQA prediction. This complements Figure 5 in the main paper, where consistency is instead measured relative to the open-ended answer. Across models and datasets, the “Roman Numeral” format is often the most stable, as it preserves the underlying option mapping (e.g., A→I, B→II), allowing models to reason over equivalent symbolic representations. The “None Provided” perturbation is again consistently the most disruptive, inducing the largest consistency drops (e.g., GPT-5-mini falls below 60% on MedMCQA, while MedGemma drops below 35%). A clear split emerges between “Open” and “Select Incorrect”: smaller models are more consistent in the “Open” setting, while larger models handle task inversion more reliably, indicating that smaller systems struggle to invert their original decision. Overall, larger reasoning models (GPT-5-mini, GPT-OSS-120B) remain the most consistent, whereas smaller and domain-specialized models

(MediPhi, MedGemma) exhibit steeper drops, reinforcing that robustness scales with model capacity rather than medical tuning.

Impact of Removing the “None Provided” Perturbation The analyses in the main paper, as well as in Section E, show that model inconsistencies are most frequently triggered by the “None Provided” perturbation, which disproportionately penalizes both ReAcc and ReCon. To better isolate model behavior under less adversarial yet still semantic-preserving conditions, we report in Table 8 the results excluding this configuration. Although all models benefit substantially in terms of both reliability metrics, overall scores remain far from the accuracy peaks observed in the easiest MCQA variants. In particular, small models still struggle, with ReAcc often below 30% and ReCon below 40%, highlighting persistent difficulties even under comparatively simple perturbations. Interestingly, removing “None Provided” alters relative model rankings: Llama-3.3-70B surpasses GPT-OSS-20B, and Gemini-2.5-Flash outperforms GPT-OSS-120B despite a smaller reasoning budget. This suggests that the “None Provided” setting requires high reasoning capabilities. Similarly, medically specialized models consistently outperform their base counterparts, suggesting that medical training may induce a bias toward assuming the existence of a correct answer—an assumption that is explicitly violated by the “None Provided” setting.

Perturbation	MedQA	MedMCQA	<i>Pro. Med.</i>	<i>College Med.</i>	<i>Anatomy</i>	<i>Clinical</i>	<i>Biology</i>	<i>Genetics</i>	MMLU Avg	Avg
GPT-5-mini 🧠										
Open	90.1	75.8	94.1	86.4	88.8	89.1	93.5	96.3	91.4	85.8
Standard	94.4	82.5	97.2	90.2	91.2	93.8	98.1	98.8	94.9	90.6
Select Incorrect	93.9	81.5	96.9	89.4	91.2	92.7	99.1	100.0	94.9	90.1
Roman Numerals	90.3	67.8	94.5	77.3	80.0	76.2	86.2	91.5	84.3	80.8
Fixed Pos.	94.5	81.4	96.1	88.6	92.0	91.2	99.1	98.8	94.3	90.1
No Labels	93.1	80.9	96.5	90.9	90.4	90.2	97.2	98.8	94.0	89.3
None Provided	72.4	45.0	80.3	68.9	68.8	69.4	74.3	87.8	74.9	64.1
Avg	89.8	73.6	93.7	84.5	86.1	86.1	92.5	96.0	89.8	84.4
Gemini-2.5-Flash 🧠										
Open	88.5	74.3	94.5	87.0	88.8	90.2	94.5	91.5	91.1	84.6
Standard	91.4	82.1	94.9	90.1	89.6	90.7	98.2	96.3	93.3	88.9
Select Incorrect	83.0	76.0	83.9	81.8	86.4	85.5	92.7	95.1	87.6	82.2
Roman Numerals	92.5	81.3	94.9	89.4	88.0	89.6	99.1	98.8	93.3	89.0
Fixed Pos.	91.2	80.5	95.7	88.6	88.8	90.7	99.1	96.3	93.2	88.3
No Labels	89.6	79.1	93.3	78.0	87.2	88.6	91.7	97.6	89.4	86.0
None Provided	58.1	41.6	71.7	61.4	61.6	54.9	68.8	76.8	65.9	55.2
Avg	84.9	73.6	89.8	82.3	84.3	84.3	92.0	93.2	87.7	82.0
GPT-OSS-120B 🧠										
Open	89.6	69.8	92.9	90.2	84.0	87.6	92.7	95.1	90.4	83.3
Standard	91.8	76.6	95.7	90.2	88.0	86.5	99.1	98.8	93.0	87.1
Select Incorrect	90.4	75.3	94.5	90.2	86.4	87.0	97.2	98.8	92.4	86.0
Roman Numerals	88.6	69.5	92.5	81.1	80.0	80.8	93.6	96.3	87.4	81.8
Fixed Pos.	92.0	74.7	95.7	87.9	85.6	89.1	98.2	97.6	92.4	86.4
No Labels	88.3	73.6	89.8	84.8	83.2	85.0	94.5	96.3	88.9	83.6
None Provided	73.0	43.5	80.3	68.9	66.4	65.3	76.1	86.6	73.9	63.5
Avg	87.7	69.0	91.6	84.8	81.9	83.0	93.1	95.6	88.3	81.7
GPT-OSS-20B 🧠										
Open	83.9	65.2	89.7	84.7	75.2	82.9	93.6	93.8	86.6	78.6
Standard	85.8	70.8	93.7	88.6	81.6	84.5	97.2	96.3	90.3	82.3
Select Incorrect	84.2	68.3	93.7	88.6	79.2	85.0	91.7	95.1	88.9	80.5
Roman Numerals	85.6	61.8	92.5	79.5	80.0	72.5	86.2	90.2	83.5	77.0
Fixed Pos.	87.0	67.7	94.9	87.9	81.6	82.9	91.7	93.9	88.8	81.2
No Labels	81.7	66.5	91.3	78.8	84.0	80.3	93.6	87.8	86.0	78.1
None Provided	61.9	34.6	74.4	57.6	64.0	59.1	68.8	76.8	66.8	54.4
Avg	81.4	62.1	90.0	80.8	77.9	78.2	89.0	90.6	84.4	76.0
Llama-3.3-70B										
Open	83.1	68.3	92.1	84.1	85.6	85.5	92.7	89.0	88.2	79.9
Standard	83.1	74.0	93.7	84.8	85.6	87.0	93.6	95.1	90.0	82.4
Select Incorrect	73.7	65.2	89.8	70.5	80.0	78.8	86.2	89.0	82.4	73.8
Roman Numerals	83.9	75.5	95.7	83.3	84.8	87.0	93.6	93.9	89.7	83.0
Fixed Pos.	84.6	76.1	95.7	86.4	84.8	88.1	95.4	96.3	91.1	83.9
No Labels	82.1	72.5	94.1	82.6	84.0	85.5	91.7	93.9	88.6	81.1
None Provided	41.7	29.3	56.3	44.7	51.2	48.2	60.6	68.3	54.9	42.0
Avg	76.0	65.8	88.2	76.6	79.4	80.0	87.7	89.4	83.6	75.2

Table 5: Accuracy of large models across perturbations and each subset of ReMedQA. 🧠 = Reasoning models.

Perturbation	MedQA	MedMCQA	<i>Pro. Med.</i>	<i>College Med.</i>	<i>Anatomy</i>	<i>Clinical</i>	<i>Biology</i>	<i>Genetics</i>	MMLU Avg	Avg
Med42-Llama3-8B ♥										
Open	74.4	58.6	85.0	76.2	76.0	77.7	89.9	92.7	82.9	72.0
Standard	62.9	59.9	74.4	68.5	72.0	69.9	77.1	78.0	73.3	65.4
Select Incorrect	45.8	42.3	49.6	47.7	55.2	56.0	55.0	63.4	54.5	47.5
Roman Numerals	63.7	59.6	78.3	65.9	65.6	71.0	80.7	78.0	73.2	65.5
Fixed Pos.	62.0	56.5	78.7	68.9	64.0	73.1	80.7	80.5	74.3	64.3
No Labels	60.7	55.2	76.8	64.4	68.8	66.8	74.3	69.5	70.1	62.0
None Provided	26.9	27.5	30.3	34.1	28.0	33.7	31.2	45.1	33.7	29.4
Avg	56.6	51.4	67.6	60.8	61.4	64.0	69.8	72.5	66.0	58.0
Llama-3-8B										
Open	72.2	58.0	85.0	78.8	72.0	79.3	87.2	87.8	81.7	70.6
Standard	61.4	60.7	75.6	71.2	72.0	77.2	73.4	79.3	74.8	65.6
Select Incorrect	28.9	27.5	28.0	37.1	35.2	36.3	35.8	50.0	37.1	31.2
Roman Numerals	61.7	58.6	74.8	65.9	70.4	72.5	77.1	80.5	73.5	64.6
Fixed Pos.	58.3	59.3	71.7	66.7	61.6	74.6	75.2	76.8	71.1	62.9
No Labels	58.5	55.9	75.2	62.9	61.6	69.4	67.0	70.7	67.8	60.7
None Provided	24.1	30.7	31.5	31.1	32.8	36.3	33.9	47.6	35.5	30.1
Avg	52.2	50.1	63.1	59.1	57.9	63.7	64.2	70.4	63.1	55.1
MedGemma-4B ♥										
Open	73.0	56.2	84.6	73.5	76.0	77.7	87.2	84.1	80.5	69.9
Standard	65.7	57.8	78.0	58.3	67.2	68.9	73.4	76.8	70.4	64.6
Select Incorrect	42.5	34.6	50.8	35.6	44.0	40.4	47.7	51.2	44.9	40.7
Roman Numerals	64.5	57.9	79.5	58.3	62.4	66.3	69.7	72.0	68.0	63.5
Fixed Pos.	62.8	52.2	75.6	59.8	57.6	65.8	65.1	68.3	65.4	60.1
No Labels	60.4	56.0	72.0	59.1	58.4	63.7	65.1	65.9	64.0	60.1
None Provided	4.7	4.8	5.9	4.5	7.2	7.3	10.1	14.6	8.3	5.9
Avg	53.4	45.6	63.8	49.9	53.3	55.7	59.8	61.8	57.4	52.1
Gemma-3-4B										
Open	67.5	53.5	82.3	78.8	74.4	79.2	83.5	91.2	81.6	67.5
Standard	52.5	48.2	69.7	60.6	60.0	67.2	72.5	67.5	66.2	55.6
Select Incorrect	36.5	35.5	42.1	36.4	41.6	48.2	53.2	58.5	46.7	39.6
Roman Numerals	53.0	49.6	67.3	60.6	53.6	68.9	71.6	69.5	65.2	56.0
Fixed Pos.	50.8	46.1	64.2	58.3	50.4	62.7	69.7	63.4	61.4	52.8
No Labels	51.4	47.5	63.0	61.4	55.2	63.7	69.7	57.3	61.7	53.5
None Provided	11.1	13.3	14.6	19.7	19.2	16.6	20.2	34.1	20.7	15.0
Avg	46.1	42.0	57.6	53.7	50.6	58.1	62.9	63.1	57.6	48.6
MediPhi-3.8B ♥										
Open	72.3	58.4	88.6	79.5	79.2	82.9	90.8	92.7	85.6	72.1
Standard	57.1	56.7	72.0	68.2	69.6	74.6	77.1	81.7	73.9	62.6
Select Incorrect	46.4	39.0	56.7	49.2	54.4	56.5	56.0	64.6	56.2	47.2
Roman Numerals	58.5	57.0	72.0	68.9	65.6	74.1	83.5	81.7	74.3	63.3
Fixed Pos.	56.5	48.2	72.8	62.9	55.2	69.4	75.2	78.0	68.9	57.9
No Labels	55.0	50.2	66.9	57.6	56.0	63.7	73.4	67.1	64.1	56.4
None Provided	22.4	22.3	24.0	28.0	25.6	29.0	33.9	42.7	30.5	25.1
Avg	52.6	47.4	64.7	59.2	57.9	64.3	70.0	72.6	64.8	54.9
Phi-3.5-mini										
Open	73.4	61.1	86.6	78.0	77.6	80.8	90.8	95.1	84.8	73.1
Standard	60.4	56.6	76.4	72.7	66.4	78.2	78.9	81.7	75.7	64.2
Select Incorrect	41.8	31.7	57.1	45.5	48.8	51.3	55.0	62.2	53.3	42.3
Roman Numerals	60.5	57.1	76.4	72.7	75.2	78.8	85.3	80.5	78.2	65.2
Fixed Pos.	60.8	57.3	76.0	74.2	62.4	76.2	77.1	84.1	75.0	64.4
No Labels	54.7	52.0	70.9	44.7	52.0	51.8	66.1	51.2	56.1	54.3
None Provided	32.4	32.9	34.6	36.4	42.4	38.9	50.5	53.7	42.8	36.0
Avg	54.9	49.8	68.3	60.6	60.7	65.1	72.0	72.6	66.6	57.1

Table 6: Accuracy of small models across perturbations and each subset of ReMedQA. ♥ = Medical-specialized.

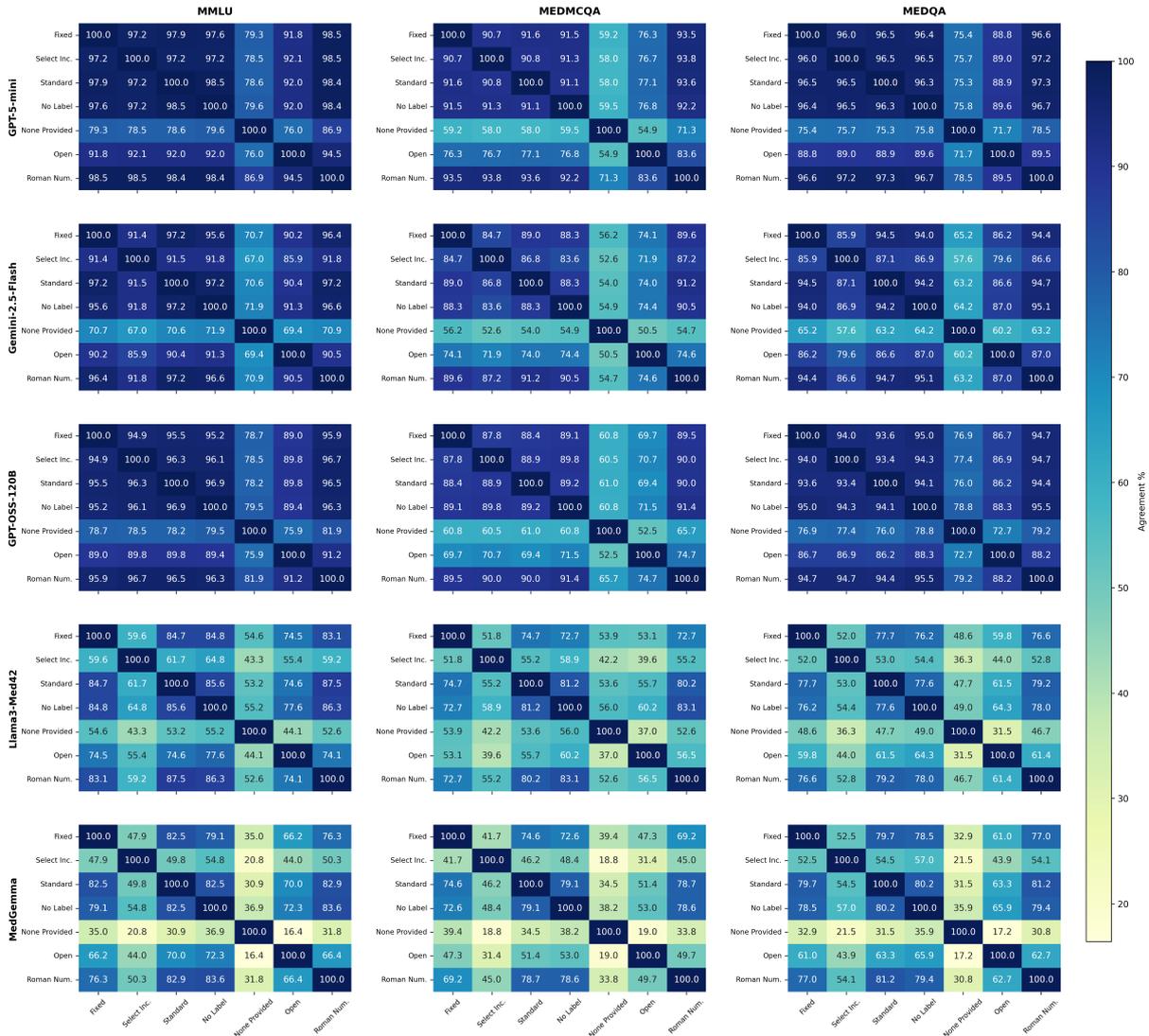


Figure 7: **Agreement grid across models and datasets.** Pair-wise agreement matrices illustrating answer consistency across different question perturbation modes for five selected models (GPT-5-Mini, Gemini-2.5-Flash, GPT-OSS-120B, MedGemma-4B, Med42-Llama3-8B) on ReMedQA subsets (MMLU, MedQA, and MedMCQA).

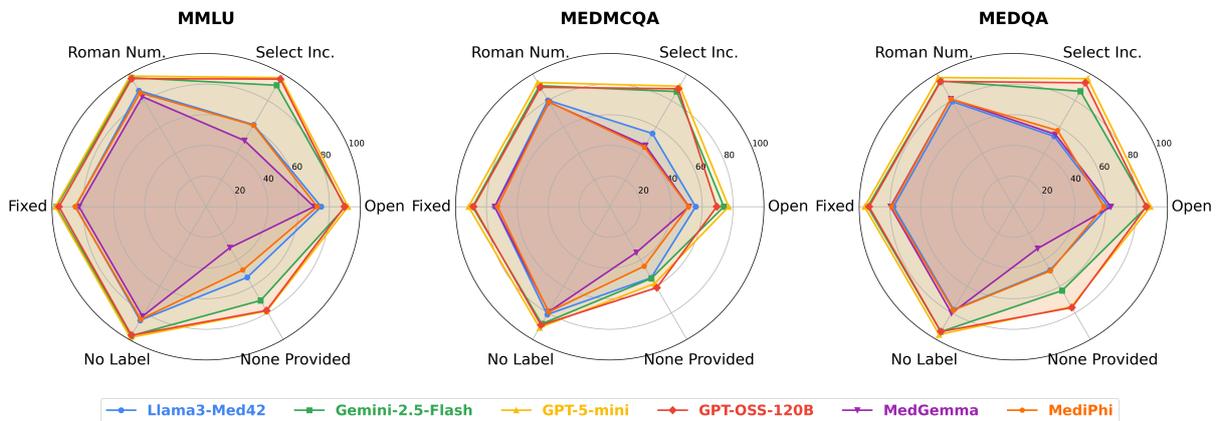


Figure 8: **Pair-wise consistency with standard MCQA answers across perturbations.** Each radar plot reports the percentage of predictions consistent with the original MCQA answer under different perturbation formats.

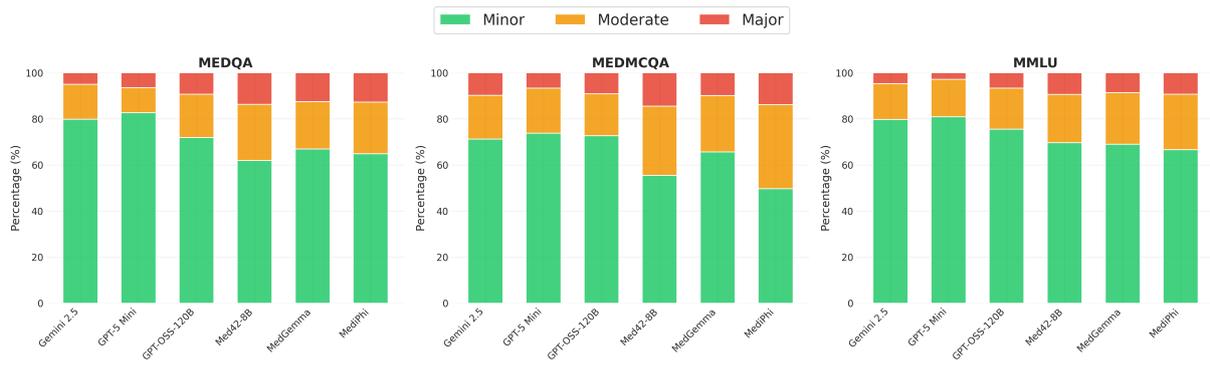


Figure 9: **Inconsistency severity.** Severity of single-flip inconsistencies (two unique answers across seven perturbations), categorized by majority agreement: *minor* ($\geq 80\%$, e.g., 6:1), *moderate* (60–80%, e.g., 5:2), and *major* ($< 60\%$, e.g., 4:3).

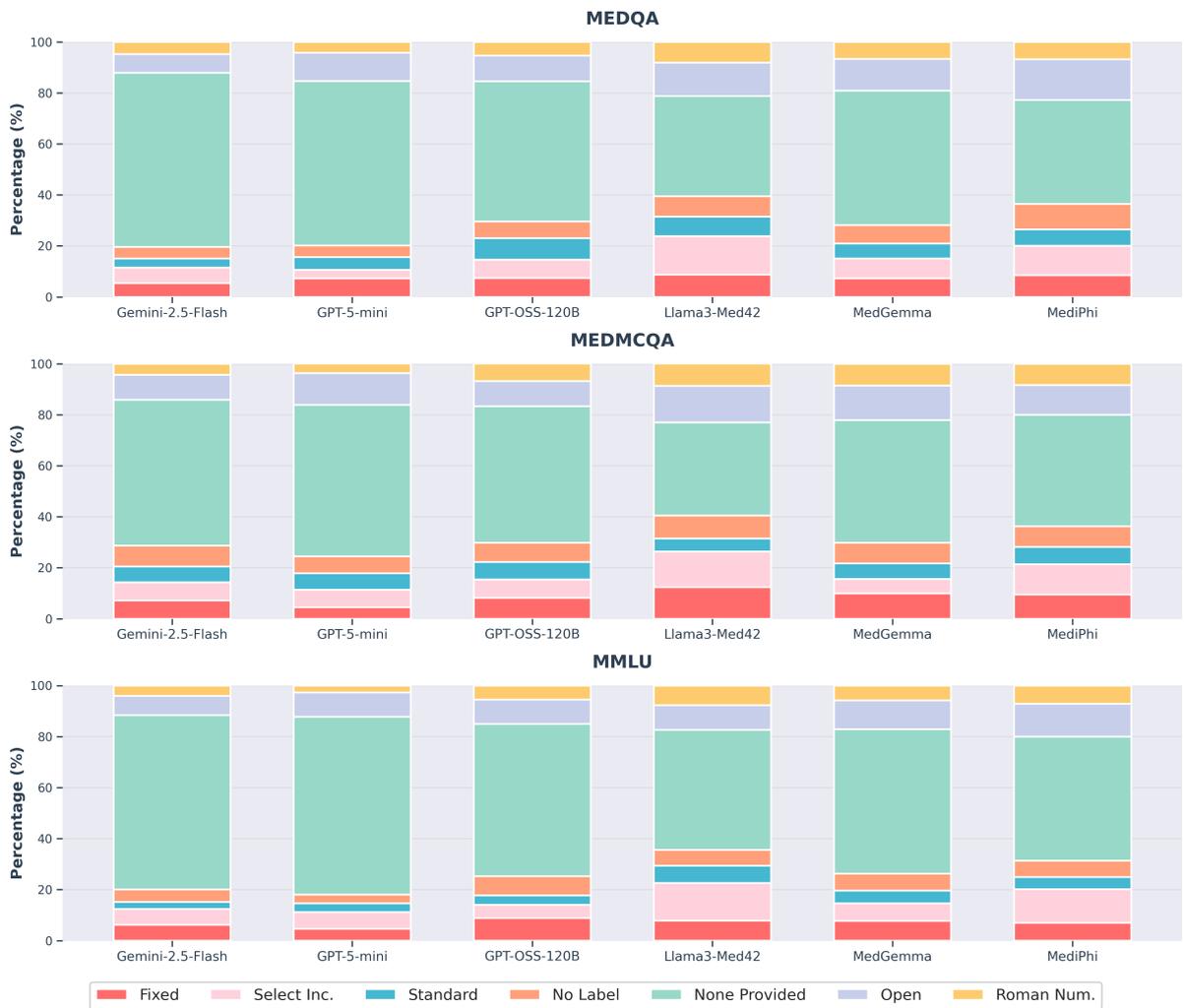


Figure 10: **Inconsistency attribution by perturbation type.** Distribution of perturbation modes responsible for single-flip inconsistencies (i.e., cases with exactly two distinct answers) across models, aggregated over the evaluated datasets. Each stacked bar represents the total percentage of inconsistent questions for a given model and dataset, with segments showing the contribution of each mode. Higher bars indicate more overall inconsistency.

Perturbation	MedQA	MedMCQA	Pro. Med.	College Med.	Anatomy	Clinical	Biology	Genetics	MMLU Avg	Avg
<i>Options Only Evaluation</i>										
GPT-5-mini										
Standard	44.1	41.4	42.5	49.2	48.8	61.1	60.6	48.8	51.8	45.8
Select Incorrect	42.3	37.2	43.3	49.2	45.6	49.2	64.2	48.8	50.1	43.2
Roman Numerals	35.1	33.0	40.6	44.7	43.2	45.6	42.2	37.8	42.4	36.8
Fixed Pos.	39.6	36.3	40.6	44.7	46.4	49.2	54.1	45.1	46.7	40.9
No Labels	42.2	42.7	46.9	50.8	46.4	56.5	56.9	45.1	50.4	45.1
None Provided	3.5	3.5	3.9	12.9	12.0	8.3	12.8	6.1	9.3	5.4
Avg	34.5	32.3	36.3	41.9	40.4	45.0	48.5	38.6	41.8	36.2
Gemini-2.5-Flash										
Standard	38.8	39.0	47.2	53.0	52.0	57.5	54.1	48.8	52.1	43.3
Select Incorrect	36.1	35.0	40.6	43.9	48.0	57.5	50.5	37.8	46.4	39.2
Roman Numerals	39.9	41.4	40.9	50.0	50.4	56.5	56.0	40.2	49.0	43.4
Fixed Pos.	42.7	41.0	44.1	52.3	45.6	54.9	58.7	56.1	52.0	45.2
No Labels	37.5	39.8	42.1	46.2	49.6	53.4	57.8	40.2	48.2	41.8
None Provided	0.8	1.2	1.6	3.0	8.8	3.1	7.3	3.7	4.6	2.2
Avg	32.6	32.9	36.1	41.4	42.4	47.2	47.4	37.8	42.0	35.8
GPT-OSS-120B										
Standard	39.0	36.6	40.2	51.5	51.2	53.4	50.5	43.9	48.4	41.3
Select Incorrect	35.7	35.4	40.6	52.3	46.4	55.4	57.8	39.0	48.6	39.9
Roman Numerals	38.5	34.8	42.1	50.8	53.6	51.8	53.2	39.0	48.4	40.6
Fixed Pos.	40.3	35.1	42.1	46.2	50.4	50.3	52.3	42.7	47.3	40.9
No Labels	38.8	37.7	42.5	44.7	44.8	57.0	63.3	46.3	49.8	42.1
None Provided	14.5	14.2	12.6	20.5	21.6	19.7	24.8	15.9	19.2	16.0
Avg	34.5	32.3	36.7	44.3	44.7	47.9	50.3	37.8	43.6	36.8

Table 7: Accuracy of reasoning LLMs across perturbations and subsets of ReMedQA, under *Options Only* evaluation.

ReMedQA	Pro Med.		College Med.		Anatomy		Clinical		Biology		Genetics		MMLU Avg		MedQA		MedMCQA		Avg		
	RA	RC	RA	RC	RA	RC	RA	RC	RA	RC	RA	RC	RA	RC	RA	RC	RA	RC	RA	RC	Acc
Large Models																					
GPT-5-mini 🧠	89.0	91.7	69.7	85.6	75.2	90.4	71.0	86.0	81.7	93.5	89.0	95.1	79.3	90.4	81.9	86.7	55.9	71.0	72.4	82.7	87.8
Gemini-2.5-Flash 🧠	74.4	77.5	65.2	78.6	74.4	86.4	72.0	80.8	80.7	89.9	87.8	89.0	75.8	83.7	71.6	75.7	57.8	65.0	<u>68.4</u>	74.8	<u>86.5</u>
GPT-OSS-120B 🧠	83.5	90.2	72.7	82.6	63.2	81.6	69.9	82.4	83.5	91.7	86.6	91.5	76.6	86.7	74.8	81.8	52.1	64.4	67.8	<u>77.6</u>	84.7
Llama-3.3-70B	84.3	87.7	62.1	72.0	70.4	75.2	64.8	70.5	78.0	79.8	82.9	84.1	73.8	78.2	61.0	67.5	47.0	54.7	60.6	66.8	80.7
GPT-OSS-20B 🧠	78.7	82.9	65.2	77.1	59.2	68.8	60.1	72.5	79.8	87.2	75.6	88.9	69.8	79.6	66.2	72.7	41.5	48.4	59.2	66.9	79.6
Small Models																					
Llama-3-8B	20.9	43.3	27.3	40.9	24.8	52.0	28.0	45.6	26.6	54.1	31.7	57.3	26.5	48.9	15.3	28.9	16.2	31.8	19.3	36.5	59.3
Llama3-Med42-8B 🧡	36.2	41.3	29.5	39.2	40.0	47.2	36.8	45.6	42.2	48.6	40.2	53.7	37.5	45.9	24.9	30.1	25.0	27.9	29.1	<u>34.6</u>	62.8
Phi-3.5-mini	44.1	48.0	18.9	39.4	32.8	44.0	31.6	45.6	41.3	46.8	29.3	53.7	33.0	46.2	25.0	28.7	19.4	24.0	25.8	33.0	60.6
MediPhi-3.8B 🧡	39.0	44.9	29.5	38.6	36.8	45.6	37.3	43.0	45.9	49.5	43.9	51.2	38.7	45.5	25.8	32.3	18.6	23.8	<u>27.7</u>	33.9	59.9
Gemma-3-4B	26.4	31.1	24.2	31.8	29.6	35.2	32.6	37.0	37.6	47.7	36.6	41.2	31.2	37.3	16.8	21.6	14.5	19.3	20.8	26.1	54.2
MedGemma-4B 🧡	37.4	42.9	24.2	31.1	27.2	33.6	24.4	30.1	31.2	38.5	29.3	35.4	28.9	35.3	26.5	33.9	19.2	23.3	24.9	30.8	59.8

Table 8: **Performance on ReMedQA using ReAcc (RA), ReCon (RC), and Accuracy (Acc), excluding the most difficult perturbation “None Provided”.** Values in columns with bolded headers are used to compute RA and RC averages. Reported Acc is obtained by averaging accuracy across all perturbation variants, excluding “None Provided” (see Tables 5 and 6 in Appendix). Dashed lines group base models with their medical-specialization (marked with 🧡). 🧠 = Reasoning models. Best scores are in bold; second-best are underlined.

Mode	Prompt
Standard	The following are multiple choice questions about medical knowledge. Solve them in a step-by-step fashion, starting by summarizing the available information. Output a single option from the four options as the final answer. Question: “<QUESTION>” Response (think step by step and then end with “Final Answer:” followed by *only* the letter corresponding to the correct answer enclosed in parentheses)
Open	The following are open-ended questions about medical knowledge. Solve them in a step-by-step fashion, starting by summarizing the available information. Output a single, concise final answer (not a letter). Question: “<QUESTION>” Response (think step by step and then end with “Final Answer:” followed by *only* the concise answer)
No Labels	The following are multiple choice questions about medical knowledge. Solve them in a step-by-step fashion, starting by summarizing the available information. Output a single option from the four options as the final answer. Question: “<QUESTION>” Response (think step by step and then end with “Final Answer:” followed by *only* the text of the correct option, without any letter or symbol)
Roman Numerals	The following are multiple choice questions about medical knowledge. Solve them in a step-by-step fashion, starting by summarizing the available information. Output a single option from the four options as the final answer. Question: “<QUESTION>” Response (think step by step and then end with “Final Answer:” followed by *only* the roman numeral corresponding to the correct answer enclosed in parentheses)
Fixed Position	The following are multiple choice questions about medical knowledge. Solve them in a step-by-step fashion, starting by summarizing the available information. Output a single option from the four options as the final answer. Question: “<QUESTION>” Response (think step by step and then end with “Final Answer:” followed by *only* the letter corresponding to the correct answer enclosed in parentheses)
Select Incorrect	The following are multiple choice questions about medical knowledge. Solve them in a step-by-step fashion, starting by summarizing the available information. Answer the given question by identifying the three incorrect options. Question: “<QUESTION>” Response (think step by step and then end with “Final Answer:” followed by *only* the list of letters corresponding to the incorrect answers, e.g., “Final Answer: [A, C, D]”)
None Provided	The following are multiple choice questions about medical knowledge. Solve them in a step-by-step fashion, starting by summarizing the available information. Output a single option from the four options as the final answer. Select the option “None of the provided options” if you think that none of the other options is correct. Question: “<QUESTION>” Response (think step by step and then end with “Final Answer:” followed by *only* the letter corresponding to the correct answer enclosed in parentheses)

Table 9: Prompts adopted for non-reasoner models.

Mode	Prompt
Standard	You are given a multiple choice question about medical knowledge. Answer by returning the correct option's letter. Question: "<QUESTION>" Enclose your final answer within \boxed{ }.
Open	You are given an open-ended question about medical knowledge. Answer by returning a concise answer. Question: "<QUESTION>" After you have finished your thinking process, please show your response with **only** the concise final answer, e.g., "Final Answer: <your concise answer>".
No Labels	You are given a multiple choice question about medical knowledge. Answer by returning the correct answer. Question: "<QUESTION>" After you have finished your thinking process, please show your response with *only* the text of the correct answer, without any letter or symbol, e.g., "Final Answer: <text of the correct answer>".
Roman Numerals	You are given a multiple choice question about medical knowledge. Answer by returning the correct option's roman numeral. Question: "<QUESTION>" Enclose your final answer within \boxed{ }.
Fixed Position	You are given a multiple choice question about medical knowledge. Answer by returning the correct option's letter. Question: "<QUESTION>" Enclose your final answer within \boxed{ }.
Select Incorrect	You are given a multiple choice question about medical knowledge. Answer by returning the three incorrect option letters, separated by commas. Question: "<QUESTION>" Enclose your final answer within \boxed{ }.
None Provided	You are given a multiple choice question about medical knowledge. Answer by returning the correct option's letter. Question: "<QUESTION>" Enclose your final answer within \boxed{ }.

Table 10: Prompts adopted for reasoner models.

Mode	Prompt
Options Only Standard	<p>You will be shown only the answer options for multiple-choice questions from the <DATASET> test set. The original question text is intentionally hidden. You were trained on this dataset, so you should recall the correct answer from memory. Otherwise, you should guess. Answer by returning the correct option's letter.</p> <p>Original question: <the text of the question is hidden> Options: <OPTIONS></p> <p>Enclose your final answer within \boxed{ }.</p>
Options Only No Labels	<p>You will be shown only the answer options for multiple-choice questions from the <DATASET> test set. The original question text is intentionally hidden. You were trained on this dataset, so you should recall the correct answer from memory. Otherwise, you should guess. Answer by returning the correct answer.</p> <p>Original question: <the text of the question is hidden> Options: <OPTIONS></p> <p>After you have finished your thinking process, please show your response with <i>*only*</i> the text of the correct answer, without any letter or symbol, e.g., "Final Answer: <text of the correct answer>".</p>
Options Only Roman Numerals	<p>You will be shown only the answer options for multiple-choice questions from the <DATASET> test set. The original question text is intentionally hidden. You were trained on this dataset, so you should recall the correct answer from memory. Otherwise, you should guess. Answer by returning the correct option's roman numeral.</p> <p>Original question: <the text of the question is hidden> Options: <OPTIONS></p> <p>Enclose your final answer within \boxed{ }.</p>
Options Only Fixed Position	<p>You will be shown only the answer options for multiple-choice questions from the <DATASET> test set. The original question text is intentionally hidden. You were trained on this dataset, so you should recall the correct answer from memory. Otherwise, you should guess. Answer by returning the correct option's letter.</p> <p>Original question: <the text of the question is hidden> Options: <OPTIONS></p> <p>Enclose your final answer within \boxed{ }.</p>
Options Only Select Incorrect	<p>You will be shown only the answer options for multiple-choice questions from the <DATASET> test set. The original question text is intentionally hidden. You were trained on this dataset, so you should recall the correct answer from memory. Otherwise, you should guess. Answer by returning the three incorrect option letters, separated by commas.</p> <p>Original question: <the text of the question is hidden> Options: <OPTIONS></p> <p>Enclose your final answer within \boxed{ }.</p>
Options Only None Provided	<p>You will be shown only the answer options for multiple-choice questions from the <DATASET> test set. The original question text is intentionally hidden. You were trained on this dataset, so you should recall the correct answer from memory. Otherwise, you should guess. Answer by returning the correct option's letter.</p> <p>Original question: <the text of the question is hidden> Options: <OPTIONS></p> <p>Enclose your final answer within \boxed{ }.</p>

Table 11: Prompts adopted for *Options Only* mode.

Dataset	Original Question (Hidden) + Options	Answer	Comment
MedQA	<p>A 3-week-old male newborn is brought to the physician because of an inward turning of his left forefoot. He was born at 38 weeks' gestation by cesarean section because of breech presentation. The pregnancy was complicated by oligohydramnios. Examination shows concavity of the medial border of the left foot with a skin crease just below the ball of the great toe. The lateral border of the left foot is convex. The heel is in neutral position. Tickling the lateral border of the foot leads to correction of the deformity. The remainder of the examination shows no abnormalities. X-ray of the left foot shows an increased angle between the 1st and 2nd metatarsal bones. Which of the following is the most appropriate next step in the management of this patient?</p> <p>(A) Foot abduction brace (B) Arthrodesis of the forefoot (C) Reassurance (D) Tarsometatarsal capsulotomy</p>	<p>I'm recalling a MedQA USMLE question with options like foot abduction brace, arthrodesis of the forefoot, reassurance, and tarsometatarsal capsulotomy. I think this likely relates to the treatment of metatarsus adductus versus clubfoot. The foot abduction brace is used for clubfoot (Ponseti method), while reassurance is typically for flexible metatarsus adductus, which usually resolves on its own. Given a baby with flexible inward deviation of the forefoot, it seems the answer is C: reassurance.</p>	<p>The model reconstructs the topic (metatarsus adductus) without reading the question: "I think this likely relates to the treatment of metatarsus adductus versus clubfoot". The reasoning shows context reconstruction from the options and retrieves the correct clinical interpretation despite the question being hidden: "Reassurance is typically for flexible metatarsus adductus, which usually resolves on its own". Both reasoning and answer match the ground truth.</p>
MedMCQA	<p>Alginat at 20 degrees, the gel is formed in: (A) 3-4 mins (B) 6-8 mins (C) 15-16 mins (D) 1-2 min</p>	<p>**Deciding on inhaled medication effects** I'm thinking about the best timing for maximum effect of inhaled beta2 agonists. I'm not entirely sure, but it reminds me of some training data regarding medication. I recall a specific question about how long suxamethonium lasts—it's about 3-4 minutes. So, I'm going to choose option A based on that. It feels like a logical decision in this context!</p>	<p>In this case, its thought process clearly shows it has no understanding of the actual question and is just pattern-matching from its training data.</p>
MMLU (Anatomy)	<p>In relation to the rib, the corresponding intercostal nerve lies: (A) deep to its superior border. (B) superficial to its superior border. (C) deep to its inferior border. (D) superficial to its inferior border.</p>	<p>**Clarifying anatomical positioning** The subclavian artery is indeed located deep to the middle third of the clavicle, which is the same as being deep to its superior border. I realize that the superior border of the clavicle faces upward, while the inferior border faces downward. So, if something is beneath the clavicle, it is considered deep to that superior surface. Given this perspective, I conclude that the subclavian artery lies posterior and inferior to the clavicle, leading me to choose answer (A).</p>	<p>The model managed to produce an answer and thinking by successfully answering a different question that it assumed was the correct one. It applied logical reasoning (though incorrect for the subclavian artery's precise location) to this unseen, assumed question and mapped the result to the provided options.</p>

Table 12: **Example of GPT-5-mini producing a correct answer without access to the original question.** The question (highlighted in red) was omitted within the input prompt to investigate possible data contamination.

Mode	Original Question (Hidden) + Options	Answer	Comment
Options Only Select Incorrect	<p>A 4-year-old previously healthy boy presents with 4 days of intermittent vomiting and 5-6 daily loose stools. His mother noted bloody stools and decreased oral intake of food and water over the last 24 hours. He is normally in daycare; however, he has been home for the past 3 days. On physical exam his temperature is 102.2F (39C), blood pressure is 140/90 mmHg, pulse is 120/min, respirations are 22/min and O2 saturation is 99% on room air. He has dry mucous membranes. On abdominal exam you note diffuse tenderness to palpation without rebound or guarding. There are no masses, hepatosplenomegaly, and bowel sounds are hyperactive. Ultrasound of the right lower quadrant is negative for appendicitis. Stool is guaiac positive. He receives 15mg/kg acetaminophen and fluids are started. The next day, he complains of lower extremity weakness and tingling. On repeat exam, lower extremity strength is 3/5 with diminished patellar deep tendon reflexes. Which of the following lab findings would most likely be seen in this patient?</p> <p>(A) Gram stain positive CSF (B) Peripheral eosinophilia (C) Xanthochromia on cerebrospinal fluid analysis (D) Increased cerebrospinal fluid protein with normal cell count</p>	<p>**Figuring out viral meningitis criteria** I'm considering the signs of viral meningitis, which include normal glucose, lymphocytes, and a negative Gram stain, along with normal or slightly elevated protein levels. There's no xanthochromia or peripheral eosinophilia involved, which complicates things a bit. MedQA often highlights that "albuminocytologic dissociation", which shows increased CSF protein with a normal cell count, is linked to Guillain-Barré syndrome. So, if D is the correct answer, then the wrong options would be A, B, and C.</p>	<p>The model demonstrated backward inference by recognizing option D's description of albuminocytologic dissociation (elevated CSF protein with normal cell count) as pathognomonic for Guillain-Barré syndrome. Using domain knowledge alone, it eliminated options A, B, and C by identifying they represent incompatible diagnoses (bacterial meningitis, parasitic infection, and subarachnoid hemorrhage respectively), successfully answering without the clinical vignette.</p>
Options Only None Provided	<p>A 56-year-old man is brought to the emergency room after a motor vehicle accident. The patient's vitals are as follows: blood pressure 80/40 mm Hg, heart rate 111/min, respiratory rate 39/min, and temperature 37.1C (98.8). On physical examination, the patient is unconscious with a GCS of 9/15 and is cyanotic. There are open fractures of the left femur and left tibia, a likely shoulder dislocation, multiple contusions on the limbs and thorax, and a puncture wound on the left side of his chest. There are no breath sounds on the left side and there is hyperresonance to percussion on the left. Preparations are made for an emergency needle thoracostomy to be performed to treat this patient's likely tension pneumothorax. Which one of the following is the best choice to provide informed consent for this procedure?</p> <p>(A) The patient's closest relatives (spouse, child or parent), who must be brought to the hospital as fast as possible (B) One of the witnesses at the scene of the car accident, who should be brought to the hospital as fast as possible (C) The hospital ethics committee (D) None of the provided</p>	<p>**Considering treatment consent** I need to think through the options regarding consent for treating an unconscious patient. Should I contact the closest relatives, a witness, or the ethics committee? However, it seems that I might not need any of those options. Under the emergency doctrine, treatment can be administered without consent, which means there's no requirement to involve relatives or witnesses. So, I think the answer should be D: None of the provided options.</p>	<p>The model correctly selected option D by applying the emergency doctrine principle in medical ethics. Recognizing the life-threatening nature of tension pneumothorax from clinical signs (hypotension, tachycardia, absent breath sounds, hyperresonance), it determined that immediate intervention is required without delay for informed consent. The model eliminated options A-C by reasoning that waiting for relatives, witnesses, or ethics committee approval would be inappropriate when emergency treatment is medically necessary to prevent imminent death.</p>

Table 13: **Example of GPT-5-mini producing a correct answer without access to the MedQA's original question in most challenging variants.** The question (highlighted in red) was omitted within the input prompt to investigate possible data contamination.

Dataset	Original Question + Options	Judge Explanation
MedQA	<p>A 59-year-old overweight woman presents to the urgent care clinic with the complaint of severe abdominal pain for the past 2 hours. She also complains of a dull pain in her back with nausea and vomiting several times. Her pain has no relation with food. Her past medical history is significant for recurrent abdominal pain due to cholelithiasis. Her father died at the age of 60 with some form of abdominal cancer. Her temperature is 37°C (98.6°F), respirations are 15/min, pulse is 67/min, and blood pressure is 122/98 mm Hg. Physical exam is unremarkable. However, a CT scan of the abdomen shows a calcified mass near her gallbladder. Which of the following diagnoses should be excluded first in this patient?</p> <p>(A) Acute cholecystitis (B) Gallbladder cancer (C) Choledocholithiasis (D) Pancreatitis</p>	The question uses “should be excluded first,” which is an exclusion-based phrasing and falls under rule 3A.
MedMCQA	<p>Biofilm forming bacteria causes antimicrobial resistance by all of the following except:</p> <p>(A) Mechanical barrier (B) Increased excretion of antibiotics (C) Altered metabolism (D) Adherence</p>	The question uses negation (“all of the following except”) and requires options to determine which mechanism is not involved, making it ambiguous without options.
MMLU	<p>Insensible loss does not come from:</p> <p>(A) pyrexia (B) using unhumidified oxygen (C) excessive weight loss (D) diarrhoea</p>	The question uses negation (“does not come from”) and requires the options to determine the correct answer.

Table 14: **Non-convertible MCQs.** Examples of multiple-choice questions that cannot be reformulated into open-ended form due to exclusion- or negation-based dependency on the provided options. Judge explanations are provided according to GPT-4.1.

Question	Answer	● Mapping Judge	◎ Validation Judge
<p>[...] Which of the following is the most appropriate treatment?</p> <p>(A) Erythromycin ointment (B) Ketotifen eye drops (C) Warm compresses (D) Fluorometholone eye drops</p>	<p>Reasoning: <i>Identifying allergic conjunctivitis</i> I need to recognize allergic conjunctivitis, especially seasonal types that cause itchy, watery eyes and sneezing, typically recurring in spring. Treatment options include topical antihistamines or mast cell stabilizers, like olopatadine, plus oral antihistamines for systemic symptoms. The most appropriate choice here would be topical antihistamine drops, specifically olopatadine. It's also important to avoid using topical steroids. So, my final answer is the topical antihistamine/mast cell stabilizer (olopatadine) ophthalmic drops. Final Answer: Topical antihistamine/mast cell stabilizer eye drops (eg, olopatadine ophthalmic drops) for seasonal allergic conjunctivitis.</p>	<p>Mapped Option: B Justification: The student's answer correctly identifies allergic conjunctivitis and recommends a topical antihistamine/mast cell stabilizer (e.g., olopatadine). Among the options, Ketotifen (B) is a mast cell stabilizer/antihistamine used for allergic conjunctivitis. Fluorometholone (D) is a steroid, which the student explicitly advises against. Erythromycin (A) is an antibiotic, and warm compresses (C) are irrelevant here.</p>	N/A
<p>Which of the following additional findings is most likely in this patient?</p> <p>(A) Silvery plaques on extensor surfaces (B) Flesh-colored papules in the lumbosacral region (C) Erosions of the dental enamel (D) Holosystolic murmur at the left lower sternal border</p>	<p>Reasoning: <i>Making assumptions from a missing image</i> I need to infer findings from an image, but since it's not provided, I have to guess common nail abnormalities for a young woman. For instance, koilonychia might suggest iron deficiency anemia, while pitting might indicate psoriasis. Terry's nails could signal liver problems, and yellow nails might be linked to respiratory issues. Given she feels embarrassed, clubbing could be possible, yet I lean toward koilonychia reflecting iron deficiency from heavy menstrual bleeding. My final answer is focused on low hemoglobin and ferritin linked to fatigue or pallor. Final Answer: Findings of iron-deficiency anemia — microcytic, hypochromic RBCs with low hemoglobin and low ferritin (e.g., pallor, fatigue).</p>	<p>Mapped Option: No clear match Justification: The student's answer focuses on lab findings (microcytic/hypochromic RBCs, low hemoglobin/ferritin) and symptoms (pallor, fatigue) of iron-deficiency anemia. However, the provided options (A-D) describe physical signs (e.g., silvery plaques, papules, enamel erosion, heart murmur) rather than anemia-related lab results or systemic symptoms. None of the options align with the student's conclusion.</p>	<p>Reasoning: The student explicitly states that the image is missing and attempts to infer common nail abnormalities for a young woman. Koilonychia (spoon nails) is a classic finding of iron deficiency anemia, which is a common condition in 23-year-old women (often due to heavy menstrual bleeding) and can cause nail changes leading to embarrassment. The additional findings listed (microcytic, hypochromic RBCs, low hemoglobin and ferritin, pallor, fatigue) are all highly relevant and correct findings associated with iron-deficiency anemia. Therefore, without the visual information, the student's deductive reasoning and proposed diagnosis are medically sound and represent a valid alternative interpretation of the question's textual information. Valid Alternative: Yes</p>
<p>[...] Which of the following anticoagulants did this patient most likely consume?</p> <p>(A) Warfarin (B) Enoxaparin (C) Fondaparinux (D) Apixaban</p>	<p>Reasoning: <i>Evaluating anticoagulant elimination</i> I need to think about the elimination rate of anticoagulants, aiming for 0.5 mg every 4 hours. Is this referring to heparin, warfarin, or maybe LMWH? Warfarin, with a half-life of around 40 hours, might indicate this elimination rate, but it typically follows first-order kinetics. If the drug eliminates a constant amount over time, it suggests zero-order kinetics, which heparin could fit. I need to clarify which anticoagulant follows zero-order kinetics. Final Answer: Aspirin (salicylate)</p>	<p>Mapped Option: No clear match Justification: The student's answer incorrectly identifies aspirin (not an anticoagulant) as the drug, and none of the provided options (Warfarin, Enoxaparin, Fondaparinux, Apixaban) exhibit zero-order kinetics as implied by the constant elimination rate (0.5 mg/4h). All listed anticoagulants follow first-order kinetics, and the student's reasoning does not align with any of the options.</p>	<p>Reasoning: The student correctly identifies that the phrase 'eliminating 0.5 mg of the drug every 4 hours' suggests zero-order kinetics [...]. They correctly note that heparin could fit this description [...]. However, the final answer, 'Aspirin (salicylate)', while exhibiting zero-order kinetics at high doses, is not an anticoagulant in the same class as the other drugs in the question [...]. The original question focuses on anticoagulant medications, and Aspirin's antiplatelet action is distinct from the anticoagulant mechanism [...]. More critically, typical anticoagulants such as Warfarin, Enoxaparin, Fondaparinux, and Apixaban do not primarily follow zero-order kinetics [...]. Therefore, the student's answer of Aspirin is not valid for the category of anticoagulants addressed in the question. Valid Alternative: No</p>

Table 15: **Examples of the open-ended answer mapping process.** We show the original multiple-choice questions (MCQs) alongside answers generated by the LLM (GPT-5-mini) to their open-ended reformulations (the questions are omitted for space). The first judge model (Qwen3-32B) attempts to map each answer to one of the original MCQ options. If no clear match is found, a second judge (Gemini-2.5-Flash) evaluates whether the answer can be considered a valid alternative to the gold option.

Open-ended conversion prompt

You are given a MedQA multiple-choice question. Rewrite it into an open-ended question using the following rules:

1. Keep the clinical vignette exactly as it is. Do not shorten, paraphrase, or change it in any way.
2. Only replace the final question sentence with an equivalent open-ended version that asks directly for the answer.
 - The rewritten question must allow a clear answer that can be directly mapped to one of the original options without introducing new information.
 - Phrases like “most likely” or “most appropriate” are allowed if they still point to a single correct answer.
 - Be sure to remove the original final question sentence
 - Enclose the rewritten final question in « ».
3. Before rewriting, check if the question CANNOT be adapted into open-ended form. If ANY of the following is true, output only “Not possible”:
 - A. The question uses NOT, EXCEPT, or asks for exclusions.
 - B. If the correct answer can only be determined by subtle differences among the listed options
 - C. The question requires comparing multiple options, ranking, or ordering (e.g., “first-line vs second-line,” “most/least effective among multiple options”).
 - D. The question requires matching items, relative likelihoods, or prioritization across multiple options.
 - E. The question would become vague or ambiguous if rewritten.
 - F. The question would allow more than one correct answer if the options were removed.
4. Return ONLY the following, in order, using these exact prefixes:
 - Rationale: <short 1–2 sentence explanation of whether the question can be rewritten>
 - Output: <either the original clinical vignette plus the rewritten final sentence in « », or “Not possible”>Do NOT add any text, comments, or formatting beyond these two lines.

Examples

Original:

A junior orthopaedic surgery resident is ... out of the operative report. Which of the following is the correct next action for the resident to take?

- (A) Disclose the error to the patient and put it in the operative report
- (B) Tell the attending that he cannot fail to disclose this mistake
- (C) Report the physician to the ethics committee
- (D) Refuse to dictate the operative report

Rationale: The question asks for a single best answer and does not require ranking or exclusions.

Output: A junior orthopaedic surgery resident is ... out of the operative report. «What is the correct next action for the resident to take?»

Original:

A 24-year-old G2P1 woman ... Which of the following is most true about the confirmatory test?

- (A) It is a Southwestern blot, identifying the presence of DNA-binding proteins
- (B) It is a Northern blot, identifying the presence of RNA
- (C) It is a Northern blot, identifying the presence of DNA
- (D) It is an HIV-1/HIV2 antibody differentiation immunoassay

Rationale: The question requires comparing multiple options. Also, it can only be determined by the subtle differences among the listed options. So it cannot be rewritten clearly.

Output: Not possible

Figure 11: **Prompt for converting MedQA into open-ended format.** The system instruction used to rewrite multiple-choice MedQA questions into open-ended form while ensuring clarity and preserving clinical vignettes.

Open-ended mapping

You are a medical expert and evaluation judge. Your task is to map an open-ended student answer to the most appropriate multiple-choice option from the given question.

Instructions:

1. Carefully read the question.
2. Compare the student's open-ended answer with the provided multiple-choice options.
3. Select the option that best matches in meaning, even if the wording differs.
4. If multiple options are similar, choose the one that is most precise and medically correct.
5. If the open-ended answer clearly corresponds to more than one option, list all possible mappings (e.g., "A or B").
6. If the answer is too vague or does not match any option, output "No clear match" with a short explanation.
7. Output only the option letter (e.g., "A"), or a list in case of multiple mappings, along with a short justification.

{user_input}

Your final answer must follow this Output Format:

- Mapped Option: [Letter or list of letters]
- Justification: [Brief reasoning]

Figure 12: Prompt for open-ended answer mapping.

Valid Alternative Check

You are an expert examiner. You will receive:

- The original multiple-choice question (MCQ) with options and the correct option.
- A student's open-ended answer, given without being shown the options and which does not clearly match any of the provided options.

Your task is to evaluate whether the student's answer is a ****valid alternative answer**** to the question, based on standard knowledge.

Ignore the MCQ options — focus only on whether the answer could reasonably be accepted as correct in concept, even though it is not one of the listed options.

Return your evaluation in JSON format as follows:

```
```json
{{
 "valid_alternative": "Yes" | "Partially" | "No",
 "justification": "A brief explanation why the student's answer is or is not a valid alternative answer."
}} ```
```

Now evaluate the following:

Original Question: {question}  
Options: {options}  
Correct Option: {gold\_answer}  
Student Answer: {given\_answer}

Figure 13: Prompt for valid alternative.

### Medical Expert Guidelines

**Objective:** You are asked to assess the reliability of Gemini-2.5-Flash in verifying answers that were initially mapped as “no clear match” in the open-answer alignment step. Your role is to determine whether Gemini’s judgment is factually correct, partially acceptable, or incorrect.

**Procedure:**

1. You will be provided with:

- The original multiple-choice question and its answer options.
- The model’s generated free-form answer.
- Gemini-2.5-Flash’s internal reasoning (thinking process).
- Gemini’s final decision regarding the correctness of the answer.

2. Carefully review the question, options, and generated answer. Compare Gemini’s reasoning with the factual correctness of the response.

3. You may consult authoritative sources such as:

- Standard medical textbooks,
- Peer-reviewed medical references,
- Trusted online medical documentation (e.g., PubMed, NIH, WHO).

**Evaluation Criteria:**

- Mark as **CORRECT** if Gemini’s verification aligns with established medical knowledge and the generated answer is factually valid in the context of the question.
- Mark as **PARTIALLY ACCEPTABLE** if Gemini’s reasoning is broadly correct but imprecise (e.g., overly general, incomplete, or lacking specificity), yet not misleading or factually wrong.
- Mark as **INCORRECT** if Gemini’s verification or the generated answer is factually wrong, misleading, or contradicts established medical knowledge.

**Notes:**

- Focus only on the medical validity of the answer in relation to the question.
- Do not penalize minor differences in wording or phrasing if the meaning is correct.
- Be consistent in applying the criteria across all cases.

Figure 14: Medical expert guidelines.

### General-domain MCQ (ARC-Challenge)

**Original Question (hidden):** *Which is a fact about penguins?*

**Answer options:**

- (A) Penguins can live in climates with freezing temperatures
- (B) Penguins are fierce competitors
- (C) Penguins are some of the most beautiful birds
- (D) Penguins make great pets

**Observation:** The answer options strongly constrain the question semantics. Even without the question text, a model can plausibly infer that the task concerns factual knowledge about penguins. This pattern is common in ARC-style benchmarks (see Table 7 in the Appendix of Balepur et al. (2024)).

### Medical MCQ (MedQA)

**Original question (hidden):**

*Two weeks after undergoing an emergency cardiac catheterization with stenting for unstable angina pectoris, a 61-year-old man has decreased urinary output and malaise. He has type 2 diabetes mellitus and osteoarthritis of the hips. Prior to admission, his medications were insulin and naproxen. He was also started on aspirin, clopidogrel, and metoprolol after the coronary intervention. His temperature is [...] Which of the following is the most likely cause of this patient's symptoms?*

**Answer options:**

- (A) Renal papillary necrosis
- (B) Cholesterol embolization
- (C) Eosinophilic granulomatosis with polyangiitis
- (D) Polyarteritis nodosa

**Observation:** The answer choices are highly specialized and correspond to subtle differential diagnoses. Without the clinical vignette, reconstructing the underlying question—or identifying the correct answer—is significantly more difficult.

Figure 15: **Contrast between general-domain and medical MCQs under *Options Only* evaluation.** While ARC-style options often reveal the question semantics, medical answer choices remain opaque without the clinical context, making option-based inference harder.