MedXpertQA: Benchmarking Expert-Level Medical Reasoning and Understanding

Yuxin Zuo [*1] **Shang Qu** [*1 2] **Yifei Li** [1] **Zhangren Chen** [1] **Xuekai Zhu** [1] **Ermo Hua** [1] **Kaiyan Zhang** [1]
Ning Ding [✉ 1 2] **Bowen Zhou** [✉ 1 2]

Leaderboard: `https://medxpertqa.github.io`
Code: `https://github.com/TsinghuaC3I/MedXpertQA`

Abstract

We introduce MedXpertQA, a highly challenging and comprehensive benchmark to evaluate expert-level medical knowledge and advanced reasoning. MedXpertQA includes 4, 460 questions spanning 17 specialties and 11 body systems. It includes two subsets, Text for text evaluation and MM for multimodal evaluation. Notably, MM introduces expert-level exam questions with diverse images and rich clinical information, including patient records and examination results, setting it apart from traditional medical multimodal benchmarks with simple QA pairs generated from image captions. MedXpertQA applies rigorous filtering and augmentation to address the insufficient difficulty of existing benchmarks like MedQA, and incorporates specialty board questions to improve clinical relevance and comprehensiveness. We perform data synthesis to mitigate data leakage risk and conduct multiple rounds of expert reviews to ensure accuracy and reliability. We evaluate 18 leading models on MedXpertQA. Moreover, medicine is deeply connected to real-world decision-making, providing a rich and representative setting for assessing reasoning abilities beyond mathematics and code. To this end, we develop a reasoning-oriented subset to facilitate the assessment of *o1-like* models.

[*]Equal contribution [1]Tsinghua University, Beijing, China [2]Shanghai Artificial Intelligence Laboratory, Shanghai, China. Correspondence to: Ning Ding <dn97@mail.tsinghua.edu.cn>, Bowen Zhou <zhoubowen@mail.tsinghua.edu.cn>.

Proceedings of the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025.

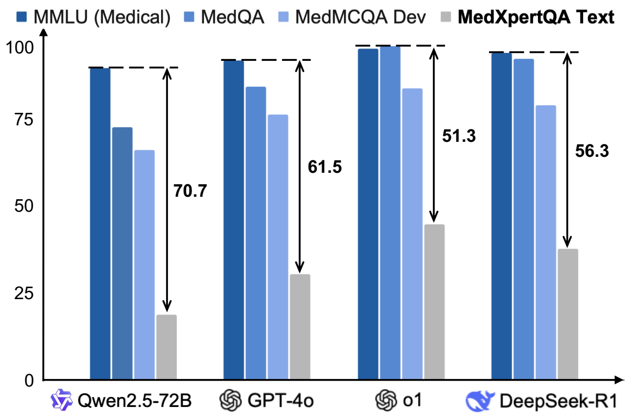


Figure 1: Performance of different models on MedXpertQA Text and other benchmarks. For other benchmarks, we report results for `o1-preview` in place of `o1`. Appendix F shows full results.

1. Introduction

Large Multimodal Models (LMMs) have demonstrated promising potential in advancing general medical AI systems for applications in clinical scenarios (Achiam et al., 2023; Liu et al., 2024b; Saab et al., 2024). However, current text and multimodal benchmarks for evaluating general medical AI capabilities have numerous limitations.

First, existing text medical benchmarks, such as PubMedQA (Jin et al., 2019), MedQA (Jin et al., 2021), MedMCQA (Pal et al., 2022), and MMLU (Medical) (Wang et al., 2024b), lack comprehensive coverage of fine-grained and diverse real-world diagnostic scenarios, including highly specialized fields such as family and addiction medicine. This lack of essential breadth limits the applicability of medical AI in thoroughly addressing realistic medical scenarios. Moreover, these benchmarks fall short of sufficiently challenging current advanced AI, hindering progress toward reliable medical AI. For instance, `o1-preview` has achieved 96% and 99% accuracy on MedQA-USMLE and MMLU Medical Genetics, respectively (Nori et al., 2024).

Second, traditional multimodal medical benchmarks, such as VQA-Rad (Lau et al., 2018), Path-VQA (He et al., 2020), Slake (Liu et al., 2021), PMC-VQA (Zhang et al., 2023), and OmniMedVQA (Hu et al., 2024), are critically inconsistent with real-world clinical scenarios due to: **1) Limited Scope and Insufficient Difficulty.** These benchmarks solely evaluate basic visual perception and medical knowledge, neglecting the complexity of real-world medical tasks across different stages of the diagnosis process. They fail to assess the expert-level knowledge and reasoning ability required for diagnostic decision-making and treatment planning. **2) Lack of Authenticity and Clinical Relevance.** Current benchmarks lack detailed clinical information and rely on automatically generated simple QAs paired with isolated medical images, diverging considerably from realistic clinical scenarios. Medical exam questions used in existing text benchmarks present a promising solution, and `Med-Gemini` (Saab et al., 2024) also demonstrates the significance of such evaluations. However, the field still lacks such a systematic and high-quality benchmark.

To address these challenges, we present MedXpertQA, a highly challenging and comprehensive medical multiple-choice benchmark. It encompasses MedXpertQA Text for text-only evaluations and MedXpertQA MM for multimodal assessments, making it suitable for a wide range of AI models. Both subsets are currently the most challenging benchmarks in their respective fields. Figure 1 presents model performance comparisons of MedXpertQA Text and other benchmarks. MedXpertQA MM includes diverse image types to simulate the wide range of visual information encountered in real-world diagnosis. Figure 2 shows an overview. Overall, MedXpertQA covers a wide range of medical specialties and systems and includes challenging real-world clinical tasks, enabling comprehensive evaluation of expert-level medical abilities.

MedXpertQA follows a rigorous and systematic construction process, including data collection, filtering, synthesis, and expert review. *We collaborate closely with medical practitioners throughout the design and review of MedXpertQA to ensure that its development reflects real-world medical practice and clinical reasoning.* Specifically, we first curate a large-scale question bank from professional exams and textbooks, ensuring difficulty and diversity. Sources include the United States Medical Licensing Examination (USMLE) and the Comprehensive Osteopathic Medical Licensing Examination of the United States (COMLEX-USA) for general medical evaluation, 17 American specialty board exams for specialized scenarios, and 3 image-rich sources, such as the NEJM Image Challenges. Thus, it primarily aligns with medical practices and standards in the United States. We subsequently perform extensive and multi-dimensional question filtering. First, we conduct hierarchical filtering using an adaptive Brier score (Zhu et al., 2024) threshold based on thousands of human responses, calibrated to the difficulty ratings annotated by human experts. We then filter questions based on 14 sampling votes from 8 AI experts for each question. Additionally, we use probabilistic semantic similarity and exact matching precision to identify and remove highly similar questions, thereby enhancing robustness. To ensure MedXpertQA is unseen for existing AI models, we use two proprietary models to rewrite questions and augment options, enhancing diversity and minimizing data leakage risk. To increase difficulty and robustness, we also eliminate low-quality distractors (*i.e.* easy and incorrect answer choices) and expand the number of options to 10 and 5 for Text and MM, respectively. Finally, to mitigate potential errors such as factual inaccuracies, we engage medical experts with physician's licenses to review MedXpertQA, annotating and correcting any errors.

We evaluate 18 proprietary and open-source LMMs and LLMs on MedXpertQA, including cutting-edge inference-time scaled models. Results demonstrate that current leading models still exhibit limited performance on MedXpertQA, particularly in complex medical reasoning tasks. In conclusion, our contributions are as follows:

- We introduce MedXpertQA, a highly challenging and comprehensive medical multiple-choice benchmark. MedXpertQA integrates specialty-specific assessments into medical benchmarking and challenging exam questions with real-world clinical information into multimodal medical benchmarking.
- After rigorous filtering and data augmentation, MedXpertQA demonstrates exceptional difficulty and robustness. We further implement data synthesis to mitigate data leakage risk and conduct multiple rounds of expert reviews to ensure accuracy and validity.
- We evaluate 18 leading models and analyze their limitations in medical capabilities. Moreover, we obtain a *Reasoning* subset specifically designed for assessing medical reasoning abilities, which is well-suited for evaluating *o1-like* reasoning models.

2. Related Work

Multimodal Medical Benchmarks Traditional multimodal medical benchmarks can be broadly categorized into two types: specialized and general purpose (Chen et al., 2024b). Specialized benchmarks focus on a specific modality or medical domain. VQA-Rad (Lau et al., 2018), VQA-Med (Ben Abacha et al., 2019), and Slake (Liu et al., 2021) are primarily centered on radiology, while Path-VQA (He et al., 2020) focuses on pathology. These benchmarks provide extensive evaluation for their intended specialties, yet have a highly constrained scope and limited general-

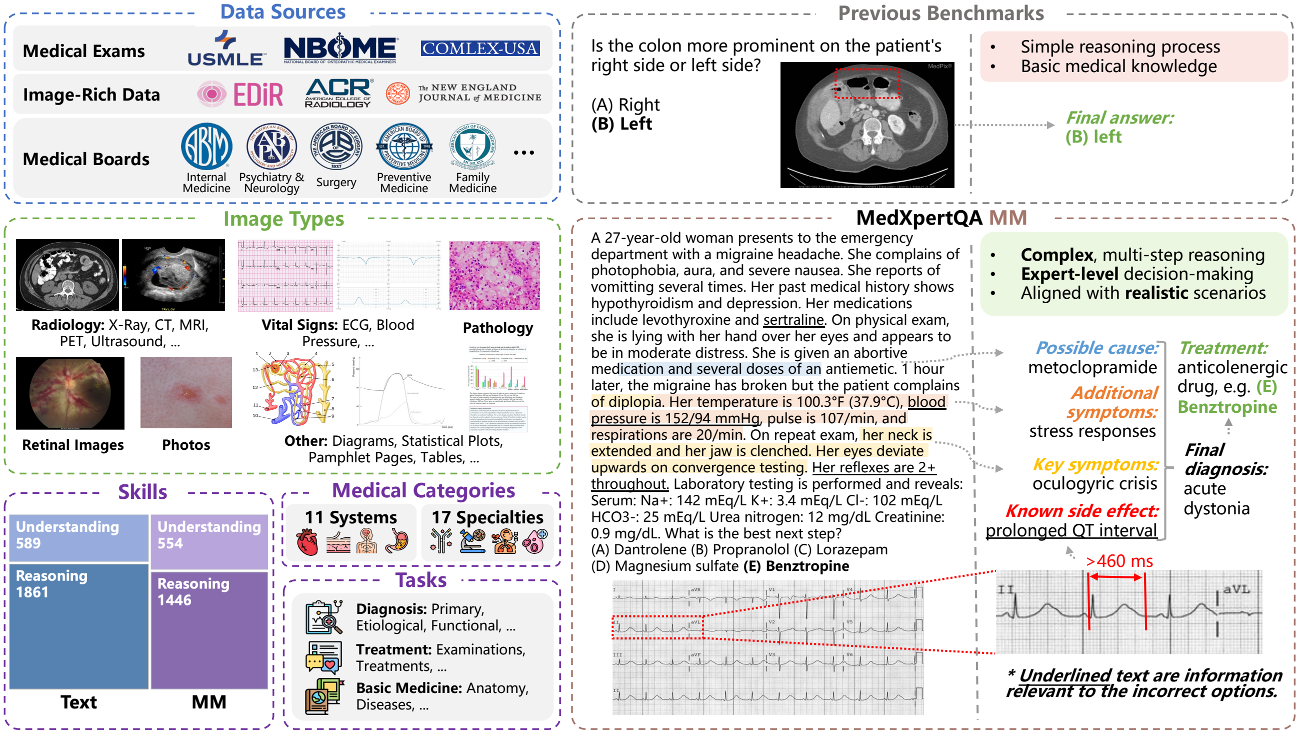


Figure 2: Overview of MedXpertQA. The left side illustrates its diverse data sources, image types, and question attributes. The right side compares typical examples from MedXpertQA MM and a traditional benchmark (VQA-Rad).

izability. With advancements in LMMs, recent general-purpose benchmarks, such as OmniMedVQA (Hu et al., 2024) and GMAI-MMBench (Chen et al., 2024b), have provided more comprehensive evaluations. However, these benchmarks cover limited medical image types, while real-world diagnostic scenarios encompass a wider variety of medical and even general images. Furthermore, all traditional multimodal medical benchmarks mentioned above are constructed automatically based on image captions, deviating from real-world medical tasks. MMMU (H & M) Series (Yue et al., 2024a;b) alleviate this issue to some extent by introducing exam questions. However, they are not specifically designed for the medical domain, and the difficulty is restricted to the university level. Moreover, these benchmarks still lack detailed clinical information commonly encountered in real-world scenarios.

Text Medical Benchmarks Existing text medical benchmarks such as MedQA (Jin et al., 2021), MedMCQA (Pal et al., 2022), MMLU (Medical) Series (Hendrycks et al., 2020; Wang et al., 2024b) are derived from medical entrance and licensing exams, which primarily emphasize general medical knowledge and evaluation. PubMedQA (Jin et al., 2019) is a closed-domain benchmark, with questions generated from PubMed articles. However, these benchmarks do not provide a thorough evaluation of highly specialized diagnostic scenarios to align sufficiently with real-world clinical practice. Their difficulty has also become inadequate.

3. MedXpertQA

3.1. Overview

We introduce MedXpertQA, a universal medical benchmark consisting of challenging text and multimodal subsets Text and MM, which are each divided into a few-shot development set with 5 questions and a test set. Figure 2 shows an overview. MedXpertQA is designed to assess expert-level medical knowledge and reasoning capabilities across 17 medical specialties, 11 human body systems, and 3 task categories. It includes a total of 4,460 questions sourced from examinations at the medical licensing level or higher, of which 2,005 are multimodal questions with a total of 2,839 images. It is the first to introduce medical specialty evaluations to better simulate realistic medical scenarios. Moreover, MedXpertQA MM is the first multimodal medical benchmark to incorporate challenging medical examination questions and real-world medical scenarios. Meanwhile, due to medicine's rigorous requirements for highly reliable outcomes across a comprehensive scope, we refrain from solely prioritizing difficulty and instead aim for question diversity and range as well. As a result, MedXpertQA both substantially challenges current models and showcases remarkable diversity across multiple dimensions. Figure 3 illustrates its wide coverage.

Medical Coverage MedXpertQA collects questions from 17/25 member board exams (specialties) of the American

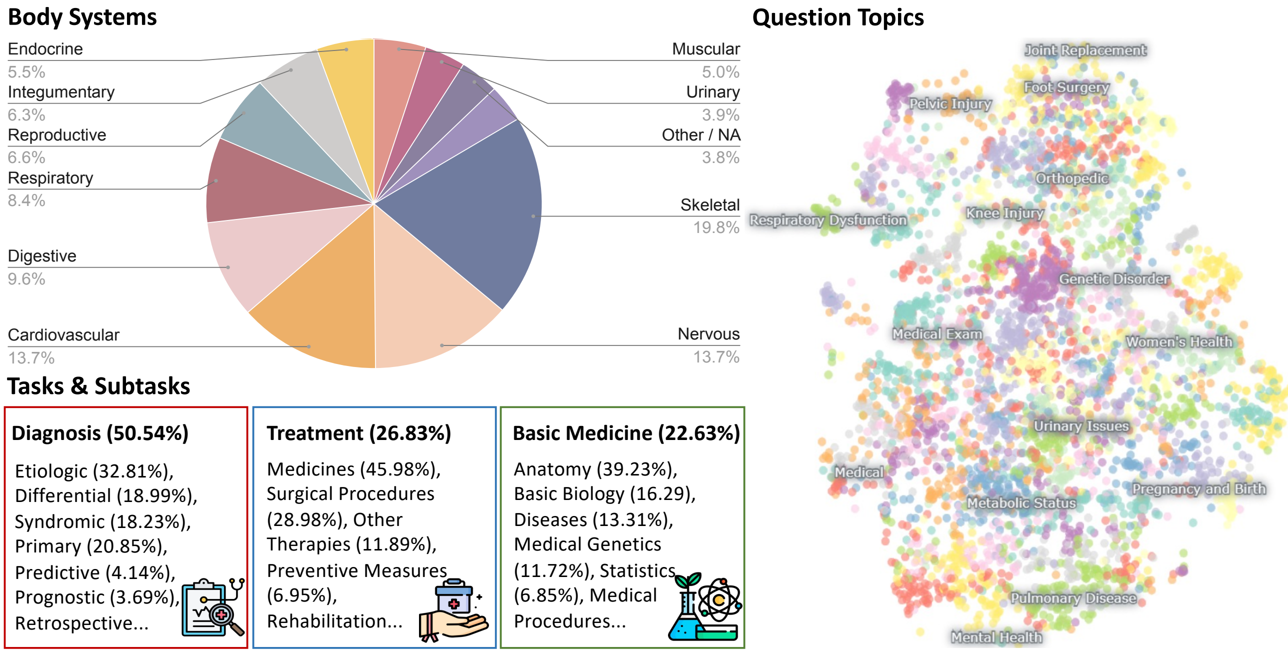


Figure 3: Attribute distributions of MedXpertQA showcase its diversity and comprehensiveness.

Board of Medical Specialties [1] to enable evaluation of highly specialized medical scenarios. Moreover, applying the categorization in Liachovitzky (2015), we instruct an LLM to annotate each question with its most relevant human body system. MedXpertQA covers all 11 systems.

Modal Diversity The data in MedXpertQA covers a broad spectrum of modalities. Its questions and answer choices both incorporate structured data, such as tables, and semi-structured documents. MedXpertQA MM's images similarly demonstrate high diversity and wide coverage. It not only encompasses medical imaging results obtained from various techniques in diverse formats, but also other image types such as diagrams, charts, and documents, fully covering the spectrum of visual information that human doctors are expected to analyze. Specifically, MedXpertQA encompasses the following image categories: *Radiology, Pathology, Medical Optical Imaging, Photos, Vital Signs, Diagrams, Documents, Charts, Tables, and Others*.

Task Diversity To enable in-depth analysis, we use `gpt-4o-2024-11-20` to annotate each question's *core medical task* (Diagnosis, Treatment Planning, or Basic Medicine) (Huang et al., 2025) and fine-grained subtask. Appendix I.1 presents relevant prompts.

3.2. Benchmark Construction

To construct MedXpertQA, we begin with data collection, followed by a carefully designed and resource-intensive procedure involving data filtering, synthesis, and expert review. Table 1 presents dataset statistics across different construction phases. After initial data collection and filtering based on question difficulty, the reductions in question count occur during the *Similarity Filtering* and *Expert Review* stages, which remove 54 and 223 questions, respectively. While the early stages primarily optimize for difficulty, these later stages focus on improving data quality. Given that MedXpertQA is constructed from authoritative and high-quality sources, filtering based on difficulty naturally has a more substantial impact on dataset size compared to quality-based filtering. Additionally, during the expert review phase, reviewers directly corrected the errors in most of the flagged questions instead of deleting them.

Phase	MM	Text	Sum
Data Collection	10,868	26,675	37,543
Difficulty-Based Data Filtering			
Human Expert Filtering	5,133	5,013	10,146
AI Expert Filtering	2,125	2,612	4,737
Quality-Based Filtering and Augmentation			
Edit Distance Filtering	2,121	2,592	4,713
Semantic Similarity Filtering	2,105	2,578	4,683
Question and Option Augmentation	2,105	2,578	4,683
Multi-Rounds Expert Review & Delete	2,005	2,455	4,460

Table 1: Data statistics after different construction phases.

Data Collection We begin by constructing a large-scale question bank, sourcing difficult multiple-choice questions from authoritative medical examinations and textbooks. Pre-

[1] https://www.abms.org

vious work primarily relied on USMLE questions for training and evaluation (Jin et al., 2021). We expand the scope by including questions from COMLEX, another major medical licensing examination in the U.S., to capture the unique challenges of medical image interpretation in orthopedic practice. To further evaluate multimodal medical capabilities, we incorporate questions from the American College of Radiology (ACR) DXIT and TXIT exams, the European Board of Radiology (EDiR) exams, and the New England Journal of Medicine (NEJM) Image Challenge. In addition, we collect questions from 17 American medical specialty board exams, spanning a wide array of common medical specialties. Ultimately, we collect $37,543$ questions. We additionally obtain human response distributions and expert annotations, including explanations and difficulty ratings.

Data Filtering We conduct *AI & Human Expert Filtering* to identify questions that challenge both humans and AI. Subsequent *Similarity Filtering* further enhances robustness.

Step 1: AI Expert Filtering. We employ 8 models, categorized as basic or advanced, as AI experts to vote on and filter questions. First, each basic AI expert performs 4 sampling attempts for each question. If any expert answers a question correctly in all attempts, the question is deemed too simple and removed. Second, questions that are answered incorrectly by all AI experts are retained. This method effectively differentiates between questions that leading AI models can solve and those that remain challenging. Appendix C.1 lists models used in this phase.

Step 2: Human Expert Filtering. We use prior and posterior human expert annotations to identify questions that challenge humans. Using human response distributions, we first assess each question's posterior difficulty by calculating its Brier score, a widely applied metric of prediction accuracy (Zhu et al., 2024). We consider the question's response distribution over all answer choices as the answer prediction. Given the prediction vector $\hat{\mathbf{y}}$ and label vector $\mathbf{y}$, the Brier score B is:

$$B = \frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2, \tag{1}$$

where N is the number of options, y_i is the label for option i (0 or 1), and $\hat{y}_i$ is the proportion of responses selecting option i. A lower Brier score indicates a more accurate overall prediction distribution, suggesting the question is easier. Compared to accuracy, the Brier score accounts for the response rates of all options, providing a more precise difficulty measurement. Subsequently, we normalize the prior difficulty ratings annotated by medical experts and categorize questions into 3 levels, each associated with an adaptive Brier score threshold for stratified sampling. Higher-rated questions are assigned higher Brier score thresholds, with the maximum threshold set at the 25th and the minimum set at the 3rd percentile of all scores. Approximately 16.78% of questions lack the annotations above and thus do not undergo human expert filtering.

Step 3: Similarity Filtering. A key factor in achieving robust evaluation is ensuring high diversity and avoiding repetitive assessments. Therefore, we filter data by identifying question pairs with extremely high semantic or edit distance similarities. For semantic similarity, we use `MedCPT-Query-Encoder` (Jin et al., 2023), a medical embedding model developed from `PubMedBERT` (Gu et al., 2020), to obtain sentence representations of all questions, then compute pairwise cosine similarities. Since similarities are one-dimensional and do not follow a Gaussian distribution, we use the Interquartile Range (IQR) to identify outliers with abnormally high similarity. For detected outlier question pairs, we compare their difficulty annotations and remove the easier question. *Similarity Filtering* plays a critical role in ensuring benchmark robustness. Traditional visual medical benchmarks, such as Slake (Liu et al., 2021), often rely on rigid, template-based question generation, which limits question diversity. This is problematic since model performance can vary greatly depending on question type, and repetitive questions allow models to exploit template-specific shortcut patterns, leading to unexpected performance gains. In contrast, since MedXpertQA was not generated from fixed templates, it exhibits high question diversity from the start, resulting in relatively few questions being removed at this stage. Nonetheless, even minimal filtering is indispensable, as it removes near-duplicate questions that could otherwise cause benchmark overfitting and distort evaluation results.

Question and Option Augmentation Since questions in MedXpertQA are from publicly available sources, we perform question and option augmentation to mitigate possible data leakage issues. Option augmentation also increases both the difficulty and robustness of MedXpertQA. For question augmentation, we instruct the LLM to rephrase the question through alternative expressions or structural adjustments while preserving all original information. We emphasize professional language style, factual accuracy, and objectiveness. For option augmentation, we first identify and remove low-quality distractors with low human response rates while retaining the correct option and at least two distractors. Most multimodal questions consist of 5 options which are often image-dependent and therefore cannot be feasibly expanded. Therefore, we set the total number of options for MedXpertQA Text and MedXpertQA MM to 10 and 5, respectively. The generated distractors should be reasonable in the question's context, challenge medical experts, and remain consistent with the original options in both language and content style. We use `gpt-4o-2024-11-20` and

Benchmark	# Size	# Images	# Rate	# Avg Lens	# Image Types	Annotation	Clinical Scenarios	Specialties
VQA-Rad (Lau et al., 2018)	451	204	0.45	14.61	1	Automatic	×	×
VQA-Med (Ben Abacha et al., 2019)	500	500	1.00	8.96	1	Automatic	×	×
Path-VQA (He et al., 2020)	6,719	858	0.13	15.38	1	Automatic	×	×
Slake-En (Liu et al., 2021)	1,061	96	0.09	13.97	1	Automatic	×	×
PMC-VQA (Zhang et al., 2023)	33,430	29,021	0.87	61.84	2	Automatic	×	×
OmniMedVQA (Hu et al., 2024)	127,995	118,010	0.92	42.40	4	Automatic	×	×
GMAI-MMBench (Chen et al., 2024b)	21,281	21,180	1.00	49.85	4	Automatic	×	✓
MMMU (H & M) (Yue et al., 2024a)	1,752	1,994	1.14	83.56	8	Expert	×	×
MMMU-Pro (H & M) (Yue et al., 2024b)	346	431	1.25	107.08	7	Expert	×	×
MedXpertQA MM	2,000	2,852	1.43	149.35	10	Expert	✓	✓

Table 2: Comparisons with existing multimodal medical benchmarks. MedXpertQA MM demonstrates high complexity through its high average question length, a large number of images, and diverse image categories. The first category includes commonly used traditional benchmarks. The second category includes recently emerging and more comprehensive benchmarks. The third category includes general multimodal benchmarks that include medicine-related subsets but are not specifically designed for medical fields.

Benchmark	# Size	# Avg Lens	Clinical Scenarios	Specialties
PubMedQA (Jin et al., 2019)	1,000	328.41	×	×
MedMCQA-Dev (Pal et al., 2022)	4,183	53.84	×	×
MedQA-USMLE (Jin et al., 2021)	1,273	215.46	✓	×
MMLU (Med.) (Hendrycks et al., 2020)	1,089	100.07	×	×
MMLU-Pro (Med.) (Wang et al., 2024b)	586	166.63	×	×
MedXpertQA Text	2,450	257.37	✓	✓

Table 3: Comparisons with text medical benchmarks.

`claude-3-5-sonnet-20241022` in this stage, and Appendix I.2 shows all prompts. Section 5.1 compares the leakage risk before and after data augmentation to verify the effectiveness of this process.

Expert Review Finally, medical experts with physician's licenses review each question in MedXpertQA, assessing the general quality of the original and augmented questions and options based on comprehensive guidelines. Objectives of the question review include identifying missing information, detecting factual inaccuracies, and removing extra content. Options review primarily emphasizes evaluating the options' validity and context-appropriateness. Appendix E provides a detailed description of this procedure.

3.3. Medical Reasoning Benchmarking

OpenAI's o1 models advance reasoning capabilities, which have been extensively validated in math and coding (Wu et al., 2024). However, evaluation in specialized domains like medical reasoning has been insufficient, as relevant benchmarks remain underdeveloped. Recent works developing *o1-like* medical reasoning models with reinforcement learning (Chen et al., 2024a) or inference-time scaling (Huang et al., 2025) similarly face limitations. Benchmarks chosen, such as MedQA, contain questions assessing medical knowledge only (see *Example of MedQA (Understanding)* in Appendix G.1), hence are suboptimal for isolating the model's medical reasoning ability. Therefore, while leading *o1-like* models show promise in medicine (Xie et al., 2024; Nori et al., 2024), systematic evaluations specifically focused on their reasoning-based medical proficiency remain lacking.

Moreover, medicine tightly interweaves professional knowledge with complex reasoning. Clinical reasoning is multifactorial (Yazdani & Hoseini Abardeh, 2019), requiring synthesizing diverse information, navigating uncertainty (Patel et al., 2024), and engaging in heterogeneous cognitive processes (Shin, 2019). These characteristics are represented in MedXpertQA through complex, information-rich tasks such as multiple-choice style differential diagnosis (Seller & Symons, 2011) (see Appendix G.1 and G.2). Therefore, we underscore the importance of identifying reasoning-oriented tasks to facilitate fine-grained performance feedback and accurate assessment. On MedXpertQA, we use `gpt-4o` to annotate questions according to whether they require complex reasoning. As seen in Appendix I.1, we instruct the LLM to categorize complicated, reasoning-heavy questions as *Reasoning*. In contrast, other questions that involve little to no reasoning and instead assess skills such as medical knowledge and image perception are categorized as *Understanding*. As shown in Figure 2, within both Text and MM, a majority of questions focus on *Reasoning*.

We initially considered having human experts annotate *Reasoning* and *Understanding* labels. However, we found that LLMs can perform this task reliably when provided with clear and precise prompt instructions. Specifically, we provide expert-written answers and explanations collected from data sources and explicitly emphasize the distinction between reasoning complexity and general difficulty in our prompt design, as shown in Table 15). This extensive guidance enables a simplified form of annotation under expert supervision. Following automatic labeling, we engage human experts to conduct a review. For a randomly selected 10% subset of the questions, 490 from Text and 400 from MM, the reviewers identified 28 and 11 questions, respectively, that were incorrectly labeled. This corresponds to an error rate of approximately only 4.3%.

3.4. Comparisons with Existing Benchmarks

Tables 2 and 3 compare two subsets of MedXpertQA with existing benchmarks. Traditional multimodal benchmarks deviate from real-world clinical tasks, reflected in the limited number of image types, low image-to-question ratios, and automatically generated questions and annotations. Meanwhile, MMMU (H & M) Series primarily based on university-level subject exams, falls short in scope, difficulty, and specificity to the medical domain.

MedXpertQA MM demonstrates advantages in length and image richness. It is the first multimodal medical benchmark to incorporate real-world clinical scenarios, resulting in a substantial increase in question length. Meanwhile, the rich image information in MedXpertQA MM can be seen from its broadest range of image types, highest number of images within same-scale benchmarks, and highest image-to-question ratio.

MedXpertQA Text is the first text medical benchmark to purposefully account for medical specialty assessment, supporting evaluations of highly specialized medical scenarios.

4. Experiments

4.1. Implementation Details

We evaluate all models using zero-shot CoT prompting (Kojima et al., 2022) unless otherwise specified. For answer cleansing, we follow the script provided by Kojima et al. (2022). Appendix G.3 shows a case of evaluation. We employ greedy decoding for output generation if available, ensuring result stability. For reasoning models with specific evaluation requirements, we follow their respective instructions. Appendix C.2 presents additional implementation details. We could not evaluate `o1` and `o3-mini` on the full MedXpertQA due to costs. Instead, for both MedXpertQA Text and MedXpertQA MM, we sample 10% of questions from the *Reasoning* and *Understanding* subsets respectively. The seed is set to 42.

4.2. Models

We evaluate leading LMMs and LLMs on full MedXpertQA and MedXpertQA Text, respectively. We include both proprietary and open-source models, and also evaluate advanced inference-time scaled models† with a focus on the latest progress in medical reasoning capabilities.

Large Multimodal Models: o1-2024-12-17† (Jaech et al., 2024), QVQ-72B-Preview† (Team, 2024a), GPT-4o-2024-11-20 (Hurst et al., 2024), GPT-4o-mini-2024-07-18 (Hurst et al., 2024), Claude-3.5-Sonnet-20241022 (Anthropic, 2024), Gemini-1.5-Pro-002 (Team et al., 2024), Gemini-2.0-Flash-Exp (Google, 2024), Qwen2.5-VL-72B (Team, 2025), Qwen2-VL-72B (Wang et al., 2024a).

Large Language Models: o3-mini† (OpenAI, 2025), DeepSeek-R1† (Guo et al., 2025), QwQ-32B-Preview† (Team, 2024b), DeepSeek-V3 (Liu et al., 2024a), Claude-3.5-Haiku-20241022 (Anthropic, 2024), LLaMA-3.3-70B-Instruct (Dubey et al., 2024), LLaMA-3-70B-UltraMedical (Zhang et al., 2024), Qwen2.5-72B-Instruct (Yang et al., 2024), Qwen2.5-32B-Instruct (Yang et al., 2024).

4.3. Main Results

Tables 4 and 5 show the main results. Overall, the low accuracies of evaluated models demonstrate MedXpertQA's ability to pose challenges to state-of-the-art models.

Human Performance Evaluation We establish a baseline denoted as `Expert (Pre-Licensed)`, referring to upper-year medical students or trainees preparing for licensing examinations such as USMLE, as shown in Table 4. This baseline is derived from aggregated response distributions based on answers provided by these examinees, collected from the original data sources. Since *Question and Option Augmentation* preserves all core clinical content, its impact on human accuracy is minimal. Accordingly, performance on the final questions closely aligns with that on the original versions. Most questions have substantial response counts, with a large portion of the dataset containing at least 500 responses per question and some reaching up to $\mathbf{238}$k. Fewer than 200 final questions lack response data. These characteristics ensure that the resulting statistics provide stable and representative estimates of human performance.

Comparison of LMMs Among vanilla LMMs, `GPT-4o` consistently performs best across all subsets. `Gemini-2.0-Flash` is the highest-scoring vanilla LMM after `GPT-4o` with an impressive performance on MedXpertQA MM, highlighting its advantage in multimodal tasks. We also observe that the highly representative open-source LMM, `Qwen2.5-VL-72B`, outperforms `GPT-4o-mini`, yet still falls behind `GPT-4o`.

Comparison of LLMs `DeepSeek-R1` shows the strongest performance among LLMs, most notably on the *Reasoning* set, where it substantially outperforms other models. The performance of `Qwen2.5-32B`, particularly on the *Reasoning* set, is nearly equivalent to random selection, so we limit the evaluation to models of this scale and above.

Model	MedXpertQA Text			MedXpertQA MM			Avg$_R$	Avg$_U$	Avg
	Reasoning	Understanding	Avg	Reasoning	Understanding	Avg			
Expert (Pre-Licensed)	41.74	45.44	42.60	45.76	44.97	45.53	43.48	45.20	43.92
Inference-Time Scaled Large Multimodal Models									
o1‡	46.24	39.66	44.67	52.78	65.45	56.28	49.09	52.21	49.89
QVQ-72B-Preview	22.08	20.71	21.76	33.54	33.57	33.55	27.09	26.95	27.06
Vanilla Large Multimodal Models									
GPT-4o	30.63	29.54	30.37	40.73	48.19	42.80	35.05	38.58	35.96
Claude-3.5-Sonnet	19.88	25.81	21.31	33.33	32.85	33.20	25.76	29.22	26.65
Gemini-1.5-Pro	19.18	21.22	19.67	32.85	37.36	34.10	25.16	29.05	26.16
GPT-4o-mini	17.09	20.20	17.84	28.22	27.62	28.05	21.95	23.80	22.43
Gemini-2.0-Flash	20.53	20.71	20.57	35.48	41.70	37.20	27.06	30.88	28.04
Qwen2.5-VL-72B	17.89	18.17	17.96	29.53	31.05	29.95	22.98	24.41	23.35
Qwen2-VL-72B	16.39	18.68	16.94	25.86	34.84	28.35	20.53	26.51	22.07

Table 4: Performance of LMMs on MedXpertQA. ‡ indicates the result is evaluated on a sampled subset. **Avg$_R$** and **Avg$_U$** indicate the average performance on the **Reasoning** and **Understanding** subsets, respectively.

Model	Reasoning	Understanding	Avg
Expert (Pre-Licensed)	41.74	45.44	42.60
o1‡	46.24	39.66	44.67
QVQ-72B-Preview	22.08	20.71	21.76
Inference-Time Scaled LLMs			
o3-mini‡	37.63	36.21	37.30
DeepSeek-R1	37.88	37.35	37.76
QwQ-32B-Preview	18.70	15.79	18.00
Vanilla LLMs			
DeepSeek-V3	23.91	24.96	24.16
Claude-3.5-Haiku	16.71	21.05	17.76
LLaMA-3.3-70B	23.86	26.49	24.49
LLaMA-3-70B-UltraMedical	20.03	21.05	21.80
Qwen2.5-72B	18.54	20.03	18.90
Qwen2.5-32B	14.02	18.34	15.06

Table 5: Performance of LLMs on MedXpertQA Text.

Phase	Perplexity ↑	Rouge-L ↓	Edit Distance Similarity ↓
Before	$1.03E + 218$	0.1893	0.2691
After	$1.35E + 247$	0.1664	0.2416

Table 6: Data leakage analysis results.

Medical Reasoning Performance Table 4 indicates that the textual reasoning performance of vanilla LMMs other than `GPT-4o` is relatively close, while their visual reasoning capabilities show marked differences, suggesting that visual perception plays a crucial role in approaching MedXpertQA questions. Moreover, vanilla LLMs and LMMs consistently perform worse on the *Reasoning* subset than on the *Understanding* subset. However, this gap noticeably narrows for *o1-like* inference-time scaled models. *This underscores the inherent challenges associated with medical reasoning, as well as the high quality of our Reasoning subset.* In other words, our annotations effectively produce a challenging reasoning-oriented subset that is well-suited for evaluating *o1-like* reasoning models. Additionally, while models generally score higher on MedXpertQA MM, this can be attributed to the smaller number of answer choices for MedXpertQA MM and more rigorous filtering parameters used for constructing MedXpertQA Text. Therefore, direct performance comparisons between the two subsets are not informative.

5. Analysis

5.1. Data Leakage

To evaluate whether *Question and Option Augmentation* reduces leakage risk, we follow Xu et al. (2024) to use perplexity (PPL) and N-gram accuracy (ROUGE-L and edit distance similarity) as metrics. We make certain adaptations and simplifications to assess leakage risk at the instance level before and after data synthesis. After concatenating the original question with a specific prompt such as "Answer:" to form the input, we calculate the PPL of the model's output. Moreover, to assess whether the evaluated model's rationale is similar to the corresponding explanations we collected, we compute the ROUGE-L and edit distance similarity between the output and the explanation. We analyze the outputs of `GPT-4o`, since it is the most effective among all vanilla LMMs, possibly reflecting a higher risk of data leakage.

Table 6 shows results. Based on the threshold set by Xu et al. (2024), the data leakage risk before synthesis is already relatively low, which can be attributed to two factors: 1) Questions in MedXpertQA are derived from difficult medical exams that may have not yet been considered for training; 2) Even questions already used for training may be difficult for the model to learn due to their complexity (Lin et al., 2024). Leakage risk is further reduced following data synthesis, particularly indicated by a notable increase in perplexity. *Results demonstrate that MedXpertQA has a low risk of data leakage and can objectively assess model ability.*

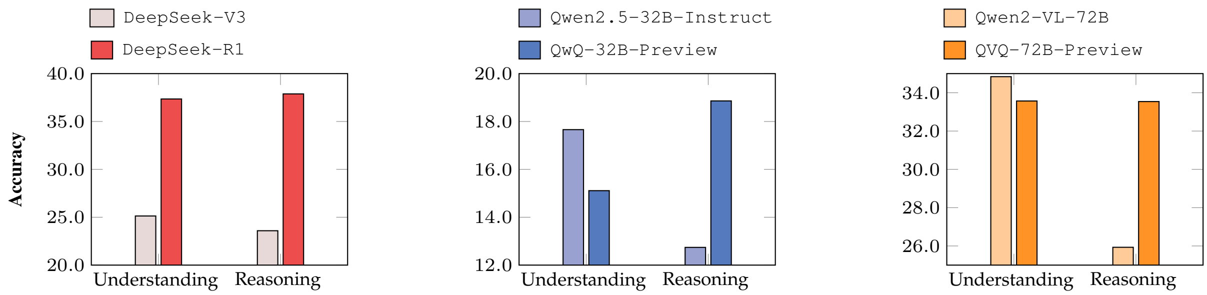


Figure 4: Performance of three groups of models on MedXpertQA.

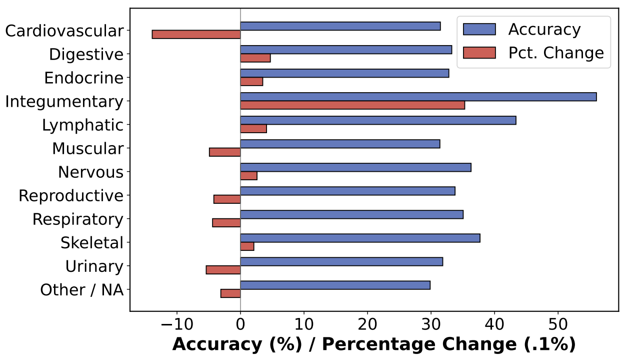


Figure 5: System analysis results. "Accuracy" bars show accuracies on each system's questions. "Percentage Change" bars show the gap between the percentages of each system's questions in the correctly answered set and the full set.

5.2. Impact of Inference-Time Scaling

We employ two groups of Qwen-Series models for comparative analysis to investigate the impact of inference-time scaling on the models' performance on challenging medical tasks. Specifically, they represent two text and multimodal *o1-like* models, along with their corresponding backbones. We similarly assess and compare `DeepSeek-R1` and `DeepSeek-V3`.

- `DeepSeek-R1` ***vs*** `DeepSeek-V3`: Evaluation on MedXpertQA Text.
- `QwQ-32B-Preview` ***vs*** `Qwen2.5-32B-Instruct`: Evaluation on MedXpertQA Text.
- `QVQ-72B-Preview` ***vs*** `Qwen2-VL-72B`: Evaluation on MedXpertQA MM.

Figure 4 illustrates the performance of the three groups across MedXpertQA Text and MedXpertQA MM. The accuracy variation observed in each subplot is consistently more pronounced in the *Reasoning* set. Furthermore, accuracy on the *Reasoning* set exhibits a clear upward trend, while performance on the *Understanding* set fluctuates, with occasional declines. This suggests that, even without additional medical training data, inference-time scaling can improve complex medical reasoning skills.

5.3. Medical Insights

We analyze `GPT-4o`'s performance based on the annotated system labels to derive fine-grained medical insights. Figure 5 shows that the proportion of *Integumentary* questions in the correctly answered set markedly increases compared with the full question set. We also observe a much higher accuracy on these questions than questions on other systems. This suggests that `GPT-4o` has a deeper understanding and stronger performance on *Integumentary*-related questions. In contrast, it exhibits lower accuracy on the *Cardiovascular* subset, with a noticeable proportion decline in the correct set, indicating suboptimal capability in this regard.

6. Conclusion

We introduce MedXpertQA, a highly challenging and comprehensive medical multiple-choice benchmark evaluating expert-level knowledge and reasoning in real-world clinical scenarios. MedXpertQA encompasses diverse medical specialties, body systems, and clinical tasks. It addresses critical gaps in current benchmarks, including inadequate coverage of medical specialties, insufficient difficulty, and limited clinical relevance. By incorporating expert-level medical examination questions rooted in comprehensive clinical information, MedXpertQA MM marks a crucial advancement in multimodal medical benchmarking. We mitigate data leakage risk through data synthesis and perform expert review to ensure accuracy and validity. We benchmark 18 leading models and analyze their system-specific medical capabilities. Additionally, we construct a reasoning-oriented subset, which demonstrates that current AI systems still face challenges when tackling complex medical reasoning tasks.

Acknowledgements

We thank anonymous reviewers for their insightful comments and suggestions. This work is supported by National Science and Technology Major Project (2023ZD0121403) , Young Elite Scientists Sponsorship Program by CAST (2023QNRC001), and National Natural Science Foundation of China (No. 62406165).

Impact Statement

This paper introduces MedXpertQA, a comprehensive medical benchmark to advance the field of healthcare application-driven machine learning.

The deployment of AI in healthcare raises ethical concerns, encompassing issues such as data privacy, algorithmic biases, and the potential for excessive dependence on automated systems. To mitigate these risks, it is crucial to prioritize transparency, explainability, and continuous validation by healthcare professionals. Furthermore, AI systems must undergo ongoing assessment to prevent the perpetuation of existing healthcare disparities.

While MedXpertQA demonstrates considerable promise for advancing medical AI, it also underscores the importance of responsible development and oversight to ensure that these technologies are applied ethically and equitably within healthcare settings.

References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Anthropic. The claude 3 model family: Opus, sonnet, haiku. 2024. URL https://api.semanticscholar.org/CorpusID:268232499.

Ben Abacha, A., Hasan, S. A., Datla, V. V., Demner-Fushman, D., and Müller, H. Vqa-med: Overview of the medical visual question answering task at imageclef 2019. In *Proceedings of CLEF (Conference and Labs of the Evaluation Forum) 2019 Working Notes*. 9-12 September 2019, 2019.

Cai, Z., Cao, M., Chen, H., Chen, K., Chen, K., Chen, X., Chen, X., Chen, Z., Chen, Z., Chu, P., Dong, X., Duan, H., Fan, Q., Fei, Z., Gao, Y., Ge, J., Gu, C., Gu, Y., Gui, T., Guo, A., Guo, Q., He, C., Hu, Y., Huang, T., Jiang, T., Jiao, P., Jin, Z., Lei, Z., Li, J., Li, J., Li, L., Li, S., Li, W., Li, Y., Liu, H., Liu, J., Hong, J., Liu, K., Liu, K., Liu, X., Lv, C., Lv, H., Lv, K., Ma, L., Ma, R., Ma, Z., Ning, W., Ouyang, L., Qiu, J., Qu, Y., Shang, F., Shao, Y., Song, D., Song, Z., Sui, Z., Sun, P., Sun, Y., Tang, H., Wang, B., Wang, G., Wang, J., Wang, J., Wang, R., Wang, Y., Wang, Z., Wei, X., Weng, Q., Wu, F., Xiong, Y., Xu, C., Xu, R., Yan, H., Yan, Y., Yang, X., Ye, H., Ying, H., Yu, J., Yu, J., Zang, Y., Zhang, C., Zhang, L., Zhang, P., Zhang, P., Zhang, R., Zhang, S., Zhang, S., Zhang, W., Zhang, W., Zhang, X., Zhang, X., Zhao, H., Zhao, Q., Zhao, X., Zhou, F., Zhou, Z., Zhuo, J., Zou, Y., Qiu, X., Qiao, Y., and Lin, D. Internlm2 technical report, 2024.

Chen, J., Cai, Z., Ji, K., Wang, X., Liu, W., Wang, R., Hou, J., and Wang, B. Huatuogpt-o1, towards medical complex reasoning with llms. *arXiv preprint arXiv:2412.18925*, 2024a.

Chen, P., Ye, J., Wang, G., Li, Y., Deng, Z., Li, W., Li, T., Duan, H., Huang, Z., Su, Y., et al. Gmai-mmbench: A comprehensive multimodal evaluation benchmark towards general medical ai. *arXiv preprint arXiv:2408.03361*, 2024b.

Chen, Z., Wu, J., Wang, W., Su, W., Chen, G., Xing, S., Zhong, M., Zhang, Q., Zhu, X., Lu, L., et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24185–24198, 2024c.

Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Yang, A., Fan, A., et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Google. Gemini 2.0 flash. https://gemini.google.com, 2024. Accessed: 12/2024.

Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., Naumann, T., Gao, J., and Poon, H. Domain-specific language model pretraining for biomedical natural language processing, 2020.

Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

He, X., Zhang, Y., Mou, L., Xing, E., and Xie, P. Pathvqa: 30000+ questions for medical visual question answering. *arXiv preprint arXiv:2003.10286*, 2020.

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., and Steinhardt, J. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

Hu, Y., Li, T., Lu, Q., Shao, W., He, J., Qiao, Y., and Luo, P. Omnimedvqa: A new large-scale comprehensive evaluation benchmark for medical lvlm. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22170–22183, 2024.

Huang, Z., Geng, G., Hua, S., Huang, Z., Zou, H., Zhang, S., Liu, P., and Zhang, X. O1 replication journey–part 3: Inference-time scaling for medical reasoning. *arXiv preprint arXiv:2501.06458*, 2025.

Hurst, A., Lerer, A., Goucher, A. P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Jaech, A., Kalai, A., Lerer, A., Richardson, A., El-Kishky, A., Low, A., Helyar, A., Madry, A., Beutel, A., Carney, A., et al. Openai o1 system card. *arXiv preprint arXiv:2412.16720*, 2024.

Jin, D., Pan, E., Oufattole, N., Weng, W.-H., Fang, H., and Szolovits, P. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421, 2021.

Jin, Q., Dhingra, B., Liu, Z., Cohen, W. W., and Lu, X. Pubmedqa: A dataset for biomedical research question answering. *arXiv preprint arXiv:1909.06146*, 2019.

Jin, Q., Kim, W., Chen, Q., Comeau, D. C., Yeganova, L., Wilbur, W. J., and Lu, Z. Medcpt: Contrastive pre-trained transformers with large-scale pubmed search logs for zero-shot biomedical information retrieval. *Bioinformatics*, 39(11):btad651, 2023.

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., and Iwasawa, Y. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35: 22199–22213, 2022.

Lau, J. J., Gayen, S., Ben Abacha, A., and Demner-Fushman, D. A dataset of clinically generated visual questions and answers about radiology images. *Scientific data*, 5(1):1–10, 2018.

Liachovitzky, C. *Human Anatomy and Physiology Preparatory Course*. 2015.

Lin, Z., Gou, Z., Gong, Y., Liu, X., Shen, Y., Xu, R., Lin, C., Yang, Y., Jiao, J., Duan, N., et al. Rho-1: Not all tokens are what you need. *arXiv preprint arXiv:2404.07965*, 2024.

Liu, A., Feng, B., Xue, B., Wang, B., Wu, B., Lu, C., Zhao, C., Deng, C., Zhang, C., Ruan, C., et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024a.

Liu, B., Zhan, L.-M., Xu, L., Ma, L., Yang, Y., and Wu, X.-M. Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pp. 1650–1654. IEEE, 2021.

Liu, H., Li, C., Wu, Q., and Lee, Y. J. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024b.

Nori, H., Usuyama, N., King, N., McKinney, S. M., Fernandes, X., Zhang, S., and Horvitz, E. From medprompt to o1: Exploration of run-time strategies for medical challenge problems and beyond. *arXiv preprint arXiv:2411.03590*, 2024.

OpenAI. Gpt-o3-mini. https://openai.com/index/openai-o3-mini/, 2025. 2025-01-31.

Pal, A., Umapathi, L. K., and Sankarasubbu, M. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on health, inference, and learning*, pp. 248–260. PMLR, 2022.

Patel, B., Gheihman, G., Katz, J. T., Begin, A. S., and Solomon, S. R. Navigating uncertainty in clinical practice: A structured approach. *Journal of General Internal Medicine*, pp. 1–8, 2024.

Phan, L., Gatti, A., Han, Z., Li, N., Hu, J., Zhang, H., Zhang, C. B. C., Shaaban, M., Ling, J., Shi, S., et al. Humanity's last exam. *arXiv preprint arXiv:2501.14249*, 2025.

Qwen Team. Qwen2.5: A party of foundation models, September 2024. URL https://qwenlm.github.io/blog/qwen2.5/.

Saab, K., Tu, T., Weng, W.-H., Tanno, R., Stutz, D., Wulczyn, E., Zhang, F., Strother, T., Park, C., Vedadi, E., et al. Capabilities of gemini models in medicine. *arXiv preprint arXiv:2404.18416*, 2024.

Seller, R. H. and Symons, A. B. *Differential Diagnosis of Common Complaints E-Book*. Elsevier Health Sciences, 2011.

Shin, H. S. Reasoning processes in clinical reasoning: from the perspective of cognitive psychology. *Korean journal of medical education*, 31(4):299, 2019.

Team, G., Georgiev, P., Lei, V. I., Burnell, R., Bai, L., Gulati, A., Tanzer, G., Vincent, D., Pan, Z., Wang, S., et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

Team, Q. Qvq: To see the world with wisdom, December 2024a. URL https://qwenlm.github.io/blog/qvq-72b-preview/.

Team, Q. Qwq: Reflect deeply on the boundaries of the unknown, November 2024b. URL https://qwenlm.github.io/blog/qwq-32b-preview/.

Team, Q. Qwen2.5-vl, January 2025. URL https://qwenlm.github.io/blog/qwen2.5-vl/.

Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., Fan, Y., Dang, K., Du, M., Ren, X., Men, R., Liu, D., Zhou, C., Zhou, J., and Lin, J. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024a.

Wang, Y., Ma, X., Zhang, G., Ni, Y., Chandra, A., Guo, S., Ren, W., Arulraj, A., He, X., Jiang, Z., et al. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *arXiv preprint arXiv:2406.01574*, 2024b.

Wu, J., Deng, W., Li, X., Liu, S., Mi, T., Peng, Y., Xu, Z., Liu, Y., Cho, H., Choi, C.-I., et al. Medreason: Eliciting factual medical reasoning steps in llms via knowledge graphs. *arXiv preprint arXiv:2504.00993*, 2025.

Wu, S., Peng, Z., Du, X., Zheng, T., Liu, M., Wu, J., Ma, J., Li, Y., Yang, J., Zhou, W., et al. A comparative study on reasoning patterns of openai's o1 model. *arXiv preprint arXiv:2410.13639*, 2024.

Xie, Y., Wu, J., Tu, H., Yang, S., Zhao, B., Zong, Y., Jin, Q., Xie, C., and Zhou, Y. A preliminary study of o1 in medicine: Are we closer to an ai doctor? *arXiv preprint arXiv:2409.15277*, 2024.

Xu, R., Wang, Z., Fan, R.-Z., and Liu, P. Benchmarking benchmark leakage in large language models. *arXiv preprint arXiv:2404.18824*, 2024.

Yang, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Li, C., Liu, D., Huang, F., Wei, H., et al. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Yazdani, S. and Hoseini Abardeh, M. Five decades of research and theorization on clinical reasoning: a critical review. *Advances in medical education and practice*, pp. 703–716, 2019.

Yue, X., Ni, Y., Zhang, K., Zheng, T., Liu, R., Zhang, G., Stevens, S., Jiang, D., Ren, W., Sun, Y., et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9556–9567, 2024a.

Yue, X., Zheng, T., Ni, Y., Wang, Y., Zhang, K., Tong, S., Sun, Y., Yu, B., Zhang, G., Sun, H., et al. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. *arXiv preprint arXiv:2409.02813*, 2024b.

Zhang, K., Zeng, S., Hua, E., Ding, N., Chen, Z.-R., Ma, Z., Li, H., Cui, G., Qi, B., Zhu, X., et al. Ultramedical: Building specialized generalists in biomedicine. *Advances in Neural Information Processing Systems*, 37: 26045–26081, 2024.

Zhang, X., Wu, C., Zhao, Z., Lin, W., Zhang, Y., Wang, Y., and Xie, W. Pmc-vqa: Visual instruction tuning for medical visual question answering. *arXiv preprint arXiv:2305.10415*, 2023.

Zhu, K., Zheng, Y., and Chan, K. C. G. Weighted brier score–an overall summary measure for risk prediction models with clinical utility consideration. *arXiv preprint arXiv:2408.01626*, 2024.

A. Leakage Prevention Statement

- The data of MedXpertQA is strictly intended for model or human medical evaluation purposes only. **Please DO NOT use it for any form of training**, including training on synthesized or rephrased versions of the data.
- To mitigate the potential data leakage risks, we refrain from releasing the data sources and request that you **do not share any example of MedXpertQA online**, whether in plain text, image, or any other format.
- We obtained all data from **freely and publicly accessible** sources. We only retained a small percentage of source data, and all questions were subject to question rephrasing and option shuffling, fully ensuring that MedXpertQA complies with **U.S. fair use laws**.

B. Discussions

B.1. Ethical Considerations

To evaluate the reliability of MedXpertQA under real-world clinical variability, we consider the impact of input incompleteness on model performance. Specifically, we propose future assessments using semantically equivalent clinical scenarios with systematically omitted salient information to quantify the sensitivity of model predictions to such variations. This is motivated by the stochastic nature of LLMs and their reliance on prompt phrasing. Furthermore, we highlight that benchmark performance does not imply clinical readiness. MedXpertQA, while constructed through rigorous expert filtering and augmentation, remains a synthetic evaluation framework. Safe use of medically oriented language models requires external safeguards beyond benchmark scores, including uncertainty estimation, structured oversight, and integration with clinical workflows. These considerations are essential for aligning benchmark utility with responsible use in healthcare contexts.

B.2. Comparison with HLE (Med)

To illustrate the difficulty of MedXpertQA, we compare it against HLE (Phan et al., 2025). HLE is derived from original questions contributed by nearly $1,000$ experts representing over 500 institutions across 50 countries. It has attracted considerable attention and is now regarded as one of the most challenging benchmarks for evaluating current state-of-the-art models. MedReason (Wu et al., 2025) specifically evaluated a fine-grained medical subset of HLE, referred to as HLE (Med), which is distinct from the B/M subset introduced in Phan et al. (2025). This medical subset provides new opportunities for detailed comparisons in terms of dataset statistics and evaluation difficulty.

Notably, MedXpertQA demonstrates a higher level of challenge compared to HLE (Med) as shown in Wu et al. (2025). MedXpertQA is not only more comprehensive but also presents significantly greater difficulty. Despite the extensive human effort involved in the construction of HLE (Med), MedXpertQA is approximately **24**x larger and imposes a more demanding evaluation. These characteristics position MedXpertQA as the most comprehensive and challenging medical question-answering benchmark currently available.

C. Additional Implementation Details

C.1. AI Expert Models

Specifically, the basic AI experts include `Qwen2.5-7B` (Qwen Team, 2024) , `InternLM2.5-7B` (Cai et al., 2024), and `LLaMA-3.1-8B` (Dubey et al., 2024) for text-only medical questions filtration, `Qwen2-VL-7B` (Wang et al., 2024a), `InternVL2-8B` (Chen et al., 2024c), and `LLaMA-3.2-11B-Vision` (Dubey et al., 2024) for multi-modal medical questions filtration. The advanced AI experts include proprietary models `gpt-4o-2024-08-06` and `claude-3-5-sonnet-20241022`.

C.2. Evaluation Settings

For both `o1` and `DeepSeek-R1`, we follow the corresponding recommended prompting guidelines to remove the system prompt. For `DeepSeek-R1`, we also include the prompt: *Put your final answer within* $\backslash boxed\{\}$. Additionally, the API parameter `temperature` for these two models is unsupported. For `QVQ-72B-Preview`, since it only supports single-round conversation, we remove the second round of zero-shot CoT and adjust the answer extraction script to handle choices within $\boxed{}$.

D. Error Analysis

In this section, we analyze the reasons behind model errors by classifying different models' incorrect answers into several error types. We consider the following error types:

- ***Reasoning Process Error:*** The model's prediction rationale indicates errors in key reasoning steps, which led to the incorrect answer.
- ***Perceptual Error (for MedXpertQA MM only):*** The incorrect answer stems from a misunderstanding or misinterpretation of the image or images provided in the question.
- ***Question Understanding Error:*** The answer shows an incorrect understanding of the original question.
- ***Lack of Medical Knowledge:*** The model's prediction shows a lack of medical knowledge necessary for arriving at the correct answer.
- ***Formatting Error:*** The answer includes the correct content but is formatted improperly and causes the answer extraction process to fail.

For each model analyzed below, we sample 200 incorrectly answered questions from the MedXpertQA Text and MedXpertQA MM subsets, respectively. We use `gpt-4o-2024-11-20` to label each error type based on the question, correct answer, correct explanation, incorrect answer, and incorrect prediction rationale produced by the model.

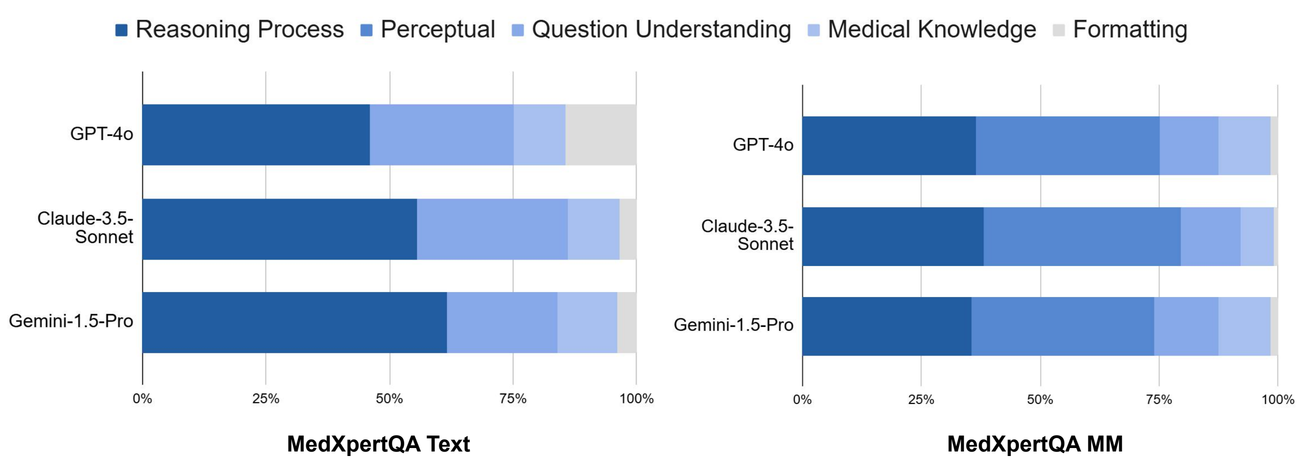


Figure 6: Error type distribution of three models on MedXpertQA.

Figure 6 compares error type distributions across the following models: `gpt-4o-2024-11-20`, `claude-3-5-sonnet-20241022`, `gemini-1.5-pro`. Across all models, the high percentage of *Reasoning Process Errors* on both MedXpertQA Text and MedXpertQA MM reflects the difficulty of our benchmark in terms of medical reasoning. *Perceptual Errors* are also prominent in MedXpertQA MM, demonstrating not only room for improvement in medical image perception for leading models, but also that image interpretation is a core part of answering MedXpertQA MM questions, as expected.

In Appendix G.2, we provided examples of 4 error types with detailed analysis.

E. Expert Review

We summarize the error types found and corrected during the expert review of augmented questions and options. Table 7 provides a statistical overview.

Subset	Error Type	# Questions
MedXpertQA Text	Question Formatting Error	15
	Narrative Inconsistency	5
	Information Inconsistency	73
	Option Formatting Error	21
	Inapplicable Options	399
	Unreasonable Options	15
MedXpertQA MM	Question Formatting Error	100
	Contained Question Source Information	53
	Narrative Inconsistency	70
	Information Inconsistency	217

Table 7: Statistics of errors identified during the human expert review process.

Appendix H lists reviewer guidelines for each phase of expert review in our benchmark construction and annotation.

F. Full Results of Performance Comparison

Table 8 presents the results of different models on MedXpertQA Text and other text benchmarks.

Benchmark	Qwen2.5-72B	GPT-4o	o1	DeepSeek-R1
MMLU (Medical)	89.62	91.90‡	95.30‡	94.12
MedQA-USMLE	72.70†	84.40‡	96.00‡	92.38
MedMCQA-Dev	66.20†	76.40‡	83.90‡	79.12
MedXpertQA Text	18.90	30.37	44.67	37.76

Table 8: Performance of different models on MedXpertQA Text and other benchmarks. † indicates the result is derived from Chen et al. (2024a). ‡ indicates the result is derived from Nori et al. (2024).

G. Cases

G.1. Question Cases

G.1.1 Example of MedQA

Reasoning

Question: A 3-year-old boy presents to the emergency department with a 'cough-like-a-seal bark' and a high-pitched inspiratory noise that is audible without a stethoscope. His mother reports that his cough has worsened over the last few hours. The patient's blood pressure is 118/78 mm Hg, pulse is 90/min, respiratory rate is 35/min, and temperature is 38.3°C (101.1°F). On physical examination, the boy is sitting and leaning forward in apparent respiratory distress with suprasternal and intercostal retractions. Auscultation reveals inspiratory stridor without wheezing. He has a frequent barking cough and a hoarse voice when he speaks. What is a chest X-ray likely to show?
Answer Choices: (A) Increased interstitial markings (B) Lobar consolidation in the lingual (C) Thumbprint sign on the lateral image (D) Steeple sign

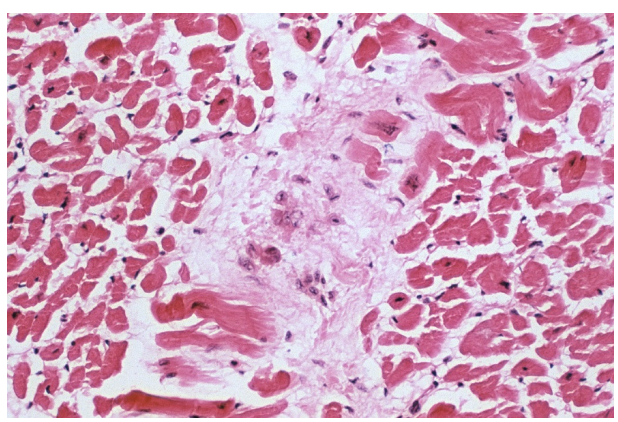
Understanding

Question: A 3-month-old boy is brought the emergency department by his parents after an episode of cyanosis and muscle hypotonia that resolved after 2 minutes. Diagnostic evaluation fails to discover an exact etiology of the boy's symptoms and the episode is classified as a brief resolved unexplained event (BRUE). The risk profile for BRUE in infants remains largely unknown. The pediatrician who saw the boy in the emergency department is trying to identify risk factors for BRUE. She is aware of several confounders, including age, socioeconomic background, and family history of medical illness. She recruits 75 infants under 1 year of age with BRUE and 75 infants without BRUE of the same age, socioeconomic background, and family history of medical illness. She then compares the two groups with regard to history of feeding problems and history of recent upper respiratory infection. Which of the following methods was conducted to control confounding bias in the study?
Answer Choices: (A) Blinding (B) Restriction (C) Randomization (D) Matching

G.1.2 Example of MedXpertQA MM

Reasoning

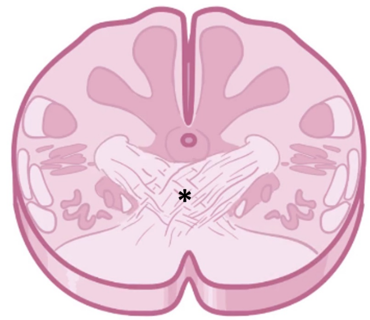
Question: A 30-year-old male presents to primary care with complaints of gradually worsening fatigue and shortness of breath over several months. His medical history reveals no smoking, alcohol use, or illicit drug use, and his family history is noncontributory. He reports travel to South America one year prior but cannot recall any illness afterward. His vital signs show a temperature of 37.0°C (98.6°F), heart rate of 75/min, and blood pressure of 131/80 mmHg. Clinical examination reveals visible respiratory distress, a systolic murmur at the cardiac apex, and bilateral diffuse rales on lung auscultation. A cardiac biopsy is performed. Based on the biopsy findings shown, what is the most probable cause of the patient's condition?



Answer Choices: (A) Autoimmune granulomatous disease (B) Viral infection (C) Bacterial infection (D) Fungal infection (E) Parasitic infection

Understanding

Question: A pathologist views cross-sectional slides of the brainstem and identifies the following structure (denoted by an asterisk).
Which of the following best characterizes the function of the fibers passing through this structure?
Answer Choices: (A) Transmission of pain signals to the brain (B) Transmission of unconscious proprioceptive sensation (C) Conduction of temperature sensation (D) Initiation of voluntary arm movements (E) Regulation of muscles of mastication



G.1.3 Example of MedXpertQA Text

Reasoning

Question: A 68-year-old right-handed male is brought to the emergency department for evaluation of sudden-onset weakness. He woke up with clumsiness in his left hand, which quickly progressed to weakness on his left side, more pronounced in the arm than the leg. His medical history includes hypertension, type 2 diabetes mellitus, and hyperlipidemia, but he has not refilled his medications in a year. He has a 40-year history of smoking one pack of cigarettes daily. Vital signs reveal a temperature of 37.3°C (99.1°F), pulse of 98/min, respirations of 20/min, and blood pressure of 200/110 mm Hg. Physical examination shows slurred speech, left-sided facial weakness sparing the forehead, 2/5 strength in the left upper limb, and 4/5 strength in the left lower limb, with normal strength and sensation on the right side. MRI reveals a right-sided infarct in the paramedian pontine base. Which additional finding is most likely in this patient?
Answer Choices: (A) Ipsilateral tongue deviation (B) Contralateral loss of proprioception (C) Contralateral hemiparesis sparing the face (D) Eye deviation to the left (E) Vertical gaze palsy (F) Ipsilateral facial weakness involving the forehead (G) Eye deviation to the right (H) Loss of pupillary light reflex on the right (I) Bitemporal hemianopia (J) Contralateral hemianopia with macular sparing

Understanding

Question: A 25-year-old previously healthy man presents to the clinic with cervical pain and stiffness that developed after a long-duration flight from California to New York the previous day. His symptoms have progressively worsened since onset. He reports no trauma to the area and denies other symptoms. Physical examination reveals a patient in mild discomfort with normal vital signs and no fever. The paraspinal region from C5-C7 demonstrates hypertonicity. Which motion represents the primary movement of the affected cervical segments?
Answer Choices: (A) Coupled rotation (B) Extension (C) Flexion (D) Sidebending (E) Lateral flexion (F) Axial compression (G) Rotation (H) Retraction (I) Translation (J) Protraction

G.2. Error Cases

Reasoning Process Error

Question: A 64-year-old man is brought to the emergency room following a fall at his farm, presenting with a shoulder injury shown in Figure A. After successful reduction, post-reduction images are obtained (Figures B and C). During his follow-up appointment 10 days later, he demonstrates an inability to perform shoulder abduction. Which diagnostic test would be most appropriate to confirm the suspected diagnosis?

Answer Choices: (A) CT-angiogram of the affected extremity (B) EMG (C) MRI of the brachial plexus (D) MRI of the shoulder (E) Repeat shoulder x-rays

Correct Answer: (D)

Model Answer: (B)

Prediction Rationale:

The presented scenario involves a 64-year-old man who sustained a shoulder injury after a fall and subsequently shows an inability to perform shoulder abduction following the incident. This could suggest nerve injury, specifically to the axillary nerve, which is responsible for innervating the deltoid muscle, crucial for shoulder abduction.

Here's the step-by-step breakdown:

1. Injury and Reduction: The man had a shoulder injury, likely a dislocation, based on initial x-rays and description. The reduction was successful, as indicated by follow-up x-rays.
2. Symptoms Post-Reduction: After 10 days, the patient demonstrates an inability to abduct the shoulder.
3. Common Cause of Inability to Abduct: Injury to the axillary nerve can cause an inability to abduct the shoulder, as it innervates the deltoid muscle.
4. Most Appropriate Diagnostic Test: Electromyography (EMG) is often used to assess the function of muscles and nerves. If there is a suspicion of nerve injury, such as the axillary nerve in this case, an EMG would be helpful in confirming nerve dysfunction or damage.

Considering the options: (A) CT-angiogram of the affected extremity would assess vascular issues, not nerve. (B) EMG would assess nerve injury. (C) MRI of the brachial plexus could be used for structural visualization of the nerve, but isn't typically the first choice for function assessment. (D) MRI of the shoulder focuses on soft tissue and structural injury, not specific nerve function. (E) Repeat shoulder x-rays assess bone placement, not nerve or muscle function.

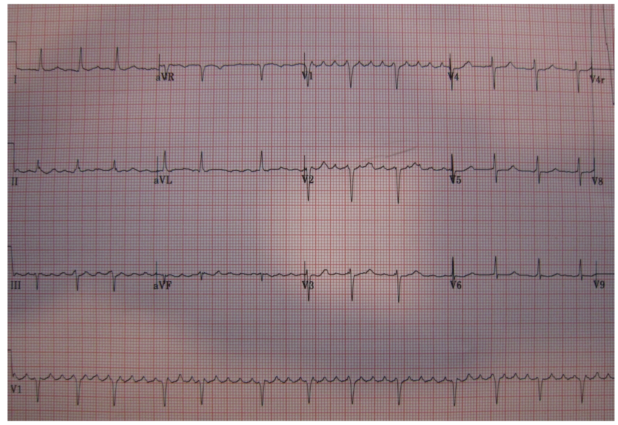
Based on this analysis, the most appropriate diagnostic test to confirm the suspected axillary nerve injury would be (B) EMG.

Analysis:

Upon approaching this question, the model correctly analyzed the provided X-ray images to determine that reduction for the patient had been successful and that X-rays showed no bone injuries. However, it proceeded to conclude that injury to the axillary nerve must have caused the patient's inability to abduct the shoulder, and followed this reasoning process to select the corresponding test, (B) EMG. Although the model was correct in stating that axillary nerve injuries can impact the deltoid muscle, demonstrating correct medical knowledge, it does not sufficiently compare the likelihoods of several candidate diagnoses. Axillary nerve injury is relatively unlikely to fully account for the severe shoulder injury reported, and in comparison, a rotator cutoff tear is likely to occur in this demographic, and also better explains the patient's symptoms. This conclusion leads to the correct diagnostic test, (D) MRI of the shoulder. This case reveals that the model followed the first plausible explanation it found and could not perform the nuanced reasoning necessary for weighing several possibilities. Multi-step reasoning is a notable challenge in this question - the answerer needs to first determine the most likely hypothetical diagnosis, then select the medical examination accordingly. The candidate diagnoses not being explicitly listed likely contributed to the model's oversimplified reasoning process.

Perceptual Error

Question: A 57-year-old man is being monitored in the cardiothoracic intensive care unit following an uncomplicated coronary artery bypass graft surgery. During overnight monitoring, the patient's telemetry alerts staff to a tachyarrhythmia, prompting a rapid response call. His vital signs show: temperature 37.0°C (98.6°F), heart rate 140/min, respiratory rate 14/min, and blood pressure 134/72 mmHg. The patient reports only post-surgical chest pain at the incision site. Given the electrocardiogram findings shown below and the patient's clinical presentation, what is the most probable diagnosis?



Answer Choices: (A) Atrial flutter (B) Multifocal atrial tachycardia (C) Atrial fibrillation with rapid ventricular response (D) Paroxysmal supraventricular tachycardia (E) Sinus tachycardia

Correct Answer: (A)

Model Answer: (C)

Prediction Rationale:

To determine the most probable diagnosis, we need to assess the clinical information alongside the ECG findings.
1. Heart Rate: The patient has a heart rate of 140 beats per minute, which indicates tachycardia.
2. ECG Analysis: The ECG shows an irregularly irregular rhythm. There are no distinct P waves visible before each QRS complex; instead, there is a disorganized electrical activity, which is typical of fibrillatory waves.
3. Differential Diagnosis: (A) Atrial flutter: Typically gives a "saw-tooth" pattern and usually has a more regular rhythm. (B) Multifocal atrial tachycardia (MAT): Characterized by an irregular rhythm with at least three different P wave morphologies which isn't apparent here. (C) Atrial fibrillation with rapid ventricular response: Characterized by an irregularly irregular rhythm without distinct P waves, which fits the ECG findings. (D) Paroxysmal supraventricular tachycardia (PSVT): Usually presents with a regular rhythm and wouldn't have the irregular characteristics seen in this case. (E) Sinus tachycardia: Would have a regular rhythm and identifiable P waves.
4. Post-Surgical Context: Atrial fibrillation (AF) is relatively common after cardiac surgery, such as coronary artery bypass grafting, due to inflammation, electrolyte imbalances, and other stressors on the heart.
Considering the irregularly irregular rhythm and absence of distinct P waves, the most probable diagnosis is: (C) Atrial fibrillation with rapid ventricular response.

Analysis:

The provided ECG demonstrates characteristic flutter waves with a 3:1 and 4:1 AV nodal block, indicative of (A) atrial flutter. The model correctly states that atrial flutter typically gives a "saw-tooth" pattern, but fails to determine that such patterns are present in the image. Instead, it incorrectly interprets the ECG as having an irregularly irregular rhythm indicative of (C) atrial fibrillation. The main distinction between these two diagnoses is the regularity of atrial activations seen in the ECG, a detailed piece of visual information that is challenging for models to accurately interpret.

Question Understanding Error

Question: A 60-year-old male with multiple comorbidities including coronary artery disease (status post stent placement 3 months ago), hypertension, diabetes, and mild to moderate asthma presents with a two-day history of cough and wheeze. His medication regimen includes aspirin, atorvastatin, atenolol, metformin, and hydrochlorothiazide. He recently added salbutamol inhaler and prednisolone for his respiratory symptoms. The patient reports experiencing occasional low blood glucose readings over the past few weeks. Physical examination reveals stable vital signs with no respiratory distress. Which medication in his current regimen is the most likely cause of his hypoglycemic episodes?
Answer Choices: (A) Aspirin (B) Ibuprofen (C) Enalapril (D) Naproxen (E) Atorvastatin (F) Prednisolone (G) Atenolol (H) Hydrochlorothiazide (I) Phenformin (J) Metformin
Correct Answer: (A)
Model Answer: (G)
Prediction Rationale:
When analyzing this case, we need to identify which medication in the patient's regimen is the most likely cause of his hypoglycemic episodes. Here is a step-by-step approach:
Step 1: Review the patient's medications: 1. Aspirin – Antiplatelet therapy. 2. Atorvastatin – Lipid-lowering agent (statin). 3. Atenolol – A beta-blocker used for hypertension. 4. Metformin – A first-line diabetes medication. 5. Hydrochlorothiazide – A diuretic used for blood pressure control. 6. Salbutamol (recently added) – A short-acting beta-agonist for asthma. 7. Prednisolone (recently added) – A corticosteroid for respiratory inflammatory issues.
Step 2: Focus on the patient's problem – hypoglycemia: The patient reports low blood glucose readings. Hypoglycemia can be caused by certain medications, especially those that: 1. Lower blood glucose directly (e.g., diabetes medications). 2. Affect glucose metabolism indirectly.
Step 3: Evaluate each medication for its relationship to hypoglycemia
- Aspirin: At high doses, it can rarely cause hypoglycemia. However, at low doses (commonly used for antiplatelet therapy), this is unlikely. Not the most likely cause.
- Ibuprofen and Naproxen (if taken, not explicitly listed in the regimen): NSAIDs may not be directly relevant to hypoglycemia unless used at high doses affecting insulin function. Likely not relevant here.
- Enalapril: ACE inhibitors can rarely cause hypoglycemia by increasing insulin sensitivity. However, this medication is not listed in his current regimen. Irrelevant.
- Metformin: Unlikely to cause hypoglycemia as it does not stimulate insulin secretion. Its primary effect is reducing hepatic gluconeogenesis. Not the cause.
- Atorvastatin: Statins are not associated with significant risks of hypoglycemia. Not the cause.
- Prednisolone: Corticosteroids typically increase blood glucose (hyperglycemia), not decrease it. Not the cause.
- Atenolol: Beta-blockers, particularly non-cardioselective ones, can mask the symptoms of hypoglycemia (e.g., tachycardia) and impair the response to hypoglycemia (e.g., by reducing glycogenolysis). Cardioselective beta-blockers, such as atenolol, are less likely to do this but can still blunt the counterregulatory response during hypoglycemia. Possible contributor.
- Hydrochlorothiazide: Thiazides are more typically associated with hyperglycemia. Unlikely to cause hypoglycemia.
Step 4: Consider the most likely cause: The most likely cause of the hypoglycemia is Atenolol because: 1. Beta-blockers impair the body's natural response to a drop in glucose by reducing glycogenolysis and gluconeogenesis. 2. They also mask typical warning signs of hypoglycemia (e.g., palpitations or tremors), which may lead to unrecognized or prolonged hypoglycemic episodes.

Analysis:
The above response fails to fully grasp the question's implications in several ways. For the incorrect answer (G) Atenolol, the model generally states that some beta-blockers reduce glycogenolysis and gluconeogenesis, but does not consider the specificity of atenolol - as a selective beta-1 blocker, it has less effect on glucose metabolism. While the claim that it masks warning signs of hypoglycemia is more accurate, this does not align with the question's emphasis on finding a cause of hypoglycemia. For the correct answer (A) Aspirin, the model's judgment is only accurate regarding non-diabetic individuals, while ample evidence supports aspirin's role in decreasing glucose levels in diabetic patients. The remaining parts of the model's analysis similarly disregard the context of the patient's diabetes. The response is therefore inconsistent with the patient's general background presented in the question.

Lack of Medical Knowledge

Question: A 69-year-old male visits his primary care physician for a routine checkup. He maintains an active lifestyle since retirement, engaging in daily exercise and home projects, and follows a balanced diet with regular evening time with his spouse. However, he reports feeling less refreshed upon waking compared to his younger years, when he slept 9 hours and felt excellent, versus his current 7 hours of sleep. His wife reports that his sleep appears peaceful. His medical conditions include hypertension and diabetes, managed with lisinopril and metformin. Physical examination reveals: temperature 98.6°F (37°C), blood pressure 131/85 mmHg, pulse 71/min, respirations 12/min, normal neurological findings, good muscle tone, healthy weight, and pleasant affect. He denies current fatigue or tiredness. Which physiological alteration would most likely be observed in this patient?
Answer Choices: (A) Reduced acetylcholine (B) Increased serotonin (C) Decreased orexin (D) Altered circadian rhythm proteins (E) Elevated histamine levels (F) Decreased melatonin (G) Elevated cortisol levels (H) Increased dopamine (I) Decreased growth hormone (J) Increased adenosine
Correct Answer: (F)
Model Answer: (I)
Prediction Rationale:
This scenario describes an older adult experiencing reduced restorative sleep, which is a common part of the aging process. To determine the most likely physiological alteration, it is important to consider how aging impacts sleep. Key details from the case:
1. Age-related sleep changes: The patient is 69 years old. Aging is associated with alterations in sleep architecture, which may lead to reductions in total sleep duration, slow-wave (deep) sleep, and sleep efficiency. Older adults often feel less refreshed upon waking, even when sleep appears otherwise normal.
2. Patient's sleep specifics: He now sleeps 7 hours and wakes feeling less refreshed compared to his younger years when sleeping 9 hours. His sleep appears peaceful, and he experiences no disruptive symptoms such as snoring or awakenings.
3. Ruling out other causes: Good management of hypertension and diabetes, normal vitals, a well-rounded active lifestyle, and no reported fatigue or significant tiredness suggest no acute medical condition contributing to his symptoms.
4. Physiological changes associated with aging: Aging often leads to a decrease in growth hormone, which is secreted primarily during slow-wave sleep and stimulates restorative processes during sleep. This can result in less restorative sleep and a less refreshed feeling upon waking.
5. Other options to consider: Decreased melatonin (option F): Melatonin levels decline with age, but melatonin primarily regulates sleep initiation and circadian rhythm rather than sleep quality or restoration. Altered circadian rhythm proteins (option D): Circadian changes occur with aging, but his symptoms are more closely linked to sleep restoration rather than timing. Other neurotransmitters like acetylcholine, orexin, serotonin, histamine, cortisol, and adenosine are not directly relevant to the described problem.
Correct Answer: (I) Decreased growth hormone. This is the most likely physiological change contributing to the patient's symptoms, as reduced growth hormone secretion in older adults is linked to decreased restorative properties of sleep.

Analysis:
The question generally describes a patient experiencing reduced sleep quality, commonly associated with aging. The model's response reflects its insufficient medical knowledge on the different effects of (D) decreased growth hormone and (F) decreased melatonin on sleep. While both are indeed correlated with aging, growth hormones primarily promotes deep sleep, or slow-wave sleep (SWS), while decreased melatonin more directly accounts for the subjective lack of restfulness upon waking referred to in the question.

G.3. Evaluation Case

System: You are a helpful medical assistant.
User: Q: A 57-year-old male presents to the emergency department with a one-week history of weakness and subjective fever. His medical history includes IV drug use, hepatitis C, atrial fibrillation, cirrhosis, alcohol dependence, obesity, and depression. On examination, his temperature is 102°F (38.9°C), blood pressure is 157/98 mmHg, pulse is 110/min, respirations are 17/min, and oxygen saturation is 98% on room air. The patient appears fatigued with diffuse jaundice. Cardiopulmonary examination reveals bibasilar crackles, and abdominal examination shows distension, dullness to percussion, a fluid wave, and generalized tenderness. The patient is admitted to the hospital and started on piperacillin-tazobactam. By day 4, he is afebrile, with a pulse of 92/min, reduced abdominal tenderness, and persistent jaundice. Nurses report bleeding from two peripheral IV sites requiring pressure to control, along with new bruises on the patient's arms and legs. What is the most likely explanation for this patient's condition?
Answer Choices: (A) Fibrinolysis activation due to liver failure (B) Vitamin K deficiency due to liver dysfunction (C) Diffuse activation of the coagulation cascade (D) Disseminated intravascular coagulation (DIC) (E) Platelet dysfunction secondary to sepsis (F) Decreased production of coagulation factors (G) Immune-mediated platelet destruction (H) Hypersplenism causing thrombocytopenia (I) Decreased metabolism of an anticoagulant (J) Bacterial destruction
A: Let's think step by step.
Assistant: Analysis:
This patient has advanced liver disease (cirrhosis) and is hospitalized on broad-spectrum antibiotics for four days. Although his infection is improving (afebrile, lower heart rate, less abdominal tenderness), he has developed new signs of a bleeding diathesis (bleeding from IV sites, new bruises). In a patient with chronic liver disease, an "acute on chronic" coagulopathy like this most often points to additional impairment in clotting-factor synthesis, above and beyond what is already chronically reduced by cirrhosis.
Two common acute drivers of coagulopathy in cirrhotic patients are:

- Vitamin K deficiency (especially when on broad-spectrum antibiotics that eradicate gut flora responsible for synthesizing vitamin K).
- Disseminated intravascular coagulation (DIC), usually in the setting of sepsis or shock, but here the patient's sepsis is improving, making DIC less likely.

Thus, the best explanation in a cirrhotic patient, newly on broad-spectrum antibiotics, who suddenly develops worse bleeding tendencies, is vitamin K deficiency superimposed on his baseline liver dysfunction.
Answer: (B) Vitamin K deficiency due to liver dysfunction.
User: Therefore, among A through J, the answer is
Assistant: The correct answer is (B) Vitamin K deficiency due to liver dysfunction.

H. Expert Review Guideline

GUIDELINES FOR EXPERT REVIEW OF AUGMENTED QUESTIONS.

```
**Intended Audience:** Human medical expert reviewers
**Objective:** Conduct a comprehensive review and proofreading of medical questions rewritten by large language models to ensure that the questions are accurate, clear, scientific, practical, and free from hallucinations or factual errors.

## 1. Review of Original Question

**Objective:** Verify whether the original question contains any logical or factual errors, is incomplete, or unanswerable. If such issues exist, no further steps should be taken, and the specific situation should be recorded.

**Requirements:** Clearly indicate the problems in the original question. Mark the question as invalid.

## 2. Review of Rewritten Question

### 2.1 Content Accuracy Check

**Objective:** Ensure the rewritten question maintains medical knowledge accuracy and is free from factual errors.

**Key Points:** Check whether any key information from the original question is missing. Verify that no inaccurate or deviated content has been introduced in the rewrite. Ensure that no content contradicts medical facts or deviates from the original context.

### 2.2 Language Clarity and Expression Quality Check

**Objective:** The language of the rewritten question should be precise, clear, and professional.

**Key Points:** Ensure the language is professional and avoids overly casual or colloquial expressions. Confirm that medical terms are used accurately and scientifically, clearly conveying the meaning of the original question. Verify that the expression is concise and clear, avoiding ambiguity or unclear phrasing.

### 2.3 Logical Consistency Check

**Objective:** Ensure the question follows a logical flow, has a reasonable structure, and is answerable.

**Key Points:** Check whether the question fits into real-world clinical or research context. Look for logical errors such as contradictions, lack of causality, reasoning mistakes, or illogical phrasing. Ensure the question structure is clear, complete, coherent, and has a clear focus.

### 2.4 Other Issues

**Tablular Data:** If the original question includes tabular data, verify that the rewritten question retains the original data's content and format. The rewritten table should not omit any data or alter the format.

**Medical Terminology Abbreviations:** If the rewritten question replaces the full medical term with an abbreviation, verify whether the abbreviation is standard and in line with medical conventions. If professional medical practitioners can understand the abbreviation in the context of the question, it is considered acceptable.
```

Table 9: Guidelines for expert review of augmented questions (Part 1).

GUIDELINES FOR EXPERT REVIEW OF AUGMENTED QUESTIONS.

```
## 3. Summary of Requirements

**Strict Adherence to Standards:** All medical experts must strictly follow these
↪  guidelines for evaluation and proofreading to ensure the accuracy, scientific
↪  integrity, and practicality of the questions.

**Quality of Evaluation Results:** The content must be accurate, with reliable facts. The
↪  language expression should be clear and meet professional medical standards.
The logic should be sound, and the questions should be answerable.

**Maintain Objectivity:** Maintain objectivity and fairness throughout the evaluation and
↪  proofreading process, avoiding subjective judgment or misleading rewrites.
```

Table 10: Guidelines for expert review of augmented questions (Part 2).

GUIDELINES FOR EXPERT REVIEW OF AUGMENTED OPTIONS.

```
**Intended Audience:** Human medical expert reviewers

**Objective:** Conduct a comprehensive review and proofreading of the expanded options generated by large language models for medical multiple-choice questions, ensuring the options are accurate, clear, reasonable, and challenging.

## 1. Option Inspection Standards

### 1.1 Content Rationality Check

**Objective:** Ensure the expanded options are reasonable and acceptable within the context of the original question.

**Key Points:** Avoid adding options that do not align with the question's context. For example, if there are 4 images in the question, generating an option for \Image 5" would be unacceptable. Avoid adding options that do not correspond to the topic of the question. For example, if the question asks to choose from several time periods, do not generate options with unrelated content.

### 1.2 Language Expression and Professionalism Check

**Objective:** Ensure the language of the options is accurate, clear, and professional.

**Key Points:** The language style of the expanded options should match that of the original options. The language should be professional, avoiding overly casual or colloquial expressions. Medical terminology should be used accurately and in line with current medical standards. The expression should be clear, concise, and avoid ambiguity or overly complex phrasing.

### 1.3 Hallucination Check

**Objective:** Ensure that the original options do not contain obvious hallucinations or errors.

**Key Points:** The original options should not be changed.
The generated options should not contain hallucinated information, i.e., content that is disconnected from the original question context or lacks scientific basis. There should be no repetition of content between options, and no redundant information in the options.

## 4. Summary and Requirements

**Strict Adherence to Standards:** All medical experts must strictly follow these guidelines for evaluation and proofreading to ensure the accuracy, scientific integrity, and practicality of the options.

**Quality of Evaluation Results:** Content must be accurate and reliable. Language expression should be clear and meet professional medical standards. Option design should be logically sound and have a high degree of differentiation.

**Maintain Objectivity:** Maintain objectivity and fairness throughout the evaluation and proofreading process, avoiding subjective judgment or misleading expansions.
```

Table 11: Guidelines for expert review of augmented options.

I. Prompts

The following lists all prompts used in this work.

I.1. Attribute Annotation Prompts

PROMPT FOR TASK ANNOTATION.

```
You are an experienced medical doctor and independent practitioner. You will be asked to
↪  label a medical question according to its task.

You will be given a list of tasks and their corresponding explanations and descriptions.
↪  You will then be given a medical examination question.
Please determine which task the question best pertains to.

If the question is related to multiple tasks, only select the most relevant one.
Directly output the task name you have selected. Do not provide any additional
↪  explanations. Ensure that it is a valid task within the list of available tasks
↪  provided.

**Tasks:**

Diagnosis: Tasks that occur in the diagnosis stage of a medical examination. These include
↪  identifying diseases based on symptoms, deducing the causes of symptoms or diseases,
↪  predicting disease prognosis or additional symptoms, and proposing intermediate steps
↪  (such as further examinations) to facilitate diagnosis.

Treatment: Tasks that occur in the treatment stage of a medical examination. These include
↪  selecting treatments for diseases, proposing preventive measures for diseases, and
↪  suggesting lifestyle changes for the patient.

Basic Medicine: Tasks that focus on basic scientific principles that may be applied in
↪  medical practice. These include understanding anatomy, diseases and medical
↪  procedures, and general STEM principles.

**Question:**

{question_text}

**Output:**
```

Table 12: Prompt for Task Annotation.

PROMPT FOR SUBTASK ANNOTATION.

```
You are an experienced medical doctor and independent practitioner. You will be asked to
↪  label a medical question's subtask according to its task.

You will be given a medical examination question and its task type. You will then be given
↪  a list of subtypes within the predefined type. Please determine which subtask the
↪  question best pertains to.
If the question is related to multiple subtasks, only select the most relevant one.
↪  Directly output the subtask name you have selected. Do not provide any additional
↪  explanations. Ensure that it is a valid task within the list of available tasks
↪  provided.

**Task:** Diagnosis

**Subtasks:**

Primary, Differential, Etiologic, Prognostic, Retrospective, Syndromic, Predictive,
↪  Intermediate

**Question:**

{question_text}

**Output:**
```

Table 13: Prompt for Subtask Annotation. We use the prompt for diagnosis tasks as an example. Each question's prompt will contain the specific subtasks within the question's labeled task.

PROMPT FOR SYSTEM ANNOTATION.

```
You are an experienced medical doctor and independent practitioner. Your task will be to
↪  label a medical question according to the human body system it corresponds to.

You will be given a list of human body systems, followed by a medical question. Please
↪  determine which system the question best pertains to. If the question is related to
↪  multiple systems, only select the most relevant one.
Directly output the name of the final system you selected from th list of available
↪  systems.

**Systems:**
Integumentary, Skeletal, Muscular, Nervous, Endocrine, Cardiovascular, Lymphatic,
↪  Respiratory, Digestive, Urinary, Reproductive

**Question:**

{question_text}

**Output:**
```

Table 14: Prompt for System Annotation.

PROMPT FOR REASONING AND UNDERSTANDING ANNOTATION.

```
You are an experienced medical doctor and independent practitioner. Your task will be to
↪  determine whether a medical question primarily challenges the answerer's medical
↪  knowledge understanding or medical reasoning skills.

You will be given a challenging medical exam question designed for professional medical
↪  practitioners, followed by its correct answer and explanation to help you understand
↪  the question.
Please determine whether the question primarily challenges the answerer's medical
↪  knowledge understanding or medical reasoning ability.
Consider the amount of information provided in the question, as well as the number and
↪  complexity of reasoning steps required to answer it.
You will be given the question you will label with its answer and explanation. Directly
↪  output either "Understanding" Or "Reasoning" as your answer without any additional
↪  information or explanations.

**Question:**

{question_text}

**Explanation:**

{explanation_text}

**Output:**
```

Table 15: Prompt for annotating questions based on their focus on either reasoning or understanding skills, used for constructing a reasoning subset of MedXpertQA.

I.2. Data Augmentation Prompts

PROMPT FOR QUESTION REPHRASE.

```
### Task

Rephrase a given medical multiple-choice question according to the following guidelines:

### Suggested Approaches

Consider the following strategies to rephrase the question:

1. **Language Rewriting**: Use different expressions and sentence structures that are equivalent in meaning to the original.

2. **Information Restructuring**: Alter the order that information is presented in the original question without disrupting the logical coherence of the question.

### General Requirements

1. **Maintain the core content:** Ensure that the rephrased question retains the core content, reasoning logic, and correct answer of the original question.

2. **Ensure a professional language style:** Maintain a professional, formal, and clear language style similar to the original question. Rigorously ensure clarity and avoid ambiguity. Feel free to copy parts of the original question if alternative appropriate phrasing is not possible.

3. **Maintain factual consistency:** Ensure that the rewritten question retains every piece of information in the original. Do not change, add, or delete any factual information.

4. **Imitate original formatting:** Keep any special formatting in the original question unchanged, especially regarding structured data presentation. Pay special attention to keep any tabular data in completely the same format as the original.

5. **Final output format:** Ensure that the options section of the question remains unchanged and the format remains as "Answer Choices: (A) [Option A] (B) [Option B] ...". Only output the rephrased question. Do not include any additional information or explanations.


{demonstrations}

### Input

**Original Question:**
{question}

**Correct Answer:**
{label}

**Explanation:**
The following is an explanation to help with understanding the question:
{explanation}

### Output

**Rephrased Question:**
```

Table 16: Prompt for Question Rephrase.

PROMPT FOR OPTION GENERATION.

```
### Task

Given a medical multiple-choice question designed to evaluate the capabilities of medical
↪  experts and its correct answer, please design additional incorrect options to
↪  challenge experts when they answer the question.

### Suggested Approaches

1. **Consider Errorneous Perspectives:** Include distractors that interpret key
↪  information in the question incorrectly.

2. **Leverage Common Misconceptions:** Consider designing distractors leveraging common
↪  errors or medical concepts that are frequently confused.

3. **Logical Misdirection:** Introduce distractors grounded in logical reasoning that is
↪  seemingly plausible but incorrect.

### General Requirements

1. **Maintain Consistency:** Ensure that the generated new options match the original
↪  options in terms of length, structure, word count, and grammatical form. They should
↪  be clear, concise, and professionally worded.

2. **Avoid Oversimplified Distractors:** Do not include options that can be easily
↪  dismissed based on intuition or surface-level analysis.

3. **Ensure High Plausibility:** Maintain the plausibility of each generated option. Avoid
↪  options that are overtly illogical or unsupported.

4. **Final Format:** Present the original question and options, followed by the
↪  **{generate_num}** additional options. Ensure that the generated options follow the
↪  same format as the original: "Answer Choices: (A) [Option A] (B) [Option B] ...". Do
↪  not output anything after the options.


{demonstrations}

### Input

**Original Question:**
{question}

**Correct Answer:**
{label}

### Output

**{generate_num}** options appended after the original question:
```

Table 17: Prompt for Option Generation.