

# 000 MEDREASON-DX: BENCHMARKING STEP-BY-STEP 001 REASONING OF LANGUAGE MODELS IN MEDICAL DI- 002 AGNOSIS 003

006 **Anonymous authors**

007 Paper under double-blind review

## 011 ABSTRACT

013 In high-stakes domains like medicine, **how** an AI arrives at an answer can be  
014 as critical as the answer itself. However, existing medical question answering  
015 benchmarks largely ignore the reasoning process, evaluating models only on final  
016 answer accuracy. This paper addresses the overlooked importance of reasoning path  
017 evaluation in medical AI. We introduce **MedReason-Dx**, a novel benchmark that  
018 assesses not just answers but the step-by-step reasoning behind them. MedReason-  
019 Dx provides expert-annotated step-by-step solutions for both multiple-choice and  
020 open-ended questions, spanning 24 medical specialties. By requiring models  
021 to produce and be evaluated on intermediate reasoning steps, our benchmark  
022 enables rigorous testing of interpretability and logical consistency in medical  
023 QA. We present the design of MedReason-Dx and outline diverse evaluation  
024 metrics that reward faithful reasoning. Using this benchmark, we identify critical  
025 gaps in existing large language models, including domain-specific medical LLMs,  
026 particularly in their ability to handle complex diagnostic reasoning and reliably  
027 recall relevant medical knowledge. Our analysis further highlights differences  
028 between general-purpose and medical LLMs in reasoning performance. We believe  
029 this resource will advance the development of robust, interpretable medical decision  
030 support systems and foster research into large language models that can reason as  
031 well as they respond.

## 032 1 INTRODUCTION

034 Artificial intelligence systems for healthcare must not only deliver correct answers but also provide  
035 faithful reasoning. In clinical decision support and medical question answering (QA), the reasoning  
036 path leading to an answer is critical for trust and safety. A model that arrives at a diagnosis through  
037 flawed logic or guesswork poses significant risks, even if the final answer is correct. Conversely, a  
038 model that articulates its reasoning enables practitioners to verify each step, ensuring the conclusion  
039 is sound and clinically valid.

040 Despite this importance, most existing benchmarks for medical AI evaluate models solely on whether  
041 the final answer is right, with little or no assessment of the reasoning process. This gap is problematic  
042 in high-stakes domains: evaluating only end answers may overlook dangerous reasoning errors and  
043 fails to encourage the development of models that “think” in a human-like, transparent manner.

044 Recent advances in large language models (LLMs) have brought reasoning to the forefront of AI  
045 research Guo et al. (2025); Chen et al. (2025). With proper training strategies or *chain-of-thought*  
046 (CoT) prompting, LLMs can generate step-by-step solutions to complex problems, from math and  
047 logic puzzles to medical questions. By configuring models to articulate intermediate steps, researchers  
048 have achieved improved performance on challenging tasks and gained insight into model decision-  
049 making. For example, state-of-the-art medical LLMs can now produce explanations or rationales  
050 alongside their answers, showcasing the potential of AI to handle intricate clinical reasoning. These  
051 developments underscore an urgent need for benchmarks that can evaluate not just final accuracy  
052 but the quality of reasoning LLMs employ. If a model is prompted to reason but we lack ground  
053 truth reasoning paths for comparison, we cannot rigorously assess whether the model’s reasoning is  
correct, complete, or clinically valid.

Several medical QA datasets and benchmarks have emerged, yet they predominantly focus on answer correctness. Standard benchmarks drawn from medical exams (e.g., USMLE-style question banks, MedQA (Jin et al., 2021) and MedMCQA (Pal et al. (2022))) and research datasets like PubMedQA (Jin et al., 2019) have driven progress in factual recall and question answering. Some of these resources include a short explanation or reference for the answer, but they do not provide a detailed, stepwise reasoning chain that could be used to evaluate a model’s thought process. In other words, existing benchmarks treat reasoning as an implicit skill, not an explicit target of evaluation. A model might earn full marks by selecting the correct option in a multiple-choice question, while in reality it could have arrived at that answer via incorrect assumptions or lucky guesswork. Conversely, a model might demonstrate mostly correct reasoning and make a minor error at the final step, but current benchmarks would simply mark the entire answer as wrong, offering no credit for nor analysis of the model’s reasoning ability. This limitation hampers the development of robust medical AI: it is difficult to discern whether improvements in accuracy are due to better reasoning or just better pattern matching, and it provides no incentive for models to output interpretable solutions.

To address these challenges, we propose **MedReason-Dx**, a new benchmark explicitly designed to evaluate chain-of-thought reasoning in medical question answering. MedReason-Dx (where “Dx” denotes Diagnosis) introduces several key innovations to the evaluation of medical AI:

- **Expert-annotated reasoning chains:** Each question in MedReason-Dx is accompanied by a step-by-step solution path crafted by medical experts. These reasoning chains detail the logical steps required to arrive at the correct answer, including relevant clinical facts, intermediate inferences, and elimination of distractors in the case of multiple-choice items. This provides a gold-standard trace of correct reasoning against which model-generated solutions can be compared.
- **Diverse question formats and topics:** MedReason-Dx covers a broad spectrum of medical knowledge through both multiple-choice and open-ended questions, ensuring that models are evaluated across varied response formats. The questions span 24 medical specialties, ranging from internal medicine and cardiology to pediatrics, surgery, and more. This diversity reflects the real-world breadth of medical practice and ensures that the benchmark evaluates reasoning across different sub-domains and problem types (diagnosis, treatment decisions, biomedical mechanism explanations, etc.).
- **Evaluation metrics for reasoning quality:** MedReason-Dx advances beyond simple accuracy measures by introducing multiple metrics to evaluate the fidelity and relevance of a model’s reasoning. To improve robustness across different LLM evaluators, we introduce a keypoint-based strategy: medical experts extract keypoints from ground-truth answers, and model outputs are evaluated by matching their extracted key points against these references, complementing step-wise matching. By aligning responses at the keypoint level rather than surface language, this approach reduces variability from linguistic complexity and mitigates instability across evaluators. Overall, by quantifying reasoning quality, MedReason-Dx encourages models that are not only correct, but correct for the right reasons.
- **Interpretability and robustness focus:** By requiring and evaluating intermediate reasoning, MedReason-Dx places interpretability at the core of model assessment. This is especially crucial for medical AI systems that clinicians need to trust. A model that can articulate a valid reasoning chain is inherently more transparent and easier to debug than one that only outputs an answer. Furthermore, focusing on reasoning helps reveal when a model’s knowledge is superficial. We anticipate that models performing well on MedReason-Dx will demonstrate greater robustness, as they must handle complex multi-step problems in a principled way rather than relying on shallow cues. Our benchmark thus serves as a stress test for genuine reasoning ability in medical contexts.
- **Revealing gaps in existing LLMs for medical reasoning:** MedReason-Dx exposes critical limitations of current LLMs, including domain-specific medical LLMs, in handling complex diagnostic reasoning and reliably recalling relevant medical knowledge. Our evaluations also investigate differences between general-purpose and medical LLMs, providing new insights into their respective strengths and weaknesses in clinical reasoning.

In summary, MedReason-Dx is the first benchmark to comprehensively evaluate reasoning paths in medical diagnosis from multiple perspectives. It provides the community with a testbed for developing

108 models that aspire to be not just answer engines, but reliable reasoning assistants in healthcare. By  
 109 emphasizing how answers are derived, our work addresses a critical gap for high-stakes AI: the need  
 110 for systems whose decisions can be inspected, trusted, and used responsibly in clinical practice.  
 111

## 112 2 RELATED WORKS

### 113 2.1 MEDICAL LLMS

116 Medical large language models (Med-LLMs) have  
 117 advanced rapidly through innovations in model ar-  
 118 chitectures, training paradigms, and domain-specific  
 119 adaptations, enabling applications in information ex-  
 120 traction, clinical decision support, dialogue systems,  
 121 and multimodal medical AI.

122 Early Med-LLMs such as BioBERT (Lee et al.,  
 123 2020) and PubMedBERT (Gu et al., 2021) lever-  
 124 aged large-scale biomedical corpora for pretrain-  
 125 ing, achieving strong performance in named entity  
 126 recognition and text classification. Subsequent mod-  
 127 els expanded capabilities to clinical text generation  
 128 and summarization, including ClinicalT5 (Lu et al.,  
 129 2022), GatorTron (Yang et al., 2022), and Codex-  
 130 Med (Liévin et al., 2024), while Galactica (Taylor  
 131 et al., 2022) targeted medical literature analysis.

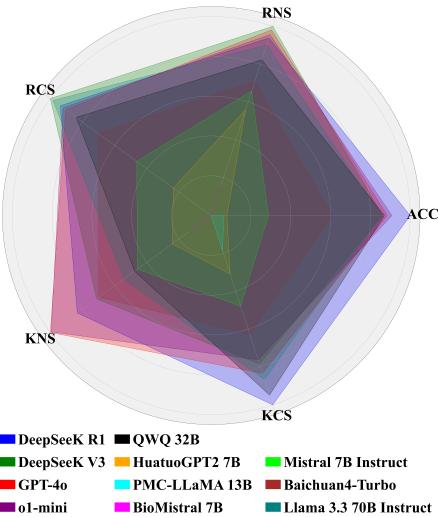
132 More recent systems adopt instruction fine-tuning  
 133 and reinforcement learning from human feedback  
 134 (RLHF) to improve reliability in clinical reason-  
 135 ing and QA. Med-PaLM (Singhal et al., 2023),  
 136 Med-PaLM 2 (Singhal et al., 2025), and Med-  
 137 Alpaca (Han et al., 2023) exemplify this trend, while  
 138 GatorTronGPT (Peng et al., 2023) adapts GatorTron  
 139 for precise report generation and ChatDoctor (Li  
 140 et al., 2023) focuses on virtual consultations. Further  
 141 domain specialization has produced mod-  
 142 els such as PMC-LLaMA (Wu et al., 2023a) for biomed-  
 143 ical literature, GPT-4-Med (Nori et al., 2023)  
 144 for advanced clinical tasks, and Taiyi-LLM (Luo et al., 2024) and Zhongjing (Yang et al., 2024b) for  
 145 Traditional Chinese Medicine. Multilingual and multimodal efforts—such as HuatuoGPT (Zhang  
 146 et al., 2023), Med-Flamingo (Moor et al., 2023), and Med-Gemini (Saab et al., 2024)—extend the  
 147 scope of Med-LLMs to global and cross-modal healthcare applications.

148 These developments underscore the rapid evolution of Med-LLMs, enhancing their ability to process  
 149 complex medical language, integrate multimodal data, and support diverse clinical tasks. As these  
 150 models become more powerful and widely deployed, it is increasingly important to evaluate not only  
 151 their final predictions but also the reasoning processes behind them. Without explicit reasoning-path  
 152 evaluation, it remains unclear whether models reach conclusions through sound clinical logic or  
 153 superficial correlations. Thus, rigorous assessment of reasoning fidelity is essential for ensuring  
 154 safety, trust, and real-world reliability in medical AI.

### 155 2.2 MEDICAL BENCHMARKS

156 The development of standardized datasets and robust evaluation platforms is crucial for advancing AI  
 157 in medicine. Existing efforts can be broadly grouped into two directions: (1) datasets tailored for  
 158 specific medical AI tasks, and (2) automated benchmarks for assessing the capabilities of LLMs.

159 The first category includes datasets supporting tasks such as information extraction, QA, and natural  
 160 language inference. Widely used corpora like CADEC (Karimi et al., 2015) and BC5CDR (Li et al.,  
 161 2016) target biomedical NER and relation extraction. QA datasets such as MedQA (Jin et al., 2021)  
 162 and PubMedQA (Jin et al., 2019) emphasize clinical knowledge retrieval and reasoning. Large-  
 163 scale resources like MIMIC-III (Johnson et al., 2016), HealthSearchQA (Singhal et al., 2023), and



154 Figure 1: Medical reasoning performance of advanced LLMs in MedReason-Dx. Our  
 155 benchmark evaluates LLMs’ medical reasoning-  
 156 capabilities across five dimensions: ACC,  
 157 RCS, RNS, KNS and KCS. Results indicate  
 158 that medical LLMs don’t outperform general  
 159 LLMs on complex medical reasoning tasks.

160

162 Table 1: Comparison with existing Medical QA benchmarks.  
163

Benchmark	CoT Evaluation	No. Domains	reasoning intensive	MCQ	OEQ	Expert Annotation
MMedBench	✗	21	✓	✓	✗	✓
MedQA	✗	-	✗	✓	✗	✗
MedMCQA	✗	21	✗	✓	✗	✓
MMLU	✗	6	✗	✓	✗	✗
Medbullets	✗	-	✓	✓	✗	✓
JAMA Challenge	✗	13	✓	✓	✗	✓
LiveQA	✗	-	✗	✓	✓	✓
ClinicBench	✗	-	✗	✓	✓	✓
<b>Ours</b>	✓	24	✓	✓	✓	✓

175 CORD-19 (Wang et al., 2020) support report summarization, while MedNLI (Romanov & Shivade,  
176 2018) focuses on natural language inference. Recently, MedReason (Wu et al., 2025a) addresses  
177 the lack of high-quality reasoning data by constructing 32,682 QA pairs with detailed, knowledge-  
178 graph-guided Chain-of-Thought explanations, further validated by expert review. Similar efforts  
179 such as MedCaseReasoning (Wu et al., 2025b) also highlights reasoning evaluation in clinical cases.

180 Table 2: Model Comparison.  
181

Models	Params
<i>General Large Language Models</i>	
GPT-4o (OpenAI, 2024a)	-
GPT-01 mini (OpenAI, 2024b)	-
DeepSeek R1 (Guo et al., 2025)	671B
DeepSeek V3 (Liu et al., 2024)	671B
Mistral (Jiang et al., 2023)	7B
LLAMA 3.3 (Dubey et al., 2024)	70B
QWQ (Yang et al., 2024a)	32B
<i>Medical Large Language Models</i>	
HuatuGPT2 (Chen et al., 2023)	7B
BioMistral (Labrak et al., 2024)	7B
PMC-LLaMA (Wu et al., 2024)	13B
Baichuan4-Turbo (Yang et al., 2023)	-

182 Table 3: Statistics of MedReason-Dx.  
183

Statistic	Number
Total questions number	1170
- Multiple-choice questions	592
- Open-ended questions	578
Maximum question length	578
Maximum answer length	1028
Average question length	166.9
Average answer length	320.0
Maximum number of steps	13
Maximum number of key points	118
Minimum number of steps	3
Minimum number of key points	3
Average number of steps	6.4
Average number of key points	27.1

197 The second category emphasizes automated evaluation frameworks that reduce reliance on costly  
198 expert review. MedBench (Cai et al., 2024) provides a broad platform with 40,041 questions spanning  
199 multiple fields. AutoEval (Liao et al., 2023) reformulates USMLE problems into multi-turn dialogues  
200 for assessing information coverage and task accuracy. LLM-Mini-CEX (Shi et al., 2023) employs  
201 patient simulators and ChatGPT to evaluate diagnostic dialogue quality, while MedGPTEval (Xu et al.,  
202 2023) integrates Chinese datasets and public benchmarks with 16 expert-curated indicators. LLM-  
203 Human Evaluation (Chiang & Lee, 2023) further explores the feasibility of automated assessment,  
204 demonstrating alignment with human experts on adversarial and open-ended tasks. Collectively, these  
205 resources provide systematic, low-cost ways to benchmark AI models and support optimization.

206  
207 

### 3 MEDICAL BENCHMARK WITH STEP-WISE EVALUATION

208  
209 

#### 3.1 DATA CURATION

210  
211 

##### 3.1.1 DATA CURATION FOR MULTIPLE-CHOICE QUESTION

212 The data collection for MedReason-Dx is designed to construct a challenging reasoning dataset that  
213 goes beyond typical knowledge-recall question answering. Our goal is to curate problems that require  
214 complex, multi-step reasoning, reflecting the intricate workflows of real-world clinical diagnostics.  
215 To this end, we employ a rigorous data selection strategy, filtering questions from well-established  
medical datasets that encompass authentic clinical cases across multiple disciplines.

216	<b>Question type:</b> multichoice question <b>Specialty:</b> <b>Nephrology</b>	<b>Question:</b> ... Serum creatinine is 2.0 mg/dL, and urinalysis shows 1+ proteinuria ... What is the most likely explanation for this patient's deterioration in renal function?		<b>Question type:</b> open-ended question <b>Specialty:</b> <b>Pediatrics</b>	<b>Question:</b> A newborn boy is evaluated in the hospital nursery 2 hours after birth ... the patient has a wide nasal bridge, down slanting palpebral fissures, and widely spaced eyes. He has good respiratory effort with a high-pitched cry. This patient is most likely to have experienced a deletion on which chromosomes?	
217	<b>Options:</b> [..., Toxic tubulointerstitial nephritis, ...]	<b>Answer steps:</b> [..., Analyze laboratory and biopsy findings: The patient has deteriorating renal function with a serum creatinine of 2.0 mg/dL, 1+ proteinuria. Consider potential causes of renal deterioration: The key findings to focus on are the eosinophilic infiltration and parenchymal inflammation, ...]		<b>Answer steps:</b> [..., Cri-du-chat syndrome is characterized by a distinctive high-pitched cry resembling a cat's mew, along with the other physical anomalies observed, ...]	<b>Key points:</b> [..., mew-like cry, microcephaly, ...]	
218	<b>Key points:</b> [..., elevated serum creatinine, proteinuria, ...]					
219						
220	<b>Question type:</b> multichoice question <b>Specialty:</b> <b>Hematology</b>	<b>Question:</b> ... studies show a prothrombin time of 12 seconds, an activated partial thromboplastin time of 35 seconds...		<b>Question type:</b> open-ended question <b>Specialty:</b> <b>Obstetrics and Gynecology</b>	<b>Question:</b> A 26-year-old woman presents to the emergency department with 1 day of vaginal bleeding and lower abdominal pain....Which is the most appropriate next step in management?	
221	<b>Options:</b> [..., Rough endoplasmic reticulum, ...]	<b>Answer steps:</b> [..., Note the laboratory findings: normal prothrombin time, normal activated partial thromboplastin time, prolonged bleeding time, and bloody synovial fluid from arthrocentesis, ...]		<b>Answer steps:</b> [..., In such cases, immediate surgical intervention is necessary to address potential intra-abdominal bleeding and prevent further deterioration, ...]	<b>Key points:</b> [..., ruptured ectopic pregnancy, anemia, ...]	
222	<b>Key points:</b> [..., bloody synovial fluid, deficiency in coagulation factors, proteinuria, ...]					
223						
224						
225						
226						
227						
228						
229						
230						
231						
232						

Figure 2: Dataset overview of MedReason-DX.

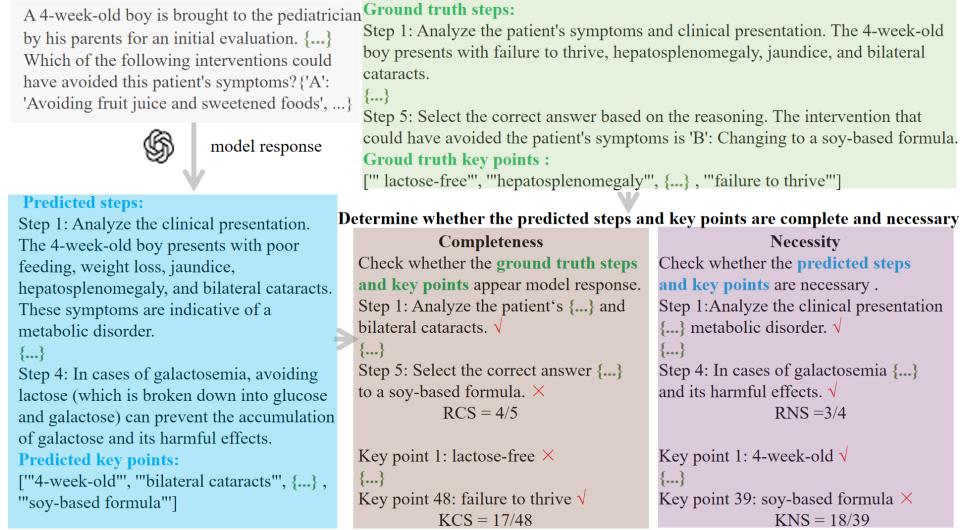
Table 4: (COT evaluation) Benchmarking the medical reasoning performance of existing models

	multiple-choice				open-ended			
	RNS	RCS	KNS	KCS	RNS	RCS	KNS	KCS
DeepSeek R1	86.11	73.46	47.62	33.90	93.04	60.97	40.68	31.20
DeepSeek V3	91.09	73.76	44.49	26.06	94.78	67.52	39.54	32.08
GPT-4o	89.92	69.76	50.72	29.78	93.88	60.97	43.53	29.93
o1-mini	89.94	69.15	51.11	29.46	91.22	63.30	42.97	27.95
Mistral 7B Instruct	72.37	38.26	40.21	24.65	76.85	43.48	34.78	23.60
Llama 3.3 70B Instruct	86.98	72.71	41.55	31.54	88.32	66.75	36.47	29.26
QWQ 32B	84.18	65.73	40.31	33.92	82.65	57.72	35.33	29.51
Baichuan4-Turbo	64.49	56.47	45.40	26.39	90.32	51.75	38.13	26.64
HuatuoGPT2 7B	62.59	24.13	35.59	20.57	76.14	32.36	31.74	22.21
PMC-LLaMA 13B	39.91	17.95	30.22	20.36	38.77	12.78	28.49	18.51
BioMistral 7B	44.72	13.97	31.58	15.60	50.83	16.43	31.96	17.23

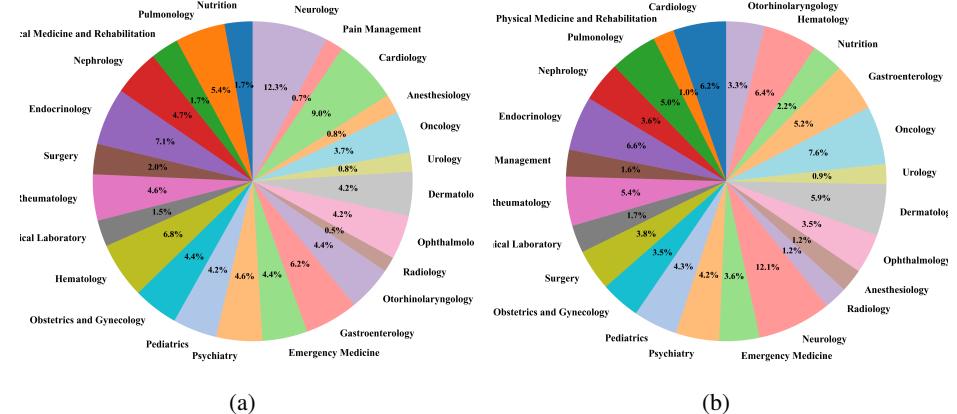
To capture the breadth of medicine, we define 24 domains aligned with common hospital departments, including "Cardiology", "Pulmonology", "Gastroenterology" and so on, as shown in Figure 4. The selection of questions prioritizes diversity in the types of clinical challenges and the reasoning methods required for problem-solving. This diversity encompasses a wide array of diagnostic tasks that span both common and rare clinical conditions. Questions are deliberately selected for their requirement of complex multi-step reasoning, including, but not limited to, physiological mechanism analysis, differential diagnosis, hypothesis testing, exclusionary reasoning, and the integration of cross-disciplinary knowledge, while simple factual recall items (e.g., "What is the normal body temperature?") are excluded. By focusing on the reasoning complexity and diversity, MedReason-Dx reflects the multifaceted nature of clinical decision-making and the diverse set of cognitive strategies employed by healthcare professionals in practice. The aim is to ensure that the dataset not only captures the breadth of medical knowledge but also challenges models to engage in higher-order reasoning reflective of real-world medical diagnostic scenarios

Following the selection of complex questions, human experts from diverse specialties annotate step-by-step solutions and key points, with LLMs providing supplementary support. This methodology enables a thorough assessment of the model's reasoning capabilities, evaluating not only the accuracy of the final answer but also the clarity and logical coherence of the reasoning process. The step-by-step solutions deconstruct the reasoning into concise, logical steps, each reflecting a critical component of the decision-making process. Key points emphasize essential information, such as clinical findings or diagnostic considerations, required for accurate diagnosis. These annotations ensure that the model's

270 response addresses all relevant aspects of medical decision-making. The objective is to rigorously  
 271 evaluate the reasoning process, confirming its completeness and logical integrity.  
 272



298 Figure 3: Illustration of the proposed evaluation metrics. For LLM responses, we match them with  
 299 annotations from human experts to evaluate the necessity and sufficiency of the reasoning process.  
 300



308 Figure 4: Medical Specialty Distribution of MedReason-Dx, including (a): multiple-choice question  
 309 and (b) open-ended question, both covering 24 departments.  
 310

### 311 3.1.2 DATA CURATION FOR OPEN-ENDED QUESTION

312 Multiple-choice questions often fail to reflect real-world clinical scenarios, where physicians make  
 313 decisions without predefined options. To address this limitation, we develop additional open-ended  
 314 reasoning questions. To ensure diversity, the dataset incorporates both multiple-choice and open-  
 315 ended sources. Multiple-choice questions are adapted into open-ended formats by using LLMs to  
 316 revise only the final question sentence, preserving the clinical context while minimizing alterations.  
 317 All modifications are reviewed by human experts to ensure accuracy. Once open-ended questions are  
 318 established, domain specialists generate step-by-step solutions using the same methodology as for  
 319 multiple-choice items. And key points are then extracted to enable systematic evaluation.  
 320

321 In summary, MedReason-Dx provides a comprehensive benchmark for evaluating LLMs on complex  
 322 medical reasoning tasks through carefully selected questions and detailed expert annotations. Detailed  
 323 statistics for MedReason-Dx are presented in Table 3 and comparisons with other relevant medical  
 benchmarks are provided in Table 1. More details are provided in Appendix A.4.

324  
325

## 3.2 EVALUATION

326  
327  
328  
329  
330  
331  
332

To rigorously evaluate the reasoning ability of LLMs, we propose a comprehensive framework that assesses three critical dimensions: the accuracy of the final answer, the quality of the reasoning process, and the incorporation of relevant key points. These dimensions are captured through five metrics: *Accuracy*, *Reasoning Completeness Score (RCS)*, *Reasoning Necessity Score (RNS)*, *Key point Completeness Score (KCS)*, and *Key point Necessity Score (KNS)*. Below, we detail each metric, its motivation, and the computational methodology, with a particular emphasis on the thoroughness and necessity of reasoning steps and key points usage.

333  
334

## 3.2.1 ACCURACY

335  
336  
337  
338  
339  
340  
341  
342  
343  
344  
345  
346

The *Accuracy* metric evaluates the correctness of the model’s final answer, a standard measure of performance in question-answering tasks. For multiple-choice questions, we compare the model’s selected option with the ground-truth answer. For open-ended questions, the model is instructed to provide answers in a standardized format: *Therefore, the answer is {your answer}*. And a LLM is employed to assess semantic equivalence between the model’s answer and the ground-truth answer, accommodating variations in expression.

347

The Accuracy is computed as:

348  
349  
350

$$\text{Accuracy} = \frac{\sum_{i=1}^N \mathbb{I}(\hat{y}_i = y_i)}{N}, \quad (1)$$

351  
352  
353  
354  
355

where  $N$  is the total number of questions,  $\hat{y}_i$  is the model’s predicted answer for the  $i$ -th question,  $y_i$  is the ground-truth answer, and  $\mathbb{I}(\cdot)$  is the indicator function, returning 1 if the predicted answer matches the ground truth and 0 otherwise. For open-ended questions, equivalence is determined by the LLM, ensuring robust evaluation across diverse answer formats.

356  
357

## 3.2.2 REASONING COMPLETENESS SCORE (RCS)

358  
359  
360

The *Reasoning Completeness Score (RCS)* quantifies the extent to which the model’s reasoning includes all critical steps required to derive the answer. This is crucial in high-stakes domains, where comprehensive reasoning ensures all relevant concepts and logical deductions are addressed.

361  
362  
363

For each question, a reference set of essential reasoning steps  $R_i = \{r_1, r_2, \dots, r_s\}$  is predefined based on expert annotations. The model’s reasoning text is evaluated to identify the presence of these steps or their semantic equivalents. The RCS for a single question is:

364  
365  
366

$$\text{RCS}_i = \frac{|\hat{R}_i \cap R_i|}{|R_i|}, \quad (2)$$

367  
368  
369

where  $\hat{R}_i$  is the set of reasoning steps in the model’s response for the  $i$ -th question, and  $|\cdot|$  denotes set cardinality. The overall RCS is:

370  
371  
372  
373

$$\text{RCS} = \frac{1}{N} \sum_{i=1}^N \text{RCS}_i. \quad (3)$$

374  
375

## 3.2.3 REASONING NECESSITY SCORE (RNS)

376  
377

The *Reasoning Necessity Score (RNS)* evaluates the conciseness of the reasoning process, penalizing extraneous or redundant steps. This ensures that the model’s reasoning is focused and interpretable, a critical requirement for practical utility in medical contexts.

Table 5: Benchmarking the accuracy (%) performance of existing models.

	Multiple-choice		Open-ended	
	CoT	Direct	CoT	Direct
DeepSeek R1	65.03	64.36	40.14	42.39
DeepSeek V3	60.47	59.79	33.56	37.02
GPT-4o	58.28	59.12	37.72	47.70
o1-mini	60.47	62.67	37.37	41.18
Mistral 7B Instruct	30.74	18.58	17.47	15.57
Llama 3.3 70B Instruct	56.42	57.60	37.02	37.89
QWQ 32B	56.08	55.91	38.75	40.83
Baichuan4-Turbo	46.62	42.74	27.16	28.37
HuatuoGPT2 7B	19.26	25.00	12.28	12.81
PMC-LLaMA 13B	20.95	23.14	9.17	6.23
BioMistral 7B	15.71	19.76	9.52	12.63

378 RNS is computed by comparing the model’s responses  $\hat{R}_i$  to the reference set  $R_i$ . Steps in  $\hat{R}_i$  that  
 379 are not in  $R_i$  or its semantic equivalents are considered unnecessary. RNS is calculated as follows:  
 380

$$381 \quad \text{RNS} = \frac{1}{N} \sum_{i=1}^N 1 - \frac{|\hat{R}_i \setminus R_i|}{|\hat{R}_i| + \epsilon}, \quad (4)$$

$$382$$

$$383$$

384 where  $\epsilon > 0$  (e.g.,  $\epsilon = 0.01$ ) prevents division by zero.  
 385

### 386 3.2.4 KEY POINT COMPLETENESS SCORE (KCS)

$$387$$

388 The *Key point Completeness Score (KCS)* assesses the model’s ability to incorporate all relevant  
 389 domain-specific key points in its reasoning, such as diagnoses, symptoms, or treatments in medical  
 390 contexts. Completeness is critical to ensure that the model captures essential domain knowledge.

391 For each question (i), a reference set of key points  $K_i = \{k_1, k_2, \dots, k_w\}$  is annotated by human  
 392 experts, representing the essential terms relevant to the correct response. The model’s response is  
 393 analyzed to extract a set of key points  $\hat{K}_i = \{\hat{k}_1, \hat{k}_2, \dots, \hat{k}_{\hat{w}}\}$ , which denotes the terms provided by  
 394 AI model. KCS is calculated as follows:

$$395 \quad \text{KCS} = \frac{1}{N} \sum_{i=1}^N \frac{|\hat{K}_i \cap K_i|}{|K_i|}. \quad (5)$$

$$396$$

$$397$$

$$398$$

### 399 3.2.5 KEY POINTS NECESSITY SCORE (KNS)

$$400$$

401 The *Key points Necessity Score (KNS)* evaluates the model’s ability to avoid irrelevant or extraneous  
 402 key points, ensuring that the reasoning remains focused and relevant.

403 The KNS is computed by identifying key points in  $\hat{K}_i$  that are not in the reference set  $K_i$ :

$$404$$

$$405 \quad \text{KNS} = \frac{1}{N} \sum_{i=1}^N 1 - \frac{|\hat{K}_i \setminus K_i|}{|\hat{K}_i| + \epsilon}. \quad (6)$$

$$406$$

$$407$$

408 The proposed evaluation framework, comprising Accuracy, RCS, RNS, KCS, and KNS, extends model  
 409 assessment beyond final answer correctness to capture reasoning quality and knowledge relevance.  
 410 Specifically, RCS measures the completeness of reasoning steps, RNS evaluates conciseness, KCS  
 411 verifies coverage of critical medical terms, and KNS assesses the relevance of recalled knowledge  
 412 within the reasoning process. Together, these metrics provide a comprehensive evaluation of both  
 413 correctness and interpretability, which is essential for trustworthy deployment in high-stakes medical  
 414 applications. An overview of the metric calculations is shown in Figure 3.

$$415$$

## 416 4 EXPERIMENTS

$$417$$

### 418 4.1 EXPERIMENTAL SETUP

$$419$$

420 **Evaluation Models.** To provide a comprehensive benchmark, we conduct evaluations on 11 advanced  
 421 LLMs, comprising 7 general LLMs and 4 medical LLMs. For general LLMs, we include DeepSeek  
 422 R1 (Guo et al., 2025), DeepSeek V3 (Liu et al., 2024), GPT-4o (OpenAI, 2024a), o1-mini (OpenAI,  
 423 2024b), Mistral 7B Instruct (Jiang et al., 2023), LlaMA3.3 70B Instruct (Dubey et al., 2024), and  
 424 QWQ 32B (Yang et al., 2024a). For medical LLMs, we include Baichuan4-Turbo (Yang et al., 2023),  
 425 HuatuoGPT2-7B (Chen et al., 2023), PMC-LlaMA 13B (Wu et al., 2024) and BioMistral 7B (Labrak  
 426 et al., 2024). A comparison of these models is provided in Table 2.

427 **Implementation Details.** Following previous work (Jiang et al., 2025), we leverage two types of  
 428 prompts to guide the model to give answers: chain-of-thought (CoT) prompts (Wei et al., 2022) and  
 429 direct prompts. When calculating RNS, RCS, KNS, and KCS scores, we utilize the model’s CoT  
 430 response to explicitly require the model to provide step-by-step answers. When calling LLMs, the  
 431 temperature is set to 0.7, Top-P is set to 0.9, and max tokens is set to 1024. In the experiment, we  
 employ gpt4o-mini (Achiam et al., 2023) as the judge. We provide detailed prompts in Appendix A.6.

432 4.2 BENCHMARKING MEDICAL LLMs  
433

434 **LLMs struggle with complex medical reasoning tasks.** As shown in Table 5, we benchmark  
435 several advanced language models on the MedReason-Dx dataset under two prompting settings:  
436 Chain-of-Thought (CoT) and Direct Answering, across both multiple-choice and open-ended formats.  
437 Our findings reveal that state-of-the-art large language models, including GPT-4o and DeepSeek-  
438 R1, continue to encounter substantial difficulties in addressing complex medical reasoning tasks.  
439 Specifically, in multiple-choice question assessments, the highest-performing model, DeepSeek-R1,  
440 achieved an accuracy of 65.03%. In contrast, performance on open-ended questions was notably  
441 lower, with the leading model attaining only 47.70% accuracy (GPT-4o). These suboptimal results  
442 underscore the persistent limitations of current large language models in navigating the intricacies of  
443 real-world medical reasoning scenarios, highlighting the critical need for further advancements and  
444 optimization in model development.

445 **LLMs face difficulties in comprehensively recalling  
446 medical knowledge.** As shown in Table 4, to thor-  
447oughly evaluate the reasoning capabilities of LLMs  
448 beyond the correctness of their final answers, we pro-  
449 pose four novel metrics: RNS, RCS, KNS and KCS.  
450 These metrics enable a comprehensive assessment of  
451 the model’s step-by-step reasoning process and the  
452 relevance of key points used during reasoning. Our results indicate that each reasoning step, along  
453 with the key points mentioned in the reasoning process, is typically critical for accurate problem-  
454 solving. However, LLMs often struggle to include all the necessary information annotated by human  
455 experts. For instance, GPT-4o achieved an RNS of 89.92% on multiple-choice questions, reflecting  
456 high precision in its reasoning process with minimal inclusion of irrelevant details. In contrast, its  
457 RCS of 69.76% highlights a significant challenge in recalling all the essential information required to  
458 solve problems as effectively as human experts. This gap is likely attributable to the complexity and  
459 irregularity of medical knowledge, which lacks the structured theorems and corollaries characteristic  
460 of fields like mathematics and physics, posing unique challenges for solving medical problems.

461 **Medical LLMs don’t outperform general LLMs in complex medical reasoning tasks.** As  
462 illustrated in Table 4 and 5, medical LLMs don’t surpass general LLMs. This observed disparity may  
463 stem from prevailing training paradigms for medical LLMs, which primarily entail the assimilation of  
464 foundational medical corpora while largely omitting the integration of sophisticated datasets designed  
465 to cultivate advanced medical reasoning competencies. Consequently, although medical LLMs may  
466 demonstrate marginal advantages over general LLMs in rudimentary medical knowledge retrieval and  
467 question-answering tasks, their proficiency in multifaceted, real-world clinical contexts frequently  
468 proves inadequate relative to anticipated benchmarks. This observation highlights the critical need  
469 for developing comprehensive training datasets specifically tailored to large-scale, complex medical  
470 reasoning tasks to comprehensively enhance the capabilities of medical LLMs.

## 471 5 EVALUATION CONSISTENCY ACROSS DIFFERENT LLM JUDGES

472 In our study, we employ GPT-4o-mini as the judge to assess the similarity between model-generated  
473 responses and the ground truth. To validate the robustness of our results, we conduct additional  
474 experiments using GPT-4.1-mini as the judge and compare the consistency across both evaluation sets.  
475 The experiments are performed on responses generated by GPT-4o and DeepSeek-R1. As reported  
476 in Table 6, the scores derived from the two independent judges exhibit a high degree of agreement,  
477 substantiating the reliability and validity of our evaluation framework.

478 6 CONCLUSION  
479

480 In this paper, we introduce MedReason-Dx, a novel benchmark designed to evaluate not only the  
481 accuracy of medical question answering systems but also the quality of their reasoning processes.  
482 By providing expert-annotated, step-by-step reasoning chains and evaluating models across multiple  
483 medical specialties, MedReason-Dx promotes the development of AI systems that can articulate  
484 their reasoning, ensuring both interpretability and robustness. Our approach highlights the critical  
485 importance of evaluating the reasoning process behind AI-generated answers, fostering the creation  
486 of transparent, reliable, and clinically valid decision support tools in healthcare.

Table 6: Evaluation using different judges.

Judge	GPT-4o		DeepSeeK R1	
	RNS	RCS	RNS	RCS
GPT-4o-mini	89.92	69.76	86.11	73.46
GPT-4.1-mini	88.47	72.63	88.08	73.83

486 REFERENCES  
487

488 Asma Ben Abacha, Eugene Agichtein, Yuval Pinter, and Dina Demner-Fushman. Overview of the  
489 medical question answering task at TREC 2017 liveqa. In Proceedings of The Twenty-Sixth Text  
490 REtrieval Conference, TREC, 2017.

491 Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman,  
492 Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report.  
493 arXiv preprint arXiv:2303.08774, 2023.

494 Yan Cai, Linlin Wang, Ye Wang, Gerard de Melo, Ya Zhang, Yanfeng Wang, and Liang He. Medbench:  
495 A large-scale chinese benchmark for evaluating medical large language models. In Proceedings of  
496 the AAAI Conference on Artificial Intelligence, volume 38, pp. 17709–17717, 2024.

497 Hanjie Chen, Zhouxiang Fang, Yash Singla, and Mark Dredze. Benchmarking large language models  
498 on answering and explaining challenging medical questions. arXiv preprint arXiv:2402.18060,  
500 2024.

501 Junying Chen, Xidong Wang, Ke Ji, Anningzhe Gao, Feng Jiang, Shunian Chen, Hongbo Zhang,  
502 Dingjie Song, Wenya Xie, Chuyi Kong, et al. Huatuogpt-ii, one-stage training for medical adaption  
503 of llms. arXiv preprint arXiv:2311.09774, 2023.

504 Michael K Chen, Xikun Zhang, and Dacheng Tao. Justlogic: A comprehensive benchmark for  
505 evaluating deductive reasoning in large language models. arXiv preprint arXiv:2501.14851, 2025.

506 Cheng-Han Chiang and Hung-yi Lee. Can large language models be an alternative to human  
507 evaluations? arXiv preprint arXiv:2305.01937, 2023.

508 Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha  
509 Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models.  
510 arXiv e-prints, pp. arXiv–2407, 2024.

511 Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann,  
512 Jianfeng Gao, and Hoifung Poon. Domain-specific language model pretraining for biomedical  
513 natural language processing. ACM Transactions on Computing for Healthcare (HEALTH), 3(1):  
514 1–23, 2021.

515 Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu,  
516 Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms  
517 via reinforcement learning. arXiv preprint arXiv:2501.12948, 2025.

518 Tianyu Han, Lisa C Adams, Jens-Michalis Papaioannou, Paul Grundmann, Tom Oberhauser, Alexander  
519 Löser, Daniel Truhn, and Keno K Bressem. Medalpaca—an open-source collection of medical  
520 conversational ai models and training data. arXiv preprint arXiv:2304.08247, 2023.

521 Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot,  
522 Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier,  
523 Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas  
524 Wang, Timothée Lacroix, and William El Sayed. Mistral 7b. 2023.

525 Dongzhi Jiang, Renrui Zhang, Ziyu Guo, Yanwei Li, Yu Qi, Xinyan Chen, Liupei Wang, Jianhan  
526 Jin, Claire Guo, Shen Yan, et al. Mme-cot: Benchmarking chain-of-thought in large multimodal  
527 models for reasoning quality, robustness, and efficiency. arXiv preprint arXiv:2502.09621, 2025.

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

531 Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W Cohen, and Xinghua Lu. Pubmedqa: A  
532 dataset for biomedical research question answering. arXiv preprint arXiv:1909.06146, 2019.

533 Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad  
534 Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a  
535 freely accessible critical care database. Scientific data, 3(1):1–9, 2016.

540 Sarvnaz Karimi, Alejandro Metke-Jimenez, Madonna Kemp, and Chen Wang. Cadec: A corpus of  
 541 adverse drug event annotations. *Journal of biomedical informatics*, 55:73–81, 2015.  
 542

543 Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and  
 544 Richard Dufour. Biomistral: A collection of open-source pretrained large language models  
 545 for medical domains. In *Findings of the Association for Computational Linguistics, ACL*, pp.  
 546 5848–5864, 2024.

547 Jinyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo  
 548 Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining.  
 549 *Bioinformatics*, 36(4):1234–1240, 2020.

550 Jiao Li, Yueping Sun, Robin J Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter  
 551 Davis, Carolyn J Mattingly, Thomas C Wiegers, and Zhiyong Lu. Biocreative v cdr task corpus: a  
 552 resource for chemical disease relation extraction. *Database*, 2016, 2016.

553 Yunxiang Li, Zihan Li, Kai Zhang, Ruilong Dan, Steve Jiang, and You Zhang. Chatdoctor: A medical  
 554 chat model fine-tuned on a large language model meta-ai (llama) using medical domain knowledge.  
 555 *Cureus*, 15(6), 2023.

556 Yusheng Liao, Yutong Meng, Hongcheng Liu, Yanfeng Wang, and Yu Wang. An automatic evaluation  
 557 framework for multi-turn medical consultations capabilities of large language models. *arXiv*  
 558 preprint [arXiv:2309.02077](https://arxiv.org/abs/2309.02077), 2023.

559 Valentin Liévin, Christoffer Egeberg Hother, Andreas Geert Motzfeldt, and Ole Winther. Can large  
 560 language models reason about medical questions? *Patterns*, 5(3), 2024.

561 Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao,  
 562 Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv* preprint  
 563 [arXiv:2412.19437](https://arxiv.org/abs/2412.19437), 2024.

564 Qiuhan Lu, Dejing Dou, and Thien Nguyen. Clinicalt5: A generative language model for clinical  
 565 text. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 5436–5443,  
 566 2022.

567 Ling Luo, Jinzhong Ning, Yingwen Zhao, Zhijun Wang, Zeyuan Ding, Peng Chen, Weiru Fu, Qinyu  
 568 Han, Guangtao Xu, Yunzhi Qiu, et al. Taiyi: a bilingual fine-tuned large language model for diverse  
 569 biomedical tasks. *Journal of the American Medical Informatics Association*, 31(9):1865–1874,  
 570 2024.

571 Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Yash Dalmia, Jure Leskovec, Cyril  
 572 Zakka, Eduardo Pontes Reis, and Pranav Rajpurkar. Med-flamingo: a multimodal medical few-shot  
 573 learner. In *Machine Learning for Health (ML4H)*, pp. 353–367. PMLR, 2023.

574 Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. Capabilities  
 575 of gpt-4 on medical challenge problems. *arXiv* preprint [arXiv:2303.13375](https://arxiv.org/abs/2303.13375), 2023.

576 OpenAI. Hello gpt-4o. <https://openai.com/index/hello-gpt-4o/>, 2024a.

577 OpenAI. Learning to reason with llms. *OpenAI Blog*, 2024b. URL <https://openai.com/index/learning-to-reason-with-llms/>.

578 Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. Medmcqa: A large-scale  
 579 multi-subject multi-choice dataset for medical domain question answering. In *Conference on*  
 580 *Health, Inference, and Learning, CHIL*, 2022.

581 Cheng Peng, Xi Yang, Aokun Chen, Kaleb E Smith, Nima PourNejatian, Anthony B Costa, Cheryl  
 582 Martin, Mona G Flores, Ying Zhang, Tanja Magoc, et al. A study of generative large language  
 583 model for medical research and healthcare. *NPJ digital medicine*, 6(1):210, 2023.

584 Pengcheng Qiu, Chaoyi Wu, Xiaoman Zhang, Weixiong Lin, Haicheng Wang, Ya Zhang, Yanfeng  
 585 Wang, and Weidi Xie. Towards building multilingual language model for medicine. *Nature*  
 586 *Communications*, 2024.

594 Alexey Romanov and Chaitanya Shivade. Lessons from natural language inference in the clinical  
 595 domain. [arXiv preprint arXiv:1808.06752](https://arxiv.org/abs/1808.06752), 2018.

596

597 Khaled Saab, Tao Tu, Wei-Hung Weng, Ryutaro Tanno, David Stutz, Ellery Wulczyn, Fan Zhang,  
 598 Tim Strother, Chunjong Park, Elahe Vedadi, et al. Capabilities of gemini models in medicine.  
 599 [arXiv preprint arXiv:2404.18416](https://arxiv.org/abs/2404.18416), 2024.

600 Xiaoming Shi, Jie Xu, Jinru Ding, Jiali Pang, Sichen Liu, Shuqing Luo, Xingwei Peng, Lu Lu,  
 601 Haihong Yang, Mingtao Hu, et al. Llm-mini-cex: Automatic evaluation of large language model  
 602 for diagnostic conversation. [arXiv preprint arXiv:2308.07635](https://arxiv.org/abs/2308.07635), 2023.

603 Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan  
 604 Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfahl, et al. Large language models encode  
 605 clinical knowledge. *Nature*, 620(7972):172–180, 2023.

606 Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou,  
 607 Kevin Clark, Stephen R Pfahl, Heather Cole-Lewis, et al. Toward expert-level medical question  
 608 answering with large language models. *Nature Medicine*, pp. 1–8, 2025.

609 Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia,  
 610 Andrew Poulton, Viktor Kerkez, and Robert Stojnic. Galactica: A large language model for science.  
 611 [arXiv preprint arXiv:2211.09085](https://arxiv.org/abs/2211.09085), 2022.

612 Lucy Lu Wang, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Douglas Burdick,  
 613 Darrin Eide, Kathryn Funk, Yannis Katsis, Rodney Kinney, et al. Cord-19: The covid-19 open  
 614 research dataset. *ArXiv*, pp. arXiv–2004, 2020.

615 Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi,  
 616 Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language  
 617 models. In *Advances in Neural Information Processing Systems 35*, 2022.

618 Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Pmc-llama: Further  
 619 finetuning llama on medical papers. [arXiv preprint arXiv:2304.14454](https://arxiv.org/abs/2304.14454), 2(5):6, 2023a.

620 Chaoyi Wu, Weixiong Lin, Xiaoman Zhang, Ya Zhang, Weidi Xie, and Yanfeng Wang. Pmc-llama:  
 621 toward building open-source language models for medicine. *J. Am. Medical Informatics Assoc.*,  
 622 2024.

623 Juncheng Wu, Wenlong Deng, Xingxuan Li, Sheng Liu, Taomian Mi, Yifan Peng, Ziyang Xu, Yi Liu,  
 624 Hyunjin Cho, Chang-In Choi, et al. Medreason: Eliciting factual medical reasoning steps in llms  
 625 via knowledge graphs. [arXiv preprint arXiv:2504.00993](https://arxiv.org/abs/2504.00993), 2025a.

626 Kevin Wu, Eric Wu, Rahul Thapa, Kevin Wei, Angela Zhang, Arvind Suresh, Jacqueline J Tao,  
 627 Min Woo Sun, Alejandro Lozano, and James Zou. Medcasereasoning: Evaluating and learning  
 628 diagnostic reasoning from clinical case reports. [arXiv preprint arXiv:2505.11733](https://arxiv.org/abs/2505.11733), 2025b.

629 Sean Wu, Michael Koo, Lesley Blum, Andy Black, Liyo Kao, Fabien Scalzo, and Ira Kurtz. A  
 630 comparative study of open-source large language models, gpt-4 and claude 2: Multiple-choice test  
 631 taking in nephrology. [arXiv preprint arXiv:2308.04709](https://arxiv.org/abs/2308.04709), 2023b.

632 Jie Xu, Lu Lu, Sen Yang, Bilin Liang, Xinwei Peng, Jiali Pang, Jinru Ding, Xiaoming Shi, Lingrui  
 633 Yang, Huan Song, et al. Medgpteval: A dataset and benchmark to evaluate responses of large  
 634 language models in medicine. [arXiv preprint arXiv:2305.07340](https://arxiv.org/abs/2305.07340), 2023.

635 Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Ce Bian, Chao Yin, Chenxu Lv, Da Pan,  
 636 Dian Wang, Dong Yan, et al. Baichuan 2: Open large-scale language models. [arXiv preprint  
 637 arXiv:2309.10305](https://arxiv.org/abs/2309.10305), 2023.

638 An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li,  
 639 Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin  
 640 Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang,  
 641 Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi  
 642 Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan,  
 643 Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. [arXiv preprint  
 644 arXiv:2412.15115](https://arxiv.org/abs/2412.15115), 2024a.

648 Songhua Yang, Hanjie Zhao, Senbin Zhu, Guangyu Zhou, Hongfei Xu, Yuxiang Jia, and Hongying  
 649 Zan. Zhongjing: Enhancing the chinese medical capabilities of large language model through  
 650 expert feedback and real-world multi-turn dialogue. In *Proceedings of the AAAI conference on*  
 651 *artificial intelligence*, volume 38, pp. 19368–19376, 2024b.

652 Xi Yang, Aokun Chen, Nima PourNejatian, Hoo Chang Shin, Kaleb E Smith, Christopher Parisien,  
 653 Colin Compas, Cheryl Martin, Mona G Flores, Ying Zhang, et al. Gatortron: A large clinical  
 654 language model to unlock patient information from unstructured electronic health records. *arXiv*  
 655 preprint [arXiv:2203.03540](https://arxiv.org/abs/2203.03540), 2022.

656 Hongbo Zhang, Junying Chen, Feng Jiang, Fei Yu, Zhihong Chen, Jianquan Li, Guiming Chen,  
 657 Xiangbo Wu, Zhiyi Zhang, Qingying Xiao, et al. Huatuogpt, towards taming language model to be  
 658 a doctor. *arXiv* preprint [arXiv:2305.15075](https://arxiv.org/abs/2305.15075), 2023.

## 661 A APPENDIX

### 662 A.1 LLM USAGE STATEMENT

663 In the preparation of this manuscript, we utilized Large Language Models (LLMs) to assist with  
 664 language polishing and refinement of the text. Specifically, the LLM was employed to enhance the  
 665 clarity, coherence, and grammatical accuracy of the writing, ensuring that the manuscript adheres to  
 666 high standards of academic communication. The LLM did not contribute to the research ideation,  
 667 methodology, data analysis, or core content development, which were entirely conducted by the  
 668 authors. All outputs generated by the LLM were carefully reviewed and edited by the authors to  
 669 ensure alignment with the intended scientific contributions and to maintain the integrity of the work.

### 670 A.2 OPEN-ENDED QUESTIONS POSE A SIGNIFICANTLY GREATER CHALLENGE TO LLMs THAN 671 MULTIPLE-CHOICE QUESTIONS.

672 In clinical practice, physicians frequently address open-ended questions reflective of real-world  
 673 medical scenarios. However, many existing medical evaluation benchmarks predominantly utilize  
 674 multiple-choice formats, which substantially lower the complexity of questions and deviate from  
 675 authentic clinical contexts. This discrepancy may result in overly optimistic evaluations of LLMs  
 676 capabilities. As evidenced in Tables 5 and 4, LLMs exhibit significantly reduced performance on  
 677 open-ended questions compared to multiple-choice questions. To elucidate the differences in model  
 678 responses across these question types, we administer identical questions in both formats to GPT-4o  
 679 and analyzed the variations in their responses. As shown in Table 7, alterations solely in the question’s  
 680 format lead to substantial variations in the model’s response.

### 681 A.3 COMPARISON OF MODEL PERFORMANCE UNDER DIFFERENT PROMPTS

682 As show in Table 5, we benchmarked several advanced language models on the MedReason-Dx  
 683 dataset under two prompting settings: Chain-of-Thought (CoT) and Direct Answering, across both  
 684 multiple-choice and open-ended formats. Overall, DeepSeek-R1 achieved the highest performance in  
 685 the multiple-choice setting, with CoT prompting slightly outperforming direct answering (65.03% vs.  
 686 64.36%). However, in the open-ended setting, its performance reversed, with direct prompting yielding  
 687 higher accuracy (42.39%) than CoT (40.14%). DeepSeek-V3 showed a similar trend with modest  
 688 gains from direct answering in open-ended questions (37.02% vs. 33.56%). Interestingly, GPT-4o  
 689 exhibited the largest gap in favor of direct prompting for open-ended questions (47.70% vs. 37.72%),  
 690 while maintaining comparable results in multiple-choice settings. o1-mini demonstrated relatively  
 691 balanced performance across settings, with a slight edge for direct prompting in both question types.  
 692 In contrast, Baichuan4-Turbo underperformed across all configurations, with particularly low scores  
 693 on open-ended questions, indicating a significant gap in step-by-step reasoning capabilities compared  
 694 to stronger models. These results suggest that while CoT prompting can provide marginal gains in  
 695 structured formats, direct answering may be more effective in complex open-ended clinical scenarios,  
 696 particularly for stronger LLMs.

702 A.4 DATA COLLECTION  
703704 To ensure data diversity, MedReason-Dx compiles data from a wide range of sources, including  
705 JAMA Clinical Challenge (Chen et al., 2024), Medbullets (Chen et al., 2024), MMedBench (Qiu  
706 et al., 2024), nephSAP (Wu et al., 2023b), LiveQA (Abacha et al., 2017) and PubMedQA (Jin et al.,  
707 2019). To confirm that the selected questions demand rigorous clinical reasoning, we employed two  
708 stringent criteria for data curation: (1) questions that advanced large language models (e.g., GPT-4o,  
709 DeepSeek R1) failed to answer correctly were designated as reasoning-intensive and retained; (2)  
710 questions necessitating more extensive and detailed responses were similarly classified as reasoning-  
711 intensive and incorporated. These criteria effectively excluded questions predicated primarily on rote  
712 recall of medical knowledge. The preliminarily screened questions subsequently underwent rigorous  
713 secondary review and expert annotation by domain specialists, culminating in the development of  
714 MedReason-Dx—a robust benchmark designed to evaluate complex medical reasoning capabilities.  
715 The distribution of questions across the various sources is detailed in Table 8.  
716717 A.5 COMPARISON OF MODEL RESPONSES  
718719 In Table 9, we present a complete data from our benchmark. In Table 10 11 12 13 we compare the  
720 differences between responses from several models. Overall, the larger the number of parameters in a  
721 model, the better its performance in generating answers.  
722  
723  
724  
725  
726  
727  
728  
729  
730  
731  
732  
733  
734  
735  
736  
737  
738  
739  
740  
741  
742  
743  
744  
745  
746  
747  
748  
749  
750  
751  
752  
753  
754  
755

756 A.6 DETAILS OF PROMPTS  
757

758

759

760 The prompt for calculating RNS is as follows:  
761

762

763

764

765

766

767

Your task is to evaluate the correctness of each step of a model's solution.

## Input Format

1. Problem: The original question/task. Answer options are provided if available.
2. Answer: The ground truth final answer
3. Ground Truth Steps: Essential steps required for deriving the correct answer
4. A solution of a model (split into predicted steps)

## Task

For each predicted step in the model solution,

- it must exactly match or be directly entailed by the ground truth.
- if there is no direct match, the step must not contradict the ground truth and must have valid logical reasoning.

### Judgement Categories

- Match: Aligns with ground truth
- Reasonable: Valid but not in ground truth
- Wrong: Invalid or contradictory
- N/A: For background information steps

### Output Requirements

1. The output must be JSON only without any other content or formatting. Do not add ““json, etc.
2. The length of the JSON list must be the same as the number of steps in model solution.

## Output Format

[

{

```
"step_{index}": <integer>, "premises": <evidence (if available)>,
"conclusion": <step conclusion>, "judgment": "Match" | "Reasonable" | "Wrong" | "N/A"
},
```

...

]

## Inputs

Here are the problem, answer, model solution (predicted steps), and ground truth steps:

[Question]

{question}

{answer\_options}

[Ground Truth Answer]

{answer}

[Ground Truth Steps]

{answer\_steps}

[Model Solution (Predicted Steps)]

{predicted\_steps}

804

805

806

807

808

809

The prompt for calculating RCS is as follows:

```

810
811 Your task is to match the ground truth steps with the provided model solution.
812 ## Input Format
813 1. Problem: The original question/task. Answer options are provided if available.
814 2. Answer: The ground truth final answer
815 3. Ground Truth Steps: Essential steps required for deriving the correct answer
816 4. A solution of a model
817 ## Matching Process
818 - You need to match each ground truth step with the model solution provided.
819 - Match Criteria:
820 - Each ground truth step should exactly match a part of the solution or is directly entailed by a
821 part of the solution
822 - All the details in a step must be matched, including the specific value and content
823 - You should judge all the ground truth steps for whether there is a match in the solution
824 ## Output Format
825 [
826 {
827 "step_index": <integer>,
828 "judgment": "Matched" | "Unmatched"
829 },
830 ...
831 ]
832 ## Additional Rules
833 1. Only output the JSON array with no additional information.
834 2. Judge each ground truth step in order without omitting any step.
835 ## Inputs
836 Here are the problem, answer, model solution, and ground truth steps:
837 [Question]
838 {question}
839 {answer_options}
840 [Ground Truth Answer]
841 {answer}
842 [Ground Truth Steps]
843 {answer_steps}
844 [Model Solution]
845 {predicted_steps}

```

```

846
847
848
849
850
851
852
853
854
855
856
857
858
859
860
861
862
863

```

The prompt for calculating KNS is as follows:

```

864
865 Your task is to match the ground truth key points with the key points provided by an AI
866 model.
867 ## Input Format
868 1. Problem: The original question/task. Answer options are provided if available.
869 2. Answer: The ground truth final answer.
870 3. Ground Truth Steps: Essential steps required for deriving the correct answer.
871 4. Ground Truth Key Points: Key points in the ground truth steps.
872 5. Key points provided by an AI model.
873 ## Matching Process
874 - You need to determine how many of the key points provided by the AI model are correct.
875 - Match Criteria:
876 - When a given key point matches a ground truth keypoint, it means that they have similar
877 meanings in the context of this issue.
878 - You should judge all the key points provided by the AI model for whether there is a match
879 in the ground truth key points.
880 ## Output Format
881 [
882 {
883 "key_point_index": <integer>,
884 "key_point": <key point>,
885 "judgment": "Matched" | "Unmatched"
886 },
887 ...
888 ]
889 ## Additional Rules
890 1. Only output the JSON array with no additional information.
891 2. Judge each key point provided by the AI model in order without omitting any key point.
892 ## Inputs
893 Here are the problem, answer, ground truth steps, ground truth key points and key points
894 provided by the AI model:
895 [Question]
896 {question}
897 {answer_options}
898 [Ground Truth Answer]
899 {answer}
900 [Ground Truth Steps]
901 {answer_steps}
902 [Ground Truth Key Points]
903 {key_points}
904 [Key points provided by the AI model]
905 {provided_key_points}
906
907
908
909
910
911
912
913
914
915
916
917

```

The prompt for calculating KCS is as follows:

```

918
919 Your task is to match the ground truth key points with the key points provided by an AI
920 model.
921 ## Input Format
922 1. Problem: The original question/task. Answer options are provided if available.
923 2. Answer: The ground truth final answer.
924 3. Ground Truth Steps: Essential steps required for deriving the correct answer.
925 4. Ground Truth Key Points: Key points in the ground truth steps.
926 5. Key points provided by an AI model.
927 ## Matching Process
928 - You need to determine how many of the key points provided by the AI model are actually
929 present in the ground truth key points.
930 - You should pay attention to how many key points are successfully found by the AI model.
931 - Match Criteria:
932 - When a given key point matches a ground truth keypoint, it means that they have similar
933 meanings in the context of this issue.
934 - You should judge all the ground truth key points for whether there is a match in the key
935 points provided by the AI model.
936 ## Output Format
937 [
938 {
939   "key_point_index": <integer>,
940   "key_point": <key point>,
941   "judgment": "Matched" | "Unmatched"
942 },
943 ...
944 ]
945 ## Additional Rules
946 1. Only output the JSON array with no additional information.
947 2. Judge each ground truth key point in order without omitting any key point.
948 ## Inputs
949 Here are the problem, answer, ground truth steps, ground truth key points and key points
950 provided by the AI model:
951 [Question]
952 {question}
953 {answer_options}
954 [Ground Truth Answer]
955 {answer}
956 [Ground Truth Steps]
957 {answer_steps}
958 [Ground Truth Key Points]
959 {key_words}
960 [Key points provided by the AI model]
961 {provided_key_words}

```

960 The prompt for get CoT response is as follows:

964 You are a medical professional. Please answer the following questions.

965 [Question]

966 {question}

967 Give your answer in the following form with clear logic:

968 Step1: Step2:.... . Therefore, the answer is \box{ }.

971 The prompt for get direct response is as follows:

972

973 You are a medical professional. Please answer the following questions.

974 [Question]

975 {question}

976 End your answer in this format:

977 Therefore, the answer is \box{ }.

978

979 The prompt for extracting key points from responses is as follows:

980

981 You are a medical professional. We want to find out how many key points the AI model got  
982 right in its answer, so please extract the key points in the answer provided by the AI model  
983 based on the question and the answer. Please use as many words from the answer as possible.

984 The original question is:

985 [question]

986 And the answer provided by the AI model is:

987 [answer]

988 Please put the found key points into a list: [key point 1, key point 2,...]

989

990

991

992

993

994

995

996

997

998

999

1000

1001

1002

1003

1004

1005

1006

1007

1008

1009

1010

1011

1012

1013

1014

1015

1016

1017

1018

1019

1020

1021

1022

1023

1024

1025

Table 7: Comparison of model responses to multiple-choice question and open-ended question

	multiple-choice question	open-ended question
1026		
1027		
1028		
1029	<b>Question</b>	A 31-year-old woman presents to her obstetrician for her first prenatal visit after having a positive home pregnancy test 1 week ago. Her last menstrual period was 8 weeks ago. The patient has a medical history of type 1 diabetes mellitus since childhood and her home medications include insulin. Her hemoglobin A1c 2 weeks ago was 13.7%. {...} Her temperature is 98.6°F (37.0°C), blood pressure is 124/81 mmHg, pulse is 75/min, and respirations are 14/min. A physical exam is unremarkable. <b>This fetus is most likely at increased risk for which of the following complications?</b> <b>["Aneuploidy", "Neonatal hyperglycemia", "Neural tube defect", "Oligohydramnios", "Post-term delivery"]</b>
1030		A 31-year-old woman presents to her obstetrician for her first prenatal visit after having a positive home pregnancy test 1 week ago. Her last menstrual period was 8 weeks ago. The patient has a medical history of type 1 diabetes mellitus since childhood and her home medications include insulin. Her hemoglobin A1c 2 weeks ago was 13.7%. ... Her temperature is 98.6°F (37.0°C), blood pressure is 124/81 mmHg, pulse is 75/min, and respirations are 14/min. A physical exam is unremarkable. <b>This fetus is most likely at increased risk for which complications?</b>
1031		
1032		
1033		
1034		
1035		
1036		
1037		
1038		
1039		
1040		
1041		
1042		
1043		
1044		
1045	<b>Response</b>	["Identify the key medical condition of the patient, which is type 1 diabetes mellitus with poor glycemic control, indicated by a hemoglobin A1c of 13.7%.", "Recognize that poor glycemic control in early pregnancy, especially in the periconceptional period, increases the risk of congenital malformations in the fetus.", "Understand that the most common anomalies in infants of diabetic mothers (IDM) with poor glycemic control include cardiovascular defects and neural tube defects, such as anencephaly and spina bifida.", "Review the options provided to determine which complication is most likely associated with poorly controlled diabetes in pregnancy:\n- Option A: Aneuploidy is related to advanced maternal age, not diabetes.\n- Option B: Neonatal hyperglycemia is incorrect as IDM typically experience hypoglycemia due to fetal hyperinsulinemia.\n- Option C: Neural tube defect is a known risk in pregnancies complicated by poorly controlled diabetes.\n- Option D: Oligohydramnios is not specifically associated with maternal diabetes; polyhydramnios is more common.\n- Option E: Post-term delivery is less likely as maternal diabetes often leads to prematurity.", "Conclude that the fetus is most likely at increased risk for neural tube defects due to the mother's poor glycemic control during early pregnancy.\n\nFinal Answer: C: Neural tube defect."]
1046		
1047		
1048		
1049		
1050		
1051		
1052		
1053		
1054		
1055		
1056		
1057		
1058		
1059		
1060		
1061		
1062		
1063		
1064		
1065		
1066		
1067		
1068		
1069		
1070		
1071		
1072		
1073		
1074		
1075		
1076		
1077		
1078		
1079		

1080  
 1081  
 1082  
 1083  
 1084  
 1085  
 1086  
 1087  
 1088  
 1089  
 1090  
 1091  
 1092  
 1093  
 1094  
 1095  
 1096  
 1097  
 1098  
 1099  
 1100  
 1101  
 1102

Table 8: Composition of MedReason-Dx.

1103  
 1104  
 1105  
 1106  
 1107  
 1108  
 1109  
 1110  
 1111  
 1112  
 1113  
 1114  
 1115  
 1116  
 1117  
 1118  
 1119  
 1120  
 1121  
 1122  
 1123  
 1124  
 1125  
 1126  
 1127  
 1128  
 1129  
 1130  
 1131  
 1132  
 1133

source	MCQ	QEQ
JAMA Clinical Challenge	210	204
Medbullets	197	125
MMedBench	164	35
nephSAP	21	7
LiveQA	0	94
PubMedQA	0	113
total	592	578

1134  
1135  
1136  
1137  
1138  
1139

Table 9: A piece of data from MedReason-Dx.

1140	<b>id</b>	9cb1a646-3e55-5b96-a769-b06289709164
1141	<b>specialty</b>	Pediatrics
1142	<b>question</b>	A 6-year-old boy presents to the emergency department after falling from his scooter. He has dull, aching pain along his left side where he hit the ground. He fell sideways off the scooter as he rounded a curve in the road. He has never had any serious injuries but that he always seems to bruise easily, especially after he started playing soccer this fall. His parents deny that he has an abnormal number of nosebleeds or bleeding from the gums. They have never seen blood in his stool or urine. His mother notes that her brother has had similar problems. His temperature is 98.6°F (37°C), blood pressure is 112/74 mmHg, pulse is 82/min, and respirations are 11/min. On physical exam, the patient has extensive bruising of the lateral left thigh and tenderness to palpation. Laboratory tests are performed and reveal the following: Hemoglobin: 14 g/dL Hematocrit: 41% Mean corpuscular volume: 89 $\mu$ m <sup>3</sup> Reticulocyte count: 0.8% Leukocyte count: 4,700/mm <sup>3</sup> Prothrombin time (PT): 13 seconds Partial thromboplastin time (PTT): 56 seconds Bleeding time (BT): 4 minutes Which of the following is the most likely underlying pathophysiology?
1143	<b>answer_options</b>	["Anti-platelet antibodies", "Factor 8 deficiency", "Factor 9 deficiency", "GP1b deficiency", "Von Willebrand factor deficiency"]
1144	<b>answer</b>	Factor 8 deficiency
1145	<b>answer_idx</b>	B
1146	<b>answer_steps</b>	["A 6-year-old boy presents with extensive bruising after falling, and has a family history of similar bleeding problems, suggesting a possible hereditary bleeding disorder.", "The laboratory findings show a prolonged partial thromboplastin time (PTT) with normal prothrombin time (PT) and bleeding time (BT), which suggests a coagulation factor deficiency rather than a platelet function disorder.", "The isolated prolongation of PTT with normal PT and BT is characteristic of hemophilia, which can be caused by either factor 8 deficiency (hemophilia A) or factor 9 deficiency (hemophilia B).", "Hemophilia A (factor 8 deficiency) and hemophilia B (factor 9 deficiency) present with similar symptoms, including hemarthrosis, frequent bruising, and prolonged PTT. However, hemophilia A is more common than hemophilia B.", "Given the epidemiological prevalence and the clinical presentation, factor 8 deficiency (hemophilia A) is the most likely diagnosis.", "Therefore, the most likely underlying pathophysiology is factor 8 deficiency, making 'B': Factor 8 deficiency the correct answer."]
1147	<b>answer_key_points</b>	["extensive bruising", "family history", "prolonged partial thromboplastin time", "normal prothrombin time", "normal bleeding time", "hemophilia A", "factor 8 deficiency", "intramuscular hematoma", "gastrointestinal bleeding", "isolated prolongation of PTT", "normal bleeding time", "factor replacement", "anti-factor antibodies"]
1148		
1149		
1150		
1151		
1152		
1153		
1154		
1155		
1156		
1157		
1158		
1159		
1160		
1161		
1162		
1163		
1164		
1165		
1166		
1167		
1168		
1169		
1170		
1171		
1172		
1173		
1174		
1175		
1176		
1177		
1178		
1179		
1180		
1181		
1182		
1183		
1184		
1185		
1186		
1187		

1188  
 1189  
 1190  
 1191  
 1192  
 1193  
 1194  
 1195  
 1196  
 1197  
 1198  
 1199  
 1200  
 1201  
 1202  
 1203  
 1204  
 1205  
 1206  
 1207  
 1208  
 1209  
 1210  
 1211  
 1212  
 1213  
 1214  
 1215  
 1216  
 1217  
 1218  
 1219  
 1220  
 1221  
 1222  
 1223  
 1224  
 1225  
 1226  
 1227  
 1228  
 1229  
 1230  
 1231  
 1232  
 1233  
 1234  
 1235  
 1236  
 1237  
 1238  
 1239  
 1240  
 1241

Table 10: Response from DeepSeek R1.

model	reasoning steps	key points
DeepSeek R1	<p>["The patient presents with **prolonged PTT (56 seconds)**, normal PT (13 seconds), and **normal bleeding time (4 minutes)**. This suggests a **factor deficiency in the intrinsic coagulation pathway** (e.g., Factor VIII, IX) rather than platelet dysfunction or extrinsic pathway issues.", **Bleeding time is normal**, ruling out primary platelet disorders (e.g., Bernard-Soulier syndrome, ITP) and mild von Willebrand disease (vWD), where bleeding time is often prolonged due to impaired platelet adhesion.", **Family history** (maternal uncle with similar symptoms) suggests an **X-linked recessive disorder**, narrowing the diagnosis to **Hemophilia A (Factor VIII deficiency)** or **Hemophilia B (Factor IX deficiency)**. Both cause prolonged PTT with normal PT and bleeding time.", **No mucosal bleeding** (denied nosebleeds, gum bleeding) and trauma-related bruising align more with **Hemophilia** than vWD, which commonly involves mucosal bleeding.", **Hemophilia A (Factor VIII deficiency)** is **more prevalent** than Hemophilia B, and the X-linked inheritance pattern fits the maternal uncle's history.\n\nTherefore, the answer is \boxed{B}."]</p>	<p>["prolonged PTT (56 seconds)", "normal PT (13 seconds)", "normal bleeding time (4 minutes)", "factor deficiency in the intrinsic coagulation pathway", "Bleeding time is normal", "ruling out primary platelet disorders", "Family history", "X-linked recessive disorder", "Hemophilia A (Factor VIII deficiency)", "Hemophilia B (Factor IX deficiency)", "prolonged PTT with normal PT and bleeding time", "No mucosal bleeding", "trauma-related bruising", "Hemophilia", "Hemophilia A (Factor VIII deficiency)", "more prevalent", "X-linked inheritance pattern"]</p>

Table 11: Response from GPT-4o.

model	reasoning steps	key points
GPT-4o	<p>["Analyze the clinical presentation. The 6-year-old boy presents with easy bruising but no significant mucosal bleeding (such as nosebleeds or gum bleeding), which suggests a clotting factor deficiency rather than a platelet function disorder.", "Review the family history. The maternal uncle has had similar bleeding issues, suggesting a potential X-linked recessive pattern of inheritance, which is typical for hemophilia.", "Examine the laboratory findings. The prolonged partial thromboplastin time (PTT) with normal prothrombin time (PT) and bleeding time (BT) suggest an intrinsic pathway clotting factor deficiency.", "Consider the specific clotting factor deficiencies that could cause the prolonged PTT. Hemophilia A (Factor 8 deficiency) and Hemophilia B (Factor 9 deficiency) are the two most common intrinsic pathway deficiencies. Both would cause a prolonged PTT without affecting PT or BT.", "Choose between Factor 8 and Factor 9 deficiency. While both conditions fit the laboratory pattern, Hemophilia A (Factor 8 deficiency) is more common than Hemophilia B (Factor 9 deficiency).\n\nTherefore, the answer is \boxed{B}."]</p>	<p>["6-year-old boy", "easy bruising", "no significant mucosal bleeding", "clotting factor deficiency", "family history", "maternal uncle", "X-linked recessive pattern", "hemophilia", "prolonged partial thromboplastin time (PTT)", "normal prothrombin time (PT)", "normal bleeding time (BT)", "intrinsic pathway", "Hemophilia A", "Factor 8 deficiency", "Hemophilia B", "Factor 9 deficiency", "prolonged PTT", "more common", "answer is B"]</p>

1242  
1243  
1244  
1245  
1246  
1247  
1248  
1249

Table 12: Response from QWQ 32B.

model	reasoning steps	key points
QWQ 32B	<p>["Step 1: **Identify key clinical findings**: The patient presents with easy bruising, a family history of similar issues (mother's brother), and normal hemoglobinhematocrit", "The labs show a normal PT, prolonged PTT (56 seconds), and a normal bleeding time (4 minutes)", "Step 2: **Analyze coagulation studies**: \n- **Normal PT** indicates the extrinsic pathway (factors VII, X, etc", ") is intact", "- **Prolonged PTT** suggests a defect in the intrinsic or common pathway", "- **Normal bleeding time** rules out platelet dysfunction (e", "g", ", von Willebrand disease, Bernard-Soulier syndrome, or anti-platelet antibodies)", "Step 3: **Consider differentials for prolonged PTT with normal PT and bleeding time**: \n- **Factor deficiencies in the intrinsic pathway**: Factors VIII (hemophilia A), IX (hemophilia B), or XI", "- **Factor VIII deficiency (hemophilia A)** and **Factor IX deficiency (hemophilia B)** are X-linked recessive disorders, fitting the family history (mother's brother affected)", "- **Factor XI deficiency** is autosomal recessive and less likely given the family history", "Step 4: **Differentiate between hemophilia A and B**: Both cause prolonged PTT and bruising/bleeding but are clinically similar", "- **Hemophilia A (Factor VIII deficiency)** is more common than hemophilia B", "- The family history (maternal uncle affected) suggests an X-linked disorder, compatible with either", "However, hemophilia A is more prevalent", "Step 5: **Exclude other options**: \n- **Von Willebrand disease (E)**: Causes prolonged PTT and bleeding time", "However, the normal bleeding time here makes this less likely", "- **GP1b deficiency (D)**: Causes prolonged bleeding time and platelet dysfunction, which is not seen here", "- **Anti-platelet antibodies (A)**: Would cause thrombocytopenia or abnormal platelet function, but bleeding time is normal", "- **Factor IX deficiency (C)** is possible but less common than hemophilia A", "Step 6: **Conclusion**: The prolonged PTT, normal PT and bleeding time, family history of X-linked bleeding, and higher prevalence of hemophilia A make **Factor 8 deficiency (hemophilia A)** the most likely diagnosis", "Therefore, the answer is \boxed{B}"]</p>	<ul style="list-style-type: none"> <li>"key clinical findings",</li> <li>"easy bruising",</li> <li>"family history",</li> <li>"normal hemoglobin",</li> <li>"hematocrit",</li> <li>"normal PT",</li> <li>"prolonged PTT",</li> <li>"normal bleeding time",</li> <li>"platelet dysfunction",</li> <li>"factor deficiencies",</li> <li>"intrinsic pathway",</li> <li>"hemophilia A",</li> <li>"factor VIII deficiency",</li> <li>"hemophilia B",</li> <li>"family history",</li> <li>"X-linked recessive disorders",</li> <li>"maternal uncle",</li> <li>"exclude other options",</li> <li>"von Willebrand disease",</li> <li>"GP1b deficiency",</li> <li>"anti-platelet antibodies",</li> <li>"factor IX deficiency",</li> <li>"conclusion"]</li> </ul>

1290  
1291  
1292  
1293  
1294  
1295

1296  
 1297  
 1298  
 1299  
 1300  
 1301  
 1302  
 1303  
 1304  
 1305  
 1306  
 1307  
 1308  
 1309  
 1310

Table 13: Response from PMC LLaMA 13B.

model	reasoning steps	key points
PMC LLaMA 13B	<p>["With these laboratory values, it is clear there is a problem with the bone marrow production of platelets, as demonstrated by an extremely low reticulocyte count and an increased mean corpuscular volume, which is indicative of a megaloblastic process", "This child likely has a disorder involving the thrombopoietin receptor, which is essential for megakaryocyte growth and maturation and for the release of platelets from the marrow", "The gene for the thrombopoietin receptor is located on chromosome 11", "Most cases of congenital amegakaryocytic thrombocytopenia result from autosomal recessive inheritance of a nonsense point mutation involving one of the thrombopoietin receptor genes", "The gene for the thrombopoietin receptor is the same as that associated with congenital amegakaryocytic thrombocytopenia", "This autosomal dominant form is extremely rare", "The other choices are not associated with a low reticulocyte count and a giant platelet", "Patients with antiplatelet antibodies, or with factor 8, 9, or Von Willebrand factor deficiencies have normal to increased platelet counts", "Patients with GP1b deficiency usually have a normal or increased platelet count, but a dysfunctional platelet population", "###Answer: OPTION D IS CORRECT"]</p>	<p>["laboratory values", "bone marrow production", "platelets", "low reticulocyte count", "increased mean corpuscular volume", "megaloblastic process", "thrombopoietin receptor", "megakaryocyte growth", "maturation", "release of platelets", "chromosome 11", "congenital amegakaryocytic thrombocytopenia", "autosomal recessive inheritance", "nonsense point mutation", "autosomal dominant form", "normal to increased platelet counts", "dysfunctional platelet population"]</p>

1337  
 1338  
 1339  
 1340  
 1341  
 1342  
 1343  
 1344  
 1345  
 1346  
 1347  
 1348  
 1349