Benchmarking LLMs on Authentic Cases from Medical Journals

Anonymous ACL submission

Abstract

In recent years, large language models (LLMs) have demonstrated remarkable capabilities in the medical domain. However, existing medical benchmarks suffer from performance saturation and are predominantly derived from medical exam questions, which fail to adequately capture the complexity of real-world clinical scenarios. To bridge this gap, we introduce **ClinBench**, a challenging benchmark based on authentic clinical cases sourced from authoritative medical journals. Each question retains the complete patient information and clinical test results from the original case, effectively simulating real-world clinical practice. Additionally, we implement a rigorous human review process involving medical experts to ensure the quality and reliability of the benchmark.

ClinBench supports both **textual and multimodal** evaluation formats, covering 12 medical specialties with over 2,000 questions, which provides a comprehensive benchmark for assessing LLMs' medical capabilities. We evaluate the performance of over 20 open-source and proprietary LLMs and benchmark them against human medical experts. Our findings reveal that human experts still retain an advantage within their specialized fields, while LLMs demonstrate superior overall performance on a broader range of medical specialties.

1 Introduction

Recent advancements in large language models (LLMs) have demonstrated a remarkable ability to understand and generate medical content, marking significant progress in the medical field (Thirunavukarasu et al., 2023; Liévin et al., 2024; Clusmann et al., 2023; Chen et al., 2024a). Their impressive performance underscores their potential to approach expert-level intelligence.

With the rapid advancement of medical LLMs, existing medical benchmarks lack sufficient challenge and face the issue of performance saturation. For instance, powerful LLMs such as GPT-4o (OpenAI, 2024), Gemini-2.5-Pro (Guo et al., 2025) have achieved accuracy approaching 90% on widely used medical benchmarks like MedQA (Jin et al., 2021), MedMCQA (Pal et al., 2022) and PubMedQA (Jin et al., 2019). To address this limitation, recent works (Zuo et al., 2025; Tang et al., 2025; McDuff et al., 2025) have attempted to introduce more challenging benchmarks by incorporating advanced, expert-level examinations, such as medical licensing tests.

However, **these benchmarks remain predominantly exam-oriented and fail to reflect realistic clinical scenarios**. Real-world clinical scenarios require physicians to integrate medical knowledge with practical experience to navigate complex and multifaceted situations, taking into account patient symptoms, medical history, imaging findings, and various diagnostic test results. In contrast, exam questions typically assess isolated pieces of medical knowledge without adequately simulating complex clinical contexts. Consequently, these questions do not sufficiently capture the intricacies and nuances of real-world clinical scenarios.

To address these limitations, we introduce ClinBench: a challenging and real-world medical benchmark for future medical LLMs. Our ClinBench has four key features: (1) **Real-world clinical scenarios:** The questions are sourced from **authoritative medical journals and based directly on real-world clinical cases**. Each question retains the complete patient information and clinical test results from the original case, effectively simulating real-world clinical practice. (2) **High challenge:** The questions are highly challenging, even for experienced physicians. Each question corresponds to a realistic clinical task, requiring specialized medical knowledge, analytical skills, medical image interpretation, and extensive clinical experience. (3) **Quality assurance:** For each question, the stem is derived from authoritative medical cases. The

Dataset	# Size	# Avg Lens	# Avg Option Num	# Saturation	Real-Med Scenario	Specialties
MedQA (USMLE) (Jin et al., 2021)	1,273	116.6	4	✓	✗ (Med Exams)	✗
PubMedQA (PQA-L) (Jin et al., 2019)	1,000	14.4	3	✓	✓ (Med Journels)	✗
MedMCQA (Pal et al., 2022)	4,183	12.8	4	✓	✗ (Med Exams)	✗
MMLU (Med) (Hendrycks et al.)	1,089	100.1	4	✓	✗ (Med Exams)	✗
MMLU-Pro (Med) (Wang et al., 2024)	586	166.6	10	✗	✗ (Med Exams)	✗
MedXpertQA$_{\text{Text}}$ (Zuo et al., 2025)	2,450	257.4	8.7	✗	✗ (Med Exams)	✓
ClinBench$_{\text{Text}}$ (Ours)	2,014	462.1	8.4	✗	✓ (Patient Cases)	✓

Table 1: Comparison with existing textual medical benchmarks. **# Saturation** indicates whether the dataset suffers from performance saturation. **Real-Med Scenario** denotes whether the questions reflect real-world clinical settings. **Specialties** indicates whether the benchmark categorizes questions by medical specialty.

Dataset	# Size	# Images	# Image Rate	# Avg Lens	# Saturation	Real-Med Scenarios	Specialties
PMC-VQA (Zhang et al., 2023b)	33,430	29,021	0.9	61.8	✓	✗	✗
OmniMedVQA (Hu et al., 2024)	127,995	118,010	0.9	42.4	✓	✗	✗
GMAI-MMBench (Ye et al., 2024)	21,281	21,180	1.0	49.9	✓	✗	✓
MMMU (H & M) (Yue et al., 2024)	1,752	1,994	1.1	83.6	✓	✗	✗
MMMU-Pro (H & M) (Yue et al., 2024)	346	431	1.3	107.1	✗	✗	✗
MedXpertQA$_{\text{MM}}$ (Zuo et al., 2025)	2, 000	2, 852	1.4	149.4	✗	✗	✓
ClinBench$_{\text{MM}}$ (Ours)	2,014	4,978	2.5	421.7	✗	✓	✓

Table 2: Comparison with existing multimodal medical benchmarks. **# Image Rate** refers to the average number of images included per question. **# Saturation**, **Real-Med Scenario**, and **Specialties** are consistent with Table 1.

golden answer is provided by an expert panel, and each question is thoroughly reviewed and validated by human experts. (4) **Comprehensive Evaluation:** ClinBench provides both textual and multimodal versions, covering 12 medical specialties and encompassing more than 2,000 questions. Additionally, the inclusion of a dedicated rare-disease track further enhances its clinical comprehensiveness, offering a more comprehensive evaluation.

We evaluate over 20 LLMs, including both open-source and proprietary LLMs. Additionally, we engage attending-level human medical experts to answer ClinBench questions, facilitating a comparison between human experts and LLMs. Our key contributions are summarized as follows:

- We propose **ClinBench**, the first medical multiple-choice benchmark focusing on realistic clinical scenarios. ClinBench has both textual and multimodal versions, with questions derived from authentic clinical cases, closely simulating the real-world scenarios.

- ClinBench is built upon authoritative medical journals with rigorous quality assurance processes. Comprehensive human checks and data leakage risk assessments are conducted to ensure the reliability and quality of the questions.

- We evaluate ClinBench across more than 20 LLMs, providing a comprehensive assessment of the current medical capabilities of existing medical LLMs. Furthermore, through the comparison between human experts and LLMs, we find that human experts still retain an advantage within their specialized fields, while LLMs demonstrate superior overall performance across a broader range of medical specialties.

2 Comparison with Existing Benchmarks

Statistic Comparison. As shown in Tables 1 and 2, traditional text medical benchmarks like MedQA and PubMedQA have short questions with limited options, lacking the challenge of complex, specialized medical tasks. Additionally, MMLU (Hendrycks et al.), MMLU-pro (Wang et al., 2024) and MedXpertQA (Zuo et al., 2025) datasets, mostly sourced from educational exams, fail to accurately represent real clinical tasks. In contrast, our ClinBench$_{\text{Text}}$ includes longer, more complex questions with multiple options, all derived from authoritative case journals, offering a better reflection of real clinical scenarios. Moreover, ClinBench$_{\text{MM}}$ incorporates more images per question compared to existing multimodal medical benchmarks, reflecting the complexity of real-world multimodal medical scenarios.

Discriminative Comparison. As LLMs continue to advance, existing medical benchmarks struggle to effectively evaluate the performance

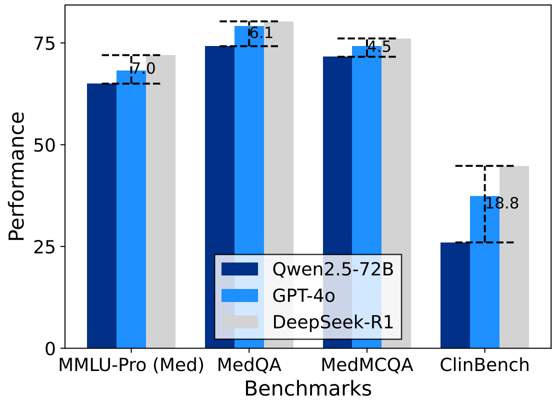


Figure 1: Performance gap among models on different benchmarks. The three exam-oriented benchmarks fail to effectively distinguish capability differences among models, while ClinBench$_{\text{Text}}$ provides clearer gap.

gap among models. As illustrated in Figure 1, compared to benchmarks such as MMLU-Pro and MedQA, our **ClinBench** demonstrates greater discriminative power, enabling clearer distinctions between model performances. This discriminative nature makes ClinBench as a more suitable benchmark for evaluating and guiding the future development of medical LLMs.

3 The ClinBench Benchmark

3.1 Overview

ClinBench consists of two versions: textual and multimodal. The textual version, ClinBench$_{\text{Text}}$, contains 2,014 multiple-choice questions covering 12 distinct medical specialties. The multimodal version, ClinBench$_{\text{MM}}$, is based on ClinBench$_{\text{Text}}$ but partially replaces textual information in question stems with medical images.

The questions in ClinBench are sourced from authoritative medical journals available on two platforms: (1) **PubMed Central**[1], an authoritative repository of English-language medical case journals; and (2) **China Medical Website**[2], a prominent Chinese medical platform hosting a wide range of high-quality medical case journals in Chinese. We construct our benchmark based on medical journals for three main reasons:

1. Medical journals provide comprehensive patient information and detailed clinical test results, closely simulating the realistic diagnostic process.

2. Diagnoses presented in these journals are validated by expert medical panels, ensuring authoritative answers.

3. Patient information in these journals is thoroughly anonymized, effectively mitigating privacy concerns.

Figure 4 provides two illustrative example from ClinBench$_{\text{Text}}$ and ClinBench$_{\text{MM}}$, demonstrating a challenging question that integrates detailed patient information with medical imaging data.

3.2 The Construction of ClinBench$_{\text{Text}}$

Figure 2 illustrates the construction process of ClinBench$_{\text{Text}}$, which consists of three steps: data preprocessing, question stem construction, and candidate options construction.

Data Preprocess. We first convert medical journal PDFs into text format using the MinerU tool[3]. Then, we apply a three-step filtering pipeline to both Chinese and English medical journals: (1) **Filtering for Diagnostic Cases:** We first select diagnostic medical cases from PubMed Central (32M) and the Chinese Medical Website (1M) using keyword tags. As a result, we obtain the full text of approximately 40K English and 6K Chinese diagnostic medical case journals. (2) **Filtering for Complete Cases:** We then apply rule-based filtering to exclude incomplete case reports, retaining only those that contain essential sections: patient information, clinical test results, diagnostic conclusions, and treatment plans. Additionally, we discard cases that lack medical images in the patient information and clinical test results, ensuring that each question includes medical images. After this step, we obtain around 4K high-quality, complete medical cases in both English and Chinese. (3) **Removing Duplicates:** Finally, we eliminate duplicate or highly similar cases to maintain the dataset's diversity and quality. A more detailed process is shown in Appendix B.

Question Stem Construction. We hired 30 undergraduate students majoring in medicine to construct the question stems. Following our detailed guidelines (see Appendix B), they extracted the *Patient Information* and *Clinical Test Results* sections from each journal to form the question stem. For Chinese cases, they used translation tools and ensured translation quality. The final question format

[1] https://pmc.ncbi.nlm.nih.gov/
[2] https://www.yiigle.com/
[3] https://mineru.net/

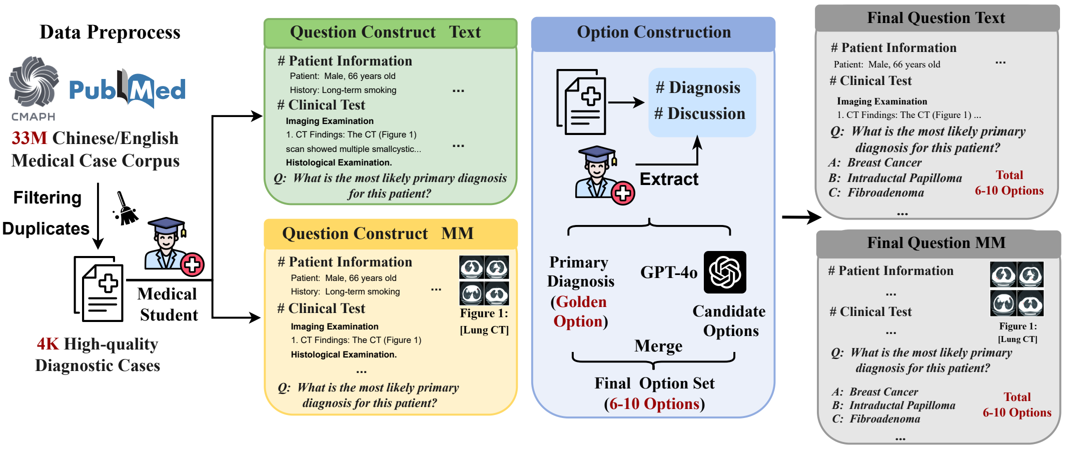


Figure 2: The pipeline for constructing ClinBench. ClinBench$_{\text{MM}}$ is built upon ClinBench$_{\text{Text}}$, with the difference being that the question stems in the multimodal version have the textual descriptions of the associated medical images removed.

for both datasets is: *What is the most likely primary diagnosis for this patient?*

Candidate Options Construction. The construction of candidate options follows four main substeps: (1) Medical undergraduate students first extract the final diagnosis from the original journals' *Diagnosis* sections as the correct answer. Potential alternative diagnoses mentioned in the discussions serve as distractors.[4] This step yields one correct option and five distractors. (2) Next, the constructed question stem is submitted to GPT-4o, which is instructed to generate the five most plausible diagnoses, ensuring that each option is clear and precise. (3) Finally, GPT-4o merges the candidate options from the previous two steps, eliminating ambiguous or duplicate entries, thus producing a final, concise set consisting of one correct option and multiple high-quality distractors.

Specialty Categorization. Finally, we classify all questions into 12 medical specialties based on the diseases indicated by the correct answers. The detailed categorization is provided in the Appendix B.3.

3.3 The Construction of ClinBench$_{\text{MM}}$

In real-world clinical scenarios, physicians rely not only on patient textual descriptions but also on medical images from clinical examinations to make informed decisions. To better reflect this multimodal scenario of medical diagnosis, we construct a multimodal version of the ClinBench benchmark dataset. Specifically, we employ undergraduate students to curate ClinBench$_{\text{MM}}$ based on questions from ClinBench$_{\text{Text}}$ by removing textual content that describes the associated medical images.

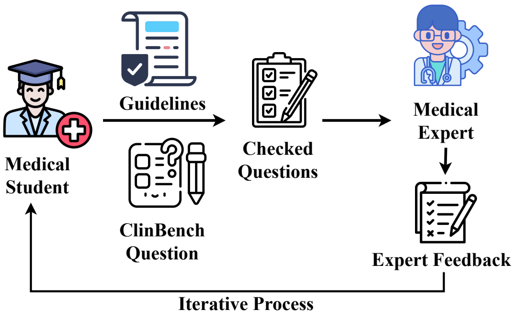
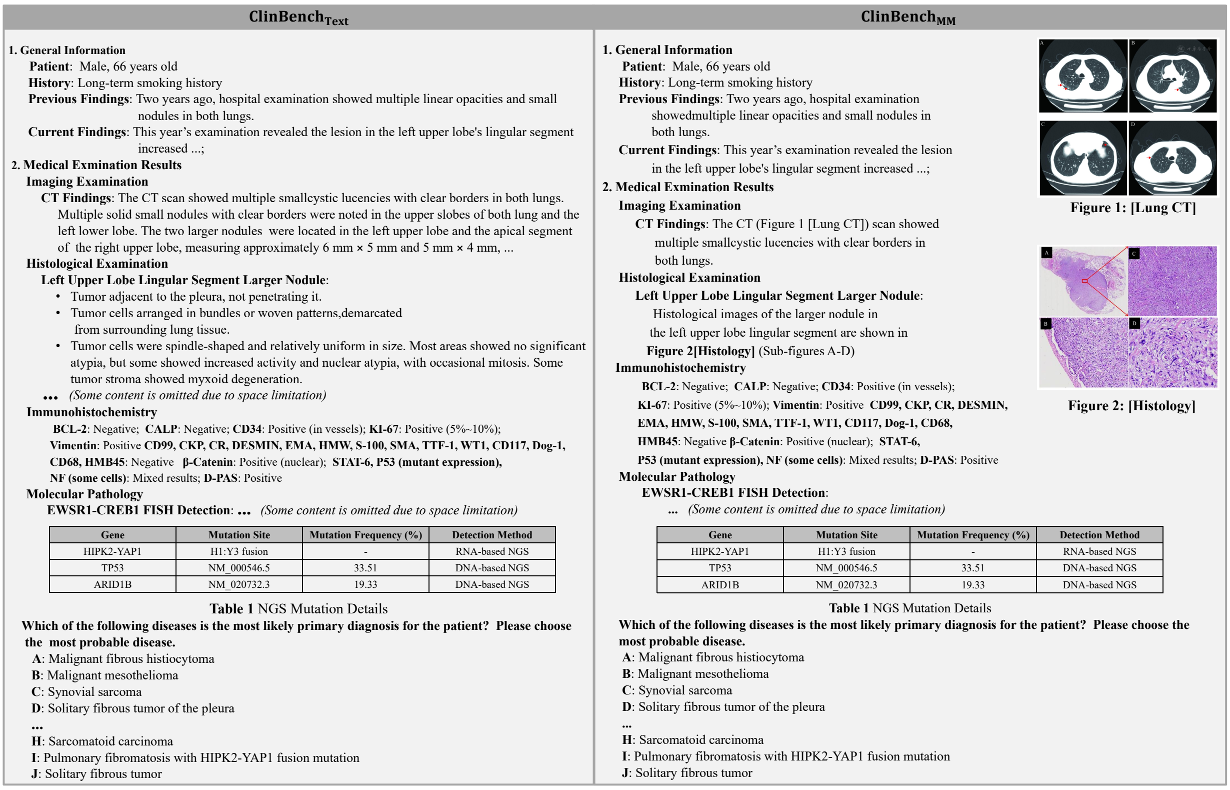


Figure 3: The human check pipeline for ClinBench.

To ensure annotation consistency and quality, we provide the annotators with detailed guidelines (see Appendix B) that instruct them on how to identify and remove image-referential text from the question stems. If a question contains a low-quality image or one that lacks clinically relevant diagnostic information, it is excluded from the dataset. Following this procedure, we obtain over 2,000 multimodal questions, forming the ClinBench$_{\text{MM}}$ benchmark. Compared to ClinBench$_{\text{Text}}$, the question stems in ClinBench$_{\text{MM}}$ exclude descriptions related to medical images. ClinBench$_{\text{MM}}$ challenges models to accurately extract and reason over visual content from medical images, providing a rigorous assessment of their multimodal understanding and diagnostic capabilities.

[4] If insufficient alternative diagnoses are mentioned, students are instructed to propose additional plausible alternatives.



ClinBench$_{\text{Text}}$

1. General Information

Patient: Male, 66 years old

History: Long-term smoking history

Previous Findings: Two years ago, hospital examination showed multiple linear opacities and small nodules in both lungs.

Current Findings: This year's examination revealed the lesion in the left upper lobe's lingular segment increased ...;

2. Medical Exmination Results

Imaging Examination

CT Findings: The CT scan showed multiple smallcystic lucencies with clear borders in both lungs. Multiple solid small nodules with clear borders were noted in the upper slobes of both lung and the left lower lobe. The two larger nodules were located in the left upper lobe and the apical segment of the right upper lobe, measuring approximately 6 mm × 5 mm and 5 mm × 4 mm, ...

Histological Examination

Left Upper Lobe Lingular Segment Larger Nodule:

- Tumor adjacent to the pleura, not penetrating it.
- Tumor cells arranged in bundles or woven patterns,demarcated from surrounding lung tissue.
- Tumor cells were spindle-shaped and relatively uniform in size. Most areas showed no significant atypia, but some showed increased activity and nuclear atypia, with occasional mitosis. Some tumor stroma showed myxoid degeneration.

... *(Some content is omitted due to space limitation)*

Immunohistochemistry

BCL-2: Negative; **CALP**: Negative; **CD34**: Positive (in vessels); **KI-67**: Positive (5%~10%); **Vimentin**: Positive **CD99, CKP, CR, DESMIN, EMA, HMW, S-100, SMA, TTF-1, WT1, CD117, Dog-1, CD68, HMB45**: Negative **β-Catenin**: Positive (nuclear); **STAT-6, P53 (mutant expression), NF (some cells)**: Mixed results; **D-PAS**: Positive

Molecular Pathology

EWSR1-CREB1 FISH Detection: **...** *(Some content is omitted due to space limitation)*

Gene	Mutation Site	Mutation Frequency (%)	Detection Method
HIPK2-YAP1	H1:Y3 fusion	-	RNA-based NGS
TP53	NM_000546.5	33.51	DNA-based NGS
ARID1B	NM_020732.3	19.33	DNA-based NGS

Table 1 NGS Mutation Details

Which of the following diseases is the most likely primary diagnosis for the patient? Please choose the most probable disease.

A: Malignant fibrous histiocytoma
B: Malignant mesothelioma
C: Synovial sarcoma
D: Solitary fibrous tumor of the pleura
...
H: Sarcomatoid carcinoma
I: Pulmonary fibromatosis with HIPK2-YAP1 fusion mutation
J: Solitary fibrous tumor

ClinBench$_{\text{MM}}$

1. General Information

Patient: Male, 66 years old

History: Long-term smoking history

Previous Findings: Two years ago, hospital examination showedmultiple linear opacities and small nodules in both lungs.

Current Findings: This year's examination revealed the lesion in the left upper lobe's lingular segment increased ...;

2. Medical Exmination Results

Imaging Examination

CT Findings: The CT (Figure 1 [Lung CT]) scan showed multiple smallcystic lucencies with clear borders in both lungs.

Histological Examination

Left Upper Lobe Lingular Segment Larger Nodule:

Histological images of the larger nodule in the left upper lobe lingular segment are shown in **Figure 2[Histology]** (Sub-figures A-D)

Immunohistochemistry

BCL-2: Negative; **CALP**: Negative; **CD34**: Positive (in vessels); **KI-67**: Positive (5%~10%); **Vimentin**: Positive **CD99, CKP, CR, DESMIN, EMA, HMW, S-100, SMA, TTF-1, WT1, CD117, Dog-1, CD68, HMB45**: Negative **β-Catenin**: Positive (nuclear); **STAT-6, P53 (mutant expression), NF (some cells)**: Mixed results; **D-PAS**: Positive

Molecular Pathology

EWSR1-CREB1 FISH Detection:

... *(Some content is omitted due to space limitation)*

Figure 1: [Lung CT]

Figure 2: [Histology]

Gene	Mutation Site	Mutation Frequency (%)	Detection Method
HIPK2-YAP1	H1:Y3 fusion	-	RNA-based NGS
TP53	NM_000546.5	33.51	DNA-based NGS
ARID1B	NM_020732.3	19.33	DNA-based NGS

Table 1 NGS Mutation Details

Which of the following diseases is the most likely primary diagnosis for the patient? Please choose the most probable disease.

A: Malignant fibrous histiocytoma
B: Malignant mesothelioma
C: Synovial sarcoma
D: Solitary fibrous tumor of the pleura
...
H: Sarcomatoid carcinoma
I: Pulmonary fibromatosis with HIPK2-YAP1 fusion mutation
J: Solitary fibrous tumor

Figure 4: Two demos of ClinBench$_{\text{Text}}$ and ClinBench$_{\text{MM}}$, respectively. More cases are shown in Appendix E.

3.4 Human Expert Check

We summarize the potential issues that may arise during the question construction process:

(1) **Incorrect Question Stem:** The constructed question stem may **omit critical medical information** present in the original case report. Additionally, when constructing ClinBench$_{\text{MM}}$ questions, students are required to manually remove textual descriptions of image-related content from the stem. This process may further result in the inadvertent loss of important information essential for accurately understanding or answering the question.

(2) **Inappropriate Candidate Options:** During the option merging process, some candidate options may **overlap with the correct answer** or be ambiguously phrased, making it difficult to ensure a clear and unambiguous set of choices.

Two-level Human Check. To ensure the quality of the ClinBench benchmark, we implement a rigorous two-level human check process, and the process is shown in Figure 3. First, medical undergraduate students review the question stems and options to ensure clarity and accuracy. Then, we invite experienced practicing physicians (e.g., attending doctors) to conduct a sampling-based inspection. If any quality issues are identified, they are systematically summarized and fed back to medical students for targeted revision. This iterative process maintains the overall reliability and quality of the benchmark. More detailed process is shown in Appendix B.

4 Experiments

4.1 Experimental Setup

We evaluate all models under a zero-shot prompt setting. Models with fewer than 32B parameters are evaluated locally using 8 * A800 GPUs. During evaluation, we set the temperature to $t = 0.6$ and report the average results over three independent runs. For models larger than 32B parameters, we use the official APIs for evaluation. Detailed evaluation prompts are provided in Appendix D.

4.2 Models

We conduct evaluations on a wide range of LLMs and large multimodal models (LMMs) using ClinBench$_{\text{Text}}$ and ClinBench$_{\text{MM}}$, respectively. Our benchmark includes both *proprietary* and *open-source models*, and additionally covers advanced large *reasoning* medical models, with a focus on capturing the latest advancements in medical reasoning capabilities. Detailed information of models is shown in Appendix D.1.

Model	GH	Surg	Neuro	Oph/ENT	DI	Resp	Dent	OG	Ortho	Cardio	Ped	Avg
General LLMs												
GPT-4o-2024-11-20	35.7	42.4	37.3	29.8	34.0	40.6	38.3	31.7	35.4	39.8	39.0	37.4
Deepseek-V3	**44.9**	**45.5**	**45.2**	**39.7**	**40.9**	**42.4**	**55.0**	36.6	43.4	**46.4**	**44.1**	**41.3**
Grok-3	38.5	38.4	40.8	33.1	35.5	35.9	40.0	**42.7**	**48.5**	40.4	28.8	38.7
Phi-4	35.2	29.3	37.0	16.5	30.5	27.1	35.0	30.5	30.3	33.8	30.5	32.3
Llama-3.1-8B-Instruct	20.8	23.2	27.4	19.8	23.7	18.8	23.3	25.6	16.2	27.3	33.9	23.7
Llama-3.1-70B-Instruct	30.0	29.3	34.3	29.8	30.1	28.2	33.3	28.1	29.3	32.0	28.8	30.9
Qwen-2.5-7B-Instruct	15.6	25.3	20.6	21.5	18.7	14.7	25.0	17.1	20.2	22.2	11.9	19.3
Qwen-2.5-32B-Instruct	27.1	24.2	31.9	22.3	18.7	25.9	23.3	23.2	27.3	29.0	32.2	26.9
Gemma-3-12B-it	12.4	25.3	21.9	18.2	16.3	15.9	25.0	23.2	2.0	19.2	15.3	17.4
Deepseek-R1	**47.4**	**46.5**	46.9	40.5	40.9	41.8	**50.0**	34.2	43.4	**47.0**	42.4	44.8
o3-mini-2025-01-31	36.6	37.4	44.5	33.1	37.4	35.9	31.7	40.4	39.5	39.0	38.8	38.8
o4-mini-2025-04-16	45.2	**46.5**	**51.4**	**41.3**	**41.9**	**42.9**	38.3	**46.5**	**45.8**	40.7	**45.4**	**45.4**
OpenAI-o1-mini	36.2	33.3	34.9	25.6	30.1	31.2	30.0	35.4	32.3	34.1	33.9	33.5
Qwen-3-235b-a22b	40.2	31.3	44.2	30.6	38.9	36.5	43.3	31.7	34.3	27.3	44.1	36.7
Llama-4-maverick	46.2	40.4	46.9	36.4	**48.3**	40.6	40.0	39.0	44.4	44.6	42.4	44.4
R1-Distill-Qwen-32B	34.0	36.4	38.4	28.1	36.0	31.2	26.7	29.3	26.3	33.5	35.6	33.9
Medical LLMs												
Llama-3-8B-UltraMedical	13.4	18.3	19.0	14.9	19.8	15.9	15.0	19.5	14.1	19.2	17.8	17.1
Llama-3-70B-UltraMedical	28.5	33.0	34.2	31.7	33.6	31.8	36.7	34.2	31.3	32.0	31.4	31.1
Llama-3-OpenBioLLM-8B	17.1	23.2	24.7	19.8	23.7	20.6	26.7	23.2	20.2	23.7	23.7	22.8
Llama-3-OpenBioLLM-70B	38.0	33.3	37.0	33.1	32.0	38.8	36.7	**40.2**	37.4	38.9	39.0	37.2
HuatuoGPT-o1-7B	24.3	25.3	27.7	21.5	26.6	21.2	26.7	23.2	26.3	34.4	20.3	26.5
HuatuoGPT-o1-70B	**39.0**	**43.4**	**42.8**	**39.7**	**38.9**	**42.4**	**46.7**	39.0	**40.4**	**43.7**	**42.4**	**39.2**

Table 3: Performance (accuracy) of various models across medical specialties on ClinBench$_{\text{Text}}$. **Bold** indicates the best performance, and underlined indicates the second best. Specialty abbreviations: GH (Gastroenterology and Hematology), Surg (Surgery), Neuro (Neurosciences), Oph/ENT (Ophthalmology and ENT), DI (Dermatology and Immunology), Resp (Respiratory and Thoracic Medicine), Dent (Dentistry), OG (Obstetrics and Gynecology), Ortho (Orthopedics), Cardio (Cardiovascular and Internal Medicine), Ped (Pediatrics).

Model	GH	Surg	Neuro	Oph/ENT	DI	Resp	Dent	OG	Ortho	Cardio	Ped	Avg
Doubao-1.5-Vision-Pro-32k	29.2	34.0	34.0	**33.3**	33.7	28.2	20.7	34.2	33.7	36.0	28.8	32.4
GPT-4o	**35.7**	**37.1**	38.9	32.5	**39.6**	**38.8**	**37.9**	**41.5**	**41.8**	37.2	33.9	**37.8**
Gemini-2.0-Flash	27.7	34.0	39.6	28.3	28.7	31.2	27.6	31.7	29.6	33.5	28.8	31.6
Gemma-3-27B-It	22.7	27.8	28.5	20.8	24.8	24.7	19.0	25.6	29.6	26.6	27.1	25.3
Internvl3-14B-It	18.0	33.3	**50.0**	28.6	20.0	23.1	20.0	30.0	26.7	**41.9**	**40.0**	30.2
Llama-3.2-11B-Vision-It	20.8	33.3	25.8	21.4	24.0	22.4	26.3	31.8	12.2	21.0	20.0	23.3
Qwen2.5-VL-72B-It	24.6	35.3	29.7	19.2	29.0	25.5	18.2	27.3	34.8	30.9	22.2	27.8
Human Experts	52.0	28.0	24.0	22.0	-	48.0	12.0	52.0	16.0	60.0	54.0	33.5

Table 4: Performance of various models across medical specialties on ClinBench$_{\text{MM}}$. **Bold** indicates the best performance, and underlined indicates the second best. Green indicates questions assessed by human experts within their own specialty, while red denotes questions out of their specialty.

4.3 Main Results

Tables 3 and 4 present the main results on ClinBench$_{\text{Text}}$ and ClinBench$_{\text{MM}}$, respectively.

Comparison of LLMs. (1) Overall, even the most advanced LLMs, such as Deepseek-R1 and OpenAI-o4-mini, achieve no more than 45% accuracy, indicating relatively poor performance on ClinBench. This highlights that ClinBench is a challenging medical benchmark, posing significant challenges for state-of-the-art LLMs. (2) Generally, large reasoning models demonstrate better performance compared to non-reasoning models. For instance, Deepseek-R1 and R1-Distill-Qwen-32B show some improvement over Deepseek-V3 and Qwen2.5-32B-Instruct. This improvement indicates the benefits of test-time scaling in complex clinical scenarios.

Comparison of LMMs. (1) Consistent with the results observed on ClinBench$_{\text{Text}}$, current LMMs demonstrate relatively low performance on ClinBench$_{\text{MM}}$. (2) On the other hand, LMMs achieve approximately 30 points (which is close to the average performance on ClinBench$_{\text{Text}}$.) even though the key information of diagnostic images is removed in the question stem. This indicates that LMMs are able to effectively utilize medical image information, which contributes to the resolution of complex clinical problems.

4.4 LLM v.s. Human Experts

In this section, we aim to assess whether the capabilities of LLMs in complex medical scenarios, such as diagnostic reasoning, have reached the level of human experts. To this end, we compare the performance of LLMs with that of human experts on the ClinBench$_{\text{MM}}$ dataset.

Setting. We invite experienced physicians from 5 different medical specialties, all of whom hold at least attending-level positions and have extensive clinical diagnostic experience. The detailed information is shown in the Appendix C. For the evaluation, we randomly select 50 questions from each medical specialty, totaling 500 questions, and assign each expert questions from two specialties: **one within their own area of expertise and another outside of their specialization.** During the answering process, experts are permitted to consult relevant medical literature and textbooks; however, the use of AI-assisted tools is strictly prohibited.

Medical Insights. From Table 3, we have the following observations: Human experts significantly outperform the strongest current LLMs, Deepseek-R1 and OpenAI-o4-mini, within their own specialized fields. On the other hand, experts perform poorly on questions outside their areas of expertise, whereas LLMs demonstrate relatively stable performance across all medical specialties, highlighting their stronger generalization capabilities.

4.5 Rare Disease Track

Rare diseases have long posed significant challenges to the medical community (Schieppati et al., 2008; Stoller, 2018), primarily due to limited clinical data, insufficient diagnostic knowledge, and a lack of effective treatments. In this work, we include a dedicated rare-disease subset within our dataset. Medical students carefully select 79 rare-disease cases from various medical specialties, strictly adhering to internationally recognized rare-disease catalogs[5]. This rare disease subset provides a specialized evaluation track for LLMs, which is beneficial for advancing LLMs to overcome the challenges of rare diseases in human medicine.

As shown in Table 5, we observe that the performance of LLMs on rare diseases is significantly lower than on non-rare diseases, highlighting the challenge that rare diseases pose to current LLMs. Additionally, medical LLMs such as Llama-3-8B-UltraMedical and HuatuoGPT-o1-7B exhibit relatively smaller performance gaps between rare and non-rare diseases. This may be attributed to their training on more medical texts, including materials related to rare diseases, enabling them to achieve better diagnostic capabilities in rare diseases.

Model	**Rare Acc**	**Non-Rare Acc**
GPT-4o	$29.1_{-9.0}$	38.1
Deepseek-R1	$36.3_{-8.5}$	44.8
Qwen2.5-7B-Instruct	$10.1_{-9.6}$	19.7
Llama-3-8B-UltraMedical	$16.5_{-6.6}$	23.1
HuatuoGPT-o1-7B	$21.5_{-5.2}$	26.7

Table 5: Comparison of model accuracy (%) for rare and non-rare diseases on ClinBench.

4.6 Data Leakage Analysis

To evaluate the potential risk of data leakage in the **ClinBench** benchmark, we follow work (Xu et al., 2024) by employing perplexity (PPL) and N-gram-based metrics (ROUGE-L and edit distance similarity) as evaluation criteria. Specifically, we concatenate the original question with a prompt such as “Answer:” as input and calculate the model’s perplexity on the generated output. Additionally, to assess the similarity between the model’s generated rationale and the reference explanations we collected, we compute both ROUGE-L scores and edit distance similarity.

We evaluate several models, including GPT-4o, LLaMA-3.1-70B-Instruct, Qwen2.5-72B-Instruct. Our analysis finds no evidence of data leakage. This result can be attributed to two main factors: (1) the questions in ClinBench are derived from professional medical case reports that have not been included in the training datasets of these models; (2) even if similar questions exist in training corpora, the inherent complexity and rich clinical context of these questions make it difficult for models to memorize or reproduce accurate answers solely based on prior exposure. Therefore, these observations support the conclusion that ClinBench poses minimal risk of data leakage, ensuring the validity and robustness this benchmark.

4.7 Error Analysis

In this section, we analyze the error cases of on ClinBench. Specifically, we choose the Deepseek-V3 and Deepseek-R1 models and investigate the

[5] https://www.who.int/standards/classifications/frequently-asked-questions/rare-diseases

Models	PPL ↑	Rouge-L ↓	EDS ↓
GPT-4o	1.18E+120	0.1712	0.2391
Qwen2.5-72B-Instruct	1.12E+115	0.1794	0.2493
LLaMA-3.1-70B-Instruct	9.73E+146	0.1597	0.2285

Table 6: Data leakage analysis results on different models. PPL denotes Perplexity, and EDS stands for Edit Distance Similarity.

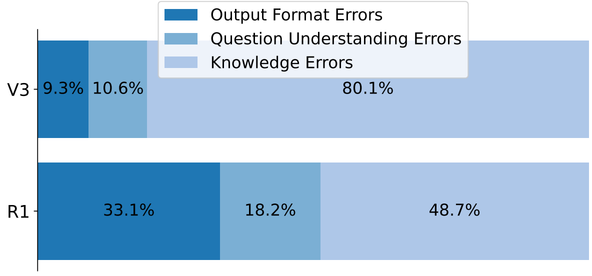


Figure 5: Error Analysis. We conduct an error analysis of Deepseek-V3 and Deepseek-R1 on ClinBench$_{\text{Text}}$.

reasons behind the models' incorrect answers by categorizing these errors into several types. The error types considered are as follows: (1) **Output Format Errors**: issues such as the model's failure to follow instructions; (2) **Question Understanding Errors**: cases where the answer reflects a misunderstanding of the question; (3) **Knowledge Errors**: the model understands the question correctly but provides an incorrect answer, for example, due to a lack of medical knowledge.

We then use GPT-4o to categorize these errors and compute the proportion of each error type. As shown in Figure 5, Deepseek-R1 exhibits a higher proportion of *Output Format Errors* compared to Deepseek-V3. We assume that this implies a limitation in the instruction-following ability of reasoning models. During evaluation, we frequently observe that the responses of reasoning LLMs do not adhere well to the given instructions, which results in many answers failing to be extracted.

5 Related Work

Medical LLMs. The success of LLMs has sparked interest in creating medical-specific models, leading to the emergence of numerous powerful medical LLMs (Nori et al., 2023; Saab et al., 2024; Li et al., 2024). For example, UltraMedical collections (Zhang et al., 2024) refine LlaMA-3 models with premium datasets, achieving top benchmark performance and advancing online preference learning. BioMistral (Labrak et al., 2024), an open-source model pre-trained on PubMed Central, excels in English medical QA tasks. HuatuoGPT series (Chen et al., 2024a,b; Zhang et al., 2023a), trained on high-quality medical data, significantly enhances automated capabilities in diagnosis, triage, and medical imaging, providing valuable support for clinical decision-making and patient care. Building on this trend, recent models such as Baichuan-M1 (Wang et al., 2025) and HealthGPT (Lin et al., 2025) further advance the field by improving medical reasoning, multimodal understanding, and have demonstrated strong empirical performance across a range of medical benchmarks.

Medical Benchmarks. With the advancement of medical LLMs, corresponding benchmarks have also evolved. Early datasets such as MedQA (Jin et al., 2021) and MedMCQA (Pal et al., 2022) focus on multiple-choice questions from USMLE and Indian medical exams, assessing models' factual knowledge across various medical domains. PubMedQA (Jin et al., 2019) emphasizes reasoning over biomedical research abstracts. On the other hand, the emergence of MMLU-Pro (Wang et al., 2024) and GPQA (Frantar et al., 2022) benchmarks provides a more effective evaluation of long-chain reasoning models like OpenAI-o1. Additionally, MedXpert-QA (Zuo et al., 2025) introduces expert-level questions derived from advanced medical exams, significantly increasing benchmark difficulty. However, these benchmarks remain predominantly exam-oriented and fail to capture the complexity of real-world clinical scenarios. In this paper, we focus on realistic medical scenarios by constructing a benchmark based on real-world clinical case questions.

6 Conclusion

In this paper, we introduce ClinBench, a medical benchmark specifically designed to simulate real clinical scenarios. This challenging benchmark originates from authoritative medical cases and incorporates detailed patient information and clinical findings, offering a more realistic assessment of LLMs' medical reasoning. Our comparative analysis of over 20 LLMs against medical experts demonstrates the continued strength of human specialists within their domains, while also highlighting the impressive ability of LLMs to generalize across a wider range of medical knowledge, suggesting their potential to complement and enhance clinical expertise.

Limitation

Our benchmark currently focuses exclusively on clinical medical diagnosis scenarios, with all data sourced solely from patient case records. While diagnosis represents one of the most critical and challenging tasks within the medical domain, it is important to acknowledge that other scenarios also play vital roles. For example, medical treatment planning, patient monitoring, and healthcare management involve complex decision-making processes that require integration of diverse data types such as longitudinal health records, medical imaging, and real-time sensor data. Furthermore, public health surveillance and preventive care demand models capable of handling population-level data and early risk detection. Therefore, although our benchmark serves as a crucial step towards evaluating AI capabilities in diagnosis, expanding it to encompass these additional healthcare domains is essential for broader applicability and impact.

References

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, and 1 others. 2024. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*.

Junying Chen, Zhenyang Cai, Ke Ji, Xidong Wang, Wanlong Liu, Rongsheng Wang, Jianye Hou, and Benyou Wang. 2024a. Huatuogpt-o1, towards medical complex reasoning with llms. *arXiv preprint arXiv:2412.18925*.

Junying Chen, Chi Gui, Ruyi Ouyang, Anningzhe Gao, Shunian Chen, Guiming Hardy Chen, Xidong Wang, Ruifei Zhang, Zhenyang Cai, Ke Ji, and 1 others. 2024b. Huatuogpt-vision, towards injecting medical visual knowledge into multimodal llms at scale. *arXiv preprint arXiv:2406.19280*.

Jan Clusmann, Fiona R Kolbinger, Hannah Sophie Muti, Zunamys I Carrero, Jan-Niklas Eckardt, Narmin Ghaffari Laleh, Chiara Maria Lavinia Löffler, Sophie-Caroline Schwarzkopf, Michaela Unger, Gregory P Veldhuizen, and 1 others. 2023. The future landscape of large language models in medicine. *Communications medicine*, 3(1):141.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. 2022. Gptq: Accurate post-training quantization for generative pre-trained transformers. *arXiv preprint arXiv:2210.17323*.

Google. 2024. Gemini 2.0 flash. https://cloud.google.com/vertex-ai/generative-ai/docs/models/gemini/2-0-flash.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

Yutao Hu, Tianbin Li, Quanfeng Lu, Wenqi Shao, Junjun He, Yu Qiao, and Ping Luo. 2024. Omnimedvqa: A new large-scale comprehensive evaluation benchmark for medical lvlm. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22170–22183.

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

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

Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. 2024. Biomistral: A collection of open-source pretrained large language models for medical domains. *arXiv preprint arXiv:2402.10373*.

Junkai Li, Siyu Wang, Meng Zhang, Weitao Li, Yunghwei Lai, Xinhui Kang, Weizhi Ma, and Yang Liu. 2024. Agent hospital: A simulacrum of hospital with evolvable medical agents. *arXiv preprint arXiv:2405.02957*.

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

Tianwei Lin, Wenqiao Zhang, Sijing Li, Yuqian Yuan, Binhe Yu, Haoyuan Li, Wanggui He, Hao Jiang, Mengze Li, Xiaohui Song, and 1 others. 2025. Healthgpt: A medical large vision-language model for unifying comprehension and generation via heterogeneous knowledge adaptation. *arXiv preprint arXiv:2502.09838*.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, and 1 others. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Daniel McDuff, Mike Schaekermann, Tao Tu, Anil Palepu, Amy Wang, Jake Garrison, Karan Singhal, Yash Sharma, Shekoofeh Azizi, Kavita Kulkarni, and 1 others. 2025. Towards accurate differential diagnosis with large language models. *Nature*, pages 1–7.

Harsha Nori, Yin Tat Lee, Sheng Zhang, Dean Carignan, Richard Edgar, Nicolo Fusi, Nicholas King, Jonathan Larson, Yuanzhi Li, Weishung Liu, and 1 others. 2023. Can generalist foundation models outcompete special-purpose tuning? case study in medicine. *Medicine*, 84(88.3):77–3.

OpenAI. 2024. Gpt-4o. https://openai.com/index/hello-gpt-4o/.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on health, inference, and learning*, pages 248–260. PMLR.

Malaikannan Sankarasubbu Ankit Pal and Malaikannan Sankarasubbu. 2024. Openbiollms: Advancing open-source large language models for healthcare and life sciences. *Hugging Face repository*.

Khaled Saab, Tao Tu, Wei-Hung Weng, Ryutaro Tanno, David Stutz, Ellery Wulczyn, Fan Zhang, Tim Strother, Chunjong Park, Elahe Vedadi, and 1 others. 2024. Capabilities of gemini models in medicine. *arXiv preprint arXiv:2404.18416*.

Arrigo Schieppati, Jan-Inge Henter, Erica Daina, and Anita Aperia. 2008. Why rare diseases are an important medical and social issue. *The Lancet*, 371(9629):2039–2041.

James K Stoller. 2018. The challenge of rare diseases. *Chest*, 153(6):1309–1314.

Xiangru Tang, Daniel Shao, Jiwoong Sohn, Jiapeng Chen, Jiayi Zhang, Jinyu Xiang, Fang Wu, Yilun Zhao, Chenglin Wu, Wenqi Shi, and 1 others. 2025. Medagentsbench: Benchmarking thinking models and agent frameworks for complex medical reasoning. *arXiv preprint arXiv:2503.07459*.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, and 1 others. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Qwen Team. 2025. Qwen2.5 models. https://huggingface.co/Qwen.

Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. 2023. Large language models in medicine. *Nature medicine*, 29(8):1930–1940.

Bingning Wang, Haizhou Zhao, Huozhi Zhou, Liang Song, Mingyu Xu, Wei Cheng, Xiangrong Zeng, Yupeng Zhang, Yuqi Huo, Zecheng Wang, and 1 others. 2025. Baichuan-m1: Pushing the medical capability of large language models. *arXiv preprint arXiv:2502.12671*.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, and 1 others. 2024. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *arXiv preprint arXiv:2406.01574*.

Ruijie Xu, Zengzhi Wang, Run-Ze Fan, and Pengfei Liu. 2024. Benchmarking benchmark leakage in large language models. *arXiv preprint arXiv:2404.18824*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Jin Ye, Guoan Wang, Yanjun Li, Zhongying Deng, Wei Li, Tianbin Li, Haodong Duan, Ziyan Huang, Yanzhou Su, Benyou Wang, and 1 others. 2024. Gmai-mmbench: A comprehensive multimodal evaluation benchmark towards general medical ai. *Advances in Neural Information Processing Systems*, 37:94327–94427.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, and 1 others. 2024. 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*, pages 9556–9567.

Hongbo Zhang, Junying Chen, Feng Jiang, Fei Yu, Zhihong Chen, Jianquan Li, Guiming Chen, Xiangbo Wu, Zhiyi Zhang, Qingying Xiao, and 1 others. 2023a. Huatuogpt, towards taming language model to be a doctor. *arXiv preprint arXiv:2305.15075*.

Kaiyan Zhang, Sihang Zeng, Ermo Hua, Ning Ding, Zhang-Ren Chen, Zhiyuan Ma, Haoxin Li, Ganqu Cui, Biqing Qi, Xuekai Zhu, and 1 others. 2024. Ultramedical: Building specialized generalists in biomedicine. *arXiv preprint arXiv:2406.03949*.

Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023b. Pmc-vqa: Visual instruction tuning for medical visual question answering. *arXiv preprint arXiv:2305.10415*.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Yuchen Duan, Hao Tian, Weijie Su, Jie Shao, and 1 others. 2025. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*.

Yuxin Zuo, Shang Qu, Yifei Li, Zhangren Chen, Xuekai Zhu, Ermo Hua, Kaiyan Zhang, Ning Ding, and Bowen Zhou. 2025. Medxpertqa: Benchmarking expert-level medical reasoning and understanding. *arXiv preprint arXiv:2501.18362*.

A Ethics Statement

All data used in this work were obtained exclusively from **freely and publicly accessible** sources. We have carefully curated the dataset by retaining only a small and representative subset of the original data. To ensure compliance with **U.S. fair use laws**, all questions underwent rephrasing, and answer options were shuffled to prevent any direct replication. Importantly, the dataset **does not contain any personal, sensitive, or identifiable information**, strictly avoiding any privacy violations or ethical concerns related to personal data. No content involving individual identities, medical records, or confidential information has been included. To mitigate the potential data leakage risks, we refrain from releasing the data sources and request that you **do not share any example of benchmark online**, whether in plain text, image, or any other format.

B Detailed Process for ClinBench Construction

In this section, we provide a detailed description of the construction process of the **ClinBench** benchmark.

B.1 Dataset Construction

We recruited 30 medical undergraduate students to assist in the dataset construction process. These students, all majoring in medicine, possess solid medical foundations. We provided them with comprehensive annotation guidelines, instructing them to carefully construct and verify question–answer pairs. Specifically, the guidelines detailed two core tasks: (1) constructing the ClinBench$_{\text{Text}}$ questions (see Table 7); and (2) constructing the multimodal ClinBench$_{\text{MM}}$ questions (see Table 8). Following these guidelines rigorously, the students successfully constructed a total of 2,014 high-quality questions.

B.2 Human Verification

Our human verification process involves two critical steps. First, medical undergraduate students carefully review the question stems and candidate options to ensure clarity and medical accuracy, following detailed guidelines as illustrated in Table 9. Subsequently, experienced practicing physicians (e.g., attending doctors) perform a sampling-based inspection of the reviewed questions. Any identified quality issues are systematically summarized and communicated back to the medical students for targeted revision, guided by the criteria provided in Table 10. This iterative feedback and revision mechanism ensures the overall reliability and high quality of the benchmark.

B.3 Data Specialty Distribution

We analyze the distribution of medical specialties covered by the ClinBench$_{\text{Text}}$ and ClinBench$_{\text{MM}}$ datasets, as illustrated in Figure 6. As shown, each medical specialty includes at least 50 questions, ensuring sufficient coverage for comprehensive evaluation.

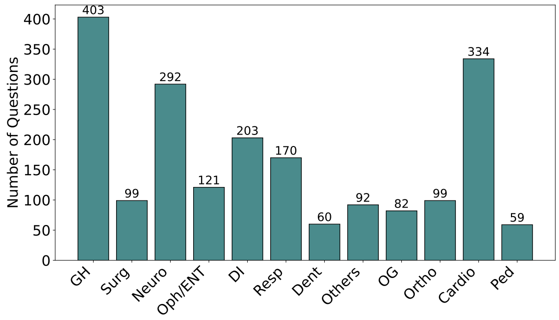


Figure 6: The pipeline for constructing ClinBench. Specialty abbreviations: GH (Gastroenterology and Hematology), Surg (Surgery), Neuro (Neurosciences), Oph/ENT (Ophthalmology and ENT), DI (Dermatology and Immunology), Resp (Respiratory and Thoracic Medicine), Dent (Dentistry), OG (Obstetrics and Gynecology), Ortho (Orthopedics), Cardio (Cardiovascular and Internal Medicine), Ped (Pediatrics).

C Expert Evaluation Details

Our experts are experienced attending physicians from Longgang People's Hospital in Shenzhen, Guangdong Province, China, specializing in Gastroenterology and Hematology, Respiratory and Thoracic Medicine, Obstetrics and Gynecology, Cardiovascular and Internal Medicine, and Pediatrics. For the evaluation, we randomly select 50 questions from each of these medical specialties, totaling 500 questions, and assign each expert questions from two specialties: **one within their own area of expertise and another outside of their specialization.** During the answering process, experts are permitted to consult relevant medical literature and textbooks; however, the use of AI-assisted tools is strictly prohibited.

For the evaluation, we developed an online assessment platform based on the streamlit framework, allowing physicians to answer questions following the provided guidelines. A screenshot of the evaluation platform is shown in Figure 9.

D Experimental Details

D.1 Models

Large Language Models: We evaluate a wide range of LLMs on ClinBench$_{\text{Text}}$. The general LLMs include GPT-4o-2024-11-20 (OpenAI, 2024), DeepSeek-V3 (Liu et al., 2024), Grok-3, Phi-4 (Abdin et al., 2024), LLaMA3.1-70B-Instruct, LLaMA3.1-8B-Instruct, Qwen2.5-32B-Instruct and Qwen2.5-7B-Instruct. Gemma-3-12B-it (Team et al., 2025) We also include general reasoning models such as DeepSeek-R1, DeepSeek-R1-Distill-Qwen-32B (Guo et al., 2025), Qwen-3-235b-a22b (Yang et al., 2025), OpenAI-o1, OpenAI-o3-mini, OpenAI-o3-mini [6], Llama-4-maverick[7]. In the medical domain, we assess medical-domain LLMs including OpenBioLLM-8B (Pal and Sankarasubbu, 2024), and UltraMedical-8B (Zhang et al., 2024), as well as the medical reasoning model HuatuoGPT-o1 (Chen et al., 2024a).

Large Multimodal Models (LMMs): We evaluate several large multimodal models on ClinBench$_{\text{MM}}$, including proprietary models such as GPT-4o-2024-11-20 (OpenAI, 2024), Gemini-2.0-Flash-001 (Google, 2024), as well as open-source models such as Doubao-1.5-Vision-Pro-32k, Internvl3-14B (Zhu et al., 2025), Qwen2.5-VL-72B (Team, 2025)and Llama-3.2-11B-Vision (Grattafiori et al., 2024)

We evaluate all models under a zero-shot prompt setting. For models such as llama-4-maverick, GPT-4o, Deepseek-V3, Gemini-2.5-Pro-Exp-03-25, Deepseek-R1, Grok-3, doubao-1-5-pro-32k-250115, o1-mini, o3-mini, o4-mini, and Qwen3-235B-a22b, we utilize the official APIs provided by the official. The remaining models are evaluated on a local setup consisting of 8 A800 80GB GPUs. During evaluation, we set the temperature to $t = 0.6$ and report the average results over three independent runs. The prompts used are illustrated in Figure 7 and Figure 8.

E Cases of ClinBench

We present two specific cases, one multimodal and one text-based, each including detailed questions, options, and the model's responses (from GPT-4o and Llama-3.2-11B-Vision) in Figure 10 and 11.

Evaluation Prompt for ClinBench$_{\text{Text}}$

`Question: {Question}`
`Options: {Options}`
Instruction: Given the following multiple-choice question and options, provide a concise answer based on accurate knowledge. Conclude your response with the correct option in the format: `The answer is [Option].`

Figure 7: Evaluation Prompt for ClinBench$_{\text{Text}}$

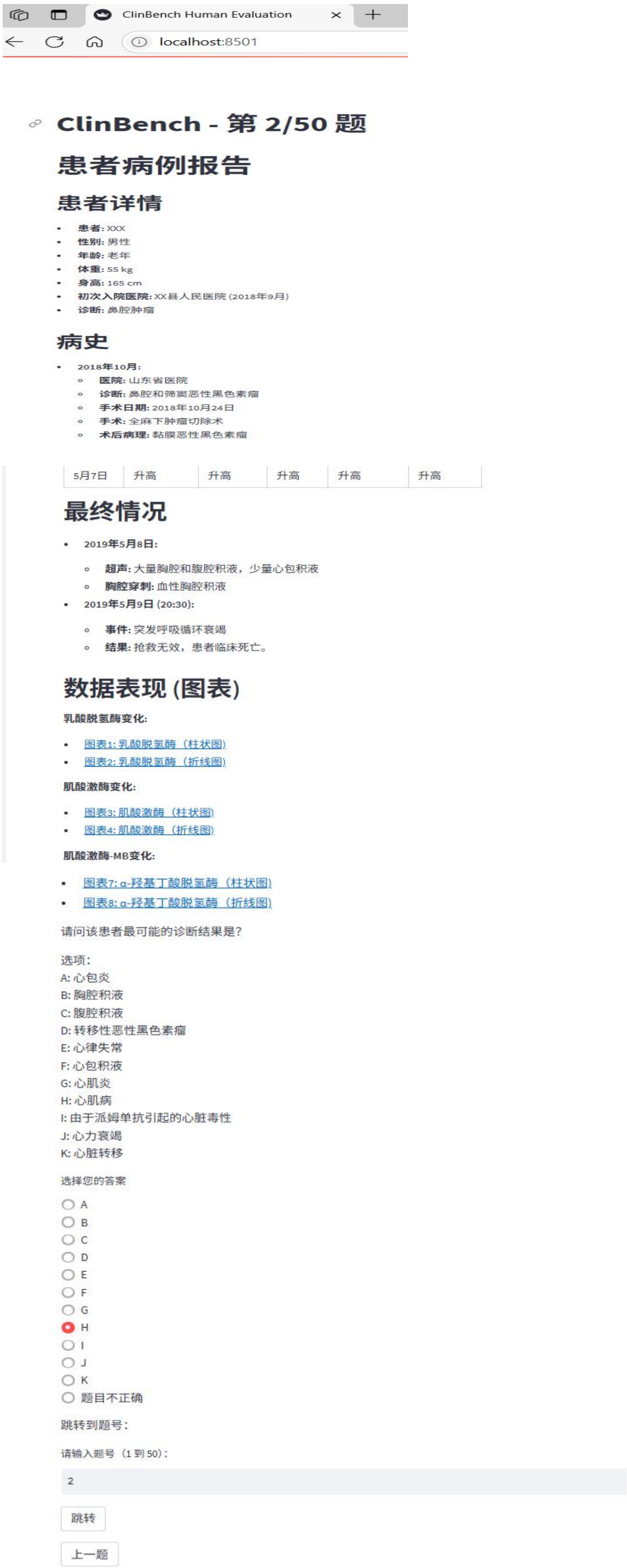
Evaluation Prompt for ClinBench$_{\text{MM}}$

`Question: {Question}`
`Images: {Images}`
`Options: {Options}`
Instruction: Given the following multiple-choice question, associated medical images, and options, analyze the images and question to provide a concise, accurate answer based on medical knowledge. Conclude your response with the correct option in the format: `The answer is [Option].`

Figure 8: Evaluation Prompt for ClinBench$_{\text{MM}}$

[6] https://openai.com/o1/

[7] https://ai.meta.com/blog/llama-4-multimodal-intelligence/



ClinBench - 第 2/50 题

患者病例报告

患者详情

- **患者:** XXX
- **性别:** 男性
- **年龄:** 老年
- **体重:** 55 kg
- **身高:** 165 cm
- **初次入院医院:** XX县人民医院 (2018年9月)
- **诊断:** 鼻腔肿瘤

病史

- **2018年10月:**
  - **医院:** 山东省医院
  - **诊断:** 鼻腔和筛窦恶性黑色素瘤
  - **手术日期:** 2018年10月24日
  - **手术:** 全麻下肿瘤切除术
  - **术后病理:** 黏膜恶性黑色素瘤

5月7日	升高	升高	升高	升高	升高

最终情况

- **2019年5月8日:**
  - **超声:** 大量胸腔和腹腔积液，少量心包积液
  - **胸腔穿刺:** 血性胸腔积液
- **2019年5月9日 (20:30):**
  - **事件:** 突发呼吸循环衰竭
  - **结果:** 抢救无效，患者临床死亡。

数据表现 (图表)

乳酸脱氢酶变化:

- 图表1: 乳酸脱氢酶（柱状图)
- 图表2: 乳酸脱氢酶（折线图)

肌酸激酶变化:

- 图表3: 肌酸激酶（柱状图)
- 图表4: 肌酸激酶（折线图)

肌酸激酶-MB变化:

- 图表7: α-羟基丁酸脱氢酶（柱状图)
- 图表8: α-羟基丁酸脱氢酶（折线图)

请问该患者最可能的诊断结果是?

选项:
A: 心包炎
B: 胸腔积液
C: 腹腔积液
D: 转移性恶性黑色素瘤
E: 心律失常
F: 心包积液
G: 心肌炎
H: 心肌病
I: 由于派姆单抗引起的心脏毒性
J: 心力衰竭
K: 心脏转移

选择您的答案

○ A
○ B
○ C
○ D
○ E
○ F
○ G
● H
○ I
○ J
○ K
○ 题目不正确

跳转到题号:

请输入题号（1 到 50）:

2

Figure 9: The online evulation of ClinBench.

Guideline for Constructing ClinBench$_{\text{Text}}$

Dear contributor, welcome to the ClinBench question construction task. You are now assigned to build high-quality samples for the ClinBench$_{\text{Text}}$ dataset. Please carefully follow the guidelines below.

1. Constructing the Question Stem

Objective: Use the original content of the medical journal to construct a question stem, while minimizing modification of the source content.

1.1 For Chinese-language journals, directly extract the *Patient Information* and *Medical Examination Results* sections.

1.2 For English-language journals, locate the *Patient Information* and *Medical Examination Results* and use a professional translation tool to translate them into fluent, accurate English. Discard journals that do not contain both sections.

1.3 If the case contains images (e.g., X-ray, MRI, CT), **remove the images**, but ensure the question stem includes a concise and accurate textual description of the image. If such a description cannot be provided, discard the journal.

1.4 The question stem should be self-contained, medically accurate, and unambiguous, providing enough context for diagnosis.

2. Constructing Candidate Answer Options

Objective: Create a set of multiple-choice options, including one correct answer and several plausible distractors.

2.1 **Identifying the Correct Answer:** Extract the final diagnosis from the Diagnosis/Conclusion section. The correct answer must be medically precise and supported by evidence from the journal.

2.2 **Selecting Distractors from the Discussion:** Extract suspected or differential diagnoses from the Discussion section to serve as medically relevant distractors.

2.3 **Generating Additional Distractors Using AI:** Use GPT-4o or a similar model to generate plausible but incorrect distractors. They should reflect realistic diagnostic confusion within the relevant medical context.

2.4 **Merging Options:** Use an AI model or human review to ensure:

(1) Exactly one correct answer is included.
(2) No ambiguity or semantic overlap among options.
(3) At least six options in total.

2.5 **Final Verification:** All options must be:

(1) Medically plausible and realistic.
(2) Mutually exclusive and clearly distinct.
(3) Professionally formatted, with consistent language and style.

3. Saving the Output

Please save each constructed question in a single-entry JSON format, as shown below:

```
{
  "id": "ClinBench-text-00001",
  "question_stem": "",
  "options": "",
  "answer": "",
 }
```

Table 7: Guideline for constructing ClinBench$_{\text{Text}}$ questions.

Guideline for Constructing ClinBench$_{\text{MM}}$

Dear contributor, welcome to the ClinBench multimodal question construction task. Your task is to build multimodal (text-image) questions based on previously constructed ClinBench$_{\text{Text}}$ questions. Please carefully follow the guidelines below.

1. Adding Medical Images

Objective: Enhance the existing ClinBench$_{\text{Text}}$ question stems by appropriately inserting medical images from the original case journals.

1.1 Identify suitable medical images (e.g., X-ray, MRI, CT scans, histological images) within the original case journal.

1.2 Insert the identified image at a relevant location within the question stem. Ensure the image directly supports and complements the textual content.

1.3 If no suitable medical images can be found or inserted into the text, discard the question.

2. Modifying the Question Stem

Objective: Adjust the existing text-based question stem by removing textual descriptions of medical examination results depicted in the inserted images.

2.1 Remove explicit textual descriptions of medical findings or examination results that are clearly illustrated by the medical image.

2.2 Retain only concise image references within the text (e.g., *"The patient's lung condition is shown in Figure 1."*).

2.3 Ensure the modified question stem remains clear, medically accurate, and contextually complete when combined with the inserted image.

3. Saving the Output

Please save each multimodal question as a JSON file following the single-entry format below:

```
{
  "id": "ClinBench-mm-00001",
  "question_stem": "",
  "options": "",
  "answer": "",
  "image": ""
}
```

Ensure each image filename corresponds clearly with the JSON entry and store images in a designated folder. Submit JSON files along with associated images.

Table 8: Guideline for constructing ClinBench$_{\text{MM}}$ multimodal questions.

Guideline for ClinBench Verification (Medical Students)

Dear contributor, welcome to the ClinBench question quality check task. Your role is critical to ensuring the high quality of constructed ClinBench$_{\text{Text}}$ and ClinBench$_{\text{MM}}$ questions. Please carefully follow the guidelines below.

1. Checking for Missing Information

Objective: Identify and correct significant information omissions in the constructed question stem compared to the original medical journal case.

1.1 Carefully review the constructed question stem against the original medical journal case to identify missing critical medical information necessary for understanding or answering the question.

1.2 Specifically, for ClinBench$_{\text{MM}}$ questions, pay special attention to information inadvertently omitted during the manual removal of image-related descriptions.

1.3 If you identify missing crucial information, supplement the question stem by adding the necessary content clearly and concisely, ensuring the revised question stem provides adequate context for accurate diagnosis.

2. Checking for Inappropriate Candidate Options

Objective: Evaluate and refine the candidate answer options to ensure clarity, distinctiveness, and medical accuracy.

2.1 Review candidate options carefully for any overlaps with the correct answer or ambiguous phrasing that may confuse test-takers or reduce question clarity.

2.2 Remove or revise any candidate options identified as overlapping with the correct answer or ambiguously phrased, ensuring the final set of options is clear, distinct, and unambiguous.

3. Saving the Corrected Output

After performing the above checks and corrections, save each question in a single-entry JSON format as follows:

```
{
  "id": "ClinBench-check-00001",
  "question_stem": "(Corrected and complete question stem)",
  "options": "(Verified and corrected candidate options)",
  "answer": "(Confirmed correct answer)",
  "image": "(Image filename if applicable, otherwise empty)"
}
```

Ensure each corrected JSON file is named appropriately and clearly linked with any associated images.

Table 9: Guideline for checking and verifying constructed ClinBench questions.

Guideline for Expert Review of ClinBench.

Dear expert reviewer, thank you for participating in the quality assurance of the ClinBench dataset. Your expertise is crucial for ensuring the accuracy and clinical relevance of our medical questions. Please carefully follow the guidelines outlined below.

Objective: Evaluate each question comprehensively from the following four perspectives: clarity of expression, clinical rationality, factual accuracy, and appropriateness of candidate options.
Evaluation Criteria:

- **Clarity of Expression:** Ensure the question stem and options are clearly phrased, understandable, and professionally articulated.
- **Factual Accuracy:** Verify that the question stem and correct answer are medically accurate, evidence-based, and free from factual errors.
- **Appropriateness of Candidate Options:** Check whether the options provided are medically plausible, clearly distinct, non-overlapping, and appropriately challenging.

Providing Feedback:
If any issues or concerns are identified during your evaluation, clearly outline your feedback as follows:

- Describe the specific issue identified (clarity, rationality, accuracy, or appropriateness).
- Provide detailed recommendations or corrections for resolving the identified issue.
- Highlight any relevant medical literature or sources that support your feedback if necessary.

Table 10: Guideline for expert review of ClinBench questions.

E.1 Example of GPT-4o

Text-Modal task

Question:

Patient Information

** General Information:

** Patient: Female, 38 years old

** Chief Complaint:Abnormal urine test for 2 months

** Symptoms: None reported (rash, purpura, joint pain, fever, Raynaud’s phenomenon, nausea, vomiting)

** Past Medical History:

- Hepatitis B carrier, no antiviral treatment,
- No history of hepatitis C
- Hypertension detected in March 2022, no antihypertensive therapy started

** Surgical History:** Cholecystectomy in March 2022

** Lifestyle:** Non-smoker, non-drinker.

** Family History: No familial hereditary disease or cancer.

** Physical Examination:

- Anemic appearance, no rashes, no enlarged lymph nodes
- Cardiovascular, respiratory, and abdominal examinations: Negative

** Electron Microscopy: - Features consistent with immune complex-mediated glomerulonephritis

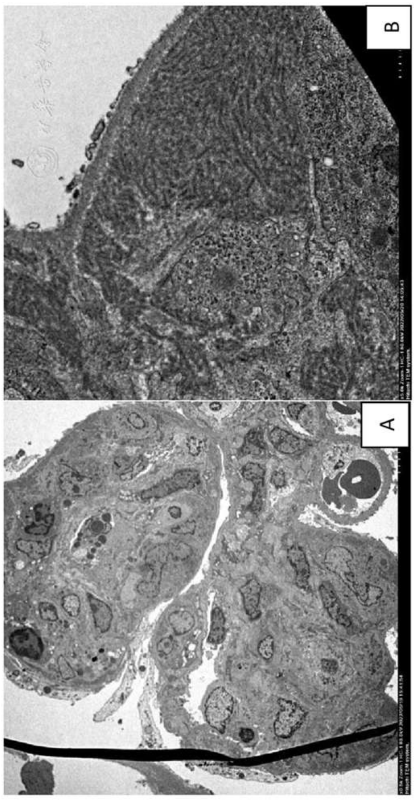


Figure 10: Electron Microscopy Image

** Genetic Test:

** MYD88 L265P Mutation: Negative (result may be influenced by low abnormal cell count)

Answer Choices: (A) Chronic Lymphocytic Leukemia (B) Essential Mixed Cryoglobulinemia (C) Heavy Chain Disease (D) Membranoproliferative Glomerulonephritis type I (E) Monoclonal Gammopathy of Renal Significance (F) Monoclonal Gammopathy of Undetermined Significance (G) Type I Cryoglobulinemic Glomerulonephritis (H) Type II Cryoglobulinemic Glomerulonephritis (I) Waldenstrom’s Macroglobulinemia

Correct Answer: (I)

Model Answer: (E)

Model Analysis:

To determine the correct answer, the patient presents with abnormal urine findings (proteinuria, hematuria, elevated ACR, and UTP) and has laboratory and biopsy results indicative of immune complex-mediated glomerulonephritis. Key findings include:

** Renal Biopsy: Immune complexes (IgM, C3, Kappa), suggest monoclonal paraprotein involvement.

** Blood and Urine Studies: Presence of monoclonal protein (IgM and Kappa light chains), increased free Kappa, positive SPE (Serum Protein Electrophoresis), and unchanged renal structure.

** Thus, the findings fulfill the criteria for **Monoclonal Gammopathy of Renal Significance (MGRS)**, wherein monoclonal proteins cause renal disease without meeting criteria for malignancy.

** The answer is E.

E.2 Example of Llama-vision

Multi-Modal task

Question:
** Clinical Case Report
** Patient Information
** Occupation: Raw material crusher in an industrial ceramics factory
** Case Details
The patient was admitted on October 16, 2018, due to recurrent cough and sputum production for over 10 years, which had worsened with chest tightness for 5 days.
** Initial Examination
** Symptoms: Lip cyanosis, reduced breath sounds in the left lung, bilateral basal lung moist rales; no pathological murmurs in valve auscultation areas.
** Supplementary Check
** Examination Findings: Palpable crepitus in the neck, left supraclavicular fossa, and left chest wall. Auscultation: Diffuse wheezing, numerous moist rales heard at lung bases.
** Chest CT Findings:
- Multiple gas shadows in mediastinum, supraclavicular fossa, and subcutaneous left chest wall.
- Small area of pneumothorax on the left side.
- Nutritional support and appropriate respiration exercises.

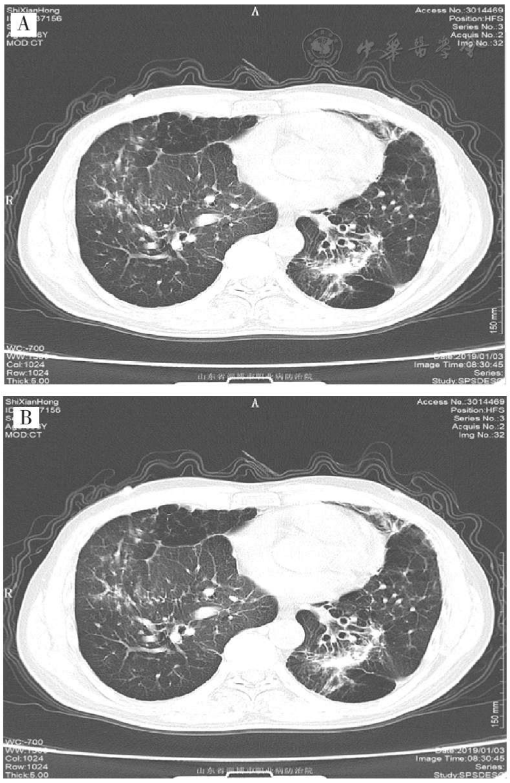


Figure 11: Chest CT Image

Answer Choices: (A) Chronic Obstructive Pulmonary Disease (B) Community-Acquired Pneumonia (C) Pneumomediastinum (D) Pneumothorax (E) Primary Spontaneous Pneumothorax (F) Pulmonary Fibrosis (G) Secondary Mediastinal Emphysema (H) Type II Cryoglobulinemic Glomerulonephritis (I) Waldenstrom's Macroglobulinemia
Correct Answer: (G)
Model Answer: (G)
Model Analysis:
** Step 1: Analyze the patient's symptoms and medical history.** The patient is a 56-year-old male with a history of recurrent cough and sputum production for over 10 years, which had worsened with chest tightness for 5 days. He works as a raw material crusher in an industrial ceramics factory.
** Step 2: Analyze the exacerbation symptoms and findings.** Aggravated frequent coughing with yellow purulent sputum, severe chest tightness, orthopnea, left-sided neck swelling and pain - Palpable crepitus in the neck, left supraclavicular fossa, and left chest wall - Diffuse wheezing, numerous moist rales heard at lung bases
** Step 3: Determine the most likely diagnosis based on the information provided.** Given the patient's occupation in an industrial ceramics factory, the presence of pneumothorax, pneumomediastinum, and subcutaneous emphysema, along with the history of recurrent respiratory issues, the most likely diagnosis is related to occupational exposure leading to lung damage.
** Answer: G. Secondary Mediastinal Emphysema.