Asclepius: A Spectrum Evaluation Benchmark for Medical Multi-Modal Large Language Models

Anonymous ACL submission

Abstract

The significant breakthroughs of Medical Multi-Modal Large Language Models (Med-MLLMs) renovate modern healthcare with robust information synthesis and medical decision support. However, these models are often evaluated on benchmarks that are unsuitable for the Med-MLLMs due to the intricate nature of the real-world diagnostic frameworks, which encompass diverse medical specialties and involve complex clinical decisions. Moreover, these benchmarks are susceptible to data leakage, since Med-MLLMs are trained on large assemblies of publicly available data. Thus, an isolated and clinically representative benchmark is highly desirable for credible Med-MLLMs evaluation. To this end, we introduce Asclepius, a novel Med-MLLM benchmark that rigorously and comprehensively assesses model capability in terms of: distinct medical specialties (cardiovascular, gastroenterology, etc.) and different diagnostic capacities (perception, disease analysis, etc.). Grounded in 3 proposed core principles, Asclepius ensures a comprehensive evaluation by encompassing 15 medical specialties, stratifying into 3 main categories and 8 sub-categories of clinical tasks, and exempting from train-validate contamination. We further provide an in-depth analysis of 6 Med-MLLMs and compare them with 5 human specialists, providing insights into their competencies and limitations in various medical contexts. Our work not only advances the understanding of Med-MLLMs' capabilities but also sets a precedent for future evaluations and the safe deployment of these models in clinical environments. We launch and maintain a leaderboard for community assessment of Med-MLLM capabilities.[1].

1 Introduction

The advent of Multi-Modal Large Language Models (MLLMs), such as GPT-4V (OpenAI, 2023), Gemini (Team et al., 2023), LLaVA (Liu et al., 2023b), and MiniGPT-4 (Zhu et al., 2023), represents a significant stride towards artificial general intelligence due to their exceptional proficiency in tackling intricate tasks. These advancements have not only expanded the capabilities of models in natural scenes but have also paved the way for specialized enhancement in healthcare, as seen with the emergence of recent Medical Multi-modal Large Language Models (Med-MLLMs) (Moor et al., 2023a; Liu et al., 2023a; Lee et al., 2023; Zhang et al., 2023c).

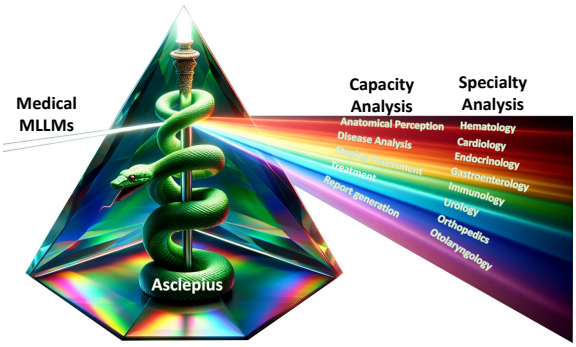


Figure 1: **Asclepius,** a spectrum evaluation benchmark for Med-MLLMs, analyzes models on the capacity dimension with 8 clinical tasks and the specialty dimension with 15 medical specialties.

Despite the promising advancements, the evaluation of these models relies predominantly on a limited set of samples (Wu et al., 2023a; Li et al., 2023b; Zhou et al., 2023; Moor et al., 2023a), which provides an incomplete picture of their capabilities. Current medical benchmarks, originally designed for traditional learning models, fall short in measuring the sophisticated capabilities of Med-MLLMs (Moor et al., 2023b; Li et al., 2023a). This misalignment highlights the necessity for comprehensive benchmarks that can systematically assess diverse perspectives of Med-MLLM. To this end, we propose a novel benchmark Asclepius for Med-MLLMs that is akin to analyzing the spectrum of

[1] https://asclepius-med.github.io/

light with a prism, as illustrated in Figure 1.

Developing such a benchmark is challenging, especially in the medical field, due to the variation in practical expertise across different domains (Wang et al., 2023). For instance, a cardiologist may require a referral to a gastroenterologist when encountering conditions outside their primary domain of expertise due to the distinct specializations within each field (Forrest et al., 2000; Forrest and Reid, 2001). Traditional model evaluation methods often overlook the need for specialty-specific assessments, resulting in limited applicability and reliability in specialized medical contexts. This drives our approach to consider the variation in expertise across different medical specialties and to assess Med-MLLMs' abilities in specialization-specific knowledge.

Furthermore, the complex clinical decision-making processes, which involve perception, cognition, and reasoning (Lyman and Kuderer, 2023; Liberatore and Nydick, 2008; Patel et al., 2002; Kassirer and Gorry, 1978), present a second challenge. Med-MLLMs should have a diverse set of capabilities, such as interpreting medical imagery and understanding pathophysiology, to mimic these processes. Current evaluation frameworks (Lau et al., 2018; Pelka et al., 2018; Liu et al., 2021), which often focus on tasks like general radiology VQA, fail to precisely capture clinical performance.

Thirdly, the evaluation data from the currently available dataset is easily cross-contaminated with training data for Med-MLLMs, thereby leading to potential data leakage. In an era where many models are trained on vast swaths of publicly available data, there is a significant risk of evaluating models on data they have previously encountered (Magar and Schwartz, 2022). The phenomenon of data leakage may artificially inflate the accuracy of these medical models, leading to misleadingly high performance (Deng et al., 2023). Thus, the benchmark of Med-MLLMs should consider this risk by avoiding the use of existing publicly available datasets as much as possible.

The philosophy to create Asclepius (1) *Multi-Specialty Coverage*: Our benchmark is meticulously designed to encompass a spectrum of 15 medical specialties. By systematically including questions from various specialties, such as cardiology, neurology, hematology, and endocrinology, the benchmark can evaluate the performance of Med-MLLMs in different medical domains. (2) *Multi-Dimensional Capacity*: Acknowledging the intricacies inherent in medical problem-solving, our benchmark is meticulously designed to evaluate a spectrum, divided into 3 main categories and 8 sub-categories. (3) *Blindness and Original*: Questions are sourced from contemporary educational materials, medical quizzes, and vision datasets not previously used in Med-MLLMs training, ensuring the original of our benchmark. Moreover, we develop a website that allows submission and server-side evaluation of results to ensure integrity.

In summary, our main contributions are summarized as follows:

- Systematically-Constructed Dataset: Our study introduces a meticulously crafted dataset designed to evaluate Med-MLLMs. This dataset encompasses a comprehensive range of 15 medical specialties, targeting 79 distinct body parts and organs. Furthermore, it is stratified into 3 main categories and 8 sub-categories, each corresponding to specific capacities within the medical domain.
- Comprehensive Benchmarking: Our study establishes a rigorous benchmark for the comprehensive assessment of 6 Med-MLLMs. In addition, 5 human doctors from varied specialties and levels of experience, ranging from junior to senior, answer the question to evaluate the human performance in this benchmark. This benchmark enables a direct comparison between Med-MLLMs and human specialists, providing valuable insights into the current state of AI in healthcare.
- Analysis and Observations: We provide several insights with corresponding suggestions (§5) based on the evaluation results, shedding light on their strengths and weaknesses for Med-MLLMs.

2 Related Work

2.1 Medical Multi-Modal Large Language Models (Med-MLLMs)

The field of medicine frequently engages with diverse data modalities, including but not limited to text, computed tomography (CT) scans, dermoscopy images, and histopathological slides. In order to effectively replicate the complex decision-making processes of healthcare professionals, Medical Multi-Modal Large Language Models (Med-MLLMs) have been developed. Initial research efforts in this area have focused on the fusion of text with single medical imaging modalities (Liu et al., 2023a; Lee et al., 2023; Zhou et al., 2023;

Table 1: **Comparison of Med-MLLMs' Datasets.** The Asclepius is categorized by both medical specialty and capability, encompassing 15 major specialties and 8 core competencies, and includes human specialists evaluation scores. B&O represents body parts and organs.

Name	Modality	B&O	Division		Original	Human Evaluation
			Specialty	Capacity		
ROCO (Pelka et al., 2018)	Radiology	/	✗	✗	✓	✗
VQA-RAD (Lau et al., 2018)	Radiology	3	✗	✗	✓	✗
SLAKE (Liu et al., 2021)	Radiology	5	✗	✗	✓	✗
PathVQA (He, 2021)	Pathology	/	✗	✗	✓	✗
MedMD (Wu et al., 2023b)	Radiology	17	✗	✗	✗	✗
PMC-VQA (Zhang et al., 2023c)	Multi-Modality	/	✗	✗	✗	✗
Asclepius	Multi-Modality	79	✓	✓	✓	✓

Thawkar et al., 2023; Wu et al., 2023b). These contributions laid the groundwork for subsequent advances in the field. Progressing further, recent studies have aimed to amalgamate a broader range of modalities (Belyaeva et al., 2023; Zhang et al., 2023a,c; Li et al., 2023a; Zhang et al., 2023b). In Asclepius, we thoroughly assess the capabilities of these models, particularly in terms of their performance in various specialties and capacities.

2.2 Benchmark for Med-MLLMs

In the rapidly evolving domain of Med-MLLMs, the development of benchmarks for evaluating these models is of paramount concern. Recent works (Zhang et al., 2023a; Belyaeva et al., 2023; Li et al., 2023a; Lu et al., 2023; Liu et al., 2023a) have endeavored to aggregate data from a variety of publicly available sources (Subramanian et al., 2020; Yang et al., 2023; Irvin et al., 2019; Johnson et al., 2019; He, 2021; Lau et al., 2018; Liu et al., 2021) to create larger and more comprehensive datasets. Subsequently, ChatGPT is employed to assist in filtering the aggregated data to ensure quality control. For instance, the MedMD (Wu et al., 2023b) collected data from existing visual-language medical datasets, such as MIMIC-CXR (Johnson et al., 2019) and PMC-OA (Lin et al., 2023), within the radiology domain. Similarly, the PMC-VQA Dataset (Zhang et al., 2023c) utilizes ChatGPT to create a Visual Question Answering (VQA) dataset based on image-text pairings from PMC-OA (Lin et al., 2023). However, all these datasets heavily rely on public benchmarks, which poses a risk of data leakage when applied to the evaluation of Med-MLLMs. This risk is compounded by the fact that many Med-MLLMs are trained on a vast array of publicly available data. Additionally, the dependency on entity extraction and ChatGPT could influence the precision and validity of the evaluations.

Different from previous works, in this paper, we propose a novel Med-MLLMs benchmark Asclepius, which is mainly built based on existing medical textbooks and medical image datasets. We aim to establish a new standard for Med-MLLM evaluation that upholds the integrity of the assessment and delivers accurate reflection of a model's true capabilities in clinical environment.

3 Asclepius Benchmark

Our Asclepius benchmark contains 3,232 original multi-modal questions, with a spectrum of 15 medical specialties and 8 capacities evaluation. This section will elaborate on the following organization of details: In Sec. 3.1 and Sec. 3.2, we discuss the design philosophy behind Asclepius and then present the specialty splitting and capacity splitting. In Sec. 3.3, we briefly introduce how we construct the Asclepius questions, and provide statistics of Asclepius.

3.1 Multi-Specialty Coverage

Medical education is characterized by significant disparities in knowledge across different specialties, necessitating rigorous training for medical students within their chosen fields (Ledford et al., 2022; Davis, 2009). Analogously, Med-MLLMs developed to support clinical decision-making should exhibit a comparable depth of knowledge in their respective specialties to be effective. To this end, we include a diverse array of specialties. For specialties concerning specific organs, we incorporate *Cardiology, Endocrinology, Obstetrics and Gynecology, Gastroenterology, Urology, Orthope-*

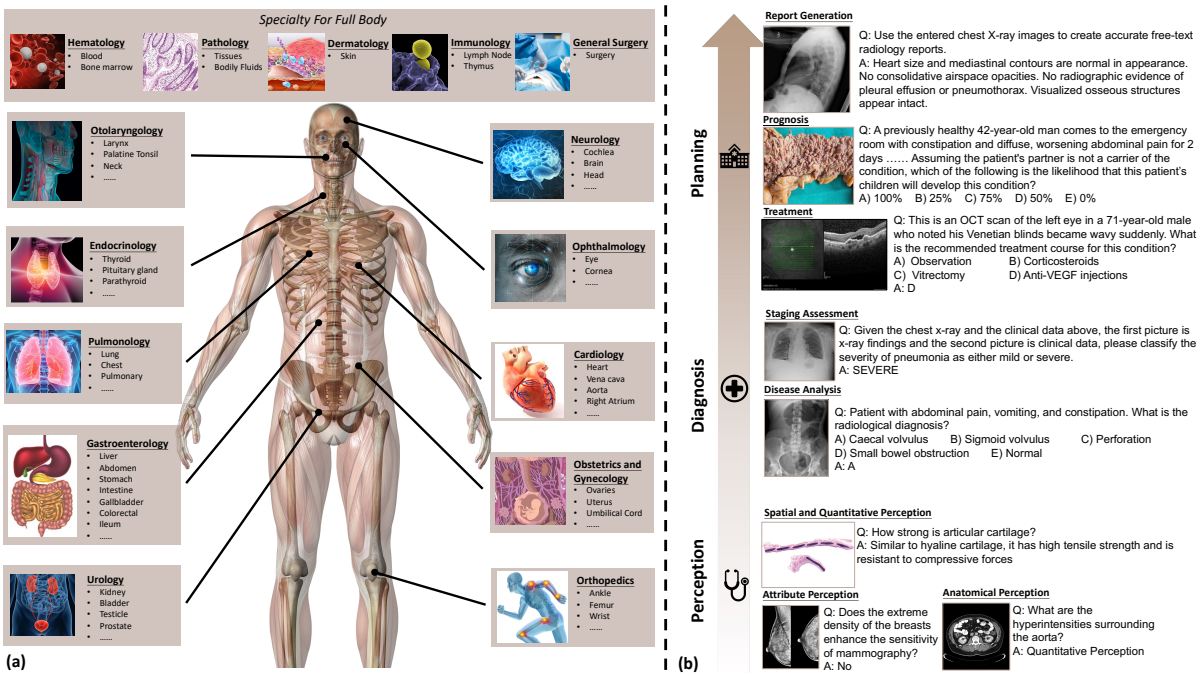


Figure 2: **Asclepius Overview.** (a) Involves 15 specialties and 79 body parts and organs in total, representing the critical component of the healthcare system. (b) Shows examples for 8 distinct capacities, offering a multifaceted evaluation of Med-MLLMs.

dics, Neurology, Otolaryngology, Pulmonology, and *Ophthalmology*. In addition, for specialties encompassing full-body considerations, we include *Hematology, Pathology, Dermatology, Immunology*, and *General Surgery*. This diversified inclusion ensures that the Med-MLLMs reflect not only technological innovation but also practical utility within the routine operations of clinical practice.

To provide a comprehensive view of the Med-MLLMs across these specialties, Figure 2 (a) offers an overview of the 15 specialties assessed in Asclepius. Each specialty represents a critical component of the healthcare system, addressing unique health concerns within its area of expertise. For additional information on the specialties included in this study, please refer to Appendix A.

3.2 Multi-Dimensional Capacity

The decision-making process in clinical practice is multifaceted and layered, encompassing a series of complex cognitive tasks (Liberatore and Nydick, 2008). The physician typically diagnoses the condition, assesses its stage, and then formulates a treatment plan and prognosis. Subsequently, these decisions and insights should be consolidated into a medical report. To mirror the decision-making intricacies in clinical settings, our benchmark incorporates three primary capacities for Med-MLLMs: *perception*, *diagnosis*, and *planning*. Furthermore, we have delineated secondary layer capacities to provide a more granular assessment. From *perception*, we derive (1) *anatomical perception*, (2) *attribute perception*, and (3) *spatial and quantitative perception*. From *diagnosis*, we extract (1) *disease analysis* and (2) *staging assessment*. Lastly, from *planning*, we have identified (1) *treatment*, (2) *prognosis*, and (3) *report generation*.

To illustrate the practical application of these capacities, Figure 2 (b) presents various case examples for each capacity within Asclepius. Asclepius currently encompasses 8 distinct sub-capacities, which offer a multifaceted evaluation of Med-MLLMs' performance in mimicking the decision-making process found in medical practice. Appendix C provides the detailed definitions of these sub-capacities.

3.3 Data Collection

To create a comprehensive benchmark with a multifaceted evaluation, our study needs to collect medical images and professional medical knowledge QA pairs, that can effectively test both the specialties and capabilities of Med-MLLMs. The Asclepius implements a tripartite strategy for the generation of these QA pairs.

The first approach within Asclepius constructs

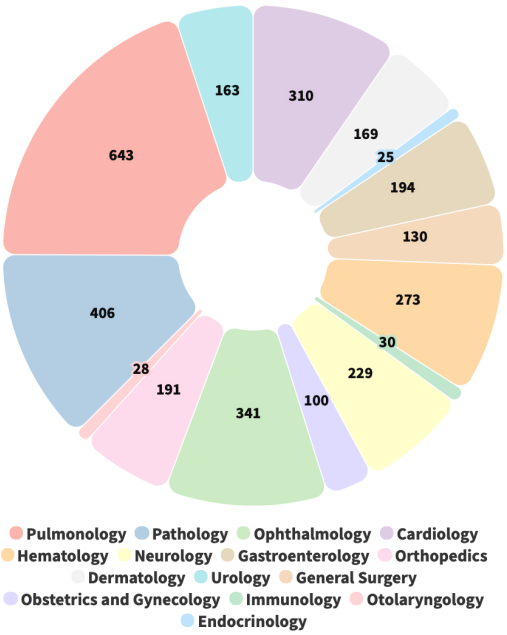


Figure 3: **Data Statistics for Specialty**. Currently, Asclepius incorporates 15 specialties with 3,232 multimodal questions.

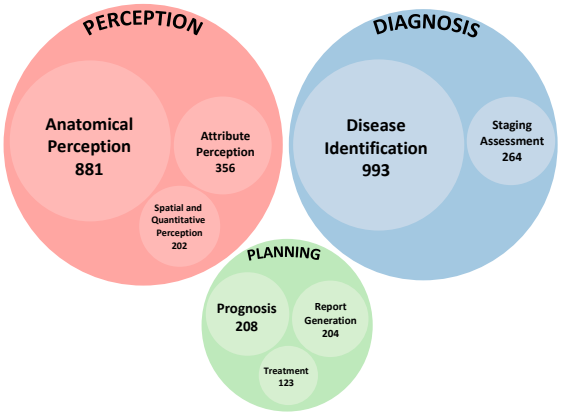


Figure 4: **Data Statistics for Capacities.** Asclepius includes two layers of capacity dimensions, which encompass 8 sub-capacities.

QA pairs from pre-existing medical image datasets. These datasets are restructured into a VQA format, with binary classification tasks formulated as yes/no questions and multi-class classification tasks presented as multiple-choice queries. Concurrently, the benchmark employs a second strategy, incorporating QA pairs with images derived from the United States Medical Licensing Examination (USMLE)[2]. The integration of this content ensures alignment with the rigorous standards required for medical licensure, establishing a high bar for the medical knowledge assessment of Med-MLLMs. The third prong of our strategy involves the inclusion of QA pairs derived from medical textbooks (Katzung et al., 2004; Pawlina and Ross, 2018; Murray et al., 2015; Barrett, 2010; Kumar et al., 2014; Snell, 2010; Sickles and D'Orsi, 2016), providing comprehensive coverage of various disciplines within the field of medicine and healthcare. Notably, the Breast Imaging Reporting & Data System (BI-RADS) atlas[3] (Sickles and D'Orsi, 2016) provides a rich variety of medical imaging types, ranging from peripheral blood cell imagery to complex radiological examinations.

Statistics. The Asclepiuscomprises a total of 3,232 data samples, which span across 15 medical specialties and 8 distinct sub-capacities. The visualizations of these two statistics are exhibited in Figure 3 and 4. Moving forward, we will preserve a balanced distribution of questions that fulfill various evaluation dimensions.

Data Split. In order to maintain the integrity and the blindness of the evaluation benchmark for Med-MLLMs, we have partitioned Asclepius into development and test subsets. The development subset is entirely accessible to the public, including the ground truth answers for all incorporated questions. Conversely, the test subset is only partially disclosed, with data samples being publicly available while the ground truth answers are withheld. To ascertain the performance on the test subset, participants are required to submit their predictions to the Asclepius server, which ensures an unbiased assessment of the Med-MLLMs.

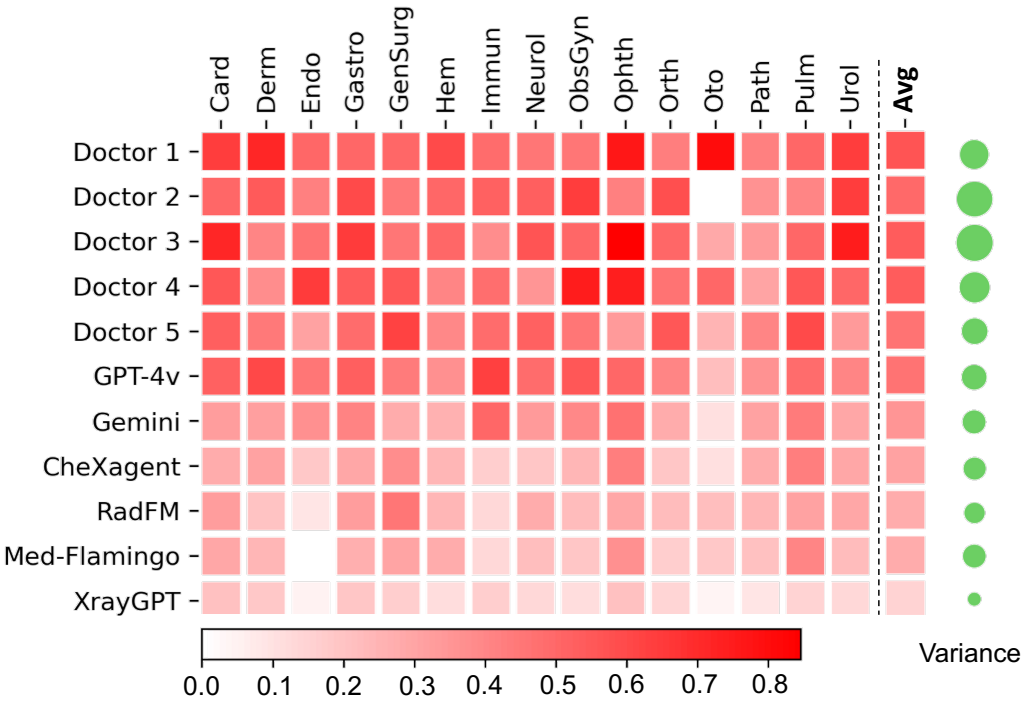
4 Experiments

4.1 Implementation

Model Evaluation. This benchmark focuses on two general MLLMs, namely GPT-4V (OpenAI, 2023) and Gemini (Team et al., 2023), alongside four specialized Med-MLLMs: CheXagent (Chen et al., 2024), RadFM (Wu et al., 2023c), Med-Flamingo (Moor et al., 2023c), and XrayGPT (Thawkar et al., 2023). The detailed prompt used for evaluation is: *You are a professional doctor. I will give you a question and one or two images. Please utilize the image given to answer the question as a medical expert would. You should only give the answer and no reason or other information. Question: <question> Image: <image>.* <ques-

[2]https://www.amboss.com/us/usmle

[3]https://www.acr.org/Clinical-Resources/Reporting-and-Data-Systems/Bi-Rads



Abbreviation: Card (Cardiology), Derm (Dermatology), Endo (Endocrinology), Gastro (Gastroenterology), GenSurg (General Surgery), Hem (Hematology), Immun (Immunology), Neurol (Neurology), ObsGyn (Obstetrics and Gynecology), Ophth (Ophthalmology), Orth (Orthopedics), Oto (Otolaryngology), Path (Pathology), Pulm (Pulmonology), Urol (Urology).

Figure 5: **The spectrum of Med-MLLMs in Specialties.** The green circle size represents the variance of the accuracy scores across all specialties. The larger circle means a larger variance. The darker color of the square indicates higher accuracy. The detailed digital values refer to Appendix Table 4.

tion> and <image> are the reserved places for question and image from the Multi-Modal QA data.

Human Study. To establish a benchmark for performance, the study also includes an evaluation of human specialists. Five clinical doctors are selected to participate, representing a mix of experience levels. The group includes senior doctors (Doctor 1, 3, and 4) and junior doctors (Doctor 2 and 5), each with expertise in distinct medical specialties. Specifically, Doctors 1 through 5 possess 4, 3, 4, 5, and 2 years of professional experience, respectively. The confidence score of these doctors for each specialty is shown in Appendix Table 3. These questions were presented to the clinical doctors, and their responses were recorded.

Evaluation Metrics. Asclepius includes a range of question types: multiple choice, yes/no, open-ended questions, and report generation tasks. We adopt accuracy as metric for multiple-choice questions and yes/no questions. Moreover, open-ended questions demand a more subtle assessment approach; here, the Exact Match metric is utilized to measure the precision of the textual responses due to its stability and rigor (Moor et al., 2023c). For the evaluation of report generation, the ROUGE-L scoring system is employed to determine the extent to which the models' generated texts align with the gold-standard reports (Wu et al., 2023c; Chen et al., 2024). Accuracy and Exact Match results can be utilized to calculate the accuracy score, and ROUGE-L scores are directly reported.

4.2 Results across Specialties

Our analysis reveals a broad spectrum of accuracy scores across the medical specialties, as depicted in Figure 5. GPT-4V generally has the highest performance among the Med-MLLMs, with its average accuracy closest to human doctors. Gemini has the second-best performance among the Med-MLLMs, with moderate accuracy, but still significantly lower than GPT-4V. CheXagent, RadFM, Med-Flamingo, and XrayGPT show lower performance across all specialties. When comparing between Med-MLLMs and human doctors, it is evident that human doctors generally outperform Med-MLLMs in all specialties. However, Med-MLLMs demonstrate a markedly lower variability in their performance across medical specialties when com-

Table 2: **The spectrum of Med-MLLMs in Capacity.** Avg* are the average accuracy of Anato, Attr, SpaQua, DisIde, Stag, Prog and Treat. Rep reports the ROUGE-L score. GPT-4V refuses to answer Rep question.

Model	Perception			Diagnosis		Planning			Avg*
	Anato	Attr	SpaQua	DisIde	Stag	Prog	Treat	Rep	
GPT-4V	0.323	**0.385**	**0.552**	0.649	**0.480**	**0.504**	**0.524**	N.A.	**0.462**
Gemini	0.285	0.292	0.342	**0.654**	0.342	0.496	0.323	0.082	0.354
CheXagent	0.238	0.253	0.321	0.524	0.252	0.451	0.315	**0.157**	0.309
RadFM	**0.344**	0.298	0.212	0.130	0.396	0.295	0.290	0.091	0.278
Med-Flamingo	0.270	0.256	0.217	0.587	0.272	0.398	0.145	0.133	0.279
XrayGPT	0.163	0.107	0.152	0.082	0.104	0.223	0.145	0.078	0.148

Abbreviation: Anato (Anatomical Perception), Attr (Attribute Perception), DisIde (Disease Identification), Prog (Prognosis), SpaQua (Spatial and Quantitative Perception), Stag (Staging Assessment), Treat (Treatment Planning), Rep (Report Generation).

pared to humans.

4.3 Results across Capacities

The spectrum of Med-MLLMs in capacity is shown in Table 2. GPT-4V leads with the highest average accuracy across the seven tasks with an average score of 46.2%. This indicates that GPT-4V is likely the most versatile and reliable model for a range of medical tasks. Notably, some Med-MLLMs perform better than GPT-4V in certain capacities. For example, RadFM has the highest score in Anatomical Perception with 34.4%, indicating that it might be the best at identifying anatomical structures. Moreover, both GPT-4V and Gemini perform equally well in Disease Identification with a score of 64.9% and 65.4%, respectively, indicating strong capabilities in diagnosing diseases. As for the report generation task, GPT-4V refused to generate the responses, which could be due to the model's guidelines or limitations in this specific task. CheXagent has the highest ROUGE-L score for Report Generation at 0.157, although this is still relatively low, suggesting room for improvement in how these models generate medical reports.

5 Discussion

From the above results, six key insights have been deduced as follows:

1) Significant variance exists in different specialties. Human doctors have strong performances in certain specialties, but weaker performances in others. As illustrated in Appendix Table 3, the confidence score assigned by Doctor3 is 5 for Ophthalmology compared to a score of 2 for Dermatology. Correspondingly, Table 4 in the Appendix indicates that Doctor3 achieved a diagnostic accuracy of 84.6% for Ophthalmology, which markedly contrasts with a reduced accuracy of 40.0% for Dermatology. These variations highlight the complexity and diversity inherent to each medical specialty. Given these differences, it is critical to establish a comprehensive Med-MLLMs benchmark to systematically evaluate the performance across various specialties.

2) Human doctors outperform Med-MLLMs. Human physicians surpass Med-MLLMs in diagnostic accuracy across all specialties. Even a junior doctor with an average accuracy of 46.5%, marginally exceeds the most proficient Med-MLLM, GPT-4V, which achieved an accuracy of 46.2%. This outcome suggests that, despite the advancements made in artificial intelligence, there is still a gap in diagnostic precision between human expertise and current Med-MLLMs. This underscores the need for continued development and specialization in the field of AI-driven medical diagnostics. On the other hand, the GPT-4V is comparable with junior human doctors, which indicates that Med-MLLMs like GPT-4V have the potential to complement the diagnostic process in practice. However, the current results indicate that the integration of such models into clinical workflows should be approached with caution, ensuring they serve as an adjunct to, rather than a replacement for, human clinical judgement.

3) Superiority of Generalist MLLMs Over Specialized Med-MLLMs. Results reveal that generalist models such as GPT-4V and Gemini outperform four specialized Med-MLLMs in a dual-spectrum evaluation. Notably, RadFM, despite being trained on 16 million multi-modal medical question-answer pairs, remains inferior to GPT-4V. This observation underscores the potential advantages of generalist MLLMs in synthesizing and applying a wide-ranging medical knowledge base.

4) Limited Long-range Instruction Capture. De-

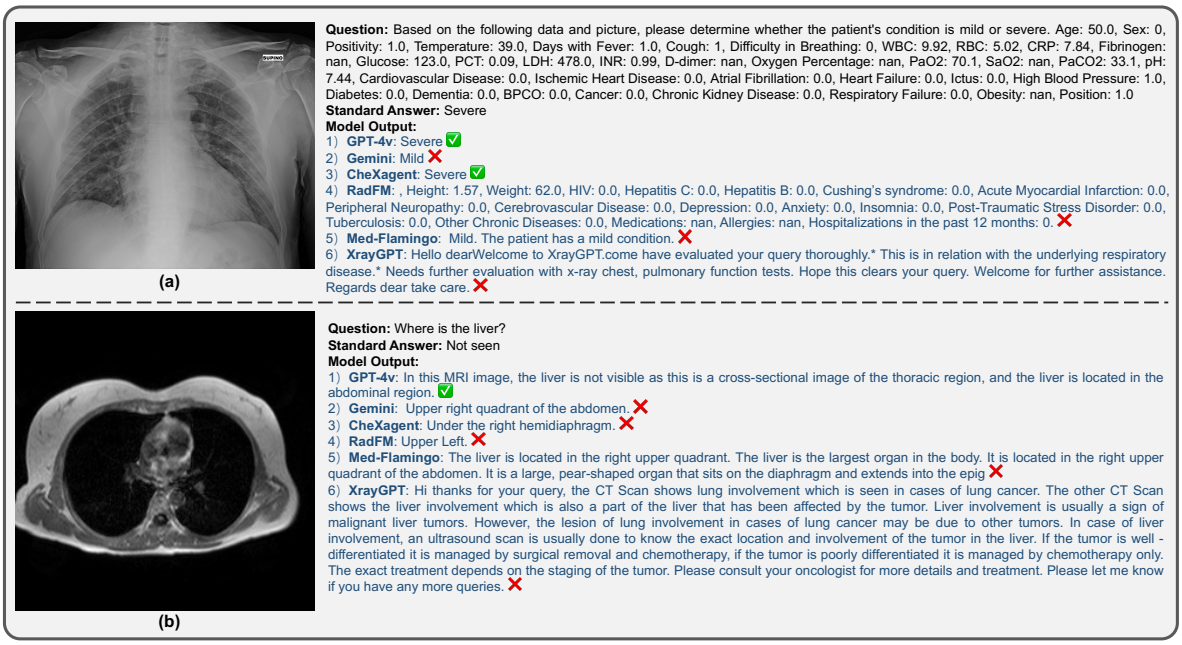


Figure 6: **Case Study for Common Problem Revealed in Evaluation**. (a) Case for Limited Instruction-following Capabilities. (b) Case for Failed Anatomical Perception.

spite the implementation of careful prompt engineering, certain Med-MLLMs exhibit a tendency to generate indirect responses. Figure 6(a) illustrates instances where, instead of answering the condition as prompted, models like RadFM and XrayGPT provide irrelevant information. We argue that an optimal Med-MLLM should adhere to such detailed instructions.

5) Limited Multi-Modality Fusion. Figure 6 (b) illustrates the case that only GPT-4V accurately incorporates image information into its response. In contrast, other Med-MLLMs simply restate the well-known fact that the liver is located in the upper right quadrant of the abdomen, neglecting to integrate the visual data presented. This pattern suggests a limited ability of most Med-MLLMs to synthesize image and text information, as they solely rely on textual prompts for their answers. Enhancing Med-MLLMs' multi-modal fusion capabilities emerges as a crucial and promising direction for future development.

6) Med-MLLMs offer the potential for integration of expansive and in-depth medical knowledge. In contrast to the wider variability observed among human physicians, Med-MLLMs exhibit more uniform performance, as depicted by the smaller green circles in Figure 5. This reduced variance signifies a standardized diagnostic capability of Med-MLLMs across different medical specialties. Such a pattern suggests that while human specialists are in-depth for particular domains, Med-MLLMs provide a more expansive knowledge across diverse medical fields, which could potentially be leveraged to augment clinical decision-making processes. This is particularly relevant for complex multi-system disorders like hypermobile Ehlers-Danlos syndrome (hEDS), where an interdisciplinary approach is paramount (Gensemer et al., 2021). The advent of Med-MLLMs makes the integration of expansive and in-depth medical knowledge feasible, offering the potential to address the multifaceted needs of patients.

6 Conclusion

In this paper, we introduce a comprehensive evaluation benchmark for Med-MLLMs, Asclepius, comprising 3,232 multi-modal questions, encompassing 15 medical specialties and 79 distinct body parts and organs, to analyze the spectrum in specialty and capacity. Additionally, we have developed a website that allows for the submission and server-side evaluation of predictive results to ensure integrity. 6 Med-MLLMs and 5 human doctors are evaluated in Asclepius. After analyzing the performance, we find that while the current Med-MLLMs have limitations, they possess the capability to supplement human clinical judgment, suggesting the possibility for integrated medical knowledge application that spans both breadth and depth.

Limitations

This paper has two primary limitations that offer avenues for future research:

- The current benchmark does not consider the long historical cases in patient records. Real-world clinical decision-making may consider the historical cases in patient records. In the future, we will expand the question set that can also measure the performance of Med-MLLMs in long sequences.

- The questions of Perception/Diagnosis/Planning are independent of each other currently, without coherence. Real clinical decision-making needs to be completed coherently from the front end to the back end. If one of them is wrong, the diagnosis will be not correct. In the future, the question sets will integrate long sequences of data in patient records and provide sequential disease questions for a patient.

References

Kim E Barrett. 2010. *Ganong's review of medical physiology*.

Anastasiya Belyaeva, Justin Cosentino, Farhad Hormozdiari, Krish Eswaran, Shravya Shetty, Greg Corrado, Andrew Carroll, Cory Y McLean, and Nicholas A Furlotte. 2023. Multimodal llms for health grounded in individual-specific data. In *Workshop on Machine Learning for Multimodal Healthcare Data*, pages 86–102. Springer.

Zhihong Chen, Maya Varma, Jean-Benoit Delbrouck, Magdalini Paschali, Louis Blankemeier, Dave Van Veen, Jeya Maria Jose Valanarasu, Alaa Youssef, Joseph Paul Cohen, Eduardo Pontes Reis, et al. 2024. Chexagent: Towards a foundation model for chest x-ray interpretation. *arXiv preprint arXiv:2401.12208*.

Dave A Davis. 2009. How to help professionals maintain and improve their knowledge and skills: Triangulating best practices in medicine. *Development of professional expertise: Toward measurement of expert performance and design of optimal learning environments*, pages 180–202.

Chunyuan Deng, Yilun Zhao, Xiangru Tang, Mark Gerstein, and Arman Cohan. 2023. Benchmark probing: Investigating data leakage in large language models. In *NeurIPS 2023 Workshop on Backdoors in Deep Learning-The Good, the Bad, and the Ugly*.

Christopher B Forrest, Gordon B Glade, Alison E Baker, Alison Bocian, Sarah von Schrader, and Barbara Starfield. 2000. Coordination of specialty referrals and physician satisfaction with referral care. *Archives of pediatrics & adolescent medicine*, 154(5):499–506.

Christopher B Forrest and Robert J Reid. 2001. Prevalence of health problems and primary care physicians' specialty referral decisions. *Journal of family practice*, 50(5):427–427.

Cortney Gensemer, Randall Burks, Steven Kautz, Daniel P Judge, Mark Lavallee, and Russell A Norris. 2021. Hypermobile ehlers-danlos syndromes: Complex phenotypes, challenging diagnoses, and poorly understood causes. *Developmental Dynamics*, 250(3):318–344.

Xuehai He. 2021. Towards visual question answering on pathology images. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing*, volume 2.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. 2019. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. *Proceedings of the AAAI conference on artificial intelligence*.

Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. 2019. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):317.

Jerome P Kassirer and G Anthony Gorry. 1978. Clinical problem solving: a behavioral analysis. *Annals of Internal Medicine*, 89(2):245–255.

Bertram G Katzung, Todd W Vanderah, and Michael J Aminoff. 2004. Basic & clinical pharmacology.

Vinay Kumar, Abul K Abbas, Nelson Fausto, and Jon C Aster. 2014. *Robbins and Cotran pathologic basis of disease, professional edition e-book*. Elsevier health sciences.

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

Christy Ledford, Esther Guard, Julie Phillips, Christopher Morley, Jacob Prunuske, and Andrea Wendling. 2022. How medical education pathways influence primary care specialty choice. *Family medicine*, 54(7):512–521.

Seowoo Lee, Jiwon Youn, Mansu Kim, and Soon Ho Yoon. 2023. Cxr-llava: Multimodal large language model for interpreting chest x-ray images. *arXiv preprint arXiv:2310.18341*.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2023a. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *arXiv preprint arXiv:2306.00890*.

Yingshu Li, Yunyi Liu, Zhanyu Wang, Xinyu Liang, Lingqiao Liu, Lei Wang, Leyang Cui, Zhaopeng Tu, Longyue Wang, and Luping Zhou. 2023b. A comprehensive study of gpt-4v's multimodal capabilities in medical imaging. *medRxiv*, pages 2023–11.

Matthew J Liberatore and Robert L Nydick. 2008. The analytic hierarchy process in medical and health care decision making: A literature review. *European Journal of Operational Research*, 189(1):194–207.

Weixiong Lin, Ziheng Zhao, Xiaoman Zhang, Chaoyi Wu, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023. Pmc-clip: Contrastive language-image pretraining using biomedical documents. *arXiv preprint arXiv:2303.07240*.

Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. 2021. Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 1650–1654. IEEE.

Fenglin Liu, Tingting Zhu, Xian Wu, Bang Yang, Chenyu You, Chenyang Wang, Lei Lu, Zhangdaihong Liu, Yefeng Zheng, Xu Sun, et al. 2023a. A medical multimodal large language model for future pandemics. *NPJ Digital Medicine*, 6(1):226.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*.

Ming Y Lu, Bowen Chen, Drew FK Williamson, Richard J Chen, Kenji Ikamura, Georg Gerber, Ivy Liang, Long Phi Le, Tong Ding, Anil V Parwani, et al. 2023. A foundational multimodal vision language ai assistant for human pathology. *arXiv preprint arXiv:2312.07814*.

Gary H Lyman and Nicole M Kuderer. 2023. Perception, cognition and thought: Part iii: Reasoning, judgement and decision-making.

Inbal Magar and Roy Schwartz. 2022. Data contamination: From memorization to exploitation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 157–165.

Michael Moor, Oishi Banerjee, Zahra Shakeri Hossein Abad, Harlan M Krumholz, Jure Leskovec, Eric J Topol, and Pranav Rajpurkar. 2023a. Foundation models for generalist medical artificial intelligence. *Nature*, 616(7956):259–265.

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

Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Cyril Zakka, Yash Dalmia, Eduardo Pontes Reis, Pranav Rajpurkar, and Jure Leskovec. 2023c. Med-flamingo: A multimodal medical few-shot learner. ArXiv:2307.15189.

Patrick R Murray, Ken S Rosenthal, and Michael A Pfaller. 2015. *Medical microbiology*. Elsevier Health Sciences.

OpenAI. 2023. Gpt-4v(ision) system card. *OpenAI*.

Vimla L Patel, David R Kaufman, and Jose F Arocha. 2002. Emerging paradigms of cognition in medical decision-making. *Journal of biomedical informatics*, 35(1):52–75.

Wojciech Pawlina and Michael H Ross. 2018. *Histology: a text and atlas: with correlated cell and molecular biology*. Lippincott Williams & Wilkins.

Obioma Pelka, Sven Koitka, Johannes Rückert, Felix Nensa, and Christoph M Friedrich. 2018. Radiology objects in context (roco): a multimodal image dataset. In *Intravascular Imaging and Computer Assisted Stenting and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis: 7th Joint International Workshop, CVII-STENT 2018 and Third International Workshop, LABELS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Proceedings 3*, pages 180–189. Springer.

Edward A Sickles and Carl J D'Orsi. 2016. Acr birads®. *ACR BI-RADS®-Atlas der Mammadiagnostik: Richtlinien zu Befundung, Handlungsempfehlungen und Monitoring*, page 474.

Richard S Snell. 2010. *Clinical neuroanatomy*. Lippincott Williams & Wilkins.

Sanjay Subramanian, Lucy Lu Wang, Sachin Mehta, Ben Bogin, Madeleine van Zuylen, Sravanthi Parasa, Sameer Singh, Matt Gardner, and Hannaneh Hajishirzi. 2020. Medicat: A dataset of medical images, captions, and textual references. *arXiv preprint arXiv:2010.06000*.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Omkar Thawkar, Abdelrahman Shaker, Sahal Shaji Mullappilly, Hisham Cholakkal, Rao Muhammad Anwer, Salman Khan, Jorma Laaksonen, and Fahad Shahbaz Khan. 2023. Xraygpt: Chest radiographs summarization using medical vision-language models. *arXiv preprint arXiv:2306.07971*.

Dequan Wang, Xiaosong Wang, Lilong Wang, Mengzhang Li, Qian Da, Xiaoqiang Liu, Xiangyu Gao, Jun Shen, Junjun He, Tian Shen, et al. 2023. Medfmc: A real-world dataset and benchmark for foundation model adaptation in medical image classification. *arXiv preprint arXiv:2306.09579*.

Chaoyi Wu, Jiayu Lei, Qiaoyu Zheng, Weike Zhao, Weixiong Lin, Xiaoman Zhang, Xiao Zhou, Ziheng Zhao, Ya Zhang, Yanfeng Wang, et al. 2023a. Can gpt-4v (ision) serve medical applications? case studies on gpt-4v for multimodal medical diagnosis. *arXiv preprint arXiv:2310.09909*.

Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023b. Towards generalist foundation model for radiology. *arXiv preprint arXiv:2308.02463*.

Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023c. Towards generalist foundation model for radiology.

Jiancheng Yang, Rui Shi, Donglai Wei, Zequan Liu, Lin Zhao, Bilian Ke, Hanspeter Pfister, and Bingbing Ni. 2023. Medmnist v2-a large-scale lightweight benchmark for 2d and 3d biomedical image classification. *Scientific Data*, 10(1):41.

Kai Zhang, Jun Yu, Zhiling Yan, Yixin Liu, Eashan Adhikarla, Sunyang Fu, Xun Chen, Chen Chen, Yuyin Zhou, Xiang Li, et al. 2023a. Biomedgpt: A unified and generalist biomedical generative pre-trained transformer for vision, language, and multimodal tasks. *arXiv preprint arXiv:2305.17100*.

Wenxuan Zhang, Sharifah Mahani Aljunied, Chang Gao, Yew Ken Chia, and Lidong Bing. 2023b. M3exam: A multilingual, multimodal, multilevel benchmark for examining large language models. *arXiv preprint arXiv:2306.05179*.

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

Juexiao Zhou, Xiaonan He, Liyuan Sun, Jiannan Xu, Xiuying Chen, Yuetan Chu, Longxi Zhou, Xingyu Liao, Bin Zhang, and Xin Gao. 2023. Skingpt-4: an interactive dermatology diagnostic system with visual large language model. *arXiv preprint arXiv:2306.00890*.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.

Appendix for Asclepius

Abstract. In this supplementary material, we provide additional information about the Asclepius. Appendix A illustrates the definition of each specialty. Appendix B provides list of involved organ. Appendix C elaborates on the capacity taxonomy. In Appendix D, we provide some examples in Asclepiusand some case stydies of Med-MLLMs for each capacity. Finally, Appendix E supplements the qualitative and quantitative results in the main paper, including the visualization of statistics for specialty and capacity, and the digital results of different specialties.

A Specialty List

Asclepius involves 15 specialties. They are Hematology, Cardiology, Endocrinology, Obstetrics and Gynecology, Gastroenterology, Immunology, Urology, Orthopedics, Neurology, Otolaryngology, Pulmonology, Dermatology, Pathology, Ophthalmology, General surgery.

- Hematology: Hematology is the branch of medicine concerned with the study of blood, the blood-forming organs, and blood diseases. It involves the diagnosis, treatment, and prevention of disorders related to the production and function of blood cells, as well as blood clotting and bleeding disorders.
- Cardiology: Cardiology is the branch of medicine that deals with the heart and its diseases. Cardiologists are medical doctors who specialize in the diagnosis, treatment, and prevention of heart disease and conditions.
- Endocrinology: Endocrinology is the branch of medicine that deals with the endocrine system, which is a network of glands that produce and release hormones into the bloodstream. Hormones are chemical messengers that travel throughout the body and control a wide range of functions, including metabolism, growth and development, reproduction, and mood.
- Obstetrics and Gynecology: Obstetrics and gynecology (OB/GYN) is the branch of medicine that deals with the female reproductive system, pregnancy, and childbirth. Obstetricians and gynecologists are medical doctors who specialize in the diagnosis, treatment, and prevention of conditions related to the female reproductive system and pregnancy.
- Gastroenterology: Gastroenterology is the branch of medicine that deals with the digestive system and its disorders. Gastroenterologists are medical doctors who specialize in the diagnosis, treatment, and prevention of diseases of the digestive tract, including the esophagus, stomach, small intestine, large intestine, and liver.
- Immunology: Immunology is the branch of medicine that deals with the immune system, which is a complex network of cells, tissues, and organs that work together to protect the body from infection. Immunologists are medical doctors who specialize in the diagnosis, treatment, and prevention of immune system disorders.
- Urology: Urology is the branch of medicine that deals with the male and female urinary tract and the male reproductive system. Urologists are medical doctors who specialize in the diagnosis, treatment, and prevention of urinary and reproductive disorders.
- Orthopedics: Orthopedics is the branch of medicine that deals with the musculoskeletal system, which includes the bones, joints, muscles, tendons, and ligaments. Orthopedic surgeons are medical doctors who specialize in the diagnosis, treatment, and prevention of musculoskeletal disorders.
- Neurology: Neurology is the branch of medicine that deals with the nervous system, which includes the brain, spinal cord, and nerves. Neurologists are medical doctors who specialize in the diagnosis, treatment, and prevention of disorders of the nervous system.
- Otolaryngology: Otolaryngology is the branch of medicine that deals with the ear, nose, and throat (ENT). Otolaryngologists are medical doctors who specialize in the diagnosis, treatment, and prevention of disorders of the ENT.
- Pulmonology: Pulmonology is the branch of medicine that deals with the respiratory system, which includes the lungs, airways, and breathing. Pulmonologists are medical doctors who specialize in the diagnosis, treatment, and prevention of respiratory disorders.
- Dermatology: Dermatology is the branch of medicine that deals with the skin, hair, and nails. Dermatologists are medical doctors who specialize in the diagnosis, treatment, and prevention of skin diseases and conditions.

- Pathology: Pathology is the branch of medicine that deals with the causes and effects of disease. Pathologists are medical doctors who specialize in the diagnosis and study of disease through the examination of tissues and bodily fluids.
- Ophthalmology: Ophthalmologists are medical doctors who specialize in the diagnosis, treatment, and prevention of eye diseases and conditions. They are also responsible for performing eye surgery.
- General Surgery: General surgery is a surgical specialty that focuses on the abdominal organs, including the esophagus, stomach, small intestine, large intestine, liver, pancreas, gallbladder, and bile ducts. General surgeons also perform surgery on the thyroid, parathyroid, and adrenal glands.

B Organ list

This section we list the involved body parts and organs in this benchmark.

53 Body Parts: Abdomen, Neck, Chest, Head, Cervical Vertebrae, Ankle, Femur, Vertebrae, Bowel, Mandible, Knee, Cochlea, Hand, Bladder, Spine, Wrist, Pelvis, Carotid Artery, Carotid Bifurcation, Trachea, Larynx, Colorectal, Blood, Forearm, Elbow, Hip, Muscle, Reproductive Systems, Gastrointestinal Tract, Ligamentum Nuchae, Small Intestine, Colon, Seminal Vesicle, Duodenum, Anterior Pituitary, Parathyroid, Vena-cava, Right Atrium, Left Ventricle, Muscular Artery, Bronchiole, Aorta, Palatine Tonsil, Pyloric Stomach, Cardiovascular, Endocrine, Musculo skeletal, Ophthalmic, Pulmonary, Blood smear, Cartilage, Adipose Tissue, Tendon, Nervous Tissue,

26 Organs: Liver, Lung, Brain, Breast, Testicle, Thyroid, Ovaries, Kidney, Heart, Uterus, Intestine, Pancreas, Pituitary gland, Stomach, Gallbladder, Skin, Eye, Blood, Ileum, Lymph Node, Umbilical Cord, Prostate, Duodenum, Parathyroid, Esophagus, Appendix

C Capacity

C.1 Perception

C.1.1 Anatomical Perception

Anatomical Perception is the ability to recognize and understand the normal structures of the body, including their locations, sizes, shapes, and the planes in which they are imaged.

Anatomical Perception is foundational to medical imaging analysis, as it allows for the accurate identification of body parts and serves as a basis for detecting abnormalities. It involves discerning the detailed anatomy within complex images and is essential for any subsequent diagnostic or therapeutic action.

C.1.2 Attribute Perception

Attribute Perception is the capacity to discern various attributes of tissues and structures, such as their density, texture, composition, and presence of pathological signs, along with recognizing instruments, modalities, and colors when applicable.

Attribute Perception focuses on the finer details that characterize tissues and help differentiate between normal and abnormal findings. It is vital for quality imaging interpretation and aids in the detailed understanding of pathologies and their implications on health.

C.1.3 Spatial and Quantitative Perception

Spatial and Quantitative Perception encompasses the skill to evaluate the spatial relations and quantitative aspects within medical images, such as counting entities and understanding their three-dimensional positions and relationships.

Spatial and Quantitative Perception is crucial for tasks that require an understanding of the geometry and distribution of anatomical structures and pathological findings, which is important for accurate diagnosis, surgical planning, and treatment evaluation.

C.2 Diagnosis

C.2.1 Disease Analysis

Disease Analysis is the ability to accurately detect and diagnose diseases from medical data, which may include imaging, laboratory results, patient history, and physical examination findings.

Disease Identification is the cornerstone of clinical practice, where the model's prowess in recognizing patterns and correlating them with potential diseases is essential. This involves not only identifying the presence of a disease but also understanding its nature.

C.2.2 Staging Assessment

Staging Assessment is the process of determining the extent or severity of a disease, especially cancer, within the body, which is crucial for choosing the most appropriate treatment strategy.

Staging Assessment evaluates how far a disease has progressed and is a critical step in the treatment

planning process. It requires a comprehensive analysis of data to accurately classify the stage, which can significantly affect prognosis and treatment choices.

C.3 Planning

C.3.1 Treatment

Treatment is the selection and administration of the appropriate therapeutic interventions tailored to the individual patient's disease and condition.

The Treatment aspect involves creating a detailed plan for managing the patient's condition, which may include medication, surgery, lifestyle changes, or other therapies. The objective is to select the most effective and least invasive treatment options while considering the patient's preferences and overall health status.

C.3.2 Prognosis

Prognosis is the prediction of the likely course and outcome of a disease, taking into account the nature of the disease, the patient's general physical condition, and the treatment options available.

Prognosis is about looking ahead to predict possible outcomes for the patient. This includes estimating survival rates, potential complications, and the likelihood of disease recurrence. It is a vital part of patient counseling and informs decision-making for both clinicians and patients.

C.3.3 Report generation

Report generation is the synthesis of medical data and analytical findings into coherent, standardized, and actionable reports for use by healthcare providers.

Report Generation combines all the collected information into a format that is understandable and useful for guiding clinical decisions. It ensures that the insights gained from the model's analysis are communicated effectively, serving as a bridge between the model's output and clinical action steps.

D Examples in Asclepius

In Figures 7 to 11, we illustrate several examples in different specialties in Asclepius. Also, we provide some case studies for various capacities in figures 12 to 14.

E Qualitative and Quantitative Results

This section supplements the results of evaluation:

Table 3 shows the confidence of doctors in various specialties.

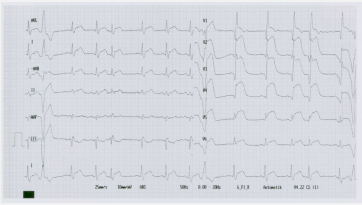
As shown in Table 4, the human doctors' performance varies widely across specialties and individuals, with accuracy scores ranging from 0.308 to 0.846 in the different fields. The average accuracy of human doctors (calculated from the "Avg" column) ranges from 0.465 to 0.564, with Doctor1 having the highest average accuracy. Comparing the Med-MLLMs to human doctors, all the Med-MLLMs have a lower average accuracy than the human doctors. Among the Med-MLLMs, GPT-4V's performance is closest to that of human doctors, although it still falls short of the lowest average accuracy of humans (Doctor5, 0.465).

Figure 15 visualizes the performance comparison of Med-MLLMs on the benchmark. Each model has its strengths and weaknesses, with GPT-4V showing the most robust performance across most areas. Gemini seems to be a decent second choice, particularly in Disease Identification. The other models have niche areas where they perform well but are generally outperformed by the top two models.

F Evaluation Models

Asclepiusfocuses on two general MLLMs, including GPT-4V (OpenAI, 2023) and Gemini (Team et al., 2023), along with four specialized Med-MLLMs, i.e. CheXagent (Chen et al., 2024), RadFM (Wu et al., 2023c), Med-Flamingo (Moor et al., 2023c), and XrayGPT (Thawkar et al., 2023). We evaluate the GPT-4V and Gemini through the official API. We test the rest of the specialized Med-MLLMs through the released code and pre-trained model.

Cardiology



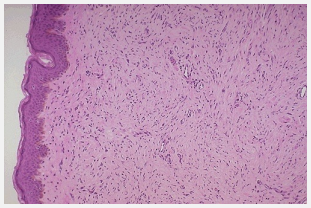
Question: A 57-year-old man is brought to the emergency department for crushing substernal chest pain at rest for the past 2 hours. The pain began gradually while he was having an argument with his wife and is now severe. He does not take any medications. He has smoked one pack of cigarettes daily for 35 years. He is diaphoretic. His temperature is 37.1°C (98.8°F), pulse is 110/min, respirations are 21/min, and blood pressure is 115/65 mm Hg. Pulse oximetry on room air shows an oxygen saturation of 97%. Cardiac examination shows an S4 gallop. The lungs are clear to auscultation. An ECG is shown. Which of the following is the most likely underlying cause of this patient's condition?
A) Occlusion of the right coronary artery
B) Inflammation of the pericardium
C) Occlusion of the left anterior descending artery
D) Thromboembolism to the right interlobar pulmonary artery
E) Diffuse coronary vasospasm
F) Occlusion of the left circumflex artery
G) Tear in the intimal lining of the aorta
Standard Answer: C) Occlusion of the left anterior descending artery

Question: A 65-year-old man has had increasing dyspnea and orthopnea for the past year. On physical examination there are rales in all lung fields. A chest x-ray shows cardiomegaly and pulmonary edema. An echocardiogram shows reduced cardiac output. The gross appearance of his heart shown here is most consistent with which of the following underlying conditions?
A) Amyloidosis
B) Systemic hypertension
C) Diffuse scleroderma
D) Atherosclerosis
E) Viral myocarditis
Standard Answer: D) Atherosclerosis

Figure 7: **Examples for Cardiology.**

Dermatology

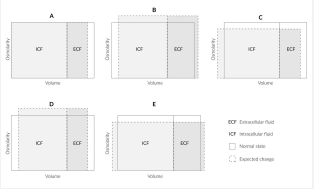
Question: A 10-year-old child complains of intense itching of his hands. Physical examination reveals several linear gray-brown lesions a few millimeters in length that appear in the skin between his fingers. Some of these lesions have been excoriated by scratching. The finding from scraping these lesions is shown here. Which of the following infectious agents is the most likely cause?
A) Epidermophyton floccosum
B) Herpes zoster
C) Sarcoptes scabiei
D) Staphylococcus aureus
E) Malassezia furfur
F) Papillomavirus
Standard Answer: C) Sarcoptes scabiei



Question: A 25-year-old man has multiple firm nodules, some of which are up to 1 cm in size, that are disfiguring because they are located on his hands and arms. Also noted on physical examination are scattered light brown macules with smooth borders, averaging about 2 to 3 centimeters in size, over the skin of his trunk and arms. The microscopic appearance of one of the nodular lesions is seen here. His father and brother have similar lesions. Which of the following is the most likely diagnosis?
A) Neurofibromatosis
B) Kaposi sarcoma
C) Nevocellular nevus
D) Actinic keratosis
E) Impetigo
Standard Answer: A) Neurofibromatosis

Figure 8: **Examples for Dermatology.**

Endocrinology

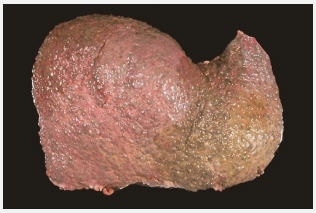


Question: A 32-year-old woman comes to the physician for increasing fatigue over the past 3 months. During this period, her appetite has decreased and she has lost 10 kg (22 lb). She also reports decreased sexual desire. She has a history of type 1 diabetes. Her pulse is 105/min and blood pressure is 95/63 mm Hg. Physical examination shows darkening of the skin and oral mucous membranes. Serum studies show:
Na+ 130 mEq/L; K+ 5.2 mEq/L; Glucose (fasting) 74 mg/dL
Which of the following changes to body fluid volume and osmolarity would most likely be expected?
A) A
B) B
C) C
D) D
E) E
Standard Answer: C) C

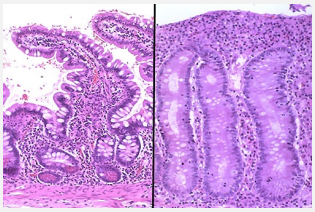
Question: A 23-year-old woman comes to physician for an annual health maintenance examination. She feels well. She is 155 cm (5 ft 1 in) tall and weighs 79 kg (174 lb); BMI is 33 kg/m2. Examination shows a skin rash over both axillae. A photograph of her left axilla is shown. This patient's skin finding is most likely associated with which of the following conditions?
A) Niacin deficiency
B) Gastric malignant lymphoma
C) Primary adrenal insufficiency
D) Polycystic ovary syndrome
E) Hashimoto thyroiditis
F) Pregnancy
Standard Answer: D) Polycystic ovary syndrome

Figure 9: **Examples for Endocrinology.**

Gastroenterology



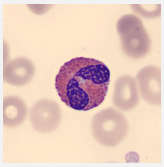
Question: A 45-year-old woman developed dyspnea with fever over the past week. Sputum culture grows Streptococcus pneumoniae and a chest radiograph shows extensive infiltrates of the right lower lobe and a nodule with air-fluid level. Her serum AST is 277 U/L and ALT 183 U/L. This is the gross appearance of her liver. Which of the following is most likely to produce this appearance of her liver?
A) Hepatitis A infection
B) Colonic adenocarcinoma
C) Chronic alcohol abuse
D) Oral contraceptive use
E) Dominant polycystic kidney disease
F) Alpha-1-antitrypsin deficiency
Standard Answer: C) Chronic alcohol abuse



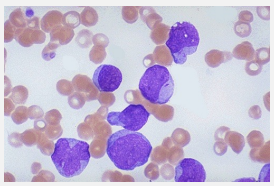
Question: A 27-year-old man has had a chronic diarrhea with weight loss of 5 kg over the past year. On physical examination there are no abnormal findings. His d-xylose absorption is abnormally low. He has elevated anti-endomysial and anti-transglutaminase antibody titers. A jejunal biopsy is performed and shown here in the right panel (compared to the normal jejunum in the left panel). He improves with a change in diet. Which of the following is the most likely diagnosis?
A) Adenocarcinoma
B) Cholera
C) Pseudomembranous enteritis
D) Celiac disease
E) Carcinoid tumor
F) Whipple disease
Standard Answer: D) Celiac disease

Figure 10: **Examples for Gastroenterology.**

Hematology



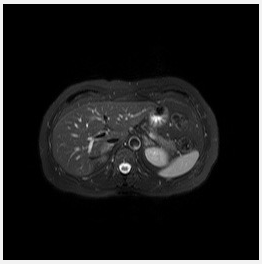
Question: Which group of peripheral blood cells is visible in this image?
a) Neutrophils
b) Eosinophils
c) Basophils
d) Lymphocytes
Standard Answer: b) Eosinophils



Question: A 30-year-old man has noted a progressively worsening cough for one month. On physical examination, a few small lymph nodes are palpable in his axillae, and the tip of the spleen is palpable. Laboratory studies show Hgb 10.2 g/dL, Hct 31.1%, MCV 90 fL, WBC count 67,400/microliter, and platelet count 36,220/microliter. His peripheral blood smear is shown here. Which of the following is the most likely diagnosis?
A) Acute myelogenous leukemia
B) Plasma cell leukemia
C) Chronic lymphocytic leukemia
D) Acute lymphoblastic leukemia
E) Hairy cell leukemia
F) Leukoerythroblastosis
Standard Answer: A) Acute myelogenous leukemiaimg1536.jpg

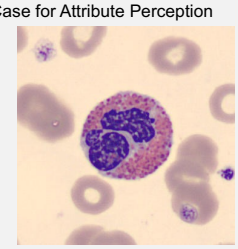
Figure 11: **Examples for Hematology.**

(a) Case for Anatomical Perception



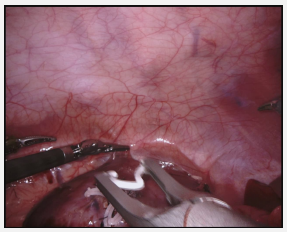
Question: What is the largest organ in the picture?
Standard Answer: Liver
Model Output:
1) **GPT-4V**: The largest organ in this MRI image appears to be the liver. ✅
2) **Gemini**: The liver. ✅
3) **CheXagent**: The largest organ in the picture is the heart. ❌
4) **RadFM**: Liver. ✅
5) **Med-Flamingo**: Liver. The largest organ in the picture is the liver. The liver is the largest organ in the body. It is located in the upper-right abdominal cavity. It is the largest organ in the body and is the second largest organ in the body. It is the largest organ in the body ✅
6) **XrayGPT**: The large organ that is visible in the picture is the stomach. It appears that there are two ultrasound images of the stomach in the picture, but it is not clear whether they are real or are part of an imaging test. If you can provide more information about the patient and the specific location of the ultrasound images, I can try to give you a more detailed answer. Hope this answers your question. If you have any further questions, please feel free to write. ❌

(b) Case for Attribute Perception



Question: Select the group that best describes the peripheral blood cell image:
a) Lymphocytes
b) Monocytes
c) Platelets
d) Basophils
Standard Answer: a
Model Output:
1) **GPT-4V**: b) Monocytes ❌
2) **Gemini**: b) Monocytes ❌
3) **CheXagent**: b) Monocytes ❌
4) **RadFM**: a) Lymphocytes ✅
5) **Med-Flamingo**: a) Lymphocytes ✅
6) **XrayGPT**: c) Platelets ❌

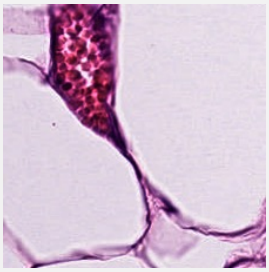
(c) Case for Spatial and Quantitative Perception



Question: Where is monopolar_curved_scissors located?
Standard Answer: bottom
Model Output:
1) **GPT-4V**: The monopolar curved scissors are located at the bottom right of the image, with black and silver handles and tips pointing to the left side of the image. ✅
2) **Gemini**: The monopolar_curved_scissors are located on the left side of the image. ❌
3) **CheXagent**: The monopolar_curved_scissors is located at the distal end of the esophagus. ❌
4) **RadFM**: Proximal. ❌
5) **Med-Flamingo**: In the right lower quadrant. ❖ Image 10.1. 10.1 Monopolar Curved Scissors 10.1.1 Indications • Dissection of the colon and rectum. • Dis ❌
6) **XrayGPT**: Hi, the monopolar curved scissors are used for cutting off excess fluid that is collected in the pleural cavity. This can be caused due to an infection in the lungs. The fluid may also be due to accumulation in the sacs, called pericardium. The fluid in the pericardium is usually not a cause of concern.However, if there is too much fluid in the pericardium it can compress the heart leading to dysfunction. In some cases it can even lead to cardiac rupture. Therefore, a Chest X-ray is done to rule out any fluid collection. Hope I have answered your query. Let me know if I can assist you further. ❌

Figure 12: **Case study for perception.**

(a) Case for Disease Identification



Question: Does the image show Cancer-associated Stroma (STR)?
Standard Answer: No
Model Output:
1) **GPT-4V**: I will review the image you've uploaded to determine if it shows Cancer-associated Stroma (STR). Let's proceed to examine the image. No, the image does not show Cancer-associated Stroma (STR). ✅
2) **Gemini**: Yes ❌
3) **CheXagent**: No ✅
4) **RadFM**: no ❌
5) **Med-Flamingo**: Yes ❌
6) **XrayGPT**: Hi, the image shows STR and a proliferation of stroma cells that are associated with malignant growth and are responsible for its development and maintenance. Stroma cells are cells that encapsulate and protect growing tumor cells, and are thought to be responsible for the initiation and proliferation of cancer. A variety of stroma cells with a variety of functions have been identified and described. ❌

(b) Case for Spatial and Staging Assessment

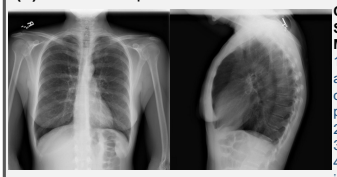
Question: What is the severity level of diabetic retinopathy in this retinal fundus image?
a) Level 0
b) Level 1
c) Level 2
d) Level 3
Standard Answer: a
Model Output:
1) **GPT-4V**: a) Level 0 ✅
2) **Gemini**: c) Level 2
3) **CheXagent**: b) Level 1 ❌
4) **RadFM**: a) Level 0 ✅
5) **Med-Flamingo**: c) Level 2 ❌
6) **XrayGPT**: a) Level 0 ✅

Figure 13: **Case study for diagnosis.**

(a) Case for Prognosis

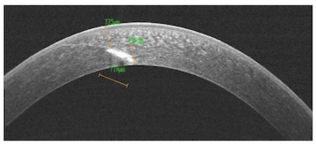
Question: A 42-year-old patient is referred for recurrent erosion syndrome after several bouts of painful epithelial erosion. His refraction has been changing. His anterior segment photo is provided. Which is most likely true about his refraction?
a) It will likely stabilize without any surgery if the corneal epithelium stops eroding.
b) It will likely get progressively hyperopic without any surgery.
c) It will likely get progressively myopic without any surgery.
d) Central photo therapeutic keratectomy (PTK), if performed, should contribute to further myopia.
Standard Answer: c
Model Output:
1) **GPT-4V**: a) It will likely stabilize without any surgery if the corneal epithelium stops eroding. ✗
2) **Gemini**: b) It will likely get progressively hyperopic without any surgery. ✗
3) **CheXagent**: c) It will likely get progressively myopic without any surgery. ✓
4) **RadFM**: His refraction has been changing. His anterior segment photo is provided. ✗
5) **Med-Flamingo**: a. 10. A 42-year-old patient is referred for recurrent erosion syndrome after several bouts of painful epithelial erosion. His refraction has been changing. His anterior segment photo is provided. Which is most likely true about his ref ✗
6) **XrayGPT**: Hello, I have studied your case and would recommend a detailed eye examination. The best treatment for your condition will be to take a close look at your cornea and prescribe you the right lenses. If you do not have any vision loss then you need not worry much. A change in refraction can happen in patients with long term use of certain types of contact lenses. So please take care of your lenses and get them checked for any scratches or dirt. If you have any visual impairment then you need to correct the same as soon as possible. Hope this answers your query. If you have additional questions or follow-up queries then please do not hesitate in writing to us. I will be happy to answer your queries. Wishing you good health. Take care. ✗

(h) Case for Report Generation



Question: Generate free-text radiology reports for the entered chest X-ray images.
Standard Answer: a
Model Output:
1) **GPT-4V**: The cardiomediastinal silhouette is normal in size and contour. The lungs are clear bilaterally without any evidence of focal airspace disease, pleural effusion, or pneumothorax. The bony structures and soft tissues are unremarkable. There is no evidence of anterior or posterior mediastinal widening. The lung fields are clear without any evidence of consolidation or mass. No pleural effusions or pneumothoraces are seen. The bony structures are intact without acute abnormalities. **(R-L: 0.235)**
2) **Gemini**: No obvious acute cardiopulmonary process. **(R-L: 0.074)**
3) **CheXagent**: No acute cardiopulmonary process. **(R-L: 0.077)**
4) **RadFM**: The system does not provide a specific interpretation or diagnosis. The physician can browse through a series of chest X-ray images and select one that best represents the patient's current condition. The physician can also provide her own description of the X-ray findings. **(R-L: 0.036)**
5) **Med-Flamingo**: The chest X-ray image shows a normal cardiac silhouette. The lungs are clear. The heart size is normal. **(R-L: 0.216)**
6) **XrayGPT**: The radiology reports can be generated in minutes. Get an instant PDF or HTML output. No software required. Get your radiology reports today with our online radiology services. Get the best radiology reports with our affordable online services. Our reports can be customized to meet your specific requirements. Contact us today to get the best online radiology services. **(R-L: 0.034)**

(g) Case Failed Prognosis



Question: A 41-year-old patient with a refraction of -4.00 diopters sphere OU desires refractive surgery. He reports an eye injury from cutting ceramic tile 10 years ago. The eye is normal other than a sliver of white foreign body in the posterior, paracentral cornea. The foreign body is 719 um long, and there are 335 microns between the epithelium and anterior point of the foreign body. Descemet's is intact.
Which method of refractive surgery would be the best option?
a) PRK
b) LASIK
c) SMILE
d) Femtosecond laser-assisted refractive lens exchange
Standard Answer: a
Model Output:
1) **GPT-4V**: a) PRK ✓
2) **Gemini**: b) LASIK ✗
3) **CheXagent**: b) LASIK ✗
4) **RadFM**: (FS-LARS) would be the best option. ✗
5) **Med-Flamingo**: a) PRK 1. The patient has a history of a foreign body in the cornea. ✓
6) **XrayGPT**: Hello, Since there is a foreign body lodged in the cornea, PRK and LASIK are out of the question. A femtosecond laser should be used to remove the foreign body in the posterior and paracentral region of the cornea. Then SMILE can be done for the rest of the cornea. Hope I have answered your query. Let me know if I can assist you further. ✗

Figure 14: **Case study for planning.** R-L represents the ROUGE-L score.

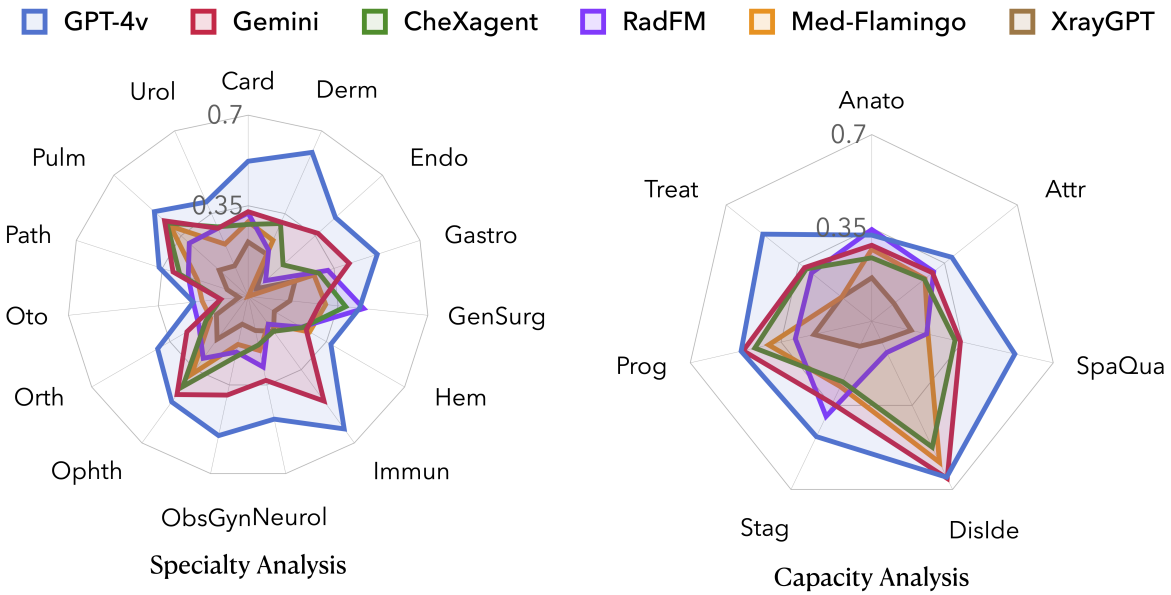


Figure 15: **Comparison of Med-MLLMs on our benchmark.**

Table 3: **The specialty confidence of each doctor.** Use $1 \sim 5$ to represent the confidence score for each specialty. The larger number means more confidence.

Specialty	Dr. 1	Dr. 2	Dr. 3	Dr. 4	Dr. 5
Card	2	2	5	4	5
Derm	5	4	2	3	3
Endo	3	3	4	4	4
Gastro	4	3	4	4	4
GenSurg	4	4	3	4	3
Hem	2	3	4	3	4
Immun	5	4	4	4	4
Neurol	4	3	4	4	4
ObsGyn	2	5	2	3	4
Ophth	5	3	2	5	2
Orth	4	3	3	4	2
Oto	3	3	2	3	3
Path	4	3	2	3	3
Pulm	2	4	4	4	4
Urol	4	5	3	4	3

Table 4: **Accuracy of models across different specialties.**

Model	Card	Derm	Endo	Gastro	GenSurg	Hem	Immun	Neurol
Doctor1	0.636	0.711	0.500	0.500	0.500	0.600	0.483	0.450
Doctor2	0.500	0.542	0.421	0.600	0.440	0.500	0.517	0.526
Doctor3	0.714	0.400	0.462	0.647	0.450	0.500	0.379	0.560
Doctor4	0.550	0.375	0.650	0.536	0.550	0.400	0.482	0.350
Doctor5	0.529	0.444	0.308	0.484	0.625	0.391	0.483	0.517
GPT-4V	0.522	0.609	0.455	0.526	0.434	0.370	0.633	0.485
Gemini	0.327	0.314	0.364	0.412	0.279	0.260	0.500	0.332
CheXagent	0.278	0.308	0.182	0.289	0.380	0.242	0.167	0.192
RadFM	0.322	0.195	0.091	0.325	0.450	0.238	0.133	0.279
Med-Flamingo	0.288	0.237	0.000	0.268	0.302	0.275	0.133	0.214
XrayGPT	0.210	0.178	0.045	0.191	0.163	0.117	0.167	0.135

Model	ObsGyn	Ophth	Orth	Oto	Path	Pulm	Urol	Avg	Var
Doctor1	0.455	0.767	0.423	0.800	0.421	0.500	0.636	0.564	0.014
Doctor2	0.639	0.421	0.579	0.000	0.357	0.400	0.636	0.498	0.021
Doctor3	0.500	0.846	0.500	0.286	0.333	0.500	0.750	0.538	0.022
Doctor4	0.750	0.737	0.464	0.500	0.300	0.550	0.500	0.536	0.015
Doctor5	0.450	0.333	0.556	0.250	0.400	0.600	0.333	0.465	0.011
GPT-4V	0.550	0.504	0.405	0.214	0.363	0.489	0.399	0.462	0.010
Gemini	0.390	0.469	0.274	0.107	0.306	0.436	0.288	0.354	0.008
CheXagent	0.240	0.431	0.189	0.107	0.279	0.424	0.294	0.309	0.008
RadFM	0.220	0.296	0.221	0.214	0.244	0.307	0.288	0.278	0.006
Med-Flamingo	0.190	0.372	0.168	0.179	0.207	0.405	0.221	0.279	0.009
XrayGPT	0.110	0.205	0.142	0.036	0.086	0.151	0.129	0.148	0.002

Table 5: **The Analysis of Human Doctor in Capacities.**

Human Doctor	Perception			Diagnosis		Planning		Avg
	Anato	Attr	SpaQua	DisIde	Stag	Prog	Treat	
Doctor1	0.541	0.554	0.567	0.554	0.567	0.554	0.567	0.564
Doctor2	0.460	0.469	0.490	0.483	0.490	0.483	0.490	0.498
Doctor3	0.542	0.502	0.512	0.502	0.512	0.502	0.512	0.538
Doctor4	0.545	0.536	0.550	0.543	0.550	0.543	0.550	0.536
Doctor5	0.420	0.545	0.548	0.545	0.548	0.545	0.548	0.465