PatientQA: A Question Answering Benchmark for Patient Diagnosis

Anonymous Authors[1]

Abstract

This paper identifies two key areas for improvement in existing medical benchmark research. In text-based medical question answering (QA), open-ended QA (OpenQA) has not been thoroughly explored, with current evaluations relying heavily on machine translation metrics. For medical visual QA (VQA), previous studies have predominantly focused on knowledge-based QA or pattern recognition, often derived from literature or textbooks, lacking benchmarks that consider patient background diagnoses. To address these gaps, we propose a Chinese medical benchmark PatientQA, which comprises two components: an OpenQA segment that employs a step-by-step evaluation scheme for assessing LLM responses, and a multiple-choice VQA segment centered on patient diagnosis. Experiments involving over 10 LLMs reveal that many models struggle with diagnosing patients based on medical backgrounds. We believe our research provides essential benchmarks and insights that can guide the future development of multimodal medical QA systems, particularly in enhancing evaluation methods and addressing the complexities of medical contexts.

1. Introduction

Assessing model performance in medical text generation tasks has been an area of active investigation for many years, encompassing a range of response formats (Lau et al., 2018; Jin et al., 2021; Liu et al.), from multiple-choice questions (MCQ) (Pal et al., 2022) to open-ended questions (OpenQA) (Liu et al., 2021), as well as various modalities, including text-only question answering and visual question answering (VQA) (He et al., 2021). We summarize a selection of representative medical evaluation tasks alongside the current state-of-the-art (SOTA) in Table 1.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

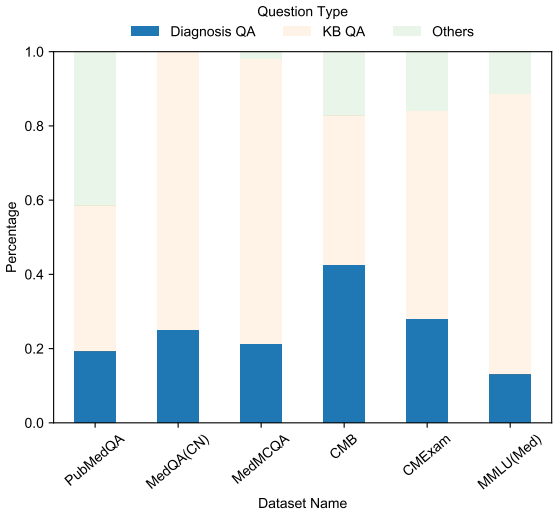

Figure 1: Text-based question type distribution of representive benchmarks.

As shown in the table, most text-based questions fall into either MCQ or OpenQA yielding short answers. This preference arises from the reason that simple responses like "A", "B", "C", "Yes", "No", "On the left" are much easier to evaluate when comparing a LLM output to the correct answer. For OpenQA with longer responses, such as the CMB evaluation (Wang et al., 2024), researchers leverage GPT's capabilities to assess fluency, relevance, completeness, and proficiency. While text-based QA for evaluating LLMs have developed in the direction that aligns with human standards - evident in using evaluations like the Medical License Examination (MLE) (Jin et al., 2021; Hendrycks et al., 2021; Pal et al., 2022; Wang et al., 2024; Liu et al.; Yue et al., 2024) - it remains underdeveloped in evaluating LLMs' ability to generate comprehensive medical texts related to *patient diagnosis*.

VQA that tends to elicit short answer can be attributed to two main reasons. First, medical VQA datasets sourced from hospitals are primarily designed to train medical assistants in identifying and interpreting medical images. In this context, concise questions and answers effectively facilitate the learning process. Second, VQA derived from textbooks often involves extracting questions from image-caption pairs using tools like ChatGPT (He et al., 2021;

Table 1: Representative medical evaluations over the years. We use overall accuracy as the measure of accuracy. Amount only refer to test set if applicable. Response type is devided into three MCQ, short and long for OpenQA. MLE abbreviates for Medical License Examination. PMC abbreviates for PubMed Central.

Dataset	Year	Amount	Source	SOTA	Modality	Response Type
VQA-RAD (Lau et al., 2018)	2018	3515 QA ,315 image	MedPix	81.9 PeFoMed (Liu et al., 2024a)	VQA	short
ImageCLEF-2019 (Ben Abacha et al., 2019)	2019	12792 QA, 3200 image	hospital	62.4 Hanlin (Yan et al., 2019)	VQA	short
PubMedQA (Jin et al., 2019)	2019	500 testset	PMC	82 GPT-4(Medprompt) (Nori et al., 2023)	text	short
PathVQA (He et al., 2021)	2020	6012 QA, 1000 image(test)	textbook	82.75 LLaVA-Med++ (Xie et al., 2024)	VQA	short
MedQA (Jin et al., 2021)	2021	6112 testset	MLE	91 Med-Gemini (Saab et al., 2024)	text	MCQ
SLAKE (Liu et al., 2021)	2021	14028 QA, 642 image	mixed	87.8 LLaVA-Med++ (Xie et al., 2024)	VQA	short
MMLU(Med) (Hendrycks et al., 2021)	2021	499 QA	textbook	88.7 Claude 3.5 Sonet (Anthropic, 2024)	text	MCQ
MedMCQA (Pal et al., 2022)	2022	6150 testset	MLE	72.3 Med-PaLM 2 (Singhal et al., 2023)	text	MCQ
PMC-VQA (Zhang et al., 2023b)	2023	227k QA, 149k image	PMC	42.3 medVInT (Zhang et al., 2023b)	VQA	MCQ
ImageCLEF-2023(Ionescu et al., 2023)	2023	36683 QA, 2000 image	hispital	-	VQA	
CMB (Wang et al., 2024)	2023	11200 test MCQ	MLE	74.38 (Bai et al., 2023)	text	MCQ,long
CMExam (Liu et al.)	2023	60000 MCQ	MLE	61.7 GPT-4 (Achiam et al., 2023)	text	MCQ
MMMU(Med) (Yue et al., 2024)	2024	907 MCQ	textbook	59 Gemini (Team et al., 2023)	VQA	MCQ
Ours	**2025**	**1151 QA, 815 images**	**MLE**	**-**	**VQA**	**MCQ,long**

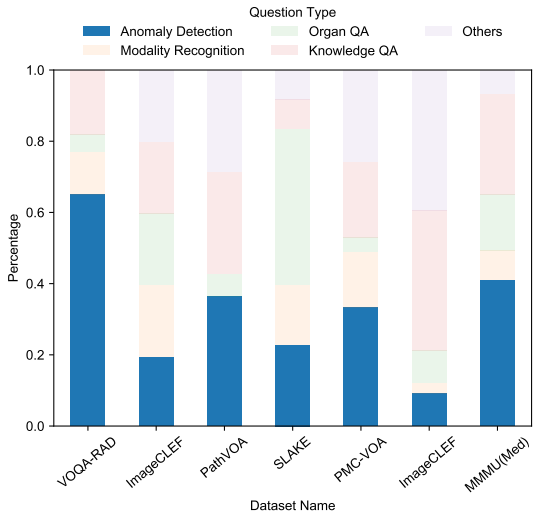

Figure 2: Multimodal question type distribution of representive benchmarks.

Zhang et al., 2023b; Yue et al., 2024). Hence, short answers that are directly pulled from the original text will help maintain a professional tone and ensure clarity. The emphasis on brevity makes it challenging to develop medical VQA evaluations that focus on patient conditions and diagnosis, as few cases can be fetched from textbooks or literature.

We analyse previous evaluation work by counting the distribution of question types, both text-based and visual-based, as illustrated in Figures 1 and 2. As observed, for text-based evaluations, we compare datasets including PubMedQA (Jin et al., 2019), the MedQA China Mainland subset (denoted as MedQA(CN)) (Jin et al., 2021), MedMCQA (Pal et al., 2022), CMB(Wang et al., 2024), CMExam (Liu et al.), and the MMLU medical related subsets (denoted as MMLU(Med)) (Hendrycks et al., 2021). Our findings reveal that a significant portion of the questions remains knowledge-based, with less than 40% pertaining to diagnostic inquiries. In the case of visual-based evaluations, we examine all multimodal datasets listed in Figure 2. It can be seen that, aside from anomaly detection, previous Med-VQA efforts have predominantly focused on image modality recognition, organ identification, and general knowledge. Quite few questions fall into the "Others" category, which includes direct diagnostic inquiries. *These observations underscore the pressing need to enhance the evaluation of LLMs in their capacity to facilitate patient diagnosis.*

The skills of give diagnostic texts and being able to read medical images during diagnosis are both essential for a clinical physician. Evaluating these gaps between an expert and current LLMs inspires our study. Hence, we introduce PatientQA, designed to assess how well LLMs perform in patient-related QA, in order to pave the way for the development of more advanced models and more rigorous evaluation sets.

In a nutshell, our paper makes the following contributions:

1. We propose PatientQA, a new benchmark for patient diagnosis that encompasses both OpenQA and MCQ components.

2. Results from the OpenQA subset suggest a more effective metric for evaluating medical long text generation.

3. Findings from the MCQ subset indicate that current models still struggle to capture essential medical features.

2. Related Work

The medical evaluation of LLMs in the context of text generation can be broadly categorized into multiple-choice questions (MCQ) and open-ended question answering (OpenQA).

MCQ for Medical Evaluation To the best of our knowledge, MedQA (Jin et al., 2021) is the seminar work to propose a popular medical and clinical MCQ evaluation based on real-world scenarios. In this study, researchers utilized the National Medical License Examination (MLE) to assess language models, alongside a test for document retrieval. Following this work, MMLU (Hendrycks et al., 2021) introduced multi-level tests covering 57 subjects, including those related to medicine. They source the data from online documents, supplemented by undergraduate and graduate-level human assistance, to evaluate the capabilities of the GPT-3 series. Building on MedQA, MedMCQA (Pal et al., 2022) developed an Indian version of the assessment, featuring a broader range of classified subjects and a higher volume of questions. While the aforementioned evaluations predominantly focus on English, researchers subsequently proposed CMB (Wang et al., 2024) and CMExam (Liu et al.), both centered on Simplified Chinese. In the realm of multimodal MCQs, PMC-VQA (Zhang et al., 2023b) is the first to construct a large-scale medical domain Visual Question Answering (VQA) dataset, utilizing PubMed Central image-caption pairs and ChatGPT to formulate MCQs. MMMU (Yue et al., 2024) represents a multimodal evaluation encompassing 30 subjects, annotated by 50 university students, with questions originally sourced from textbooks and online materials. While there are no exceptions regarding dataset collection and accuracy testing methods for MCQs, our multimodal MCQ tests for LLMs are specifically designed with patient-centered objectives in mind.

OpenQA for Medical Evaluation In the domain of multimodal OpenQA, VQA-RAD (Lau et al., 2018) proposed a framework that was manually constructed with short answers and verified for accuracy through manual review. They adopt a one-image, one-question pair format to ensure compatibility with current algorithms. VQA-Med-2019, introduced at ImageCLEF 2019, evaluates multimodal language models primarily using BLEU metrics, which also includes short answers. PathVQA (He et al., 2021) sourced image-caption pairs from pathology textbooks and employed NLP tools to generate questions and short answers. For this work, they utilized BLEU-(1,2,3), exact match scores, and F1 scores to assess accuracy. SLAKE (Liu et al., 2021) proposed a fine-grained, manually labeled OpenVQA, dedicating over half a year focusing on CT, MRI, and X-ray images, with content provided in both Chinese and English. For text-only OpenQA, PubMedQA (Jin et al., 2019) generated a large-scale "Yes/No" question set using articles from the PubMed database, requiring models to answer after reading lengthy paragraphs. CMB also offers a small subset of OpenQA containing over 70 questions for diagnosing patients. Given the challenges associated with evaluating long-generated medical diagnosis text, we propose a new OpenQA approach to address this issue by providing a step-by-step rating guide.

内容	得分
一、初步诊断	**3 分**
1. 支气管扩张	2 分
2. 双下肺肺炎	1 分
二、诊断依据(初步诊断错误，诊断依据不得分；未分别列出各自诊断依据扣一分。)	**5 分**
1. 支气管扩张	
(1)老年男性，慢性病程，反复咳嗽、咳脓痰，伴痰中带血。	0.5 分
(2)查体：双下肺湿啰音，杵状指。	0.5 分
(3)胸部CT：双肺多发囊状、斑片状影。	1 分
2. 双下肺肺炎	
(1)发热，痰量增加、脓性痰。	1 分
(2)查体：双下肺湿啰音。	1 分
(3)血常规：白细胞总数及中性粒细胞比例明显增高。	0.5 分
(4) 胸部 CT：双下肺斑片状阴影。	0.5 分
三、鉴别诊断	**4 分**
1. 慢性阻塞性肺疾病	2 分
2. 肺结核	1 分
3. 支气管肺癌	1 分
四、进一步检查	**5 分**
1. 肝、肾功能，肿瘤标志物。	1 分
2. 痰病原学检查 (细菌培养 + 药敏试验、痰涂片抗酸染色)。	1 分
3. 动脉血气分析。	1 分
4. 肺功能检查 (病情控制后)。	1.5 分
5. 必要时支气管镜检查。	0.5 分
五、治疗原则	5 分
1. 休息、吸氧、营养支持。	1.5 分
2. 应用广谱抗生素+抗厌氧菌药物。	1.5 分
3. 应用支气管舒张剂、祛痰药物。	1 分
4. 病情缓解后行肺炎球菌疫苗、流感疫苗接种。	1 分

Figure 4: An example of OpenQA reference answer.

The current research trends indicate a growing focus on multilingual and multimodal evaluations in the medical domain, emphasizing the need for patient-centered approaches. Our work serves as a further exploration in this direction.

3. The Proposed Benchmark – PatientQA

In this section, we provide a detailed description of how the evaluation dataset was constructed, the metrics we employed for evaluation, and statistical information regarding our evaluation set.

Data Source Our datasets comprise printed exercises containing both MLE and OpenQA provided directly by partner clinical doctors. While multiple-choice questions (MCQs) are generally easy for humans to read and interpret alongside the accompanying images, we must ensure that the questions are also readable for large language models (LLMs). Figure 3 illustrates our approach to extracting MCQs. Initially, we have over 10,000 MCQs, which we categorize

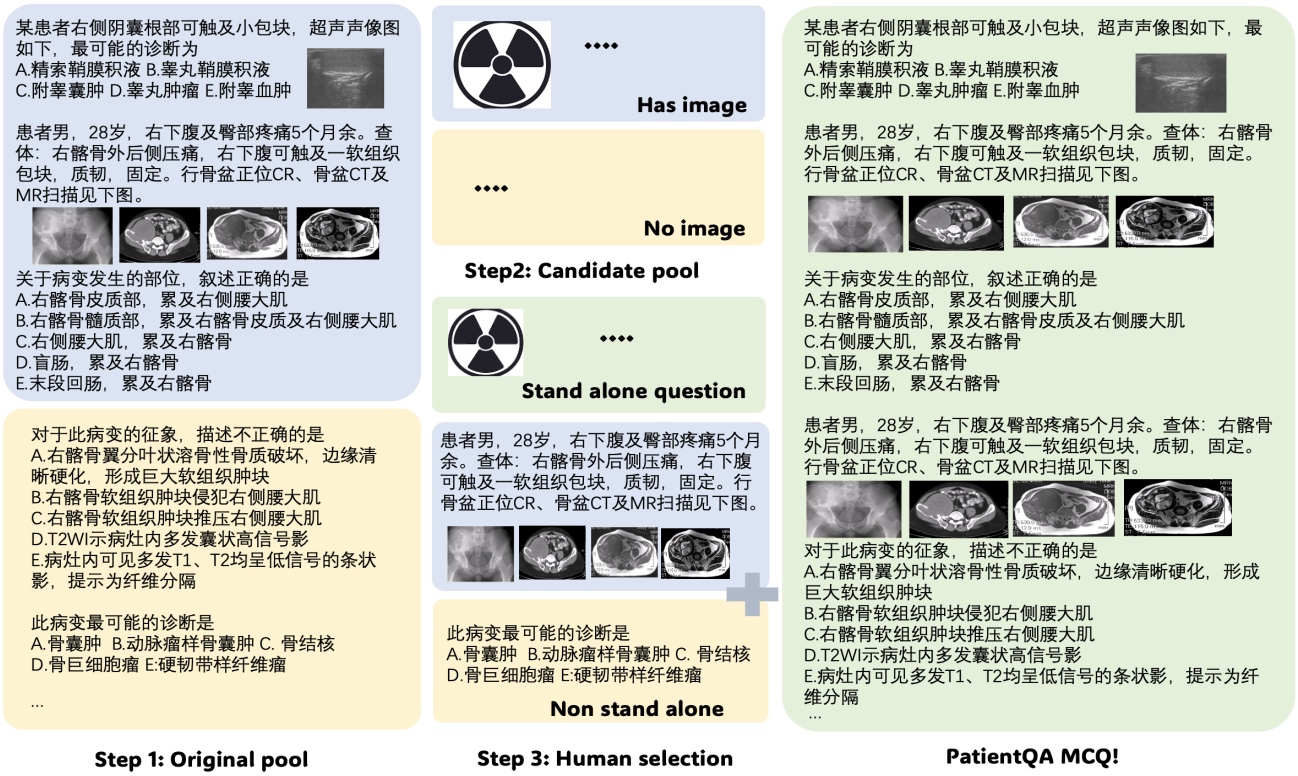

Figure 3: PatientQA MCQ subset collection process.

based on whether they mention an image. This results in two candidate pools. Questions that contain standalone images (i.e., those that do not require additional context from previous questions) are directly assigned to the PatientQA subset. For questions that share one or more images, or that may require an image or reference to a previous question for answering undergo manual verification to ensure they are appropriately framed as multimodal questions. This process yields approximately 665 distinct questions. In the OpenQA section, the layout is well-formatted, and we use it as is. This subset includes two parts, medical history collection and medical summary diagnosis, resulting in 487 questions with corresponding reference answers, serving as a guide for teachers to evaluate students' responses, as an example shown in Figure 4. Additionally, for each subject, there is a cheat sheet containing extensive content outlining key diagnostic points, which we will use as knowledge augmentation during testing.

Evaluation Scheme In the MCQ section, we instruct GPT to extract the corresponding candidate letter from the LLM's output, irrespective of the chain-of-thought reasoning provided. In the OpenQA section, we first guide GPT to compare the LLM's output step by step with the reference answer. This process enables GPT to determine the appropriate score to assign to the LLM's output. Following this evaluation, we prompt GPT a second time to extract the total score generated by the model.

Statistic Information We employ two different terminology sets to classify these two subsets. The distribution of both MCQ questions and OpenQA questions by subject is illustrated in Figures 5 and 6. For MCQs that include image input, the majority pertain to IM (Internal Medicine), Neuro (Neurology), Pulm (Pulmonology), and Surg (Surgery). This trend is attributed to the prevalence of medical imaging examinations conducted in these departments. As for OpenQA, the subjects with the highest response rates include Digestive, Respiratory, Female Reproductive (Fem Reproductive), and Circulatory systems. Overall, regardless of the classification method used, these observations are consistent with distributions in other medical evaluations.

4. Experiments and Discussions

4.1. Experimental Setup

Model Selection We select representative open-source models at 10B levels and API-based models for evaluation. For text-based tasks, we choose the following models: Qwen-2.5-7B-Instruct (Bai et al., 2023), ChatGLM-2-6B (GLM et al., 2024), Llama-3.1-8B-Instruct (Dubey et al., 2024), Vicuna-7B-v1.5 (Zheng et al., 2023), GPT-3.5-Turbo-0125 (Brown, 2020), GPT-4o-mini (Achiam et al., 2023), Gemini-

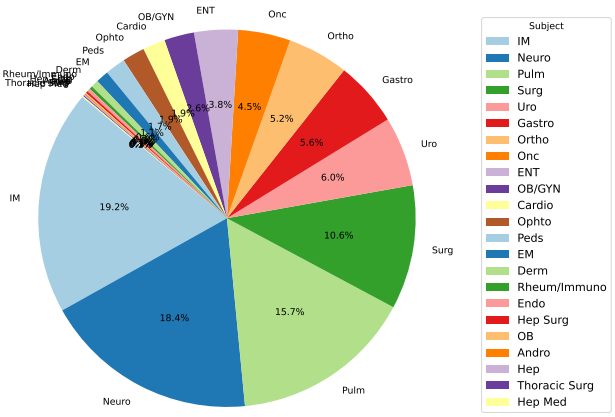

Figure 5: MCQ subject distribution.

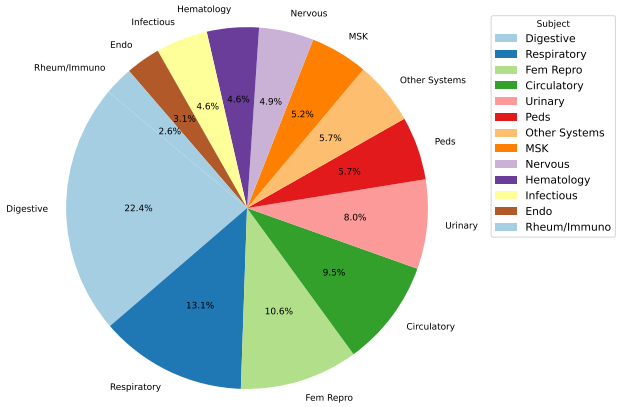

Figure 6: OpenQA subject distribution.

1.5-Flash-8B (Team et al., 2024), PULSE-7Bv5 (Xiaofan Zhang, 2023) and HuatuoGPT (Zhang et al., 2023a). For multimodal tasks, the following models are considered: Qwen2-VL-7B (Bai et al., 2023), InstructBlip-Vicuna-7B (Dai et al., 2023), Llama3.2-11B-Vision (Inc), GPT-4o-mini, LLaVA-v1.6-Vicuna-7B (Liu et al., 2024b), VisualGLM-6B (Du et al., 2022), CogVLM (Ding et al., 2021), and LLaVA-Med-7B (Li et al., 2023). For models that do not support the placement of multiple images, we will concatenate the images vertically.

Evaluation Setting The experimental settings for the two subsets differ significantly, allowing us to explore various dimensions of model performance.

For the MCQ subset, we first assess the zero-shot performance on PatientQA against other well-known multimodal medical benchmarks. This initial evaluation helps us determine whether this subset presents greater difficulty compared to existing benchmarks. Following this, we evaluate the one-shot performance in relation to the zero-shot results of the models, providing insights into how additional context influences performance. To further explore multimodal LLM's behavior, we conduct visualization to analyze the *attention distribution* differences between natural images and medical images.

In the case of the OpenQA subset, we also compare the zero-shot results with other popular text-based medical benchmarks. After establishing the baseline with zero-shot results, we compare the performance of the models in one-shot and Knowledge-Augmented Generation (KAG) settings, using the provided cheat sheets as input. We use two rounds chat to evaluate answers. The first round prompts GPT with reference answers and candidate answers, focusing mainly on step by step score assigning. The second round extract total scores using outputs of the first round. Additionally, we compare our metric with BLEU, ROUGE and METEOR. This comprehensive evaluation allows us to draw meaningful conclusions about the models' performance and their potential for improvement in future research endeavors.

4.2. Results

Baseline Comparison The performance of various models on text-based medical benchmarks is summarized in Figure 7, which includes MedQAcn, PubMedQA, CMB, CMExam, MMLU (med), and MedMCQA, alongside our OpenQA subsets, namely "Ours_collect" for medical history collection task, and "Ours_analysis" for patient summary diagnosis task. It is shown that most of the models score lower on our tasks compared to other benchmarks. Notably, to pass the corresponding medical exam, candidates must achieve over 60% on the OpenQA subsets. This suggests that there is still a room for improvement in apply LLMs for medical-related instructions for providing diagnoses.

For multimodal MCQ, we illustrate the zero-shot performance compared with other popular multimodal medical benchmarks in Figure 8. From the result, we can conclude that, with the exception of the LLaVA model, most models perform only marginally better or worse than a random selection strategy across our benchmarks. In addition, while evaluations like VQA-RAD and PathVQA may show some competitive scores, they still fall short of the expert performance level. The results indicate that the current SOTA multimodal LLM on medical benchmark remains underwhelming, suggesting the challenges faced by existing models in effectively integrating and interpreting multimodal medical information.

Zero-shot, One-shot, KAG and COT For OpenQA, the performance between one-shot and KAG is shown in Figure 9 for medical history collection task and 10 for patient summary diagnosis task. In general, LLMs gain from in context learning. Nearly all of them shows best performence in one-

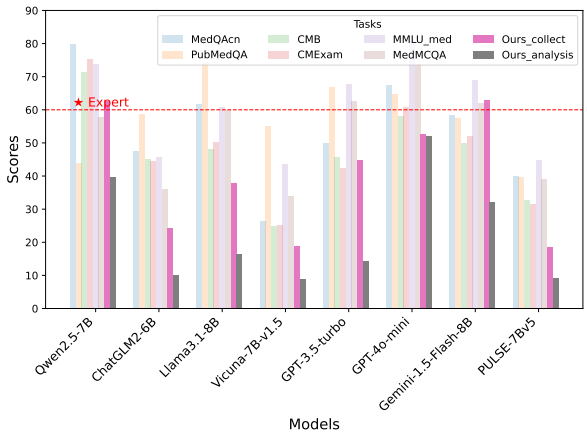

Figure 7: Zero-shot performance of popular text-based medical benchmarks

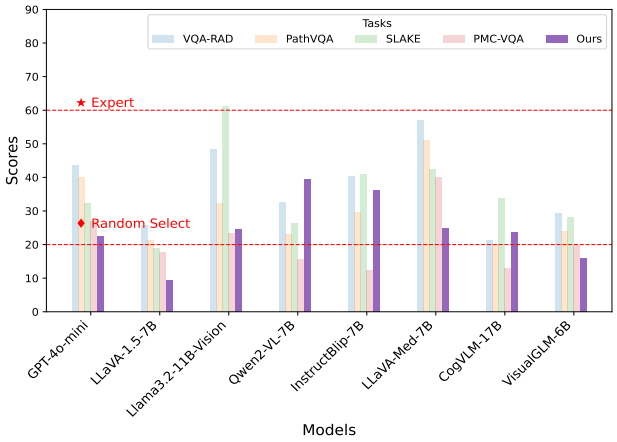

Figure 8: Zero-shot performance of popular multimodal medical benchmarks

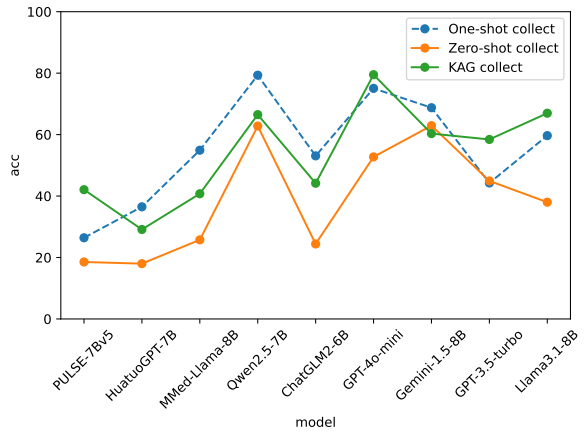

Figure 9: One-shot and KAG performance on medical history collection.

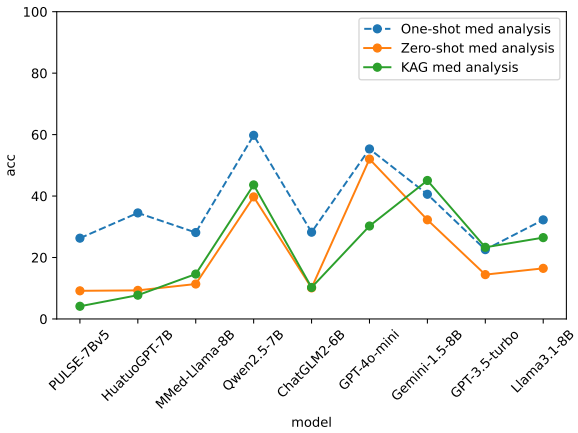

Figure 10: One-shot and KAG performance on summary diagnosis.

shot setting compared with zero-shot and KAG setting in both text generation task, which is consistent with previous research (Brown, 2020), as one-shot help models better formulate answers following given reference, and recall related medical knowledge. However, even the best models struggle to achieve scores above 60 in patient summary diagnosis task, indicating that they still lag behind human doctors in exam performance. We also find that cheat sheet usually helps models do better in medical history collection task, but there shows no obvious sign that it would help the same in summary diagnosis task. The unstable performance for LLMs when providing them with KAG content to do diagnosis, is because our KAG content has very long text in the format of cheat sheet, which may exceeds the context length of the models, therefore result in LLM undefined behavior. However, students in open book exam are often expected to make good use of cheat sheet, as they can quickly identify and concentrate on relevant paragraphs based on the given questions, therefore help them formulate their answers.

For MCQ, we compare model performance using zero-shot, one-shot, and chain-of-thought (COT) settings, as illustrated in Figure 11. It is evident that, among all of them, gpt-4o-mini, LLaVA-1.5 and Llama-3.2-Vision gets obvious performance lift using COT, but for the rest, the chain-of-thought strategy yields slightly performance lift compared to both zero-shot and one-shot settings, which is quite different from test results of OpenQA subset. Only GPT-4o-mini and LLaVA-Med benifits from one-shot setting in the MedVQA test. This implies that current multimodal LLMs are still facing problem of learning from given examples in the medical domain. Moreover, although Llama3.2-11B-Vision has the highest number of parameters, it does not consistently outperform Qwen2-VL, InstructBlip, and LLaVA-Med across all settings.

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE	ROUGE-L	METEOR	Ours
Pulse-7bv5	0.1331	0.0567	0.0251	0.0137	0.1733	0.1627	0.1149	9.16
HuatuoGPT-7B	0.1800	0.0810	0.0381	0.0191	0.1966	0.1856	0.1393	9.3
MMed-Llama-8B	0.1114	0.0422	0.0153	0.0069	0.1485	0.1359	0.1014	11.37
Qwen2.5-7B-instruct	0.1717	0.0652	0.0245	0.0118	0.1629	0.1525	0.1426	39.73
ChatGLM-2-6B	0.1991	0.0981	0.0497	0.0278	0.2456	0.2295	0.1925	10.11
GPT-4o-mini	0.1818	0.0669	0.0240	0.0111	0.1659	0.1551	0.1471	52.04
Gemini-1.5-flash-8b	0.1720	0.0677	0.0239	0.0108	0.1765	0.1650	0.1655	32.27
GPT-3.5-turbo-0125	0.1069	0.0391	0.0141	0.0071	0.1502	0.1407	0.0981	14.43
Llama3.1-8B-instruct	0.1697	0.0683	0.0281	0.0123	0.1674	0.1575	0.1286	16.47

Table 2: Model performance comparison using different metrics on OpenQA task2. Best results for each metric are highlighted in lightgray.

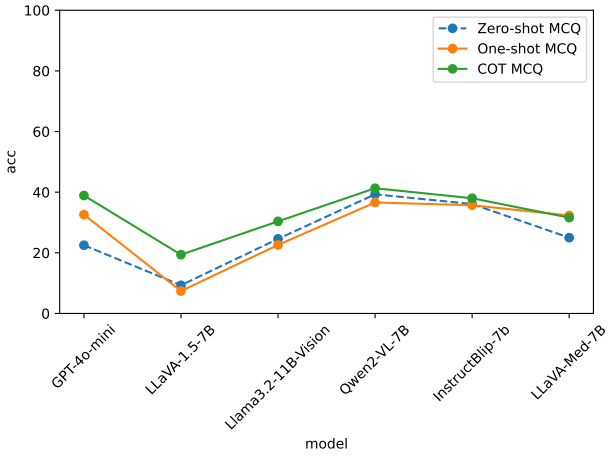

Figure 11: Model performance on multimodal MCQ.

Metric Study We also calculate the BLEU, ROUGE, METEOR and Ours metrics using zero-shot setting with answers translated into English, as shown in Table 2. From the table we can find that ChatGLM-2-6B achieve the highest BLEU and ROUGE sores, but using Ours metric it only gets 10.11 out of 100, which is far worse than GPT-4o-mini getting 52.04 out of 100. This can be because ChatGLM tends to generate more words repeating from past information, while GPT-4o-mini focus more on diagnosis itself. As BLEU, ROUGE and METEOR metrics are for NLP tasks mostly, we suppose it may not be able to find the best answers nor accurately reflect best model performance in measuring long context generation task.

Attention Visualization To investigate why the performance of multimodal LLMs is generally not satisfying, we conducte an attention visualization experiment using the VLM-Visualizer tool (Zhang, 2024) to extract attention during token generation process. We only prompt the model one sentence "Describe the image, paying attention to any special details." and observe the results.

Firstly, we display attention maps over images during token generation in Figure 13, which gives an example of how LLaVA focuses differently between natural images and medical images. For natural images, it always can focus on key object on the image, such as the eye of a dog, the leather in the background, etc. But for medical images, LLaVA sometimes is confusing and cannot focus as detailed as natural image, therefore only output vague description of given medical images.

To investigate the different behavior when facing between natural images and medical images, we further conduct attention vectors t-SNE analysis. For natural images, we randomly selected 1,000 images from the ImageNet-1K (Deng et al., 2009) test set, while for medical images, we used images in the PatientQA MCQ subset. The t-SNE of attention vectors figure across two different groups of images is illustrated in Figure 12.

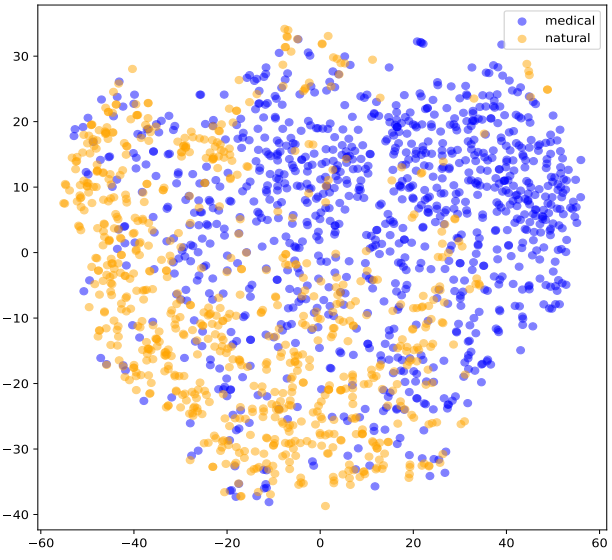

Figure 12: T-SNE attention visualization for visual tokens of natural images and medical images.

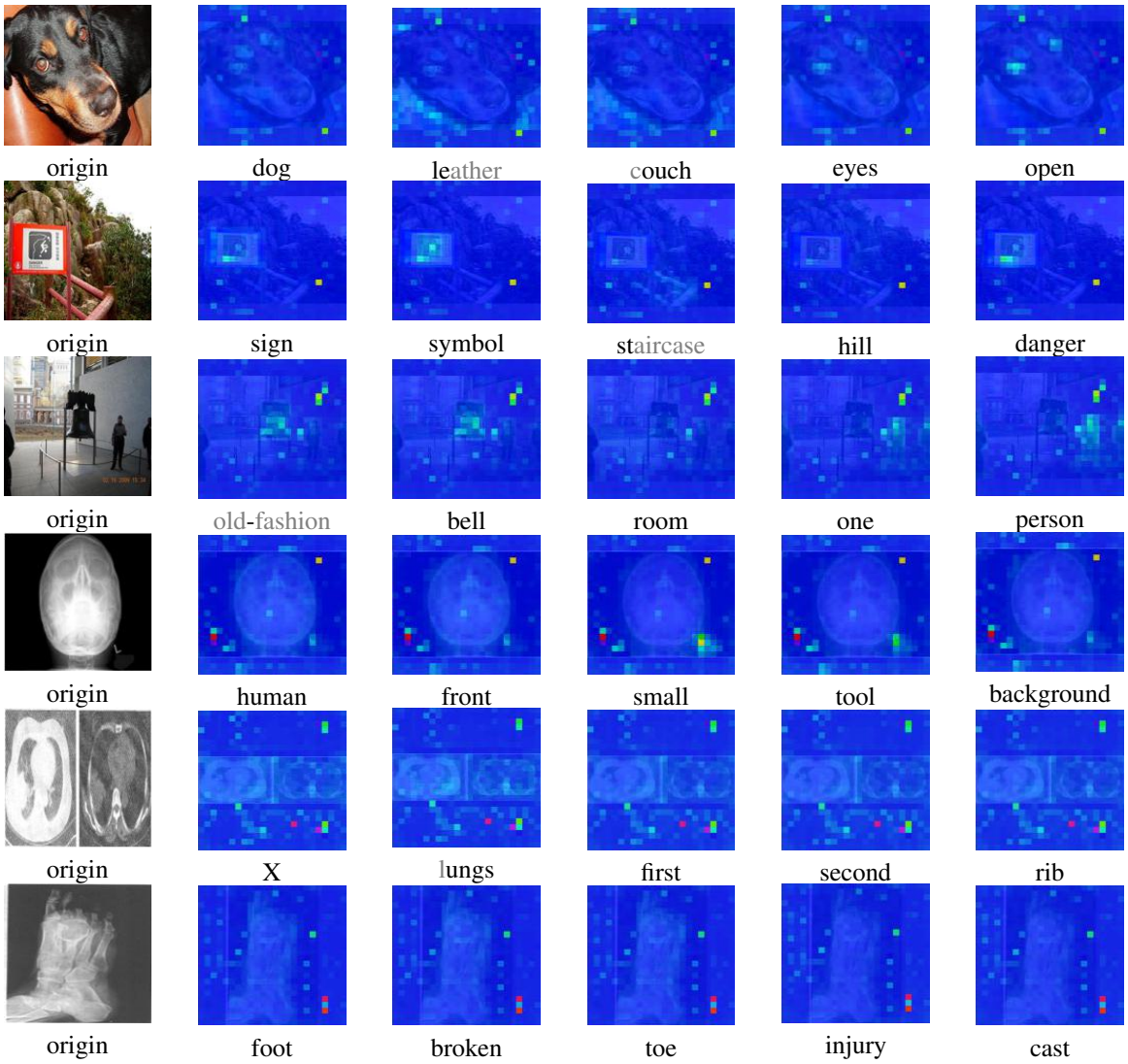

Figure 13: Attention map illustration of LLaVA in different scenes. We use gray letter to indicate tokens surrounding the generated ones for clarity.

As observed, there is an obvious boundary in the attention vectors distribution across the two groups. This may just reflect the behavior of LLaVA describing many details in natural images but mostly vague attention in medical images. As describing medical findings necessitates specialized knowledge, we believe that Figure 12 indicates the need for new developments in medical large vision-language models (LVLMs), including enhancements in datasets or architectural approaches.

5. Conclusions

In this paper, we introduce PatientQA as a complementary to existing medical benchmarks. The OpenQA subset presents a step-by-step reasoning approach for evaluating long-generated diagnostic texts, and MCQ subset provides a multimodal version of patient diagnosis.

Experiments on OpenQA subset show that LLMs still struggle to achieve expert level scores, and long cheat sheet may just disorient them. Metric comparison shows that when evaluating long context OpenQA in knowledge specific domain, we may consider step-by-step reasoning rating scheme other than NLP metrics. Experiments on MCQ reveal that LLMs are still not able to earn over 60% in medical VQA benchmarks, nor focus on medical images like they do on natural images.

Our findings underscore the complexities involved in evaluating and enhancing model performance in medical contexts, paving the way for future research and advancements in this critical field.

References

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

Anthropic, A. Claude 3.5 sonnet model card addendum. *Claude-3.5 Model Card*, 2024.

Bai, J., Bai, S., Chu, Y., Cui, Z., Dang, K., Deng, X., Fan, Y., Ge, W., Han, Y., Huang, F., Hui, B., Ji, L., Li, M., Lin, J., Lin, R., Liu, D., Liu, G., Lu, C., Lu, K., Ma, J., Men, R., Ren, X., Ren, X., Tan, C., Tan, S., Tu, J., Wang, P., Wang, S., Wang, W., Wu, S., Xu, B., Xu, J., Yang, A., Yang, H., Yang, J., Yang, S., Yao, Y., Yu, B., Yuan, H., Yuan, Z., Zhang, J., Zhang, X., Zhang, Y., Zhang, Z., Zhou, C., Zhou, J., Zhou, X., and Zhu, T. Qwen technical report, 2023. URL https://arxiv.org/abs/2309.16609.

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

Brown, T. B. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

Dai, W., Li, J., Li, D., Tiong, A. M. H., Zhao, J., Wang, W., Li, B., Fung, P., and Hoi, S. Instructblip: Towards general-purpose vision-language models with instruction tuning, 2023. URL https://arxiv.org/abs/2305.06500.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Ding, M., Yang, Z., Hong, W., Zheng, W., Zhou, C., Yin, D., Lin, J., Zou, X., Shao, Z., Yang, H., et al. Cogview: Mastering text-to-image generation via transformers. *Advances in Neural Information Processing Systems*, 34: 19822–19835, 2021.

Du, Z., Qian, Y., Liu, X., Ding, M., Qiu, J., Yang, Z., and Tang, J. Glm: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 320–335, 2022.

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

GLM, T., Zeng, A., and et al., B. X. Chatglm: A family of large language models from glm-130b to glm-4 all tools, 2024.

He, X., Cai, Z., Wei, W., Zhang, Y., Mou, L., Xing, E., and Xie, P. Towards visual question answering on pathology images. In Zong, C., Xia, F., Li, W., and Navigli, R. (eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pp. 708–718, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-short.90. URL https://aclanthology.org/2021.acl-short.90.

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., and Steinhardt, J. Measuring massive multitask language understanding, 2021. URL https://arxiv.org/abs/2009.03300.

Inc, M. meta-llama/llama-3.2-11b-vision. https://huggingface.co/meta-llama/Llama-3.2-11B-Vision.

Ionescu, B., Müller, H., Drăgulinescu, A.-M., Yim, W.-W., Ben Abacha, A., Snider, N., Adams, G., Yetisgen, M., Rückert, J., García Seco de Herrera, A., et al. Overview of the imageclef 2023: Multimedia retrieval in medical, social media and internet applications. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pp. 370–396. Springer, 2023.

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

Jin, Q., Dhingra, B., Liu, Z., Cohen, W., and Lu, X. Pubmedqa: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2567–2577, 2019.

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

Li, C., Wong, C., Zhang, S., Usuyama, N., Liu, H., Yang, J., Naumann, T., Poon, H., and Gao, J. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *arXiv preprint arXiv:2306.00890*, 2023.

Liu, B., Zhan, L.-M., Xu, L., Ma, L., Yang, Y., and Wu, X.-M. Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In *2021*

IEEE 18th International Symposium on Biomedical Imaging (ISBI), pp. 1650–1654. IEEE, 2021.

Liu, G., He, J., Li, P., He, G., Chen, Z., and Zhong, S. Pefomed: Parameter efficient fine-tuning of multimodal large language models for medical imaging. *arXiv e-prints*, pp. arXiv–2401, 2024a.

Liu, H., Li, C., Li, Y., and Lee, Y. J. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26296–26306, 2024b.

Liu, J., Zhou, P., Hua, Y., et al. Benchmarking large language models on cmexam–a comprehensive chinese medical exam dataset. published online june 8, 2023. *Advances in Neural Information Processing Systems*, 36.

Nori, H., Lee, Y. T., Zhang, S., Carignan, D., Edgar, R., Fusi, N., King, N., Larson, J., Li, Y., Liu, W., et al. Can generalist foundation models outcompete special-purpose tuning? case study in medicine. *arXiv preprint arXiv:2311.16452*, 2023.

Pal, A., Umapathi, L. K., and Sankarasubbu, M. Medmcqa : A large-scale multi-subject multi-choice dataset for medical domain question answering, 2022. URL https://arxiv.org/abs/2203.14371.

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

Singhal, K., Tu, T., Gottweis, J., Sayres, R., Wulczyn, E., Hou, L., Clark, K., Pfohl, S., Cole-Lewis, H., Neal, D., et al. Towards expert-level medical question answering with large language models. *arXiv preprint arXiv:2305.09617*, 2023.

Team, G., Anil, R., Borgeaud, S., Alayrac, J.-B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A. M., Hauth, A., Millican, K., et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

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

Wang, X., Chen, G., Dingjie, S., Zhiyi, Z., Chen, Z., Xiao, Q., Chen, J., Jiang, F., Li, J., Wan, X., Wang, B., and Li, H. CMB: A comprehensive medical benchmark in Chinese. In Duh, K., Gomez, H., and Bethard, S. (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 6184–6205, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.343. URL https://aclanthology.org/2024.naacl-long.343.

Xiaofan Zhang, Kui Xue, S. Z. Pulse: Pretrained and unified language service engine. 2023. URL https://github.com/openmedlab/PULSE.

Xie, Y., Zhou, C., Gao, L., Wu, J., Li, X., Zhou, H.-Y., Liu, S., Xing, L., Zou, J., Xie, C., et al. Medtrinity-25m: A large-scale multimodal dataset with multigranular annotations for medicine. *arXiv preprint arXiv:2408.02900*, 2024.

Yan, X., Li, L., Xie, C., Xiao, J., and Gu, L. Zhejiang university at imageclef 2019 visual question answering in the medical domain. *CLEF (working notes)*, 85, 2019.

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

Zhang, H., Chen, J., Jiang, F., Yu, F., Chen, Z., Li, J., Chen, G., Wu, X., Zhang, Z., Xiao, Q., et al. Huatuogpt, towards taming language model to be a doctor. *arXiv preprint arXiv:2305.15075*, 2023a.

Zhang, J. Vlm-visualizer. GitHub, 2024. URL https://github.com/zjysteven/VLM-Visualizer.

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

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E., et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36: 46595–46623, 2023.