Converting Open-ended Questions to Multiple-choice Questions Simplifies Biomedical Vision-Language Model Evaluation

Abstract

Vision-language models (VLMs) show promise in medicine, but their evaluation remains challenging due to their open-ended nature. Current metrics often fail to capture nuances in human judgment, while model-based evaluations are computationally expensive and unstable. We propose converting open-ended questions into multiple-choice format to address these limitations. Using an agent-based framework with GPT-4, we transform questions through iterative refinement. Our results demonstrate strong correlation between multiple-choice and open-ended performance across three datasets. We evaluate 18 models on these converted datasets, showing improved capability discrimination. Case studies illustrate our approach's success where rule-based evaluations fail. This work contributes a novel evaluation framework, aiming to enable easier and more consistent VLM evaluation in medicine.

Keywords: Vision-language models, model evaluation, multiple-choice questions

Data and Code Availability We use publicly available datasets: SLAKE, VQA-RAD, PathVQA. We will make code and data publicly available.

Institutional Review Board (IRB) Our research does not require IRB approval.

1. Introduction

Vision-language models (VLMs) have emerged as powerful tools in medicine (Zhang et al. (2023)), offering potential to enhance clinical decision-making by integrating visual and textual data. However, evaluating these models poses significant challenges, particularly in capturing the nuanced understanding required in medical contexts.

Current evaluation methods face critical limitations. Quasi-exact match metrics like BLEU (Papineni et al. (2002)) and accuracy often fail to capture semantic nuances, leading to discrepancies between automated and human evaluations. Alternatively, model-based evaluations using large language models like GPT-4 (OpenAI et al. (2024)) offer more flexibility but are computationally expensive and suffer from instability and unfairness across different candidate systems and model versions, undermining long-term comparability (Shen et al. (2023)).

To address these issues, we propose a novel approach: converting open-ended questions into multiple-choice format. This method aligns with established practices in standardized medical testing (e.g., USMLE) and offers several advantages. It provides a more straightforward and cost-effective evaluation, correlates strongly with open-ended question performance, and maintains consistency across evaluations, independent of underlying language model versions.

Our work introduces an agent-based framework utilizing GPT-4 to transform open-ended questions into high-quality multiple-choice questions. Inspired by MoA (Wang et al. (2024)) and self-reflection (Shinn et al. (2024), we create choices through refinement with simulated teacher and student agents to ensure the quality and validity of the resulting questions.

This study contributes a novel benchmark for evaluating VLMs in the medical domain, along with an agent-based framework for generating high-quality multiple-choice questions. We provide empirical evidence demonstrating strong correlation between multiple-choice and open-ended performance across three diverse datasets, offer a comprehensive evaluation of over 10 models on our converted datasets, and present case studies illustrating the effectiveness of our approach compared to rule-based evaluations.

By providing a standardized, efficient evaluation method, this benchmark aims to accelerate advancements in vision-language understanding and its practical applications in medicine. Our approach not only addresses the limitations of current evaluation methods but also paves the way for more reliable and consistent assessment of VLMs in specialized fields.

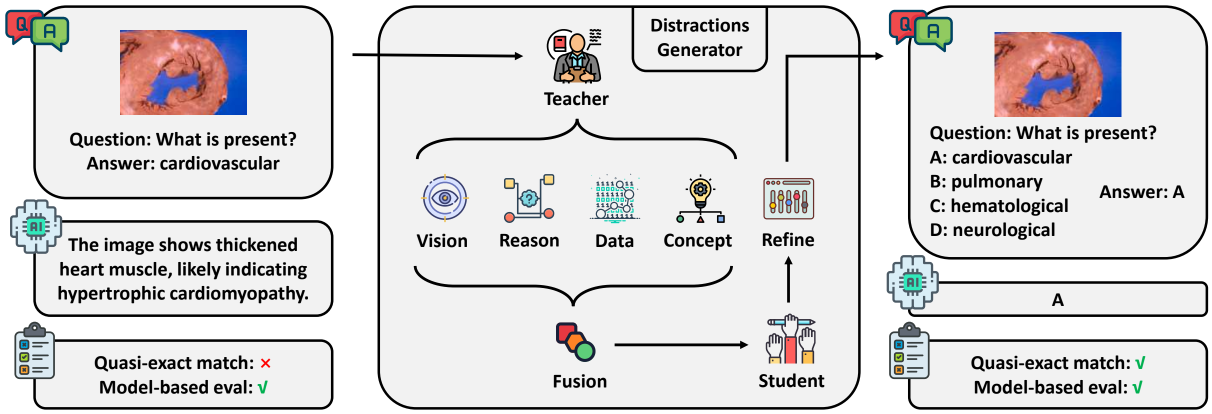

Figure 1: The challenge of open-ended VLM evaluation in medicine and our solution. Left: Traditional metrics and model-based evaluations face issues of inaccuracy, high cost, and instability. Right: Our approach converts open-ended questions to multiple-choice format, offering a more reliable, efficient, and consistent evaluation method.

2. Challenges in Evaluating Open-ended Questions

Evaluating open-ended questions, particularly in specialized domains like medicine, has long challenged natural language processing (Jin et al. (2020)). Traditional quasi-exact match metrics such as BLEU, which measure lexical similarity between ground-truth and model predictions, often fail to capture the nuanced understanding required in medical contexts. For example, two semantically equivalent medical descriptions may receive vastly different BLEU scores due to minimal word overlap, highlighting the inadequacy of quasi-exact match metrics in this domain.

Recent research has shown that model-based evaluation using large language models (LLMs) like GPT-4 strongly correlates with human expert judgments (Liu et al. (2023)). To quantify the limitations of traditional metrics, we conducted an experiment comparing BLEU scores with model-based evaluation scores for VLM-generated medical descriptions. The results revealed a near-zero correlation between BLEU and the model-based scores, which serve as a proxy for human evaluation (Figure 3). This finding underscores the severity of the problem with quasi-exact match metrics in specialized domains.

However, model-based evaluation, despite its promise, faces significant challenges. The computational cost of using LLMs for evaluation is substantial, especially for large datasets. Moreover, updates to the underlying LLM can lead to inconsistent evaluations over time, a phenomenon known as version instability (Figure 2). These issues complicate long-term comparisons and benchmarking efforts.

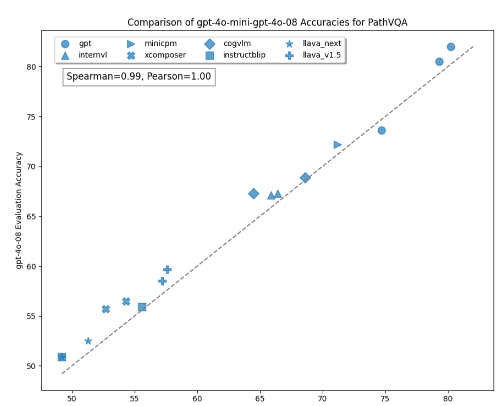

Figure 2: While model-based eval has a strong correlation across time, the absolute numbers changed substantially, making it hard to compare methods against time. Moreover, it costs more than $50 to run this evaluation.

Our findings highlight a crucial dilemma in evaluating open-ended responses in specialized domains: traditional metrics are inadequate, but more effective

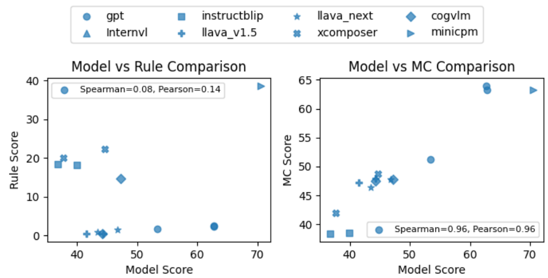

Figure 3: Comparison of the correlation between different evaluating metric. We have zero correlation between quasi-exact match metric BLEU and GPT-4o model-based evaluation reveals severe problems of quasi-exact match metric. However, the correlation between model-based evaluation and multi-choice score is much higher.

model-based evaluations are both resource-intensive and potentially inconsistent. This situation underscores the urgent need for new evaluation methodologies that can offer the semantic understanding of model-based approaches while maintaining consistency and computational efficiency.

3. Our Solution: Convert Open-ended Questions to Multi-choice Questions

To address the limitations of both quasi-exact match metrics and model-based evaluation, we propose a novel approach: converting open-ended questions to multiple-choice format for VLM evaluation in the medical domain.

The biggest challenge of question conversion is to generate challenging yet correct distractors for each question. Our method employs an agent-based framework using GPT-4 to transform open-ended questions into high-quality multiple-choice questions (MCQs).

The process begins with a medical teacher model generating questions based on images, questions, and answers. Four specialized generation agents are involved: a visual interpretation agent focusing on medical image comprehension, a reasoning agent concentrating on logical deduction, a data processing agent specializing in data handling, and a concept agent focusing on medical concepts. Each agent generates 9 distractors, resulting in a total of 36 distractors. A fusion agent then selects 9 of these distractors.

Next, a simulated high-performing medical student attempts to answer these questions. Based on the student's responses, a medical refinement agent selects and modifies the three best distractors along with the correct option to form the final MCQ.

4. Discussion

We have verified that the agent-based conversion approach outperforms naive methods, such as simply generating three distractors. This approach presents a greater challenge, as evidenced by the generally lower performance of various VLMs, as shown in Figure 4.

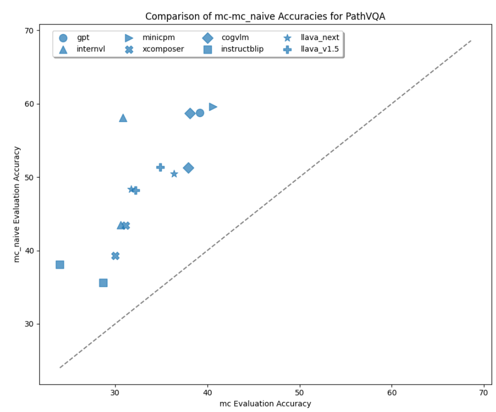

Figure 4: Our agent-based multi-choice conversion generates more challenging questions compared to the naive approach.

To validate our approach, we conducted a similar correlation study. We evaluated the performance of VLMs on both the original open-ended questions and our converted MCQs for each dataset, then calculated the Spearmann and Pearson correlation coefficient (Hauke and Kossowski (2011)) between the scores.

The results showed strong correlations across all datasets. These high correlations suggest that our MCQ format effectively preserves the discriminative power of the original open-ended questions, offering a promising solution to the challenges of evaluating VLMs in specialized domains like medicine (Figure 3). We present case studies where rule-based evaluation fails but both model-based and our MCQ

Model	Quasi-exact Match			Model-based Eval			Multi-choice Eval		
	VQA-RAD	SLAKE	PathVQA	VQA-RAD	SLAKE	PathVQA	VQA-RAD	SLAKE	PathVQA
GPT4o	2.4	8.97	1.56	62.8	75.0	28.4	63.3	72.9	48.3
GPT4o-mini	1.6	2.76	0.87	53.4	60.6	30.8	51.2	65.8	39.2
GPT4o-HIGH	2.42	9.2	1.47	62.8	75.8	30.4	64.0	74.1	51.3
Mini-InternVL-Chat-2B-V1.5	6.53	37.05	1.48	49.5	57.5	9.0	52.0	59.9	30.6
Mini-InternVL-Chat-4B-V1.5	8.06	41.83	1.65	57.1	64.7	9.8	57.5	56.8	30.8
MiniCPM-Llama3-V-2.5	38.72	22.74	2.09	70.6	66.5	14.3	63.3	61.1	40.6
XComposer2	22.41	34.43	1.92	44.6	61.0	12.3	48.7	64.0	31.1
XComposer2_1.8b	20.08	10.4	1.01	37.7	54.4	10.0	41.9	57.0	30.0
cogvlm-chat	14.66	4.9	2.06	47.3	66.0	12.9	47.7	60.9	38.1
cogvlm2-llama3-chat-19B	0.38	0.2	0.31	44.3	57.2	25.8	47.5	55.5	37.9
instructblip_13b	18.43	0.47	1.67	36.8	56.4	12.8	38.3	38.2	24.0
instructblip_7b	18.28	0.54	1.6	40.0	62.3	13.2	38.5	52.5	28.7
llava-internlm2-20b	-	44.06	0.1	-	64.9	12.8	46.5	62.9	34.6
llava_next_vicuna_13b	1.43	44.64	0.19	46.8	63.7	12.3	47.7	47.9	36.4
llava_next_vicuna_7b	0.91	44.41	0.25	43.4	61.0	13.9	46.4	54.2	31.7
llava_next_yi_34b	-	46.52	0.6	-	67.5	14.2	-	65.7	40.3
llava_v1.5_13b	0.46	44.98	0.07	44.3	63.5	12.9	47.9	59.8	34.9
llava_v1.5_7b	0.35	42.42	0.09	41.5	58.4	13.4	47.2	60.3	32.2

Table 1: Summary of Model Evaluations Across Different Datasets and Evaluation Metrics. We use different model from various families, including GPT4o (OpenAI et al. (2024)), InternVL (Chen et al. (2024)), MiniCPM-V (Yao et al. (2024)), InternLM-XComposer2 (Dong et al. (2024)), CogVLM (Hong et al. (2024)), InstructBlip (Dai et al. (2023)), LLaVA-v1.5 and LLaVA-NeXT (Liu et al. (2024))

Aspect	Content
Question	Why does this image show kidney, thickened and hyalinized basement membranes?
Model Answer	Diabetic nephropathy
Reference Answer	due to diabetes mellitus pas
Rule-based Evaluation	Score: 0.0. Explaination: Low similarity score due to different phrasing.
Model-based Evaluation	Score: 1.0. Explanation: Though the phrasing is different, the model answer successfully captures the underlying meaning.
MCQ Version	Why does this image show kidney, thickened and hyalinized basement membranes? A) due to hypertensive nephrosclerosis with fibrotic changes B) due to membranous nephropathy with subepithelial deposits C) due to amyloidosis with Congo red stain D) due to diabetes mellitus pas
MCQ Evaluation	Score: 1.0. Explaination: Correctly identifies the model's understanding.

Table 2: Case Study: Semantic Equivalence in Medical Evaluation

Dataset	MC-Model		Rule-Model	
	Spearman	Pearson	Spearman	Pearson
SLAKE	0.55	0.66	0.28	0.01
VQA-RAD	0.96	0.96	0.08	0.14
PathVQA	0.70	0.74	-0.21	-0.03

Table 3: Correlation coefficients for different datasets. Our conversion methods apparently get higher correlation in all datasets.

approach succeed in Table 2. We present results on three commonly-used medical VLM evaluation benchmarks in Table 1.

In this work, we proposed a novel approach to evaluating visual language models (VLMs) in the medical domain by converting open-ended questions to multiple-choice format. Our method addresses the limitations of both quasi-exact match metrics and model-based evaluation, offering a balanced solution that combines accuracy, consistency, and efficiency.

Our experiments across diverse medical datasets demonstrated strong correlations between MCQ and open-ended question performance, validating the effectiveness of our approach. The MCQ format captures nuanced distinctions in medical knowledge, provides consistent results independent of LLM versions, and enables efficient, large-scale model assessment.

While this method may not fully capture a model's capacity for creative responses and requires regular updates to the question bank, it represents a significant advancement in VLM evaluation for specialized domains. Future work could explore incorporating creativity assessment, developing efficient updating mechanisms, and extending this approach to other fields.

References

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, Ji Ma, Jiaqi Wang, Xiaoyi Dong, Hang Yan, Hewei Guo, Conghui He, Botian Shi, Zhenjiang Jin, Chao Xu, Bin Wang, Xingjian Wei, Wei Li, Wenjian Zhang, Bo Zhang, Pinlong Cai, Licheng Wen, Xiangchao Yan, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites, 2024. URL https://arxiv.org/abs/2404.16821.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning, 2023. URL https://arxiv.org/abs/2305.06500.

Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Bin Wang, Linke Ouyang, Xilin Wei, Songyang Zhang, Haodong Duan, Maosong Cao, Wenwei Zhang, Yining Li, Hang Yan, Yang Gao, Xinyue Zhang, Wei Li, Jingwen Li, Kai Chen, Conghui He, Xingcheng Zhang, Yu Qiao, Dahua Lin, and Jiaqi Wang. Internlm-xcomposer2: Mastering free-form text-image composition and comprehension in vision-language large model, 2024. URL https://arxiv.org/abs/2401.16420.

Jan Hauke and Tomasz M. Kossowski. Comparison of values of pearson's and spearman's correlation coefficients on the same sets of data. 2011. URL https://api.semanticscholar.org/CorpusID:5311856.

Wenyi Hong, Weihan Wang, Ming Ding, Wenmeng Yu, Qingsong Lv, Yan Wang, Yean Cheng, Shiyu Huang, Junhui Ji, Zhao Xue, Lei Zhao, Zhuoyi Yang, Xiaotao Gu, Xiaohan Zhang, Guanyu Feng, Da Yin, Zihan Wang, Ji Qi, Xixuan Song, Peng Zhang, Debing Liu, Bin Xu, Juanzi Li, Yuxiao Dong, and Jie Tang. Cogvlm2: Visual language models for image and video understanding, 2024. URL https://arxiv.org/abs/2408.16500.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *ArXiv*, abs/2009.13081, 2020. URL https://api.semanticscholar.org/CorpusID:221970190.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2024. URL https://arxiv.org/abs/2310.03744.

Yang Liu, Dan Iter, Yichong Xu, Shuo Wang, Ruochen Xu, and Chenguang Zhu. G-eval: Nlg evaluation using gpt-4 with better human alignment. In *Conference on Empirical Methods in Natural Language Processing*, 2023. URL https://api.semanticscholar.org/CorpusID:257804696.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. Gpt-4 technical report, 2024. URL https://arxiv.org/abs/2303.08774.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

Chenhui Shen, Liying Cheng, Yang You, and Lidong Bing. Large language models are not yet human-level evaluators for abstractive summarization. In *Conference on Empirical Methods in Natural Language Processing*, 2023. URL https://api.semanticscholar.org/CorpusID:258833685.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36, 2024.

Junlin Wang, Jue Wang, Ben Athiwaratkun, Ce Zhang, and James Zou. Mixture-of-agents enhances large language model capabilities. *arXiv preprint arXiv:2406.04692*, 2024.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, Qianyu Chen, Huarong Zhou, Zhensheng Zou, Haoye Zhang, Shengding Hu, Zhi Zheng, Jie Zhou, Jie Cai, Xu Han, Guoyang Zeng, Dahai Li, Zhiyuan Liu, and Maosong Sun. Minicpm-v: A gpt-4v level mllm on your phone, 2024. URL https://arxiv.org/abs/2408.01800.

Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. Vision-language models for vision tasks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46:5625–5644, 2023. URL https://api.semanticscholar.org/CorpusID:257913547.

Appendix A. 20 Questions for Dataset

Questions	A	B	C	D	Answers
what is also seen in the wall?	calcified lymph node	a cystic lesion	a calcified nodule	a partly formed unerupted tooth	D
what does this image show?	granuloma	granulomatous inflammation	vasculitis	fibromuscular dysplasia	C
what does this image show?	spleen	kidney	adrenal gland	pancreas	B
how does this image show childs hands?	with obvious clubbing	with signs of Raynaud's phenomenon	with splinter hemorrhages	with onycholysis	A
what is present?	cranial artery	neural ganglion	venous sinus	spinal artery	A
what is present?	musculoskeletal	lymphatic tissue	vascular	hematologic	D
what shows failure of normal differentiation, marked nuclear and cellular pleomorphism, and numerous mitotic figures extending toward the surface?	high-power view of a benign lesion	reactive atypia	high-power view of another region	chronic inflammation	C
where is this?	kidney	liver	pancreas	heart	D
what are present?	extremities	limbs	growths	appendages	A
what does this image show?	skin	connective tissue	blood vessels	vascular system	A
where is this?	maxillary	mandibular	oral	zygomatic	C
what is present?	urinary	endocrine	lymphatic	gastrointestinal	D
what is multilobulated with increased fat while lower part of the image shows a separate encapsulated gelatinous mass?	lipoma	lobulated mass	cystic component	main mass	D
what does this image show?	chronic fibrotic infarct	chronic infarct with complete fibrotic transformation	large hemorrhagic infarct with significant fibrosis	large and typically shaped old infarct but far from fibrotic	D
what is present?	cardiovascular	nervous system	nervous	musculoskeletal	A
what is present?	pancreatic acini	adrenal cortex	cardiovascular	hepatobiliary	D
what does this image show?	kidney	lymph node	lung	pancreas	C
what is present?	vascular	epithelial	exocrine	endocrine	D
what is present?	peritoneal fluid	pleural effusion	ascitic fluid	cerebrospinal fluid	A
what is present?	dysplasia	adenoma	hyperplasia	neurofibroma	B

Table 4: 20 MCQ of PathVQA