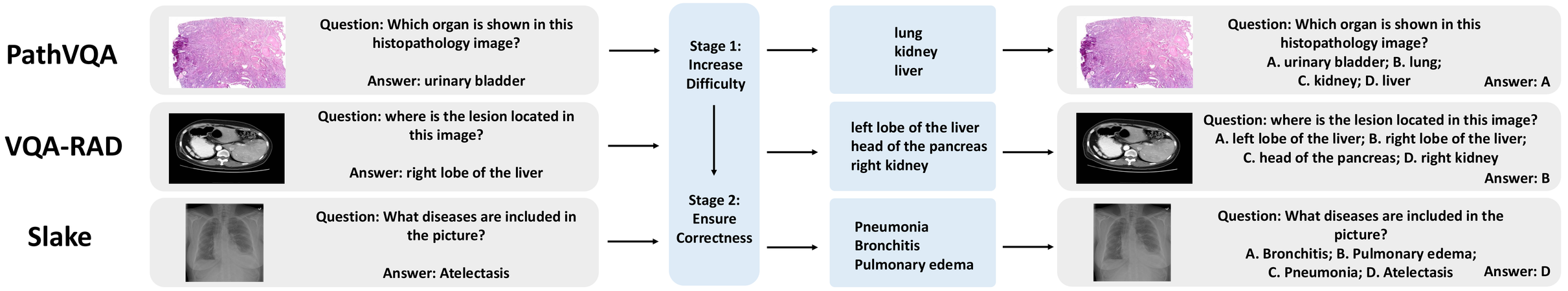
Converting Open-ended Questions to Multiple-choice Questions Simplifies Biomedical Vision-Language Model Evaluation

Yuchang Su, Yuhui Zhang, Yiming Liu, Ludwig Schmidt, Serena Yeung-Levy

Motivation: Solve the fundamental problem in Biomedical VLM Evaluation

What are we doing?

Given an open-ended format question, answer, and corresponding image, output three challenge distractors. Then, combine question, answer and distrators as a multiple-choice format question



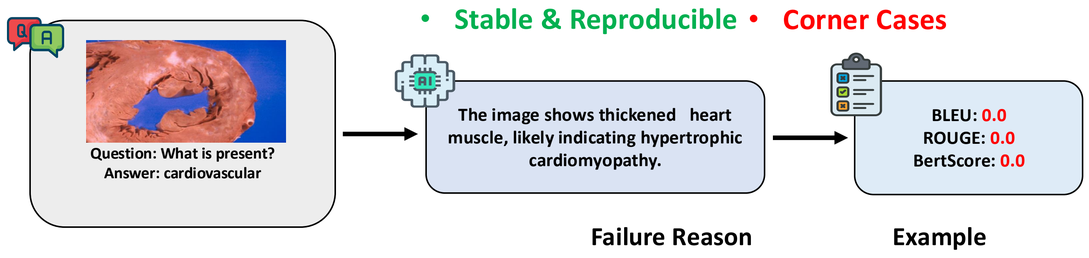
Why are we doing?

We revisit open-ended medical question evaluation, finding critical limitations of existing methods.

Rule-based Evaluation

Pros:
- Efficient & Cheap
- Stable & Reproducible

Cons:
- Low Correlation with Human
- Corner Cases



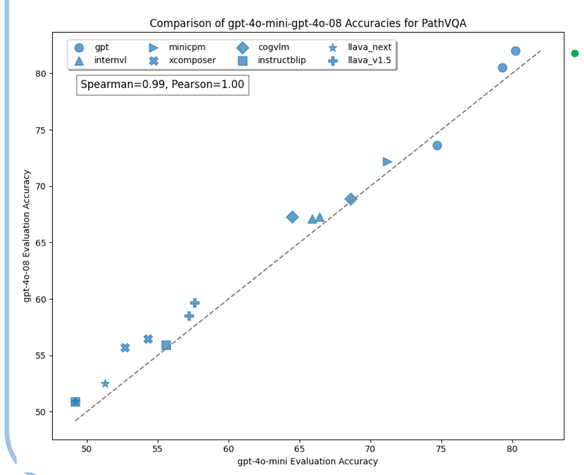
Quasi-exact is the most commonly used evaluation metric (BLEU, ROUGE, METEOR, BERTScore). However, it demonstrates enormous errors due to multiple reasons.

Failure Reason	Example
Ambiguous Questions	*Q: What is present? A: …*
Verbosity	*A: Atelectasis / The disease in the picture is Atelectasis*
Synonym	*A: cardiovascular / hypertrophic cardiomyopathy*
……	……

Model-based Evaluation

Pros:
- High Correlation with Human
- Ranking Stability

Cons:
- Time-consuming & Expensive
- Absolute Score Change
- Hard to reproduce



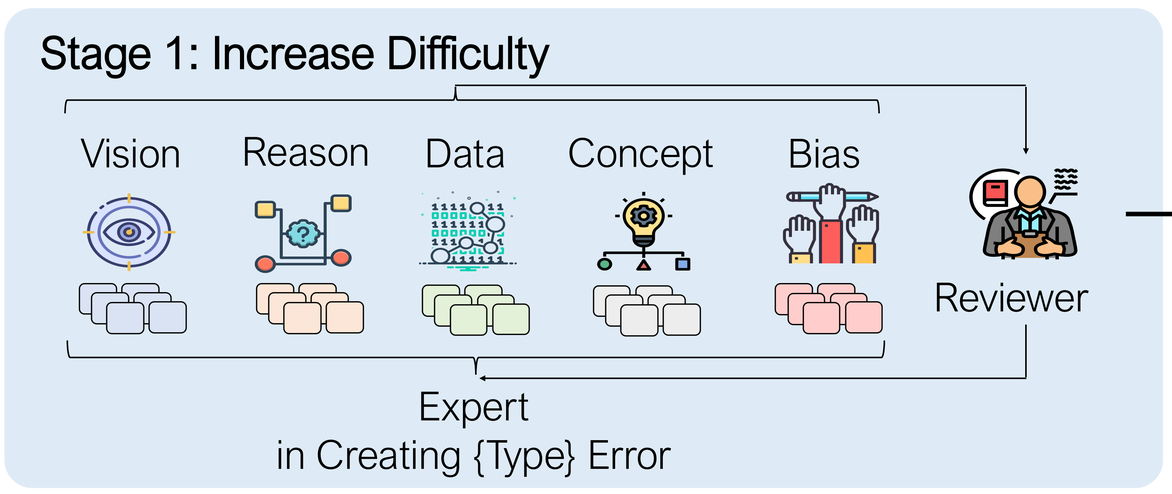
Using strong VLM(e.g.GPT-4o) as a judge

Cost: $50 for 20 model evaluations on 500 questions

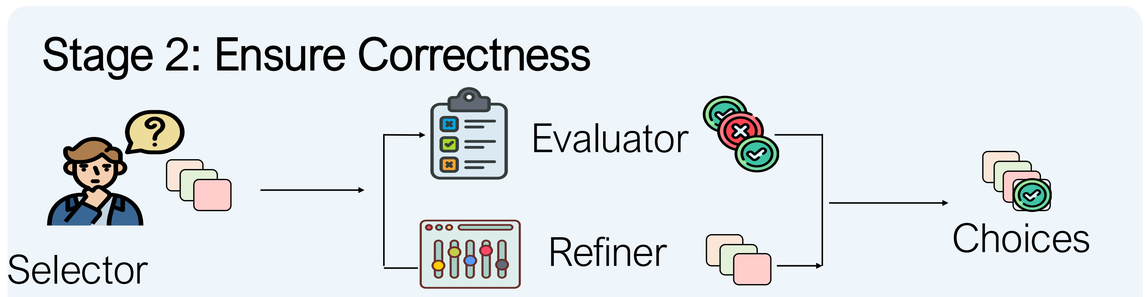
Time: Evaluation efficiency is limited due to model inference

Instability: Absolute score change for different GPT versions, e.g. 56% on GPT-4o-0816 and 53% on GPT-4o-mini

How are we doing it?



In stage 1, we define 5 different types of errors and design expert agents to create distractors based on these types. Each expert creates 6 distractors. Then, the reviewer agent comments on each distractor, and the expert develops the distractors based on these comments.



In stage 2, we first use a selector to identify the 3 hardest distractors from a pool of 30 choices. Next, we employ an evaluator to determine their accuracy and refine them iteratively. Finally, once the correctness score of the distractors meets the specified threshold, we will output the final selections.

How good can we do it?

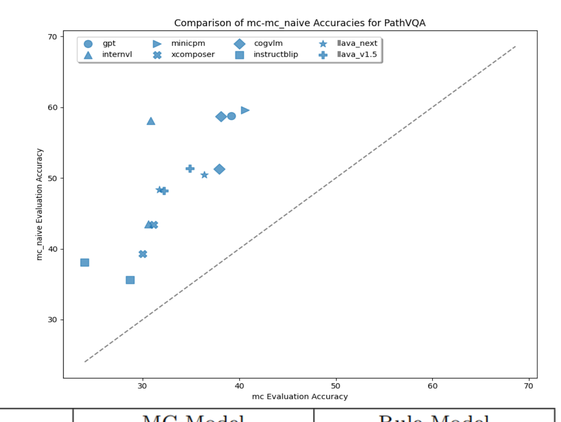
- Convert 3 open-ended datasets (VQA-RAD, PathVQA, Slake) to multiple-choice format
- Evaluate 18 models on open-ended and multiple-choice questions
- Conduct experiments on correlation and difficulty

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

Difficulty

Compare naïve version generation (“create 3 distractors for this question”) and our agentic pipeline generation. Use model accuracy as the metric.

Acc (Naïve) **>** Acc (Ours)



Correlation

Use model-based score as a substitute of human evaluation score. Compare correlation between model and rule / multi-choice.

Corr (model, rule)

>

Corr (model, multi-choice)

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