Beyond Multiple Choice: Verifiable OpenQA for Robust Vision-Language RFT

First Author
Institution1
Institution1 address
firstauthor@i1.org

Second Author
Institution2
First line of institution2 address
secondauthor@i2.org

Abstract

*Multiple-choice question answering (MCQA) has been a popular format for evaluating and reinforcement fine-tuning (RFT) of modern multimodal language models. Its constrained output format allows for simplified, deterministic automatic verification. However, we find that the options may leak exploitable signals, which makes the accuracy metrics unreliable for indicating real capabilities and encourages explicit or implicit answer guessing behaviors during RFT. We propose ReVeL (**Re**write and **Ve**rify by **LL**M), a framework that rewrites multiple-choice questions into open-form questions while keeping answers verifiable whenever possible. The framework categorizes questions according to different answer types, apply different rewriting and verification schemes, respectively. When applied for RFT, we converted 20k MCQA examples and use GRPO to finetune Qwen2.5-VL models. Models trained on ReVeL-OpenQA match MCQA accuracy on multiple-choice benchmarks and improve OpenQA accuracy by about six percentage points, indicating better data efficiency and more robust reward signals than MCQA-based training. When used for evaluation, ReVeL also reveals up to 20 percentage points of score inflation in MCQA benchmarks (relative to OpenQA), improves judging accuracy, and reduces both cost and latency. We will release code and data publicly.*

1. Introduction

As large language and multimodal models [2, 3, 8, 11, 24, 25, 27] increasingly tackle diverse real-world tasks, the demand for reliable and scalable evaluation has grown significantly. MCQA is convenient because restricting outputs simplifies scoring [22] across language [14, 34] and vision language benchmarks [13, 15, 16, 43, 44, 46].

However, MCQA departs from real-world usage where answers are usually open-ended [20], while the predefined options encourage selection heuristics [4, 47] rather than genuine understanding. To quantify the unreliability of MCQA for evaluation and verification, we conduct multiple experiments: (1) When options are added to the questions in an open-form benchmark, the accuracy metrics can be greatly boosted; (2) In MCQA benchmarks, When the ground-truth option is perturbed, or replaced with 'None of the above', model behavior degrades. These patterns indicate that the MCQA metrics are heavily dependent on the option set, rather than solely on the knowledge and skills required in the question stem. This fragility matters because many visual reasoning datasets used for outcome-based RFT have included large proportions of MCQA data. We find that training on MCQA increases multiple-choice accuracy metrics but hurts open-form generalization, widening the gap between the two evaluation settings. In other words, this reward encourages shortcuts tied to options rather than transferable knowledge or reasoning (See Fig. 1).

Therefore, we present ReVeL (Rewrite and Verify by LLM), a unified framework that rewrites MCQA into open-ended QA (OpenQA) and preserves verifiability whenever possible. ReVeL categorizes the original multiple-choice questions into numeric, keyword, per-option verification, and genuinely generative cases. The first three types can be accurately graded by deterministic rules, and only the last type may need an LLM Judge for grading. This hybrid design reduces cost and variance from a trivial solution that entirely uses an LLM judge for all problems, while maintaining reliability during evaluation. Across four benchmarks, 70–96% of items become rule-verifiable, reaching higher judging accuracy numbers than entirely using a strong LLM judge (GPT 4.1 mini).

Based on ReVeL, we also rewrite 20k MCQA examples into OpenQA and perform GRPO-based RFT on Qwen2.5-VL-3B/7B. Models trained with ReVeL-OpenQA match MCQA accuracy on choice benchmarks while improving OpenQA accuracy by about six percentage points, demonstrating higher data efficiency and stronger robustness than MCQA-based training. With a modest data, OpenQA-trained 7B models also exceed the counterparts trained on open-source data recipes such as VL-Rethinker-7B [33], R1-OneVision-7B [41], and Mixed-R1-7B [39] on open-ended evaluation. In summary,

our contributions are threefold:

- **Quantifying the non-robustness of MCQA**: We find that evaluation via MCQA not only makes benchmark scores overestimating true capabilities, but also lacks robustness to trivial modifications of the options. Furthermore, RFT on MCQA improves multiple-choice accuracy at the cost of harming open-ended generalization.
- **The ReVeL framework**: We propose a scalable framework to rewrite MCQA into OpenQA, using accurate rule-based judging whenever possible, with much less cost and variance than entirely shifting to an LLM judge.
- **Demonstration of impact on training and evaluation**: Performing RFT on 20K rewritten samples (Qwen2.5-VL-3B/7B) maintains MCQA accuracy while improving OpenQA accuracy by 6 percentage points. Rewriting four benchmarks also reveals up to 20 percentage points of score inflation when shifting from MCQA to OpenQA.

2. Fragility of MCQA

Our work is directly motivated by a series of experiments that quantitatively expose the weaknesses of the MCQA format. We describe our methodology and results here.

2.1. Adding options to open-ended benchmarks

Setup. We start from two recent benchmarks that expect free-form answers from an LLM or VLM: SimpleQA [37] and VisualSimpleQA [35]. We convert each question into an MCQA variant (SimpleQA-Choice / VisualSimpleQA-Choice) by retaining the ground-truth answer and adding five plausible distractors via a human-in-the-loop procedure with GPT-4.1. This conversion preserves the original semantics, but the metrics may be affected by random guessing. Therefore, besides accuracy, we also report a random-guessing upper bound:

$$Acc_{UB} = Acc_{Open} + (1 - Acc_{Open}) \times \frac{1}{K}, K = 6$$

i.e., the model answers correctly on items it can already solve in open-ended form and guesses uniformly on the rest.

Findings. Across both open-weight (e.g., Qwen2.5-72B, Llama-3.3-70B) and proprietary models (e.g., GPT-4.1, Gemini 2.5 Pro), converting to MCQA yields consistently large gains relative to the open-ended baseline and the random-guessing upper bound (Fig. 2). This pattern holds for both text-only (SimpleQA) and multimodal (VisualSimpleQA) settings, indicating that when a model correctly answers a multiple-choice question, it is often utilizing the information embedded in the option set even when it does not actually have the required knowledge or reasoning skills.

Implication. The presence of options supplies huge extra signal that can be exploited independent of task competence, directly leading to overestimation of model capabilities from MCQA accuracy.

2.2. Replacing GT with None-of-the-Above

Another way to test the target knowledge or reasoning skill is to replace the ground-truth option with an option to abstain: 'None of the above (is correct)' (NOTA), after shifting the remaining false options frontwards. We conduct such an experiment on MMLU-Pro [34] and MMMU [43], the most popular MCQA benchmarks for LLM and VLM evaluation.

When the correct option is replaced by NOTA, models frequently display a logical inconsistency: the chain-of-thought reasoning process sometimes correctly eliminates the incorrect options yet still selects one of them as the final answer. As shown in Fig. 3, such contradictions occur even when the model explicitly reasons towards the correct concept ("forest" in that example) but finalizes with an inconsistent choice ("C. home"). Quantitatively, mismatch rates rise from 18% in standard MCQA to 50% under NOTA, listed in Appendix Also, we notice that models often reuse the original "correct" letter position even after the content was modified (after shifting) henceforth incorrect (listed in Appendix implying potential test set contamination or shallow recall of positional cues.

Together, these effects expose how fragile MCQA could be, motivating the shift to option-free OpenQA evaluation.

2.3. Omitting the options from an MCQ

To examine the genuine reasoning ability without the aid of options, we can also remove the options for some multiple-choice questions, treating them as open-form questions. Note that after removing the options, some questions are still valid, but some would become ill-posed.[1] Based on an LLM-assisted analysis (prompt attached in appendix), we find that only about half of the questions in widely used MCQA benchmarks remain suitable using open-form evaluation: 48.9% for MMLU-Pro and 44.1% for MMMU, shown in Tab. 1.

Table 1. Proportion of open-ended questions after filtering.

Dataset	Total	Open Ratio (%)
MMLU-Pro (sampled)	1000	48.9
MMMU (validation)	900	44.1

On the same questions that are still valid without options, models achieve consistently lower accuracy than the original MCQA format, as shown in Fig. 3.

[1] For instance, "How many apples are in the basket?" is still a valid question without any options, but "Which of the following statements are true?" is not. We illustrate four primary categories of questions that cannot apply option removal in the supplementary appendix.

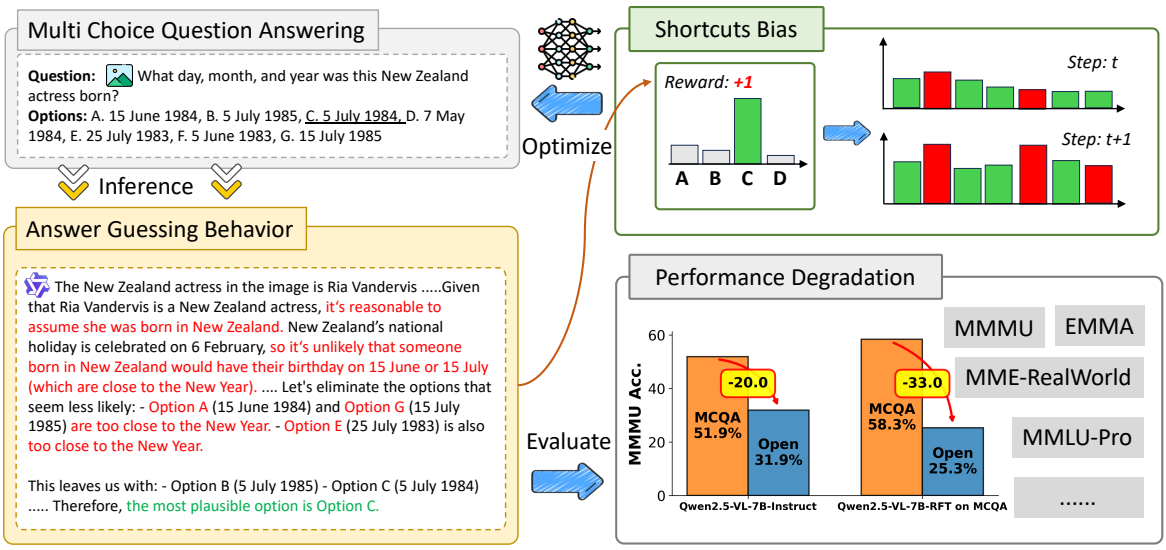


Figure 1. Illustration of MCQA fragility. The example (left) shows an unfaithful reasoning chain that eliminates distractors incorrectly yet provide a correct final answer, yielding a positive reward signal that, when used in reinforcement learning, further amplifies shortcut behavior (top right). This shortcut behavior leads to widening gap between MCQA and OpenQA. The diagram motivate us to propose ReVeL, which aligns evaluation and training with reliable OpenQA.

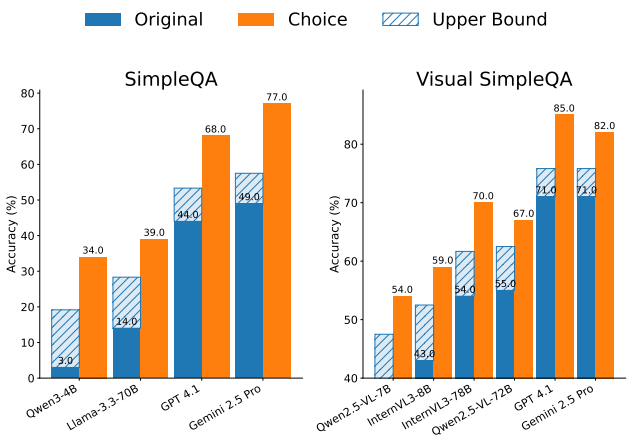


Figure 2. Performance comparison on original open-ended datasets (SimpleQA, Visual SimpleQA) and their multiple-choice versions (*-Choice, with 6 options). The *Random Guess* score is a theoretical upper bound that combines the model's actual open-ended accuracy with the probability of correctly guessing on the rest of the questions from six options.

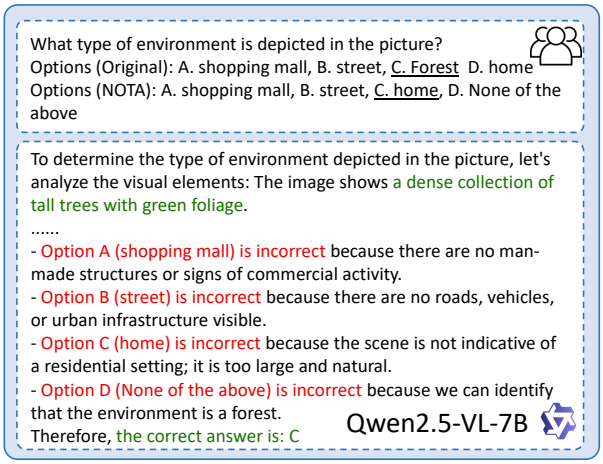


Figure 3. Reasoning and answer can mismatch after replacing the ground-truth option with NOTA.

2.4. RFT on MCQA hurts open-ended QA

Finally, we study training effects by utilizing reinforcement fine-tuning on MCQA data and evaluating on both MCQA and their open counterparts described in Sec. 2.3. We use the popular GRPO algorithm [30] in this work for RFT experiments. RFT on MCQA improves MCQA scores but degrades open-ended performance, thereby widening the MCQA–OpenQA gap. For example, on MMMU, the gap grows for both 3B and 7B models; similar trends hold on EMMA (see Table 2). This indicates that the verifiable reward under MCQA may overfit to option-specific heuristics rather than transferable reasoning.

Across settings, MCQA enables option exploitation that inflates accuracy, amplifies shortcuts tied to options during training. These findings motivate our Rewrite-and-Verify approach in Sec. 3, which mitigate these shortcuts for both evaluation and training.

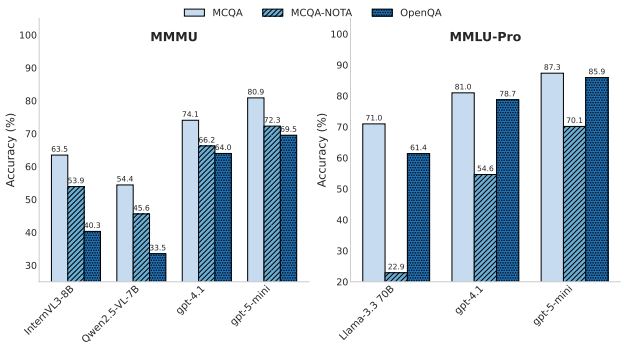


Figure 4. On the impact of options on multiple-choice benchmarks: when options are removed, accuracy is uniformly lower, especially on VQA benchmarks like MMMU.

Table 2. **Impact of RFT on ViRL MCQA data.** MCQ = multiple-choice benchmark score; Open = Open-ended benchmark score. Δ denotes the inflation gap (MCQ–Open). RFT on ViRL (5K MCQA samples) improves MCQ scores but enlarges Δ, indicating reinforced shortcut behavior.

Model	MCQA	OpenQA	Δ (Acc Drop)
MMMU			
Qwen2.5-VL-3B	46.6	11.8	34.8
+ MCQA (ViRL)	50.9	11.6	39.3 (+4.5)
Qwen2.5-VL-7B	51.6	21.4	30.2
+ MCQA (ViRL)	56.4	17.1	39.3 (+9.1)
MMLU-Pro			
Qwen2.5-VL-3B	39.5	21.1	18.4
+ MCQA (ViRL)	47.4	20.4	27.0 (+8.6)
Qwen2.5-VL-7B	53.4	27.6	25.8
+ MCQA (ViRL)	53.6	27.0	26.6 (+0.8)

3. ReVeL: The Rewrite-and-Verify framework

We have shown that MCQA suffers from several shortcomings both in evaluation and in providing reliable training signals. Transforming MCQA to open-ended QA (OpenQA) has the potential to address these issues. In this work, we introduce **ReVeL** (**Re**write-and-**Ve**rify by **L**LMs), a framework that rewrites MCQA into open ended yet verifiable formats while ensuring semantic fidelity and minimizing information loss.

3.1. Pipeline overview

As summarized in Fig. 5, ReVeL operates in three phases: (1) Triage and Classification, (2) Prompt-based Rewriting, and (3) Hybrid Evaluation and Verification. The core principle is to maximize deterministic, rule-based evaluation for questions with unambiguous answers, while reserving LLM-based judging only for cases that genuinely require

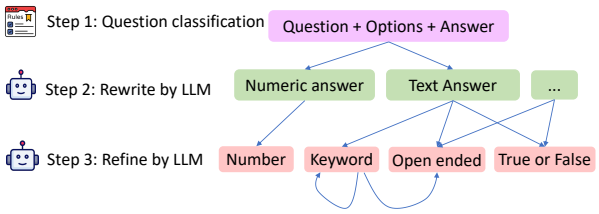


Figure 5. Illustration of the rewrite-and-verify framework

semantic understanding.

During Triage, questions are first passed through a rule-based filter to leave out those expecting numeric answers, mostly quantities or ratios such as *50kg* or $9.8\times10^{-23}m/s^2$. These will be processed via pattern matching. Remaining non-numeric questions are routed to a lightweight LLM-assisted classifier that assigns each question to one of three answer verification categories:

- Keywords matching: single or short tokens that have limited variations (e.g., names, dates).
- Open answers: short, factual or descriptive sentences that are unambiguous for a typical human or LLM grader.
- Per-option verification: questions heavily depend on the option set, such as *Which of the following statements describes the process of ....*

Each category is paired with a tailored rewriting prompt with the goal to preserve semantics while enabling deterministic verification. Examples of all four categories and their rewritten counterparts are shown in Tab. 5

- **Numeric**. ReVeL reformulates them into explicit quantitative prompts by incorporating measurement units and specifying answer format (e.g., comma separated or value–unit pairs).
- **Keywords**. The rewriting step enumerates acceptable synonyms or lexical variants to permit flexible but rule consistent matching.
- **Open answers**. These are rephrased into concise free form queries that solicit factual, non subjective responses without relying on the original options.
- **Per-option verification**. Each option is converted into a declarative statement, and models output a comma separated list of True/False judgments, enabling structured verification and preserving the discriminative intent of MCQA.

3.2. Benchmarks and rewriting coverage

We evaluate ReVeL on four major multimodal benchmarks, including EMMA, MMMU, MME-RealWorld and MMLU-Pro. **EMMA** [13] targets multimodal reasoning in STEM, emphasizing visual–textual integration; we focus on the physics and chemistry subsets for domain-specific evaluation. **MMMU** [43] assesses college-level, multi-discipline reasoning across six domains with diverse im-

age types; we use its 900-question validation set. **MME-RealWorld** [46] offers large-scale, high-quality, real-world tasks with greater difficulty; we adopt its "Lite" subset of 1,700 questions. **MMLU-Pro** [34] is a more challenging variant of MMLU, incorporating reasoning-oriented questions, ten-choice answers, and cleaner data. We sample 1,000 questions for evaluation.

Table 3. Performance comparison of hybrid pipeline versus entirely using an LLM judge

Dataset	Judger	Recall ↑	PPV ↑	FPR ↓	Acc. ↑
EMMA	LLM	100	100	0.0	100
	ReVeL	100	100	0.0	100
MME-RW	LLM	93.5	98.6	1.4	95.9
	ReVeL	95.7	100	0.0	98.0
MMLU-Pro	LLM	95.1	97.5	3.2	95.8
	ReVeL	100	100	0.0	100
MMMU	LLM	100	95.0	5.4	97.3
	ReVeL	93.2	98.6	1.3	96.0
Overall	LLM	96.4	97.2	2.0	97.3
	ReVeL	**96.8**	**99.6**	**0.3**	**98.5**

3.3. Judge accuracy and efficiency

To enhance evaluation consistency and efficiency, ReVeL reclassifies the majority of tasks into deterministically verifiable categories: numeric, keyword, and per-option verification. This design substantially reduces both computational cost and subjective variance by eliminating unnecessary LLM judgment on straightforward verifiable cases.

To validate robustness, we compare ReVeL's hybrid evaluation against a pure LLM-judge baseline across 600 randomly sampled responses from GPT-4.1-mini, Qwen2.5-VL-7B, and Qwen2.5-VL-72B on four benchmarks. As shown in Tab. 3, ReVeL achieves an overall accuracy of 98.5%, exceeding the LLM judge's 97.3%, while simultaneously reducing false positive rate from 2.0% to 0.3%. These trends indicate that integrating rule-based verification improves evaluative stability by enforcing stricter decision boundaries and confirms the robustness of the hybrid verification design.

ReVeL's rewriting not only improves accuracy but also yields substantial efficiency gains. By turning many open-ended questions into structured formats, most items can now be graded automatically with simple rules. This reduces the need for costly and sometimes inconsistent LLM-based judging. As reported in Tab. 4, between 70% and 96% of questions across datasets can be evaluated through deterministic rules. For example, 95.9% of EMMA items become fully rule-checkable after rewriting, and even in MME-RealWorld's complex visual tasks, 71% are deterministically verifiable.

Table 4. **Evaluation format distribution after rewriting.** "Num", "Text", and "Opt" denote rule-based deterministic categories, while "Open" requires LLM judging. The large fraction of rule-based items demonstrates the efficiency of our hybrid evaluation design comparing to pure LLM-judge.

Dataset	LLM	Rule-based		
	Open(%)	Num(%)	Text(%)	Opt(%)
EMMA	4.1	39.0	6.6	50.3
MMMU	17.0	31.3	33.5	18.2
MME-RW	28.4	3.3	55.7	12.6
MMLU-Pro	20.8	39.7	19.6	19.9

4. Experiments

In this section, we apply our ReVeL framework to rewrite existing visual reasoning datasets for reinforcement learning. Firstly, we find that training with our new data improves both accuracy in MCQA and open-end QA format. Then we use our data for evaluation and observe that there is a large performance gap between MCQA and OpenQA across existing MLLMs.

4.1. Expertmental settings

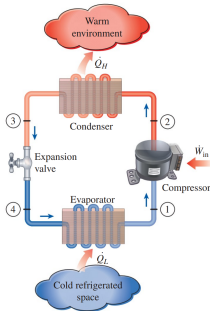
As discussed in Sec. 2.4, training with MCQA tends to reinforce option-exploiting behaviors and amplify format shortcuts, which can degrade model performance. Thus, we employ ReVeL to convert MCQA datasets into OpenQA form for training.

We train Qwen2.5-VL-3B and Qwen2.5-VL-7B with GRPO. To conduct a controlled comparison of the impact of different training data, we designed 4 training configurations based on the ViRL dataset, as shown in Tab. 6

1. **Original MCQA Only** (+*MCQA (ViRL)*): The baseline model is trained exclusively on the original ViRL MCQA data. Rewards are derived from rule-based exact match.
2. **Original MCQA & Original OpenQA** (+*OpenQA (ViRL)*): This configuration auguments (1) by further adding the original OpenQA questions from the ViRL dataset.
3. **Rewritten OpenQA Only** (+*OpenQA (ReVeL)*): The baseline model is trained exclusively on the OpenQA data rewritten by our ReVeL pipeline.
4. **Rewritten OpenQA & Original OpenQA** (+*OpenQA (ViRL)*): This configuration augments (3) by further adding the original OpenQA questions from the ViRL dataset.

This setup enables a controlled comparison between reinforcement driven by MCQA versus OpenQA by our ReVeL.

Table 5. Examples of our ReVeL Pipeline applied to different question types. Each quadrant displays an original multiple-choice question and its OpenQA counterpart.



Numeric

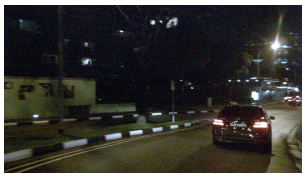
Original: An ideal vapor-compression refrigeration cycle that uses refrigerant-134a as its working fluid maintains a condenser at 800 kPa and the evaporator at 212°C. Determine this system's COP and the amount of power required to service a 150 kW cooling load.
Options: A. 4.07, 31.8 kW, B. 4.97, 33.8 kW, C. 4.87, 30.8 kW

Rewritten Question: An ideal vapor-compression refrigeration cycle that uses refrigerant-134a as its working fluid maintains a condenser at 800 kPa and the evaporator at 212°C. Determine this system's coefficient of performance (COP) and the amount of power required to service a 150 kW cooling load, in kilowatts. Provide your answer as two numbers separated by a comma: COP, power (kW).
Rewritten Answer: 4.87, 30.8

Open answer

Original: Goya created this work while
Options: A. in political exile in England B. serving as a soldier on the front lines against France C. working as the court painter to the king of Spain D. studying Classical antiquity in Rome.

Rewritten Question: Goya created this work while holding what professional position?
Rewritten Answer: Working as the court painter to the king of Spain



Per-option verification

Original: This image shows the front view of the ego car. Predict the behavior of the ego vehicle.
Options: (A) The ego vehicle is steering to the right. The ego vehicle is driving fast. (B)... (C)... (D)... (E)...

Rewritten Question: This image shows the front view of the ego car. Predict the behavior of the ego vehicle. Now, evaluate each of the following statements about the ego vehicle's behavior. (A)... Provide your answer as a single, comma separated list of True or False values corresponding to statements A through E.
Rewritten Answer: True, False, False, False, False

Keywords

Original: What is the manufacturer of the vehicle in the picture?
Options: (A) Mercedes Benz (B) FORD (C) BMW (D) HYUNDAI (E) This image doesn't feature the content.

Rewritten Question: What is the manufacturer of the vehicle in the picture?
Rewritten Answer: BMW ⟨OR⟩ Bayerische Motoren Werke ⟨OR⟩ BMW AG

Our evaluation is based on the four benchmarks mentioned above.

4.2. Training details

We implement all experiments on the VeRL framework with a near on-policy RL setup and train for up to 10 epochs. We do not use KL regularization. For ViRL-Open/MCQA-5K, we use a training batch size of 256, PPO mini-batch size 128, and rollout size 8. For Mixed-R1-Open/MCQA-15K, we use a training batch size of 512, PPO mini-batch size 256, and rollout size 8. Inference and serving for all models are done with vLLM. These settings are fixed across regimes to isolate the effect of the reward design.

4.3. Performance on rewritten training data

As shown in Tab. 6, training on OpenQA consistently produces hight overall accuracy than MCQA across both model sizes: Qwen2.5-VL-3B achieves 34.3 overall with OpenQA vs 30.1 with MCQA (+4.2), and Qwen2.5-VL-7B achieves 40.4 vs 36.3 (+4.1). Importantly, open ended accuracy im-

Table 6. Performance Comparison of MCQA vs. OpenQA Training on In-Domain and Out-of-domain Benchmarks

Model / Train	In-domain				Out-of-domain				Overall Scores		
	EMMA		MMMU		MME-RW		MMLU-Pro		MCQA	OpenQA	Overall
	MCQ	Open	MCQ	Open	MCQ	Open	MCQ	Open			
R1-Onevision-7B	28.9	4.7	42.2	23.9	44.6	31.6	42.5	32.3	39.5	23.1	31.3
Mixed-R1-7B	29.8	13.2	56.3	30.6	45.6	32.8	51.4	37.7	45.8	28.6	37.2
VL-Rethinker-7B	30.6	14.9	53.9	33.4	44.3	32.7	52.4	37.6	45.3	29.6	37.5
Qwen2.5-VL-3B	27.4	5.7	44.3	23.3	35.9	26.6	38.7	29.6	36.6	21.3	28.9
+ MCQA (ViRL)	28.2	3.1	50.2	22.0	39.7	25.6	44.0	28.0	40.5	19.7	30.1
+ OpenQA (ViRL)	31.0	4.4	50.2	23.8	42.1	28.6	43.9	30.3	41.8	21.8	31.8
+ OpenQA (ReVeL)	29.8	**18.6**	49.4	27.4	41.2	31.9	42.2	34.1	40.7	28.0	34.3
+ OpenQA (ViRL)	31.4	17.3	49.4	26.5	41.4	31.7	41.3	33.4	40.9	27.2	34.1
Qwen2.5-VL-7B	28.9	10.2	51.9	31.9	44.8	32.8	49.1	39.0	43.7	28.5	36.1
+ MCQA (ViRL)	30.2	9.1	58.3	25.3	50.1	32.0	52.8	32.4	47.8	24.7	36.3
+ OpenQA (ViRL)	**31.7**	10.4	58.2	33.4	47.6	36.3	**53.7**	37.7	47.8	29.5	38.6
+ OpenQA (ReVeL)	29.2	17.1	56.4	**37.0**	**50.6**	**38.8**	51.1	**43.0**	46.8	**34.0**	**40.4**
+ + OpenQA (ViRL)	29.8	16.9	54.3	36.8	50.3	38.4	51.5	39.9	46.5	33.0	39.8

proves on every benchmarks while MCQA scores remain competitive. Models trained with ReVeL data achieves a 40.4 overall score, compared to 31.3 for R1-OneVision-7B, 37.2 for Mixed-R1-7B, 37.5 for VL-Rethinker-7B. These results indicate that verifiable OpenQA align better with transferable reasoning and real-world usage, improving both open-ended performance and the combined overall metric.

4.4. Performance gap in MCQA and OpenQA

To further quantify the discrepancy in model capabilities between MCQA and OpenQA, we conduct a comparative analysis of model performance in MCQA and OpenQA setting with two rewritten datasets (ViRL and Mixed-R1).

The comprehensive results of this evaluation are presented in Table 7. The result reveal a consistent and substantial performance degradation across all evaluated models when transitioning from the MCQA to the OpenQA format, even strong MLLMs such as GPT-5 and Gemini-2.5 flash are not immune to this effect. For instance, GPT-5's accuracy on the MMMU benchmark drops by 19.8 points (from 79.2% to 59.5%), and Gemini-2.5 flash's accuracy on EMMA decreases by 15.7 points. This indicates that the challenge of OpenQA is a fundamental problem that affects even the most advanced models.

And we observe that the performance gap is often more pronounced for open-weight models. For example, R1-OneVision-7B exhibits a staggering 24.2-point drop on EMMA, while InternVL3-8B's performance on MMMU plummets by 27.9 points. This suggests that many open-weight MLLMs may particularly overfit the MCQA format, which is prevalent in many VQA datasets.

5. Related work

Multiple-choice question answering (MCQA) has been a popularly used assessment tool for ages due to simplified grading [1, 5, 10, 26, 28, 31]. This convenience led to its wide adoption for evaluation of large language models [14, 34], and in particular vision-language models [9, 17, 43, 44] because of more diverse wording choices in describing many visual concepts or scenes. However, MCQA has many shortcuts. Performance can drop dramatically simply from changing an option's placement[21, 47]. While mitigation strategies—such as better distractors, more options, randomized order, or 'select all that apply' formats [40, 42, 45, 47, 48] reduced some biases. And models typically cannot reject all options when the correct answer is absent [12, 32]. Some recent work has shown that reasoning models are good at exploiting the information in the options, implying the performance may be inflated [4, 29]. Recognizing these issues, the community's shift to open-ended evaluation faces its own challenges. Rule-based, short-answer benchmarks [36, 38] are limited in scope, while general open-ended formats rely on an LLM-as-a-judge. Furthermore, simply remove options must discard a significant portion of unsuitable items and still depend on an LLM-Judge for evaluation [23]. These works analyse the flaws of MCQA but do not try to propose a method to mitigate these shortcuts. These analyses focus on identifying the flaws of MCQA rather than proposing systematic mitigation strategies.

Multimodal reinforcement learning: Many visual reasoning datasets are predominantly designed in an MCQA format. For instance, earlier datasets such as ScienceQA [19], AI2D [15], Geometry3K [18], and GeoQA-Plus [7]

Table 7. Overall accuracy (%). Accuracy drop between MCQA and OpenQA is marked after ↓. Bold numbers indicate the smallest drop across open-sourced models

Model	EMMA		MMMU		MME-RealWorld		MMLU-Pro	
	MCQA	OpenQA	MCQA	OpenQA	MCQA	OpenQA	MCQA	OpenQA
Proprietary Models								
GPT-5	42.0	36.0 (↓6.0)	79.2	59.5 (↓19.8)	57.8	42.4 (↓15.4)	84.6	67.6 (↓17.0)
GPT-5 mini	42.8	35.0 (↓7.8)	75.2	55.5 (↓19.7)	58.3	43.7 (↓14.6)	78.7	63.8 (↓14.9)
GPT-4.1	36.4	27.3 (↓9.1)	71.7	56.1 (↓15.5)	52.7	39.6 (↓13.1)	81.2	67.1 (↓14.1)
GPT-4.1 mini	40.2	22.3 (↓17.9)	65.3	51.6 (↓13.7)	54.8	44.0 (↓10.9)	75.4	64.4 (↓11.0)
Gemini-2.5 flash	49.2	33.6 (↓15.7)	69.6	57.7 (↓11.9)	57.3	46.5 (↓10.8)	78.3	63.8 (↓14.5)
Open-Source Models								
InternVL3-78B	34.6	20.8 (↓13.8)	67.7	51.5 (↓**16.2**)	48.9	31.4 (↓17.5)	70.9	57.0 (↓13.9)
InternVL3-8B	32.2	14.5 (↓17.6)	60.0	32.1 (↓27.9)	49.6	33.2 (↓16.4)	55.3	39.0(↓16.3)
Qwen3-VL-8B-Instruct	42.1	23.0 (↓19.1)	68.5	46.5 (↓22.0)	51.7	41.5 (↓10.2)	74.6	60.7 (↓13.9)
R1-OneVision-7B	28.9	4.7 (↓24.2)	42.2	23.9 (↓18.3)	44.6	31.6 (↓13.0)	42.5	32.3 (↓10.2)
Mixed-R1-7B	29.8	13.2 (↓16.7)	56.3	30.6 (↓25.8)	45.6	32.8 (↓12.8)	51.4	37.7 (↓13.7)
VL-Rethinker-7B	30.6	14.9 (↓15.8)	53.9	33.4 (↓20.5)	44.3	32.7 (↓11.6)	52.4	37.6 (↓14.8)
Qwen2.5-VL-72B	35.9	20.6 (↓15.3)	68.2	47.9 (↓20.3)	48.4	37.4 (↓11.0)	70.8	57.6 (↓13.2)
Qwen2.5-VL-3B	27.4	5.7 (↓21.7)	44.3	23.3 (↓21.0)	35.9	26.6 (↓9.2)	38.7	29.6 (↓9.1)
+OpenQA(ViRL)	29.8	18.6 (↓**11.3**)	49.4	27.4 (↓22.0)	41.2	31.9 (↓9.3)	42.2	34.1 (↓**8.1**)
+OpenQA(Mixed-R1)	31.4	17.2 (↓14.1)	46.3	29.8 (↓16.5)	38.0	36.3 (↓**1.7**)	43.3	32.8 (↓10.5)
Qwen2.5-VL-7B	28.9	10.2 (↓18.7)	51.9	31.9 (↓20.0)	44.8	32.8 (↓12.0)	49.1	39.0 (↓10.1)
+OpenQA(ViRL)	29.2	17.1 (↓12.1)	56.4	37.0 (↓19.5)	50.6	38.8 (↓11.7)	51.1	43.0 (↓8.1)
+OpenQA(Mixed-R1)	29.4	15.1 (↓14.4)	56.1	34.1 (↓22.0)	51.9	39.6 (↓12.3)	53.8	40.9 (↓12.9)

are entirely formed by multiple-choice questions. This trend continues in recent MLLMs designed for general-purpose reasoning, such as Mixed-R1 [39], R1-OneVision [41], and VL-Rethinker[33], which all employ a considerable proportion of choice-based items, accounting for 43%, 80%, and 45% of their data, respectively. Our work is built on these visual reasoning datasets and explores open-form rewriting from those MCQA samples.

6. Limitations

We acknowledge several limitations in our proposed pipeline. First, the rewriting and classification phases, while highly accurate, are not perfect and may occasionally introduce errors. Hopefully such errors could diminish when the LLM components are getting stronger and stronger in the future. Second, our work focuses on converting the format of evaluation to be more robust and efficient, without addressing the inherent fallibility of the LLM-judge itself. Issues such as positional bias, verbosity bias, or factual inaccuracies within the LLM-judge [6] are orthogonal to our contribution. We deliberately sidestep some of these known issues; for instance, questions in the EMMA dataset requiring the validation of SMILES chemical structures were intentionally converted to a Per-Option Verification format. This leverages rule-based checking and avoids relying on an LLM-judge for a domain-specific task, thereby mitigating a potential failure point of LLM-based evaluation. There are several directions for future research. One key avenue is to extend our framework beyond QA to other NLP tasks, such as long-form generation, where evaluation remains a major challenge. Finally, developing adaptive evaluation systems that can dynamically choose the most appropriate and cost effective judging mechanism based on the question's complexity and the model's response would be a valuable next step.

7. Conclusions

In this work, we systematically demonstrated the fragility of MCQA format for both evaluation and reinforcement fine-tuning. We found that MCQA metrics significantly overestimate model capabilities, and RFT on MCQA data reinforces format-specific shortcuts, harming open-ended generalization. To solve this, we propose ReVeL, a framework that rewrite MCQA into verifiable OpenQA by categorizing questions for a hybird evaluation scheme. Applying ReVeL to RFT, we found that models trained on our rewritten OpenQA data achieved approximately a 6-point improvement in open-ended accuracy while maintaining performance on original MCQA benchmarks, confirming its role in fostering more robust and transferable reasoning.

References

[1] Norah Alzahrani, Hisham Alyahya, Yazeed Alnumay, Sultan AlRashed, Shaykhah Alsubaie, Yousef Almushayqih, Faisal Mirza, Nouf Alotaibi, Nora Al-Twairesh, Areeb Alowisheq, M Saiful Bari, and Haidar Khan. When benchmarks are targets: Revealing the sensitivity of large language model leaderboards. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13787–13805, Bangkok, Thailand, 2024. Association for Computational Linguistics. 7

[2] Anthropic. Claude 4. https://www.anthropic.com/news/claude-4, 2025. Accessed: 2025-07-11. 1

[3] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 1

[4] Nishant Balepur, Abhilasha Ravichander, and Rachel Rudinger. Artifacts or abduction: How do llms answer multiple-choice questions without the question? In *Annual Meeting of the Association for Computational Linguistics*, 2024. 1, 7

[5] Nishant Balepur, Rachel Rudinger, and Jordan L. Boyd-Graber. Which of these best describes multiple choice evaluation with llms? a) forced b) flawed c) fixable d) all of the above. In *Annual Meeting of the Association for Computational Linguistics*, 2025. 7

[6] Dongping Chen, Ruoxi Chen, Shilin Zhang, Yinuo Liu, Yaochen Wang, Huichi Zhou, Qihui Zhang, Pan Zhou, Yao Wan, and Lichao Sun. Mllm-as-a-judge: Assessing multimodal llm-as-a-judge with vision-language benchmark. In *International Conference on Machine Learning*, 2024. 8

[7] Jiaqi Chen, Jianheng Tang, Jinghui Qin, Xiaodan Liang, Lingbo Liu, Eric P. Xing, and Liang Lin. Geoqa: A geometric question answering benchmark towards multimodal numerical reasoning. *ArXiv*, abs/2105.14517, 2021. 7

[8] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, Lixin Gu, Xuehui Wang, Qingyun Li, Yimin Ren, Zixuan Chen, Jiapeng Luo, Jiahao Wang, Tan Jiang, Bo Wang, Conghui He, Botian Shi, Xingcheng Zhang, Han Lv, Yi Wang, Wenqi Shao, Pei Chu, Zhongying Tu, Tong He, Zhiyong Wu, Huipeng Deng, Jiaye Ge, Kai Chen, Kaipeng Zhang, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling, 2025. 1

[9] Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. *ArXiv*, abs/1803.05457, 2018. 7

[10] Robert Dufresne, William Leonard, and William Gerace. Making sense of students' answers to multiple-choice questions. *The Physics Teacher*, 40:174–180, 2002. 7

[11] Google. Gemini2.5 pro. https://deepmind.google/models/gemini/, 2025. Accessed: 2025-07-11. 1

[12] Gracjan G'oral, Emilia Wiśnios, Piotr Sankowski, and Pawel Budzianowski. Wait, that's not an option: Llms robustness with incorrect multiple-choice options. 2024. 7

[13] Yunzhuo Hao, Jiawei Gu, Huichen Will Wang, Linjie Li, Zhengyuan Yang, Lijuan Wang, and Yu Cheng. Can mllms reason in multimodality? emma: An enhanced multimodal reasoning benchmark. *ArXiv*, abs/2501.05444, 2025. 1, 4

[14] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Xiaodong Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *ArXiv*, abs/2009.03300, 2020. 1, 7

[15] Aniruddha Kembhavi, Michael Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. *ArXiv*, abs/1603.07396, 2016. 1, 7

[16] Yuanzhan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. Mmbench: Is your multi-modal model an all-around player? *ArXiv*, abs/2307.06281, 2023. 1

[17] Ziqiang Liu, Feiteng Fang, Xi Feng, Xinrun Du, Chenhao Zhang, Zekun Moore Wang, Yuelin Bai, Qixuan Zhao, Liyang Fan, Chengguang Gan, Hongquan Lin, Jiaming Li, Yuansheng Ni, Haihong Wu, Yaswanth Narsupalli, Zhigang Zheng, Chengming Li, Xiping Hu, Ruifeng Xu, Xiaojun Chen, Min Yang, Jiaheng Liu, Ruibo Liu, Wenhao Huang, Ge Zhang, and Shiwen Ni. Ii-bench: An image implication understanding benchmark for multimodal large language models. *ArXiv*, abs/2406.05862, 2024. 7

[18] Pan Lu, Ran Gong, Shibiao Jiang, Liang Qiu, Siyuan Huang, Xiaodan Liang, and Song-Chun Zhu. Inter-gps: Interpretable geometry problem solving with formal language and symbolic reasoning. In *Annual Meeting of the Association for Computational Linguistics*, 2021. 7

[19] Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and A. Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *ArXiv*, abs/2209.09513, 2022. 7

[20] Chenyang Lyu, Minghao Wu, and Alham Fikri Aji. Beyond probabilities: Unveiling the misalignment in evaluating large language models. *ArXiv*, abs/2402.13887, 2024. 1

[21] Francesco Maria Molfese, Luca Moroni, Luca Gioffre, Alessandro Sciré, Simone Conia, and Roberto Navigli. Right answer, wrong score: Uncovering the inconsistencies of llm evaluation in multiple-choice question answering. *ArXiv*, abs/2503.14996, 2025. 7

[22] Steven Moore, Huy Anh Nguyen, Tianying Chen, and John C. Stamper. Assessing the quality of multiple-choice questions using gpt-4 and rule-based methods. In *European Conference on Technology Enhanced Learning*, 2023. 1

[23] Aidar Myrzakhan, S. Mahmoud Bsharat, and Zhiqiang Shen. Open-llm-leaderboard: From multi-choice to open-style questions for llms evaluation, benchmark, and arena. *ArXiv*, abs/2406.07545, 2024. 7

[24] OpenAI. Gpt-4v(ision) system card. *OpenAI Research*, 2023. 1

[25] OpenAI. Introducing GPT-5, 2025. 1

[26] Moragh Paxton. A linguistic perspective on multiple choice questioning. *Assessment & Evaluation in Higher Education - ASSESS EVAL HIGH EDUC*, 25:109–119, 2000. 7

[27] Baoqi Pei, Yifei Huang, Jilan Xu, Yuping He, Guo Chen, Fei Wu, Yu Qiao, and Jiangmiao Pang. Egothinker: Unveiling egocentric reasoning with spatio-temporal cot. 2025. 1

[28] Pouya Pezeshkpour and Estevam Hruschka. Large language models sensitivity to the order of options in multiple-choice questions. In *NAACL-HLT*, 2023. 7

[29] Narun K. Raman, Taylor Lundy, and Kevin Leyton-Brown. Reasoning models are test exploiters: Rethinking multiple-choice. *ArXiv*, abs/2507.15337, 2025. 7

[30] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Jun-Mei Song, Mingchuan Zhang, Y. K. Li, Yu Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *ArXiv*, abs/2402.03300, 2024. 3

[31] Mark G. Simkin and William L. Kuechler. Multiple-choice tests and student understanding: What is the connection? *Decision Sciences Journal of Innovative Education*, 3:73–98, 2005. 7

[32] Zhi Rui Tam, Cheng-Kuang Wu, Chieh-Yen Lin, and Yun-Nung Chen. None of the above, less of the right: Parallel patterns between humans and llms on multi-choice questions answering. *ArXiv*, abs/2503.01550, 2025. 7

[33] Haozhe Wang, Chao Qu, Zuming Huang, Wei Chu, Fangzhen Lin, and Wenhu Chen. Vl-rethinker: Incentivizing self-reflection of vision-language models with reinforcement learning. *ArXiv*, abs/2504.08837, 2025. 1, 8

[34] Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max W.F. Ku, Kai Wang, Alex Zhuang, Rongqi "Richard" Fan, Xiang Yue, and Wenhu Chen. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *ArXiv*, abs/2406.01574, 2024. 1, 2, 5, 7

[35] Yanling Wang, Yihan Zhao, Xiaodong Chen, Shasha Guo, Lixin Liu, Haoyang Li, Yong Xiao, Jing Zhang, Qi Li, and Ke Xu. Visualsimpleqa: A benchmark for decoupled evaluation of large vision-language models in fact-seeking question answering. *ArXiv*, abs/2503.06492, 2025. 2

[36] Zirui Wang, Mengzhou Xia, Luxi He, Howard Chen, Yitao Liu, Richard Zhu, Kaiqu Liang, Xindi Wu, Haotian Liu, Sadhika Malladi, Alexis Chevalier, Sanjeev Arora, and Danqi Chen. Charxiv: Charting gaps in realistic chart understanding in multimodal llms. *ArXiv*, abs/2406.18521, 2024. 7

[37] Jason Wei, Nguyen Karina, Hyung Won Chung, Yunxin Joy Jiao, Spencer Papay, Amelia Glaese, John Schulman, and William Fedus. Measuring short-form factuality in large language models. *ArXiv*, abs/2411.04368, 2024. 2

[38] xAI. Introducing grok-1.5v and realworldqa benchmark, 2024. 7

[39] Shilin Xu, Yanwei Li, Rui Yang, Tao Zhang, Yueyi Sun, Wei Chow, Linfeng Li, Hang Song, Qi Xu, Yunhai Tong, Xiangtai Li, and Hao Fei. Mixed-r1: Unified reward perspective for reasoning capability in multimodal large language models. *ArXiv*, abs/2505.24164, 2025. 1, 8

[40] Weijie Xu, Shixian Cui, Xi Fang, Chi Xue, Stephanie Eckman, and Chandan K. Reddy. Sata-bench: Select all that apply benchmark for multiple choice questions. *ArXiv*, abs/2506.00643, 2025. 7

[41] Yi Yang, Xiaoxuan He, Hongkun Pan, Xiyan Jiang, Yan Deng, Xingtao Yang, Haoyu Lu, Dacheng Yin, Fengyun Rao, Minfeng Zhu, Bo Zhang, and Wei Chen. R1-onevision: Advancing generalized multimodal reasoning through cross-modal formalization. *ArXiv*, abs/2503.10615, 2025. 1, 8

[42] Han Cheng Yu, Yu An Shih, Kin Man Law, KaiYu Hsieh, Yu Chen Cheng, Hsin Chih Ho, Zih An Lin, Wen-Chuan Hsu, and Yao-Chung Fan. Enhancing distractor generation for multiple-choice questions with retrieval augmented pretraining and knowledge graph integration. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11019–11029, Bangkok, Thailand, 2024. Association for Computational Linguistics. 7

[43] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9556–9567, 2023. 1, 2, 4, 7

[44] Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Ming Yin, Botao Yu, Ge Zhang, Huan Sun, Yu Su, Wenhu Chen, and Graham Neubig. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. *ArXiv*, abs/2409.02813, 2024. 1, 7

[45] Yuhui Zhang, Yuchang Su, Yiming Liu, Xiaohan Wang, James Burgess, Elaine Sui, Chenyu Wang, Josiah Aklilu, Alejandro Lozano, Anjiang Wei, Ludwig Schmidt, and Serena Yeung-Levy. Automated generation of challenging multiple-choice questions for vision language model evaluation. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pages 29580–29590, 2025. 7

[46] Yi-Fan Zhang, Huanyu Zhang, Haochen Tian, Chaoyou Fu, Shuangqing Zhang, Jun Wu, Feng Li, Kun Wang, Qingsong Wen, Zhang Zhang, Liang Wang, Rong Jin, and Tien-Ping Tan. Mme-realworld: Could your multimodal llm challenge high-resolution real-world scenarios that are difficult for humans? *ArXiv*, abs/2408.13257, 2024. 1, 5

[47] Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. Large language models are not robust multiple choice selectors. *ArXiv*, abs/2309.03882, 2023. 1, 7

[48] Wenjie Zhou, Qiang Wang, Mingzhou Xu, Ming Chen, and Xiangyu Duan. Revisiting the self-consistency challenges in multi-choice question formats for large language model evaluation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 14103–14110, Torino, Italia, 2024. ELRA and ICCL. 7