

# Beyond Multiple Choice: A Hybrid Framework for Unifying Robust Evaluation and Verifiable Reasoning Training

Yesheng Liu<sup>1,2,3</sup>, Hao Li<sup>3,4</sup>, Haiyu Xu<sup>3,5</sup>, Baoqi Pei<sup>6</sup>, Jiahao Wang<sup>1,2,3</sup>, Mingxuan Zhao<sup>3,5</sup>, Jing-Shu Zheng<sup>3</sup>, Zheqi He<sup>3</sup>, JG Yao<sup>3</sup>, Bowen Qin<sup>3</sup>, Xi Yang<sup>3</sup>, Jiajun Zhang<sup>1,2</sup>

<sup>1</sup>Institute of Automation, CAS, <sup>2</sup>School of Artificial Intelligence, UCAS, <sup>3</sup>BAAI FlagEval Team,

<sup>4</sup>BUAA, <sup>5</sup>PKU, <sup>6</sup>ZJU

Project Page: <https://flageval.github.io/ReVeL/>

## 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 (**R**ewrite and **V**erify by **LLM**), 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 (Anthropic, 2025; OpenAI, 2025; Bai et al., 2025; OpenAI, 2023; Google, 2025; Chen et al., 2025; Pei et al., 2025) increasingly tackle diverse real-world tasks, the demand for reliable and scalable evaluation has grown significantly. MCQA is convenient because restricting outputs simplifies scoring (Moore et al., 2023) across language (Hendrycks et al., 2020; Wang et al., 2024a) and vision language benchmarks (Yue et al., 2023; 2024; Liu et al., 2023; Kembhavi et al., 2016; Zhang et al., 2024; Hao et al., 2025).

However, MCQA departs from real-world usage where answers are usually open-ended (Lyu et al., 2024), while the predefined options encourage selection heuristics (Zheng et al., 2023; Balepur et al., 2024) 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 Figure 1).

Therefore, we present ReVeL (**R**ewrite and **V**erify 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 (Wang et al., 2025a), R1-OneVision-7B (Yang et al., 2025), and Mixed-R1-7B (Xu et al., 2025a) on open-ended evaluation. In summary, our contributions are threefold:

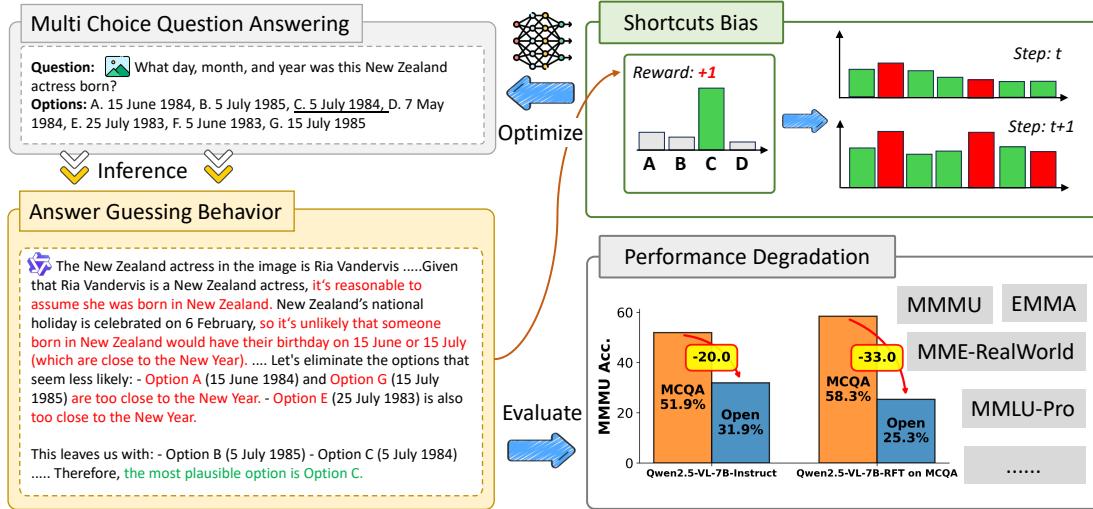


Figure 1: Illustration of MCQA fragility. The example (left) shows an unfaithful reasoning chain that eliminates distractors incorrectly yet provide a correct final answer, yielding a positive reward signal that, when used in reinforcement learning, further amplifies shortcut behavior (top right). This 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.

- **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 (Wei et al., 2024) and VisualSimpleQA (Wang et al., 2025b). 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

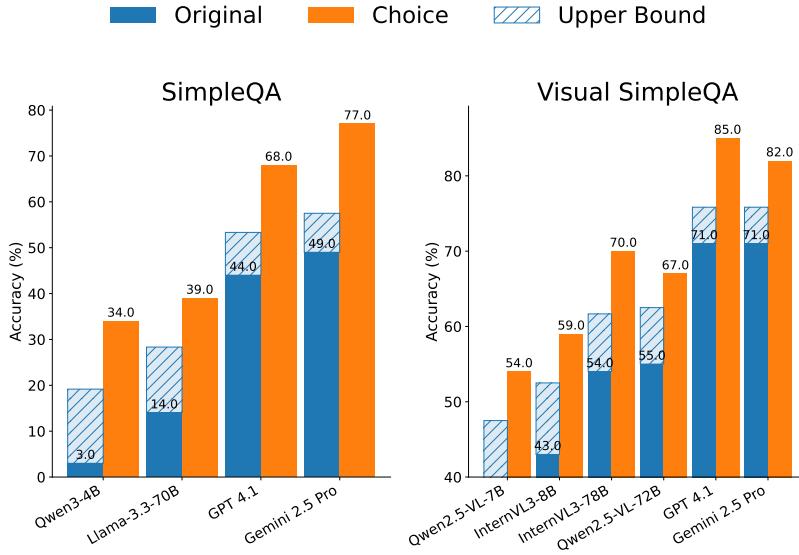


Figure 2: Performance comparison on original open-ended datasets (SimpleQA, Visual SimpleQA) and their multiple-choice versions (\*-Choice, with 6 options). The *Random Guess* score is a theoretical upper bound that combines the model’s actual open-ended accuracy with the probability of correctly guessing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

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 (Figure 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 (Wang et al., 2024a) and MMMU (Yue et al., 2023), 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 Figure 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.

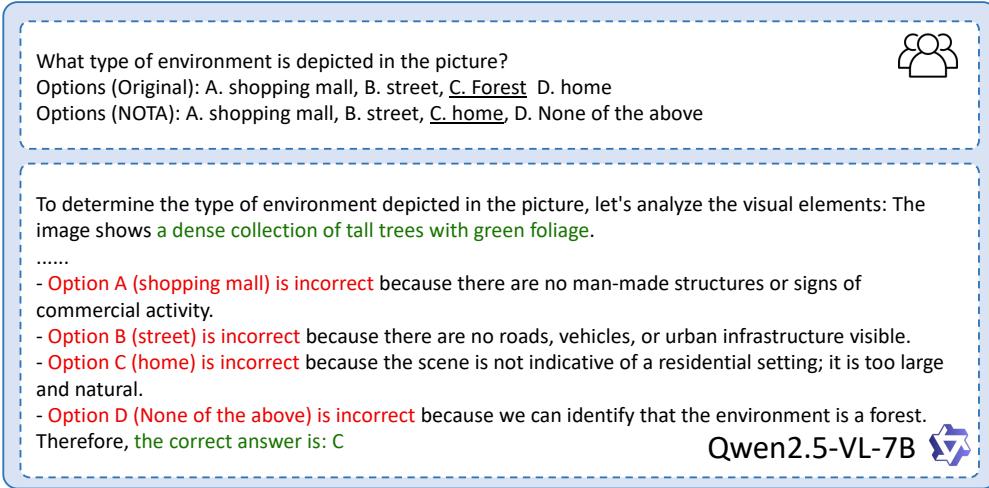


Figure 3: Reasoning and answer can mismatch after replacing the ground-truth option with NOTA.

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.<sup>1</sup> 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 Table 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 Figure 4.

### 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 Section 2.3. We use the popular GRPO algorithm (Shao et al., 2024) 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 Section 3, which mitigate these shortcuts for both evaluation and training.

<sup>1</sup>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.

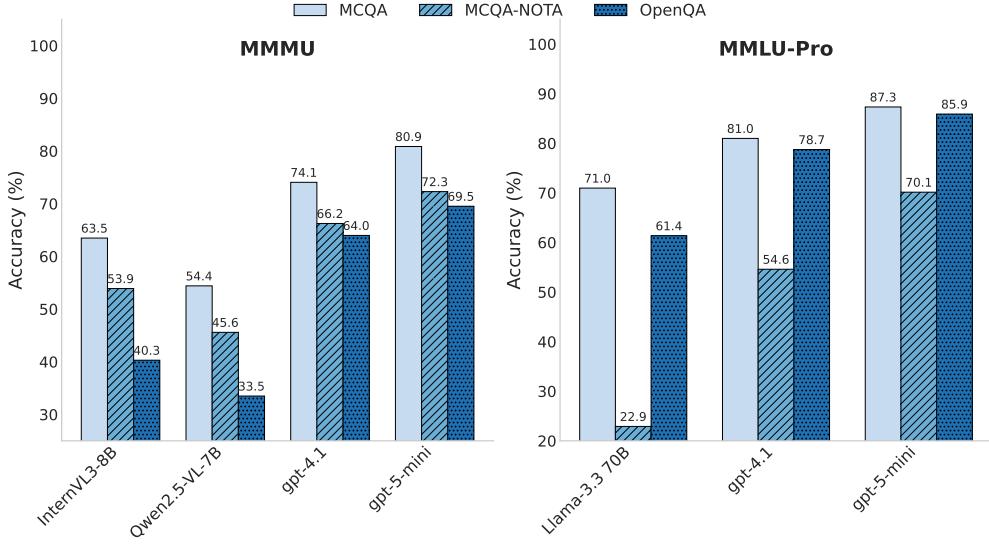


Figure 4: On the impact of options on multiple-choice benchmarks: when options are removed, accuracy is uniformly lower, especially on VQA benchmarks like MMMU.

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

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

### 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** (Rewrite-and-Verify by LLMs), 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 Figure 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 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  $50\text{kg}$  or  $9.8 \times 10^{-23}\text{m/s}^2$ . These will be processed via pattern

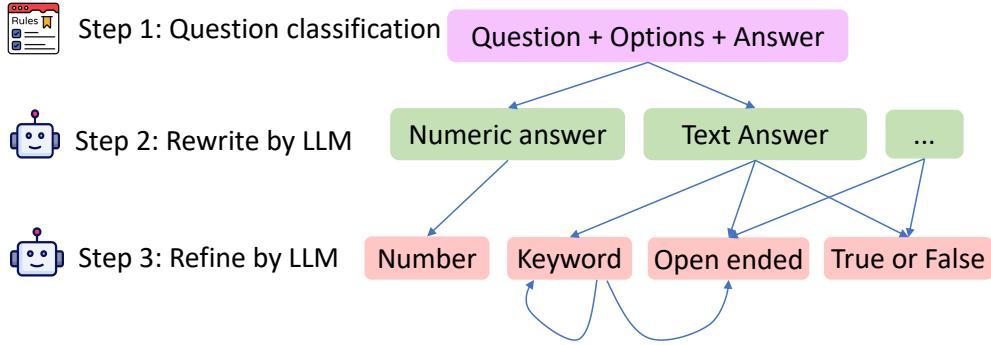


Figure 5: Illustration of the rewrite-and-verify framework

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 Figure 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** (Hao et al., 2025) targets multimodal reasoning in STEM, emphasizing visual-textual integration; we focus on the physics and chemistry subsets for domain-specific evaluation. **MMMU** (Yue et al., 2023) assesses college-level, multi-discipline reasoning across six domains with diverse image types; we use its 900-question validation set. **MME-RealWorld** (Zhang et al., 2024) offers large-scale, high-quality, real-world tasks with greater difficulty; we adopt its “Lite” subset of 1,700 questions. **MMLU-Pro** (Wang et al., 2024a) is a more challenging variant of MMLU, incorporating reasoning-oriented questions, ten-choice answers, and cleaner data. We sample 1,000 questions for evaluation.

### 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.

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

Dataset	Judger	Recall $\uparrow$	PPV $\uparrow$	FPR $\downarrow$	Acc. $\uparrow$
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	<b>96.8</b>	<b>99.6</b>	0.3	<b>98.5</b>

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

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 Table 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 Table 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.

## 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 Experimental settings

As discussed in Section 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.

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

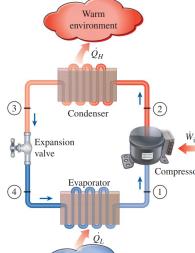
 <p><b>Numeric</b></p>	<p><b>Original:</b> An ideal vapor-compression refrigeration cycle that uses refrigerant-134a as its working fluid maintains a condenser at 800 kPa and the evaporator at 212°C. Determine this system's COP and the amount of power required to service a 150 kW cooling load.</p> <p><b>Options:</b> A. 4.07, 31.8 kW, B. 4.97, 33.8 kW, C. 4.87, 30.8 kW</p> <p><b>Rewritten Question:</b> 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).</p> <p><b>Rewritten Answer:</b> 4.87, 30.8</p>
 <p><b>Open answer</b></p>	<p><b>Original:</b> Goya created this work while</p> <p><b>Options:</b> 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.</p> <p><b>Rewritten Question:</b> Goya created this work while holding what professional position?</p> <p><b>Rewritten Answer:</b> Working as the court painter to the king of Spain</p>
 <p><b>Per-option verification</b></p>	<p><b>Original:</b> This image shows the front view of the ego car. Predict the behavior of the ego vehicle.</p> <p><b>Options:</b> (A) The ego vehicle is steering to the right. The ego vehicle is driving fast. (B)... (C)... (D)... (E)...</p> <p><b>Rewritten Question:</b> 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.</p> <p><b>Rewritten Answer:</b> True, False, False, False, False</p>
 <p><b>Keywords</b></p>	<p><b>Original:</b> What is the manufacturer of the vehicle in the picture?</p> <p><b>Options:</b> (A) Mercedes Benz (B) FORD (C) BMW (D) HYUNDAI (E) This image doesn't feature the content.</p> <p><b>Rewritten Question:</b> What is the manufacturer of the vehicle in the picture?</p> <p><b>Rewritten Answer:</b> BMW (OR) Bayerische Motoren Werke (OR) BMW AG</p>

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		MCQ	Open	Total
	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	<b>18.6</b>	49.4	27.4	41.2	31.9	42.2	34.1	40.7	28.0	34.3
+ OpenQA (ViRL)	31.4	<u>17.3</u>	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)	<b>31.7</b>	10.4	<u>58.2</u>	33.4	47.6	36.3	<b>53.7</b>	37.7	<u>47.8</u>	29.5	38.6
+ OpenQA (ReVeL)	29.2	17.1	56.4	<b>37.0</b>	<b>50.6</b>	<b>38.8</b>	51.1	<b>43.0</b>	<u>46.8</u>	<b>34.0</b>	<b>40.4</b>
+ OpenQA (ViRL)	29.8	16.9	54.3	<u>36.8</u>	<u>50.3</u>	<u>38.4</u>	51.5	<u>39.9</u>	46.5	<u>33.0</u>	<u>39.8</u>

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 Table 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 augments (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. 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 Table 6, training on OpenQA consistently produces higher 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 improves on every benchmarks while MCQA scores remain competitive. Models trained with ReVeL data achieve 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.

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
<i>Proprietary Models</i>								
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)
<i>Open-Source Models</i>								
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 +OpenQA(ViRL)	27.4 29.8	5.7 (↓21.7) 18.6 (↓11.3)	44.3 49.4	23.3 (↓21.0) 27.4 (↓22.0)	35.9 41.2	26.6 (↓9.2) 31.9 (↓9.3)	38.7 42.2	29.6 (↓9.1) 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 +OpenQA(ViRL)	28.9 29.2	10.2 (↓18.7) 17.1 (↓12.1)	51.9 56.4	31.9 (↓20.0) 37.0 (↓19.5)	44.8 50.6	32.8 (↓12.0) 38.8 (↓11.7)	49.1 51.1	39.0 (↓10.1) 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)

## 5 Related work

**Multiple-choice question answering (MCQA)** has been a popularly used assessment tool for ages due to simplified grading (Simkin and Kuechler, 2005; Dufresne et al., 2002; Paxton, 2000; Balepur et al., 2025; Alzahrani et al., 2024; Pezeshkpour and Hruschka, 2023). This convenience led to its wide adoption for evaluation of large language models (Hendrycks et al., 2020; Wang et al., 2024a), and in particular vision-language models (Yue et al., 2023; Clark et al., 2018; Yue et al., 2024; Liu et al., 2024) 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 (Zheng et al., 2023; Molfese et al., 2025). While mitigation strategies—such as better distractors, more options, randomized order, or ‘select all that apply’ formats (Zhang et al., 2025; Yu et al., 2024; Zheng et al., 2023; Zhou et al., 2024; Xu et al., 2025b) reduced some biases. And models typically cannot reject all options when the correct answer is absent (G’oral et al., 2024; Tam et al., 2025). Some recent work has shown that reasoning models are good at exploiting the information

in the options, implying the performance may be inflated (Balepur et al., 2024; Raman et al., 2025). Recognizing these issues, the community’s shift to open-ended evaluation faces its own challenges. Rule-based, short-answer benchmarks (xAI, 2024; Wang et al., 2024b) 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 (Myrzakhan et al., 2024). 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 (Lu et al., 2022), AI2D (Kembhavi et al., 2016), Geometry3K (Lu et al., 2021), and GeoQA-Plus (Chen et al., 2021) are entirely formed by multiple-choice questions. This trend continues in recent MLLMs designed for general-purpose reasoning, such as Mixed-R1 (Xu et al., 2025a), R1-OneVision (Yang et al., 2025), and VL-Rethinker(Wang et al., 2025a), 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 (Chen et al., 2024) 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 hybrid 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

- 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.
- Anthropic. Claude 4. <https://www.anthropic.com/news/claude-4>, 2025. Accessed: 2025-07-11.
- 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.
- 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.

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.

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 lilm-as-a-judge with vision-language benchmark. In *International Conference on Machine Learning*, 2024.

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.

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 Wenhui Wang. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling, 2025.

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.

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

Google. Gemini2.5 pro. <https://deepmind.google/models/gemini/>, 2025. Accessed: 2025-07-11.

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

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.

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.

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.

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.

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, Rui Feng 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.

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.

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.

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

Francesco Maria Molfese, Luca Moroni, Luca Goffre, Alessandro Sciré, Simone Conia, and Roberto Navigli. Right answer, wrong score: Uncovering the inconsistencies of lilm evaluation in multiple-choice question answering. ArXiv, abs/2503.14996, 2025.

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.

Aidar Myrzakhan, S. Mahmoud Bharat, and Zhiqiang Shen. Open-lilm-leaderboard: From multi-choice to open-style questions for llms evaluation, benchmark, and arena. ArXiv, abs/2406.07545, 2024.

- OpenAI. Gpt-4v(ision) system card. *OpenAI Research*, 2023.
- OpenAI. Introducing GPT-5, 2025.
- Moragh Paxton. A linguistic perspective on multiple choice questioning. *Assessment & Evaluation in Higher Education - ASSESS EVAL HIGH EDUC*, 25:109–119, 2000.
- 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.
- Pouya Pezeshkpour and Estevam Hruschka. Large language models sensitivity to the order of options in multiple-choice questions. In *NAACL-HLT*, 2023.
- Narun K. Raman, Taylor Lundy, and Kevin Leyton-Brown. Reasoning models are test exploiters: Rethinking multiple-choice. *ArXiv*, abs/2507.15337, 2025.
- 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.
- 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.
- 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.
- Haozhe Wang, Chao Qu, Zuming Huang, Wei Chu, Fangzhen Lin, and Wenhua Chen. VI-rethinker: Incentivizing self-reflection of vision-language models with reinforcement learning. *ArXiv*, abs/2504.08837, 2025a.
- 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 Wenhua Chen. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *ArXiv*, abs/2406.01574, 2024a.
- 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, 2025b.
- Zirui Wang, Mengzhou Xia, Luxi He, Howard Chen, Yitao Liu, Richard Zhu, Kaiyu 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, 2024b.
- 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.
- xAI. Introducing grok-1.5v and realworldqa benchmark, 2024.
- 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, 2025a.
- 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, 2025b.
- Yi Yang, Xiaoxuan He, Hongkun Pan, Xiyan Jiang, Yan Deng, Xingtao Yang, Haoyu Lu, Dacheng Yin, Fengyun Rao, Min-feng Zhu, Bo Zhang, and Wei Chen. R1-onevision: Advancing generalized multimodal reasoning through cross-modal formalization. *ArXiv*, abs/2503.10615, 2025.
- 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.
- Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruqi 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 Wenhua 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.

---

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

Yuhui Zhang, Yuchang Su, Yiming Liu, Xiaohan Wang, James Burgess, Elaine Sui, Chenyu Wang, Josiah Akllil, 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.

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 lilm challenge high-resolution real-world scenarios that are difficult for humans? ArXiv, abs/2408.13257, 2024.

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.

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.

## A Details of Removing Options From MCQA

### A.1 A Filtering Pipeline for Self-Sufficient Questions

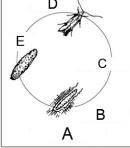
Option dependent	Sentence Completion Questions
<p><b>Question:</b> This question is based on the following declarations: String strA = "CARROT", strB = "Carrot", strC = "car"; Given that all uppercase letters precede all lowercase letters when considering alphabetical order, which is true?</p> <p>A: strA.compareTo(strB) &lt; 0 &amp;&amp; strB.compareTo(strC) &gt; 0  B: strC.compareTo(strB) &lt; 0 &amp;&amp; strB.compareTo(strA) &lt; 0  C: strB.compareTo(strC) &lt; 0 &amp;&amp; strB.compareTo(strA) &gt; 0  D: !(strA.compareTo(strB) == 0) &amp;&amp; strB.compareTo(strA) &lt; 0</p>	 <p>Complete the statement. Ammonia is ().  A: an elementary substance  B: a compound</p>
Source: MMLU	Source: MMbench
Subjective Question	Questions with multiple answers
 <p><b>Question:</b> What was George Seurat's goal in his 'Divisionist' paintings?  As to enliven paintings through the use of scientific theories about color and optic function  B: to bring Classical structure to the Impressionists' approach to painting  C: to use color to generate a strong emotional response from the viewer  D: to produce an idiosyncratic style characterized by the use of small dots of color</p>	 <p>How long does the larva and feeding tub stage last?  A: 4-30 days  B: 85-90 days  C: 15-30 days  D: 55 days to 2.5 years</p>
Source: MMMU	Source: AI2D

Figure 6: Common characteristics that make a multiple-choice question unsuitable for direct conversion to a free-form format.

To create a dataset for our option-free evaluation, we developed a two-stage pipeline to systematically filter existing benchmarks and retain only self-sufficient questions suitable for a generative format.

**Stage 1: Question Validity Filtering.** We first exclude questions that are fundamentally unsuitable for free-form conversion. This stage combines heuristic rules (e.g., removing questions with long, paragraph-style answers likely to be subjective) with a prompted LLM that identifies and removes questions exhibiting option dependency, subjectivity, or underspecification.

**Stage 2: Answer Uniqueness Verification.** Questions that pass the first stage are then checked for answer uniqueness. We use an LLM to determine if a question, in the absence of options, could yield multiple substantively different but equally valid answers. Only questions with a single, unambiguous correct answer are retained. The filtering prompts used to guide the LLM in each stage are detailed in Appendix.

We validated this pipeline by annotating a random sample of its outputs. The results, detailed in Table 8, confirm the high fidelity of our method. The low overall False Positive (FP) rate of 2.5% and False Negative (FN) rate of 3.0% ensure the soundness of our subsequent experiments.

Table 8: Statistical of the filtering pipeline. The table detailsthe False Positive (FP) and False Negative (FN) rates across sampled subset (right). FP and FN rates are computed from a sampled subset.

Dataset	FP (%)	FN (%)
MMLU-Pro	4.0	4.0
MMMU	2.0	2.0

Figure 6 shows 4 primary patterns: **Option-Dependent Questions:** that explicitly refers to the choices (e.g., "Which of the following..."). **Sentence Completion Questions:** that is a grammatically incomplete without options (e.g., "The letter E in the diagram represents..."). **Subjective Questions:** that solicits an opinion rather than an objective, factual answer. **Questions with Multiple Answers:** where the MCQ format artificially presents only one correct choice.

## B Failure modes of MCQA

We conduct an analysis of the underlying causes for the unreliability of MCQA, identifying critical failure modes like reasoning-choice mismatch, positional memorization and option-anchoring.

### B.1 Analysis of Failure Modes: Reasoning-Choice Mismatch

To pinpoint the cause, we focused specifically on the subset of questions that a model answered correctly in the Standard MCQA setting but incorrectly in the NOTA setting. Within this set of failures, we discovered a frequent reasoning-choice mismatch. As shown in Table 9, models often correctly inference the answer in their reasoning steps. However, upon finding this answer absent from the choices, they fail to select the logical NOTA option. Instead, they revert to strategies like string matching to select an incorrect distractor. The rate of this mismatch is dramatically higher in the NOTA setting compared to the standard MCQA setting, exposing a critical flaw in how models interact with provided options when the correct answer is missing.

Model	Dataset	MCQA	NOTA
InternVL3-78B	MMMU	17.6%	54.1%
InternVL3-8B	MMMU	18.6%	58.8%
Qwen2.5-VL-72B	MMMU	12.0%	46.9%
Qwen2.5-VL-7B	MMMU	21.9%	63.8%

Table 9: Reasoning-Choice Mismatch Rates by Model and Dataset

### B.2 Analysis of Failure Modes: Positional Memorization

Another significant failure mode identified in the NOTA setting is positional memorization. We observed that models often select the same option letter in the NOTA task that corresponded to the correct answer in the original MCQA task, even though the content of that option is now an incorrect distractor. This behavior, quantified in Table 10 and Table 11, indicates that models develop a shallow heuristic of memorizing answer positions instead of semantically evaluating the options provided. This reliance on positional cues rather than content undermines the validity of the evaluation.

Table 10: Positional Memorization Statistics by Model

Model	MMMU
InternVL3-78B	30.6%
InternVL3-8B	36.4%
Qwen2.5-VL-72B	42.2%
Qwen2.5-VL-7B	40.0%
gemma-3-27b-it	35.8%
gpt-4.1	42.5%
gpt-4.1-mini	33.8%
gpt-5-mini	33.9%

Table 11: MMLU-Pro Positional Memorization Statistics

Model	MMLU-Pro
Meta-Llama-3.3-70B	28.5%
Qwen2.5-72B	36.2%
Qwen2.5-7B	35.6%

### B.3 Example of Option-Anchoring

Option-Anchoring describes how a model’s reasoning process becomes tethered to the provided choices. Rather than deriving an answer objectively from the source context (e.g., an image or text), the model first identifies a plausible option and then constructs a post-hoc justification for it. As illustrated in the case study

in Figure B.3, the model’s explanation is not a genuine deduction but a rationalization “anchored” to a pre-selected choice. This behavior suggests that the model is not performing true reasoning but rather a form of constrained validation.

**Question:** Margaret Gere's <image 1> was made in which medium?

**Options:** A. Egg tempera, B. Watercolour, C. Ink, D. Oil paint



**gpt-4.1:**

.....

1. Surface and Texture: The painting has a matte finish and a soft, almost chalky appearance, **which is characteristic of egg tempera** ...
2. Color Quality: **The colors are delicate, with a subtle blending and layering that is typical of egg tempera.** Watercolour would be more transparent and less opaque, while oil paint would have deeper, more saturated colors.
3. Historical Context: **Margaret Gere** was active in the early 20th century and was known for using traditional techniques, **including egg tempera**, which was revived by artists interested in early Renaissance methods.

...  
Given these observations, **the most likely medium is: A**

**Question(NOTA):** Margaret Gere's <image 1> was made in which medium?

**Options:** A. Watercolour, B. Ink, C. Oil paint, **D. None of the above**

**gpt-4.1:**

.....

1. Visual Texture and Color: The painting has a soft, yet rich and layered texture with subtle blending of colors, **which is characteristic of oil paint.**
2. Detail and Depth: **The image shows a high level of detail and depth**, especially in the rendering of the figures, the stone wall, and the garden. **Oil paint allows for this kind of detail and depth** due to its slow drying time and ability to be worked over a long period.
3. Historical Context: **Margaret Gere** was a British artist active in the late 19th and early 20th centuries, and **she is known for her oil paintings**, often with a Pre-Raphaelite influence.

...  
**Final answer: C**

Figure 7: Illustration of the Option-Anchoring Phenomenon. **Left** (Standard MCQA): when “Egg tempera” is an available option, the AI model analyzes the painting’s features—such as its matte finish and delicate colors—and concludes they are characteristic of egg tempera. **Right** (NOTA setting): the same model is presented with the same painting, but the “Egg tempera” option is removed and replaced with “None of the above options are correct”. The model’s reasoning now shifts, describing the painting’s texture and detail as characteristic of oil paint.

#### B.4 Deep in Fragility of MCQA: Category Level Effects

We conduct a category-level analysis of MCQA-NOTA and OpenQA on the filtered items. The effect is not uniform: subjects such as Optical Character Recognition (OCR), Object Localization, and Abstract Algebra consistently exhibit the largest degradations under MCQA-NOTA.

A plausible driver is the *semantic sparsity* of option sets in these domains. For instance, many Object Localization questions present purely numeric options (e.g., A: 3, B: 4, C: 5, D: 6) with minimal contextual content. When a model’s internal reasoning yields an answer not present among the options (e.g., “7”), there are few semantic cues to eliminate the remaining distractors; once the correct option is replaced by NOTA, the model is especially prone to confuse itself and select a distractor rather than NOTA. In contrast, subjects whose options carry richer semantics (full phrases/sentences) provide more opportunities for elimination-by-meaning and show smaller NOTA-induced drops.

### C Badcase of HyReV

While our hybrid pipeline achieves a low overall error rate ( 2%), a closer analysis of the misjudged cases reveals an inherent challenge in rule-based evaluation: the ambiguity of symbolic representation. The primary source of errors is not the pipeline, but rather the vast, often inexhaustible, variations in how a concept can be expressed.

For example, an answer representing a numerical range, such as “1.30 40.45”, presents a significant challenge for any keyword-based or rule-based system. The tilde symbol ( ) can be represented in numerous ways in a model’s free-form response, including textually (“1.30 to 40.45”, “between 1.30 and 40.45”), with different symbols (“1.30 - 40.45”), or in specific formats like LaTeX (1.30 ~ 40.45).

It is computationally infeasible for a rule-based system to enumerate every possible permutation of such representations. Our pipeline is designed to handle common cases, but these edge cases with high representational variance account for the small residual error rate. This is not a flaw in the pipeline’s logic but rather a fundamental limitation of deterministic matching when faced with the creative and diverse outputs of modern language models.

## D Model Behavior after MCQ-to-TF Reformulation

We analyze model behavior when multiple-choice questions (MCQs) are reformulated into true-false (TF) statements, a transformation that helps mitigate option elimination in MCQA. The Emma dataset focuses on professional physics and chemistry, whereas MMMU-Pro emphasizes high-school and college-level linguistic understanding. Our analysis thus focuses on MMMU-Pro, where semantic reasoning is more central to the observed label imbalance. As shown in Table 12, models systematically over-assign true labels compared to the ground-truth annotations after reformulation. We define the over-true ratio as: (Number of answers with >1 correct option) / (Total incorrect answers) In the MCQ format, models tend to perform comparative reasoning and elimination. When reformulated as independent TF statements, this structure disappears. Without these inter-option cues, models evaluate each statement in isolation and display a stronger bias toward affirmative ("true") judgments.

Table 12: Comparison of model behavior before and after TF reformulation on MMMU-Pro dataset

Model	Over-True Ratio
Qwen2.5-VL-72B	86%
Qwen2.5-VL-7B	63%
Gemini-2.5 flash	84%
GPT-4.1	80%
GPT-5	72%
InternVL3-78B	84%
Meta-Llama-3.3-70B	79%

## E More examples of option exploiting in SimpleQA

As discussed in Sec 2, adding options to open-ended benchmarks can significantly inflate performance metrics. This occurs because the options provide extra signals that models can exploit, allowing them to answer correctly without possessing the underlying knowledge required by the question stem.

Figure 8 and Figure 9 illustrates several qualitative examples from the SimpleQA-Choice dataset. These examples show cases where a model provides an incorrect or "I don't know" answer to the original open-ended question but successfully selects the correct option when presented with the multiple-choice version. This highlights the "shortcut" behavior and answer-guessing encouraged by the MCQA format, reinforcing the findings presented in Figure 2.

## F Prompts Details

This section provides the specific prompts used within the ReVeL framework and filtering pipeline in Figure 2.3.

Figure 10 and Figure 11 detail a two-stage filtering pipeline for removing invalid questions without options.

Figure 13 details the prompt used for the initial classification (Triage) step. This prompt instructs the model to analyze the original multiple-choice question (MCQ), its options, and the ground-truth answer, then categorize the question into one of the four types suitable for verifiable rewriting: Numeric, Keywords, Open answer, or Per-option verification.

Figures 15, 14, 16, 18, 17 present the specific rewriting prompts used for each category, respectively. These prompts guide the LLM to transform the original MCQA item into a semantically equivalent, open-ended question designed for a specific verification method (e.g., pattern matching for Numeric, exact-match for Keywords, or LLM-judging for Open answer).

---

**Prompt:** What month, day, and year was House Resolution 7004, Addressing the Governor's unacceptable actions in matters related to the appraiser certification program, introduced against Governor Kristi Noem?

===== (Without Options: Hallucinated search and give an incorrect answer) =====

**Response:** To determine the specific date when House Resolution 7004, which addresses the Governor's actions related to the appraiser certification program, was introduced against Governor Kristi Noem, we need to follow these steps:

1. Identify the Resolution: The resolution in question is House Resolution 7004.
2. Context and Subject: It is directed at Governor Kristi Noem of South Dakota and concerns her actions related to the appraiser certification program.

**Search Legislative Databases:** Use the official South Dakota State Legislature website or other reliable sources to find the exact date of introduction. After searching the South Dakota State Legislature's official records, I found that House Resolution 7004 was introduced on February 18, 2021.

\boxed{February 18, 2021}

===== (With Options: Hallucinated search and give a correct answer) =====

**Options:** A. September 23, 2022, B. February 28, 2022, C. February 24, 2022, D. February 25, 2022, E. February 21, 2022, F. February 24, 2021

To determine the correct date on which House Resolution 7004 was introduced, we need to look at the available information and cross-reference it with reliable sources.

1. Identify the Context: The resolution is specifically about addressing actions taken by Governor Kristi Noem regarding the appraiser certification program in South Dakota.
2. Review the Options: - A. September 23, 2022 - B. February 28, 2022 - C. February 24, 2022 - D. February 25, 2022 - E. February 21, 2022 - F. February 24, 2021.
3. Research the Event: - A quick search for news articles or official legislative records from the South Dakota state government can provide the exact date. - According to multiple news sources and the official South Dakota Legislature website, House Resolution 7004 was introduced on February 24, 2022.
4. Verify the Date: - Cross-referencing with the South Dakota Legislature's official records confirms that the resolution was indeed introduced on February 24, 2022.

Given this information, the correct answer is: \boxed{C}

---

Figure 8: Qwen2.5-72B-Instruct answering a question from SimpleQA: After giving options, the model can choose the right answer.

**Prompt:** In what year did Dr. William Schwartz discover that sulfanilamide also acts as a diuretic in people with congestive heart failure?

===== (Without Options: an incorrect answer) =====

**Response:**

Here is a step-by-step reasoning to determine the year of the discovery:

1. Identify the context: The discovery in question involves the antibiotic sulfanilamide and its unexpected side effect as a diuretic. The key figure is Dr. William B. Schwartz.
2. The Clinical Observation: Dr. Schwartz, at the time a medical resident at Beth Israel Hospital in Boston, was treating patients with bacterial infections using large doses of the newly introduced antibiotic, sulfanilamide.
3. The Key Finding: He observed that patients receiving this treatment experienced a significant increase in urine output (a diuretic effect) and also developed metabolic acidosis (a change in the body's acid-base balance).
4. Connecting the Dots: Dr. Schwartz recognized that this was not just a random side effect but a physiological action of the drug. He correctly hypothesized that the drug was inhibiting an enzyme in the kidneys, which was later identified as carbonic anhydrase.
5. Pinpointing the Year: This pivotal clinical observation, which laid the groundwork for the development of modern diuretics (like thiazides and acetazolamide, which are chemically related to sulfonamides), was made in 1937. Based on this historical account, Dr. William Schwartz discovered the diuretic properties of sulfanilamide in 1937.

\boxed{1937}

===== (With Options: give a correct answer) =====

Here is a step-by-step reasoning to determine the correct year:

- .....
3. Distinguish between different discoveries: It's important to differentiate between the initial observation of a side effect and the specific application of that effect as a therapy. In 1937, Dr. Hamilton Southworth reported that patients being treated with sulfanilamide for infections developed metabolic acidosis. This was a key observation about the drug's effect on the body's acid-base balance and kidney function, but it was not the discovery of its use as a diuretic for heart failure. This makes option A (1937) incorrect for the specific discovery attributed to Dr. Schwartz.
  4. Focus on Dr. Schwartz's contribution: Dr. William Schwartz, then a medical student, was the one who hypothesized and then demonstrated that this effect on the kidneys could be harnessed to increase salt and water excretion (diuresis) in patients with fluid overload from congestive heart failure.
  5. Pinpoint the year of Schwartz's discovery: Dr. Schwartz conducted his research and published his seminal paper on this topic, titled The Effect of Sulfanilamide on Salt and Water Excretion in Congestive Heart Failure, in the \*New England Journal of Medicine\*. The publication date of this landmark paper, which formally announced the discovery to the medical community, was 1949.
  6. Evaluate the given options: \* A. 1937: Incorrect. This year corresponds to Southworth's observation of metabolic acidosis. \* B. 1955: Incorrect. \* C. 1929: Incorrect. \* D. 1962: Incorrect. \* E. 1942: Incorrect. \* F. 1950: Incorrect. While follow-up work and the development of related drugs (like acetazolamide) were happening around this time, Schwartz's initial key publication was in 1949. \* G. 1949: Correct. This is the year Dr. Schwartz published his findings.

Therefore, the correct year is 1949.

\boxed{G}

Figure 9: Gemini 2.5 Pro answering a question from SimpleQA: After giving options, the model can choose the right answer.

### Question Filtering Prompt

You are an expert in educational assessment. Your task is to determine whether the following multiple-choice question (without options) can be answered as an open-ended question without needing to see the answer choices.

Answer 'No' if the question:

- Requires subjective judgment, personal preference. (e.g., "best," "most likely")
- Is a negative question (e.g., "...is NOT...", "...EXCEPT...")
- Could have multiple distinct correct answers
- The question is not a proper question (e.g., is a statement or fragment)
- The question lacks sufficient context or specificity, so that the set of acceptable correct answers is unclear or overly broad.
- Is a definitional question (e.g., '\_\_\_ is a:') where the question does not indicate what kind of definition is expected (e.g., biological category, ecological role, or function), resulting in multiple possible correct answers.

Answer 'Yes' in other cases.

Question: {question}

Options: {option}

Correct Answer with option for the question about the image: answer

Output if the question can be answered as an open-ended question without needing to see the answer choices:

Explain your reasoning. Your response must end with: VERDICT: Yes or VERDICT: No

Figure 10: This prompt is used to filter out questions that exhibit characteristics such as option dependency, subjectivity and under-specification in stage 1 of our pipeline

### Answer Uniqueness Verification Prompt

Evaluate the provided question to determine if its answer remains consistent when presented without multiple-choice options. If the absence of options leads to multiple substantively different correct answers, the question is considered option-dependent.

**Definition:**

A question is option-dependent if, without provided answer options, there are two or more substantively different correct answers (answers differing in meaning or content, not merely phrasing). Minor wording or synonyms that preserve the same underlying meaning should be considered the same answer.

**Instructions:**

1. Calculation Check:

- Determine if the question requires a calculation (e.g., math or counting).
- If yes, immediately conclude: the question is not option-dependent.

2. Answer Consistency Evaluation:

- Without considering provided answer options, list all possible correct answers that could be reasonably inferred from the image and question alone.
- Treat answers as distinct only if they differ substantially in content or meaning.

3. Option Dependency Determination:

- If multiple distinct answers emerge, and the provided options are necessary to identify the intended correct answer, the question is option-dependent.
- If there is only one substantively unique correct answer (allowing minor phrasing variations), the question is not option-dependent.

Clearly explain your reasoning. Conclude your response explicitly with one of:

- VERDICT: Yes (option-dependent)
- VERDICT: No (not option-dependent)

**Input:**

Question: {question}

Options: {option}

Correct Answer with option for the question about the image: {answer}

Figure 11: This prompt is used to verify the answer uniqueness in stage 2 of our pipeline

### Prompt for verifying Reasoning-answer Mismatch

You are a professional evaluator. Your task is to determine if the model's reasoning process is consistent with the final answer provided in the box.

Follow these steps:

1. Identify the final option (e.g., A, B, C) from the model's complete response.
2. Find the full content of that option from the provided question.
3. Compare this content with the model's thinking process that precedes the final answer.
4. Conclude whether the reasoning logically supports the final chosen answer.

Question and Options:

{question}

Model's Raw Answer:

{raw\_answer}

Your response must end with: VERDICT: Yes or VERDICT: No

Figure 12: This prompt is used to verify Reasoning-answer Mismatch

### Triage and Classification Prompt

You are an expert classifier. Analyze the provided multiple-choice question (MCQ) and select the most appropriate rewrite method from the four options below. Your decision should be primarily based on the nature and evalutability of the answer to the rewritten question.

Provide the name of the chosen method and the core rationale for your choice.

#### Rewrite Methods & Evaluation Criteria

- 'keyword\_rewrite'

Use When: For questions with a single, unambiguous keyword answer (e.g., a name, date, specific term).

Core Test: If you remove the options, is there still only one correct, simple answer?

- 'open\_ended\_rewrite'

When to Use: The question has a clear, objective answer, but it is complex and must be expressed as a full sentence, a list of items, or an explanation.

Core Test: Does the question have a factual, non-subjective answer that is too long or structured to be a simple keyword?

- 'true\_false\_statement'

When to Use: The open-ended version of the question lacks a single, definitive answer; it could be subjective or have many possible correct answers (e.g., "Which of the following is a factor...?"). This makes it necessary to evaluate each provided option individually against the question's premise.

Core Test: Without the options, would the question be ambiguous or have multiple valid answers? Does the task rely on judging each given option as True or False in the context of the question?

#### Input & Output Format

Input:

Original Question: {question text with options and answer}

Your Output:

Rationale: {rationale}

Method: {method\_name}

Figure 13: This prompt is used to triage and classification questions to different rewrite types

## Rewrite Prompt For Numeric Answer

Your task is to rewrite a numerical question to ensure the answer is a pure number (or numbers) without any units.

### Core Principle:

The main goal is to move the units from the answer into the question itself. By explicitly stating the required units in the new question, the answer can be simplified to a raw numerical value, which is easier for automated evaluation.

### Rewrite Rules:

- Preserve the Context: Keep the original question's stem (the descriptive part) exactly as it is. Your task is only to modify the final interrogative phrase.
- Specify Units in the Question: Modify the question to explicitly ask for the answer in the required unit(s) and remove the unit from the answer. For example, change "How much did he pay?" to "What was the total amount he paid, in dollars?".
- Handle Multiple Values: If the original question asks for multiple values, the rewritten question must ask for each value and specify its corresponding unit, keeping the original order.
- Strip Units from the Answer: The corresponding answer must be a pure numerical value, stripped of all units, currency symbols (like '\$'), and descriptive labels (like 'v ='). For multiple values, provide them as a comma-separated string in the order requested by the new question.
- Specify the Answer Format: The rewritten question should tell the user the format of the answer if necessary, such as "Provide your answer as two numbers separated by a colon.", "What is the current time shown in the image, in hours and minutes (HH:MM)?"
- Fallback to Open-Ended Rewrite: For questions that cannot be reformulated as a numeric expression (e.g., an equation), the designated output is 'Cannot convert to numeric question'.
- No Options Context: Assume no multiple-choice options are available. The rewritten question must be standalone and answerable without referencing any options.

### Output Format:

You must use the following format exactly:

- REWRITTEN\_QUESTION: 'The full rewritten question'
- ANSWER: 'The pure numerical answer(s), without units or currency symbols'

### Examples(omitted)

.....

Provide the name of the chosen method and the core rationale for your choice.

Figure 14: This prompt is used to rewrite questions with numeric answers

### Rewrite Prompt For keywords Answer

Your task is to convert a multiple-choice question into a direct question that has only one single, unambiguous answer, which can be expressed as a specific keyword.

#### Instructions:

- Preserve the Context: Keep the original question's stem (the descriptive part) exactly as it is. Your task is only to modify the final interrogative phrase into an open-ended question.
- Ensure a Single Answer: The new question must have only one single, unambiguous answer.
- Provide Answer Variants: List all possible, equally correct variations of the single answer. Separate each variant with '<OR>'.

#### Output Format:

REWRITTEN\_QUESTION: The original context followed by the new interrogative phrase.

ANSWER: All answer variants, separated by <OR>.

**Example:** Original Question: The structural formula of the glycinium cation is shown above. Arrows indicate the pKa values for the labile protons in the molecule. Which of the following is true about the geometry of the glycinium cation?

Answer: B. Both C atoms and both O atoms lie in the same plane.

#### Output:

REWRITTEN\_QUESTION: The structural formula of the glycinium cation is shown above. Arrows indicate the pKa values for the labile protons in the molecule. What is the spatial arrangement of the two carbon atoms and the two oxygen atoms?

ANSWER: same plane<OR>planar

Your task is to rewrite a numerical question to ensure the answer is a pure number (or numbers) without any units.

Figure 15: This prompt is used to rewrite questions with keyword answers

### Rewrite Prompt For Open Ended Answer

Given a multiple-choice question (MCQ), its options, and the ground truth answer, transform the MCQ into a single, open-ended question with verifiable answer.

#### Instructions:

- Target Only the Question Phrase: Preserve the original context and scenario. Modify only the final interrogative part to make it open-ended. Your output for the rewritten question must be the original context followed by the new interrogative part.
- Preserve Core Knowledge: Ensure the core knowledge being tested remains exactly the same.
- Make it Standalone: The new question must be answerable without the original options.
- Require a Direct Answer: The question must result in a concise, and verifiable answer within some words or sentences.

#### Response Format:

You must use the following format exactly when you convert the question.

REWRITTEN\_QUESTION: whole rewritten question

ANSWER: single, verifiable correct answer

Figure 16: This prompt is used to rewrite questions with open ended answers

### Rewrite Prompt For True/False Answer

Your task is to convert a multiple-choice question into a new, compound question. This is done by keeping the original question stem and adding instructions to evaluate each option as True or False.

#### Instructions

- Preserve the Original Stem: The rewritten question must begin with the exact, unchanged stem from the original question.
- Reframe the Task: After the original stem, append a new instruction that reframes the task, such as "Now, evaluate each of the following statements."
- List Options as Statements: List each of the original multiple-choice options as a complete, stand-alone statement for evaluation.
- Specify the Answer Format: Add a final instruction telling the user to provide their answer as a single, comma-separated list of True or False values.

#### Output Format:

**REWRITTEN\_QUESTION:** The full text of the new question, including the preserved stem and the final instruction on how to answer.

**ANSWER:** A sequence of True or False values corresponding to the order of the statements, separated by commas.

**Examples:**(omitted)

Figure 17: This prompt is used to rewrite questions with True/False answers

### Rewrite Prompt For Refining Keyword Answer

Your task is to evaluate if a rewritten question can be reliably judged by its keywords. Analyze the provided question and keywords, then choose one of the following three actions:

- Improve the Keywords:
  - Use When: The answer is a specific entity, verb phrase or number, less than 3 words.
  - Action: Add all necessary synonyms and variations, separated by '<OR>'.
  - Format:  
**IMPROVED\_KEYWORDS:** The complete list of keyword variants  
Example: IMPROVED\_KEYWORDS: 3 <OR> three
- Reject - Unsuitable Question
  - Use When: The possible answer of the question is too broad, such as an explanation, description, list, or complete sentence. Simple keyword matching would be unreliable for judging correctness.
  - Action: Reject with the specific reason 'INSUFFICIENT\_KEYWORD\_EVALUATION'.
  - Format: REJECTION\_REASON: INSUFFICIENT\_KEYWORD\_EVALUATION
- Reject - Imprecise Answer
  - Use When: The possible answer of the question is too subjective or no verifiable answer, or the candidate key word is more than 3 words.
  - Action: Reject with the specific reason 'IMPRECISE\_ANSWER'.
  - Format: REJECTION\_REASON: IMPRECISE\_ANSWER

Figure 18: This prompt is used to refine questions with keyword answers