

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

Supplementary Material

8. Details of Removing Options From MCQA

8.1. A Filtering Pipeline for Self-Sufficient Questions

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 Tab. 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 details the 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.

9. 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.

9.1. 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

9.2. 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 Tab. 10 and Tab. 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.

9.3. Example of Option-Anchoring

Option-Anchoring describes how a model’s reasoning process becomes tethered to the provided choices. Rather

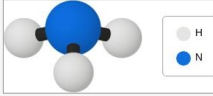

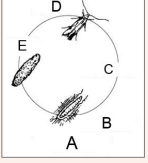
Option dependent	Sentence Completion Questions
<p>Question: This question is based on the following declarations: String strA = "CARROT"; strB = "Carrot", strC = "car"; Given that all uppercase letters precede all lowercase letters when considering alphabetical order, which is true?</p> <p>A: <code>strA.compareTo(strB) < 0 && strB.compareTo(strC) > 0</code> B: <code>strC.compareTo(strB) < 0 && strB.compareTo(strA) < 0</code> C: <code>strB.compareTo(strC) < 0 && strB.compareTo(strA) > 0</code> D: <code>!(strA.compareTo(strB) == 0) && strB.compareTo(strA) < 0</code></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>Question: What was George Seurat's goal in his 'Divisionist' paintings? A: 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: <u>to produce an idiosyncratic style characterized by the use of small dots of color</u></p>	 <p>How long does the larva and feeding tub stage last?</p> <p>A: 4-30 days B: 85-90 days C: 15-30 days D: <u>55 days to 2.5 years</u></p>
Source: MMMU	Source: AI2D

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

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%

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 Sec. 9.3, 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.

9.4. 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), Ob-

ject 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.

10. 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

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.

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.

11. 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.

12. 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 mod-

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%

els can exploit, allowing them to answer correctly without possessing the underlying knowledge required by the question stem.

Fig. 8 and Fig. 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 Fig. 2.

13. Prompts Details

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

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

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.

Fig. 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.

Figs. 14 to 18 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: 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 evaluability 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, standalone 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>'.
Example: IMPROVED_KEYWORDS: 3 <OR> three
 - **Format:**
IMPROVED_KEYWORDS: The complete list of keyword variants
- **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