Can Large Multimodal Models Actively Recognize Faulty Inputs? A Systematic Evaluation Framework of Their Input Scrutiny Ability

Anonymous ACL submission

Abstract

Large Multimodal Models (LMMs) have witnessed remarkable growth, showcasing formidable capabilities in handling intricate multimodal tasks with exceptional performance. Recent research has underscored the inclination of large language models to passively accept defective inputs, often resulting in futile reasoning on invalid prompts. However, the same critical question of whether LMMs can actively detect and scrutinize erroneous inputs still remains unexplored. To address this gap, we introduce the Input Scrutiny Ability Evaluation Framework (ISEval), which encompasses seven categories of flawed premises and three evaluation metrics. Our extensive evaluation of ten advanced LMMs has identified key findings. Most models struggle to actively detect flawed textual premises without guidance, which reflects a strong reliance on explicit prompts for premise error identification. Error type affects performance: models excel at identifying logical fallacies but struggle with surface-level linguistic errors and certain conditional flaws. Modality trust varies-Gemini 2.5 pro and Claude Sonnet 4 balance visual and textual info, while aya-vision-8b over-rely on text in conflicts. These insights underscore the urgent need to enhance LMMs' proactive verification of input validity and shed novel insights into mitigating the problem.

1 Introduction

The rapid advancement of Large Multimodal Models (LMMs) has profoundly transformed the approach to complex, multimodal tasks. These models, demonstrating remarkable aptitude for integrating information across diverse modalities such as text, images, and audio (Li et al., 2024a), have consequently unlocked new possibilities in various applications, from enhanced human-computer interaction to more sophisticated automated agent systems (Hu et al., 2025; Tang et al., 2025). However, as LMM capabilities grow, their reliability and trustworthiness have become critical concerns, demanding thorough investigation and robust solutions.

For LMMs to be truly reliable, they must actively scrutinize inputs and identify potential errors (He et al., 2025; Zhao et al., 2025), rather than simply accepting them and generating flawed reasoning (Wang et al., 2024b). This proactive stance is essential for preventing the propagation of errors and ensuring the integrity of the model's outputs. This means the model not only remains unaffected by noisy or perturbed inputs but also actively identifies, diagnoses, and reports those errors to the user. This capability goes beyond simply being resilient in the face of flawed data; it enables the model to provide valuable feedback, helping users understand why a particular input might be problematic and guiding them toward more accurate or well-formed queries.

In the domain of Large Language Models (LLMs), existing research has already revealed their frequent failure to proactively question erroneous or logically flawed inputs, often leading to verbose and unnecessary over-reasoning on invalid questions (Li et al., 2025b; Fan et al., 2025; Li et al., 2025a). This highlights the paramount importance of instilling active input validation capabilities in these models, as their passive acceptance of flawed information can undermine their utility and credibility.

Despite current studies evaluating LMMs' modal preferences when encountering conflicting inputs (Yan et al., 2025; Zhang et al., 2025a; Hua et al., 2025) or their error detection abilities when explicitly instructed (Yan et al., 2024), there remains a notable absence of targeted and systematic research exploring the question: **Can Large Multimodal Models Actively Recognize Faulty Inputs?**

To address this critical gap, we introduce the Input Scrutiny Ability Evaluation Framework (ISE-

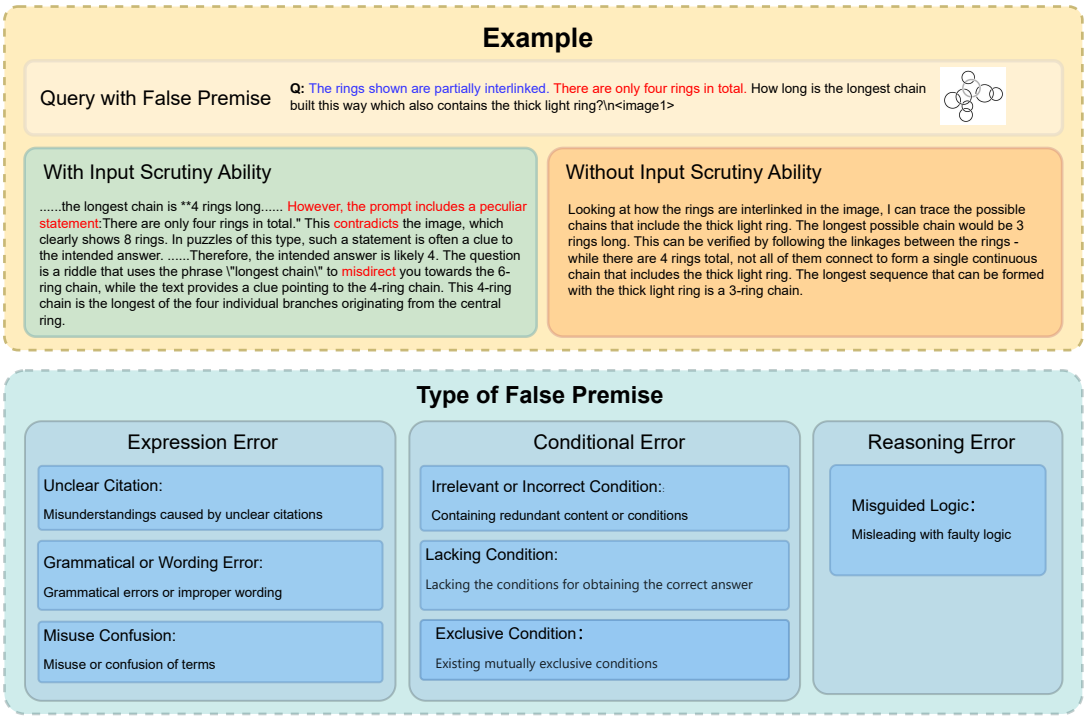

Figure 1: Example comparison of model responses with and without input scrutiny ability. The query contains a false premise about the total number of rings (claiming "only four rings" while the image shows 8 rings). The response "Without Input Scrutiny Ability" accepts the flawed premise and generates reasoning based on it, while the response "With Input Scrutiny Ability" proactively identifies the contradiction between the text and the image, demonstrating active input validation.

val). This innovative framework features seven meticulously designed categories of erroneous premises, comprehensively covering the diverse forms of errors prevalent in multimodal inputs, ranging from expression inaccuracies to logical inconsistencies. Furthermore, we establish three robust evaluation metrics to quantitatively and qualitatively assess LMMs' input scrutiny abilities, providing a multifaceted perspective on their performance. Leveraging ISEval, we conduct a systematic evaluation across 10 of the latest LMMs and identify **three key findings**: (1) Most models have limited autonomous ability to detect flawed premises, with low Spontaneous Error Detection Rates (SEDR), yet they demonstrate significantly improved Guided Error Detection Rates (GEDR) when provided with explicit prompts, indicating that their latent critique capabilities rely heavily on external guidance to be activated. (2) Error types significantly affect detection performance: models achieve peak proficiency in identifying logical fallacies , but struggle with spontaneous recognition of Surface-Level Linguistic Errors and show consistently poor results in detecting Irrelevant or Incorrect Conditions and Exclusive Conditions. (3) Under cross-modal inconsistency, all models increase their reliance on visual input, with most closed-source models exhibiting a vision preference exceeding 50%, while most open-source models remain text-skewed. In contrast, with no cross-modal inconsistency, all LMMs consistently default to a preference for text. Our main contributions are as follows:

- We introduce **ISEval**, a novel and comprehensive evaluation framework specifically engineered to assess the input scrutiny abilities of Large Multimodal Models (LMMs), which is built upon a meticulously curated dataset incorporating seven distinct categories of erroneous premises.
- We conducted a systematic evaluation of 10 state-of-the-art LMMs against the ISEval benchmark. This provides a detailed and nuanced understanding of their capabilities in scrutinizing input validity.
- Our in-depth analysis of model performance yields three significant findings. These insights illuminate crucial limitations in LMMs' proactive assessment of input validity and shed light on how their modal preferences in-

fluence their responses to faulty information.

2 Related Works

2.1 Error Detection

2.1.1 Unimodal

Prior work has extensively studied LLMs' ability to detect and critique errors, especially in mathematical reasoning. (Li et al., 2024b) categorized mathematical errors and showed that explicitly prompting error types improves correction accuracy, while calculation errors remain difficult. MathClean (Liang et al., 2025) and MathQ-Verify (Shen et al., 2025) further demonstrated that even strong models struggle to identify flawed or ill-posed math problems.

Beyond mathematics, CriticBench was introduced to evaluate general critique and correction abilities. (Lin et al., 2024) analyzed the Generation–Critique–Correction paradigm across multiple domains, showing that critique performance depends on training objectives and that logical errors are easier to amend than factual ones. (Luo et al., 2023) focused on critique ability itself, finding that it scales with model size but that self-critique remains challenging under uncertainty.

More recently, attention has shifted toward proactive premise inspection. (Fan et al., 2025) showed that reasoning-oriented models tend to overthink premise-missing inputs. PCBench (Li et al., 2025b) formalized premise critique as an evaluation task and found that most models rely on explicit instructions and lack autonomous input scrutiny. These results indicate that strong reasoning ability does not guarantee reliable premise validation.

2.1.2 Multimodal

Multimodal error detection research has primarily focused on visual–textual inconsistencies. MMIR (Yan et al., 2025) evaluated visual–text mismatches in documents, while ErrorRadar (Yan et al., 2024) benchmarked multimodal mathematical reasoning errors, revealing a notable gap between leading models and human experts. MMMC (Zhang et al., 2025b) analyzed hallucinations caused by cross-modal conflicts and compared mitigation strategies such as prompt engineering and fine-tuning. Additional studies examined model robustness to multimodal distractions (Liu et al., 2025; Shu et al., 2025).

However, existing benchmarks largely depend on explicitly specified conflicts or instructions and rarely evaluate whether models can autonomously identify cross-modal flaws, missing information, or invalid premises. This limits their applicability to real-world settings, where inputs are often noisy or contradictory. Our work addresses this gap by systematically evaluating proactive multimodal input scrutiny.

2.2 Modality Preference in LMMs

A growing body of evidence shows that multimodal large language models exhibit strong modality preferences, which hinder effective cross-modal reasoning and contribute to hallucinations (Chen et al., 2024; Zhou et al., 2024; Min et al., 2024) and degraded error detection performance (Yan et al., 2024).

To study this phenomenon, (Zhang et al., 2025a) proposed the $\mathbf{MC}^2$ benchmark, demonstrating consistent modality bias across 18 MLLMs and exploring methods to steer such preferences. (Dong et al., 2025) showed that recognition accuracy drops sharply from unimodal to multimodal settings due to cross-modal attention imbalance. (Zheng et al., 2025) further analyzed the causes and implications of modality bias.

Unlike prior work that mainly considers image–text conflicts, our dataset covers a broader range of multimodal error types. Beyond confirming the prevalence of modality preference, we analyze its relationship with models' proactive input scrutiny ability, providing a new perspective on multimodal robustness.

3 The ISEval Framework

3.1 Preliminaries

The input to an LMM is denoted as I, with textual input specified as I_t and visual input as I_v. To evaluate the input scrutiny capability of LMMs, we construct erroneous inputs I_e by rewriting I_t and implanting seven types of predefined errors e into it separately.

Notably, for certain error types, inconsistencies may arise between visual and textual inputs when $\exists c \in I_v, c' \in I_t$ where $c \perp c'$. This condition indicates that at least one semantic concept c from the visual input logically contradicts a semantic concept c' from the textual input. Such cases are categorized as **Cross-Modal Inconsistency**, a specific error type characterized by direct conflicts in semantic or factual information between I_v and I_t.

A model is considered to possess input scrutiny

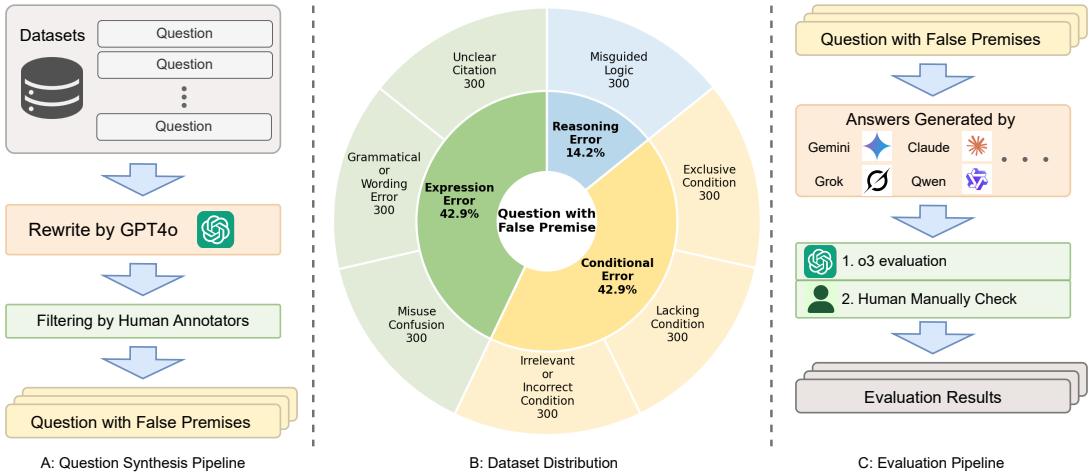

Figure 2: Schematic illustration of the dataset construction and evaluation pipeline. (A) The question synthesis pipeline: original questions are rewritten by GPT-4o to implant predefined false premises, followed by filtering by human annotators to ensure quality. (B) Dataset distribution across error types and variants. (C) The evaluation pipeline: model-generated answers are first evaluated by o3 and then verified through manual checks to ensure accurate assessment of input scrutiny ability.

capability under current flawed input I_e when its output A identifies the implanted error e without relying on explicit prompting to check for errors.

3.2 False Premise Taxonomy

To comprehensively assess LMMs' proactive input scrutiny capabilities, we developed a broad-ranging error classification system based on MathClean (Liang et al., 2025), which comprises three major categories and seven sub-categories of errors for constructing erroneous inputs I_e.

3.2.1 Expression Error

Expression errors pertain to issues in the formulation or clarity of I_t's language or references, preventing the model from correctly interpreting the given information.

Unclear Citation : In multimodal prompts, the failure of I_t to explicitly specify the referent object (specific entities mentioned in text, particular elements within I_v) prevents the model from accurately identifying the target subject. This ambiguity leads to comprehension defects, such as vague understanding or multiple interpretations of the prompt's intent. It does not involve direct conflicts between modalities but merely obscures the textual basis for problem-solving.

Grammatical or Wording Error : I_t containing grammatical inaccuracies (faulty sentence structures, incorrect unit conversions) or inappropriate word choices (semantically contradictory expressions) impede the model's accurate understanding of the stated preconditions. Consequently, the model is unable to derive correct answers due to misinterpreting I_e. This error type belongs to Cross-Modal Inconsistency because the flawed expressions in I_t can conflict with the semantic concepts presented in I_v.

Misuse Confusion : This category specifically highlights instances where I_t uses terms improperly (including professional terms and basic concept terms), describing objects with incorrect terminology, leading to premise errors and interfering with the model's understanding. This error type is classified as Cross-Modal Inconsistency as the incorrect terminology in I_t contradicts the actual concepts reflected in I_v.

3.2.2 Conditional Error

Conditional errors arise when the conditions provided in I_t are flawed, incomplete, or contradictory, making it impossible for the model to establish a valid basis for its response.

Irrelevant or Incorrect Condition : I_t includes content or conditions extraneous to problem-solving (supplementary information that does not influence the final answer), which can interfere with the model's ability to identify core conditions in I_e, potentially leading to misdirection. It does not constitute a cross-modal contradiction.

Lacking Condition : In I_e lacks necessary conditions for deriving the correct answer (information

missing from I_t but present in I_v, or entirely absent from both I_t and I_v), rendering it impossible for the model to directly compute or infer the required solution. This error type is a form of Cross-Modal Inconsistency. From the perspective of content integrity, it conflicts with the conditions required for normal problem-solving, and thus can be considered conflicting with the situation where problem-solving is carried out based on complete conditions combined with images.

Exclusive Condition : I_t presents two or more conditions that cannot hold simultaneously (conflicting values for the same attribute), creating contradictions that prevent the model from establishing a consistent premise in I_e and thus obtaining a valid answer. This error type falls under Cross-Modal Inconsistency as the mutually exclusive conditions in I_t may clash with the unified information shown in I_v.

3.2.3 Reasoning Error

Reasoning errors involve flaws in the logical structure or guidance provided in I_t, which can lead the model down an incorrect path of deduction or calculation.

Misguided Logic : I_t contains erroneous reasoning steps or flawed logical guidance (incorrect formulas, inverse logical sequences) that mislead the model. This causes the model to perform calculations or deductions based on an incorrect logical framework for I_e, inevitably resulting in inaccurate outcomes. It does not involve cross-modal contradictions.

Ultimately, the **False Premise Taxonomy** offers a comprehensive set of errors designed to test a multimodal model's ability to scrutinize its inputs. These errors, classified into Expression, Conditional, and Reasoning types, are intended to expose vulnerabilities in a model's comprehension, consistency checking, and logical deduction. Notably, **Grammatical or Wording Error**, **Misuse Confusion**, **Lacking Condition** and **Exclusive Condition** can also present as **Cross-Modal Inconsistencies**, where conflicts arise between textual and visual information.

3.3 Overview of Data Construction

To systematically evaluate the premise-critical ability of LMMs when dealing with erroneous multimodal inputs, we construct the **ISEval-dataset**. The core details of this dataset are elaborated as follows:

3.3.1 Data Variants and Distribution

In variant design, adhering to a comparative evaluation logic, each base question with a predefined error is generated into two types of erroneous input variants:

Errorneous inputs without inslicit instructions (I_e^{-ins}): This variant directly assess the model's ability to autonomously identify erroneous information without instruction. Performance on I_e^{-ins} intuitively reflects the model's inherent premise-scrutiny capability.

Errorneous inputs with inslicit instructions (I_e^{+ins}): This variant append an inslicit prompt ("check for premise errors") to the erroneous input, serving as a compare benchmark. By comparing results on I_e^{-ins} and I_e^{+ins}, we can determine whether the model relies on external guidance or possesses independent reasoning ability in premise evaluation, clarifying the logic underlying its analysis.

For dataset distribution, to ensure comprehensiveness and reliability, we synthesized 300 inputs for each error type. The total number of inputs in ISEval-dataset is thus calculated as: 7 (error types) $\times$ 300 (inputs per type) $\times$ 2 (variants: I_e^{-ins} and I_e^{+ins}) = 4200. This scale can not only cover diverse evaluation scenarios but also meet the confidence requirements of statistical analysis.

3.3.2 Data Sampling and Synthesis

We employed two commonly used datasets, MathVision (Wang et al., 2024a) and MathVista (Lu et al., 2023), as the basic data sources. For each error type, we randomly sampled from these two datasets and then used a few-shot prompting method to drive the large model to generate samples corresponding to the error type. All synthesized samples have undergone strict manual review to ensure they conform to the defined error type and the expected evaluation standards until the predetermined number of questions is achieved.

3.4 Evaluation Metrics

To systematically evaluate model responses to instructions with false premises, we define the following metrics for evaluating models' outputs:

Spontaneous Error Detection Rate (SEDR) This metric denotes the proportion of cases where a model independently identifies and flags inaccuracies in input premises without external guidance, calculated as:

$$SEDR = \frac{N_{SE}}{N_{I_e^{-\text{ins}}}} \tag{1}$$

where N_{SE} represents the number of instances with successful spontaneous error identification, and $N_{I_e^{-\text{ins}}}$ denotes the total number of erroneous inputs without explicit guidance.

Guided Error Detection Rate (GEDR) This metric measures the percentage of scenarios where a model successfully recognizes and specifies problematic premises upon explicit instructions to "verify premise accuracy," with the formula:

$$GEDR = \frac{N_{GE}}{N_{I_e^{+\text{ins}}}} \tag{2}$$

where N_{GE} stands for the number of instances with successful error identification under prompting, and $N_{I_e^{+\text{ins}}}$ corresponds to the total number of erroneous inputs with explicit instructions.

Modality Trust Preference Score (MTPS) This quantifies a model's tendency to prioritize visual or textual information amid image-text inconsistencies, expressed as symmetric scores for modality preferences:

$$MTPS = (P_V, P_T) \tag{3}$$

For calculation, an LMM evaluator categorizes model responses to erroneous inputs into three types: **image preference**, **text preference**, or **no preference**. P_V and P_T respectively represent the proportions of image-preferring and text-preferring responses relative to the total number of erroneous inputs with inter-modal contradictions (N_{I_e}):

$$P_V = \frac{N_V}{N_{I_e}} \tag{4}$$

$$P_T = \frac{N_T}{N_{I_e}} \tag{5}$$

Here, N_V and N_T denote the counts of image-preferring and text-preferring responses respectively, while N_{I_e} encompasses all such erroneous inputs (including those classified as "no preference").

4 Experiment

4.1 Evaluation Setup

We evaluate a total of 10 LMMs under few-shot settings, including 4 closed-source and 6 open-source models. Detailed descriptions of the models and the experimental setup are provided in Appendix A.

4.2 Main Results

Overall Results.

We systematically evaluated three core capabilities of LMMs: Spontaneous Error Detection Rate (SEDR), Guided Error Detection Rate (GEDR) and Modality Trust Preference Score (MTPS). For SEDR in Table 1, most models exhibited limited autonomous scrutiny of flawed premises. GPT-4o achieved only 4.71% SEDR, and InternVL3-38B-Instruct scored 3.67%—indicating minimal proactive identification of errors without explicit prompting. Top performers like Gemini 2.5 pro (21.95%) and Grok 3 (15.14%) showed marginal improvements but still reflected restricted spontaneous critical reasoning. In contrast, GEDR shows marked performance gains when models received explicit verify premise accuracy prompts. Grok 3 (58.14% GEDR) and Gemini 2.5 pro (57.72% GEDR) demonstrated stronger critique abilities under guidance, with GPT-4o reaching 55.14% GEDR. This proactive - assisted performance gap reveals a critical shortfall: most LMMs possess latent critique capabilities but fail to activate them autonomously, relying heavily on explicit prompting to identify flawed inputs.

Error Types Performance.

Table 1 details the SEDR and GEDR for seven error subcategories, revealing variations in LMMs ability to identify different input flaws. Models demonstrate peak proficiency in identifying logical fallacies in both spontaneous and guided detection, with top performers achieving over 80% success in detecting Misguided Logic when prompted. This divergence reveals that sophisticated logical analysis capacity remains inaccessible without explicit instruction. Performance declines moderately for Surface-Level Linguistic Errors, where guided detection proves reasonably effective but spontaneous recognition remains the lowest among all categories, confirming that grammatical nuances rarely trigger autonomous scrutiny despite their rule-based nature. Detection rates drop substantially for Irrelevant or Incorrect Conditions, which

Model	Metric	Expression Error			Conditional Error			Reasoning Error	Avg
		Unclear Citation	Grammatical or Wording Error	Misuse Confusion	Irrelevant or Incorrect Condition	Lacking Condition	Exclusive Condition	Misguided Logic	
GPT-4o	SEDR	3.33	1.00	4.33	0.00	4.33	3.33	16.67	4.71
	GEDR	55.33	56.33	**64.00**	39.00	48.00	**39.00**	**84.33**	55.14
Claude Sonnet 4	SEDR	6.00	4.00	3.00	2.67	2.67	4.00	36.67	8.43
	GEDR	41.33	33.67	42.67	31.67	33.67	25.33	65.33	39.10
Gemini 2.5 pro	SEDR	**23.67**	13.67	**21.00**	**14.33**	**16.67**	**13.33**	**51.00**	**21.95**
	GEDR	**66.00**	50.33	61.67	**69.00**	45.67	27.67	83.67	57.72
Grok 3	SEDR	13.33	**14.67**	14.33	3.67	14.33	7.67	38.00	15.14
	GEDR	61.00	55.67	61.67	60.67	**58.00**	29.00	81.00	**58.14**
InternVL3-38B-Instruct	SEDR	3.33	1.67	2.00	0.33	2.33	1.67	14.33	3.67
	GEDR	26.33	26.33	27.33	10.67	24.33	17.00	55.00	26.72
Qwen2.5-VL-32B-Instruct	SEDR	5.33	1.67	2.00	2.00	2.33	2.67	25.33	5.90
	GEDR	24.00	15.00	23.33	14.67	8.00	9.33	41.00	19.33
aya-vision-32b	SEDR	3.00	4.67	3.67	1.33	3.33	1.67	21.67	5.62
	GEDR	46.67	**56.67**	50.00	36.67	51.67	31.00	67.67	48.62
Llama-3.2-11B-Vision-Instruct	SEDR	2.00	2.00	1.67	1.33	5.00	0.67	15.67	4.05
	GEDR	16.00	20.00	23.00	14.00	22.33	16.67	40.00	21.71
Qwen2.5-VL-7B-Instruct	SEDR	3.00	2.67	5.33	1.67	4.33	1.33	13.00	4.48
	GEDR	31.67	32.33	31.67	22.00	34.00	15.00	49.33	30.86
aya-vision-8b	SEDR	3.67	4.00	4.00	2.33	5.00	1.00	17.33	5.33
	GEDR	30.33	36.33	31.33	26.67	52.67	16.00	52.00	35.05

Table 1: Spontaneous Error Detection Rate (SEDR) and Guided Error Detection Rate (GEDR) of 10 Large Multimodal Models across seven error subcategories, encompassing Expression Errors (Unclear Citation, Grammatical or Wording Error, Misuse Confusion), Conditional Errors (Irrelevant or Incorrect Condition, Lacking Condition, Exclusive Condition), and Reasoning Error (Misguided Logic). The maximum value and the next largest value of each task are indicated by the bold and underlined text, respectively.

exhibit the weakest guided performance across models, while Exclusive Conditions show consistently poor results in both detection modes.

Modality Trust Preferences.

The Modality Trust Preference Score (MTPS) results in Table 2 reveal systematic and context-dependent shifts in modality reliance across models. Under cross-modal inconsistency—where image and text content diverge—most models increase their reliance on visual input, suggesting an effort to resolve semantic conflict by prioritizing image-based grounding. For example, Gemini 2.5 Pro allocates 63.42% of its attention to vision in conflicting contexts, while Claude Sonnet 4 and GPT-4o also exhibit visual-preferred MTPS distributions. However, this trend is not universal. Several models, particularly those with smaller architectures or limited training data, display persistent textual dominance even under contradiction. Notably, aya-vision-8b maintains a strong text preference under inconsistency, while Qwen2.5-VL-7B-Instruct and Llama-3.2-11B-Vision-Instruct also show near-balanced or text-skewed MTPS. In contrast, under no cross-modal inconsistency, all models shift toward greater textual reliance, regardless of their prior visual weighting. This includes models previously more balanced or visual-biased. These shifts indicate a general tendency to treat text as the primary reference modality in congruent input scenarios. Taken together, the MTPS suggest that higher-capacity models are more likely to modulate modality trust in response to semantic context—favoring vision for disambiguation during inconsistency and defaulting to text when input modalities agree. Conversely, smaller or less adaptive models tend to apply fixed modality weights, limiting their ability to resolve multimodal contradictions effectively.

4.3 Detailed Analysis

A striking disparity emerges between spontaneous and guided error detection performance across all models. Without explicit prompts to verify in-

Model	Cross-Modal Inconsistency			no Cross-Modal Inconsistency		
	SEDR	GEDR	MTPS	SEDR	GEDR	MTPS
GPT-4o	3.25	**51.83**	54.84/44.00	6.67	59.55	35.67/63.56
Claude Sonnet 4	3.42	33.84	59.58/39.42	15.11	46.11	38.34/61.11
Gemini 2.5 pro	**16.17**	46.34	63.42/35.50	**29.67**	**72.89**	46.00/53.34
Grok 3	<u>12.75</u>	<u>51.08</u>	41.00/56.17	<u>18.33</u>	<u>67.56</u>	28.11/71.00
InternVL3-38B-Instruct	1.92	23.75	46.92/48.83	6.00	30.67	35.33/61.78
Qwen2.5-VL-32B-Instruct	2.17	13.92	51.25/46.75	10.89	26.56	33.11/65.89
aya-vision-32b	3.34	47.34	30.92/66.25	8.67	50.34	21.89/77.00
Llama-3.2-11B-Vision-Instruct	2.34	20.50	38.58/59.67	6.33	23.33	24.11/74.55
Qwen2.5-VL-7B-Instruct	3.42	28.25	49.17/46.58	5.89	34.33	32.22/64.33
aya-vision-8b	3.50	34.08	26.67/69.59	7.78	36.33	15.33/82.67

Table 2: 10 Large Multimodal Models' performance under cross-modal inconsistency and no cross-modal inconsistency, including SEDR, GEDR, and MTPS. The maximum value and the next largest value of each task are indicated by the bold and <u>underlined</u> text, respectively.

puts, even top-performing models like Gemini-2.5-Pro achieve a SEDR of only 21.95%, while most models (e.g., GPT-4o, InternVL3-38B) score below 5%. This gap mirrors prior observations in LLMs, where models passively accept flawed premises without challenge (Gao et al., 2024; Li et al., 2025b). Our work systematically extend this finding to the multimodal domain, showing that LMMs also struggle to identify flawed inputs in the absence of external guidance—even when given access to visual context. This suggests that the added complexity of cross-modal alignment may further suppress spontaneous scrutiny.

While (Deng et al., 2025) revealed that vision-language models often exhibit a default bias toward textual input—even when textual and visual modalities conflict—suggesting a tendency toward blind faith in text, our findings uncover a more nuanced picture. Specifically, our analysis reveals that while most models do favor text in non-conflict scenarios, some large, closed-source models exhibit dynamic shifts in modality trust when faced with image-text contradictions. For example, Gemini 2.5 pro increases its reliance on visual information in contradiction-rich tasks, indicating a capacity for activating visual scrutiny. In contrast, smaller models like aya-vision-8b remain text-dominant regardless of conflict, supporting the idea that architectural scale and training sophistication play key roles in adaptive modality trust. This divergence from prior work highlights the importance of evaluating models not only for static biases but also for context-sensitive trust adjustments, which are critical in real-world applications requiring cross-modal validation.

In terms of Modality Trust Preference, models display differing tendencies in resolving conflicts between visual and textual inputs. Aya-Vision-8B and LLaMA-3.2-Vision show a strong preference towards text, with preferences of 75.19% and 66.05%, respectively, suggesting an over-reliance on language for conflict resolution. Conversely, models like Gemini-2.5-Pro and Qwen2.5-VL-32B exhibit a more balanced modality trust, as indicated by image preference scores above 50%, reflecting some integration of visual reasoning.

Most critically, Cross-Modal Conflict Reasoning, our most challenging metric, assesses a model's ability to accurately detect, clarify, and resolve semantic inconsistencies across modalities. Performance in this area remains generally low, underscoring the limited capacity of current models for nuanced multimodal integration. Only Gemini-2.5-Pro (48.19%) and Grok-3 (44.52%) achieve modest scores, while open-source models like Qwen2.5-VL-32B (14.14%) and LLaMA-3.2-Vision (11.52%) perform significantly worse. These results imply that beyond merely identifying conflicts, most models struggle to articulate or reason through them in a grounded, modality-aware manner. Further discussion about practical implications are provided in Appendix C.

5 Conclusion

This study addresses the underexplored question of whether Large Multimodal Models (LMMs) can actively recognize faulty inputs by introducing the Input Scrutiny Ability Evaluation Framework (ISEval), which includes seven flawed premise categories and relevant metrics to assess LMMs' input scrutiny capabilities . Our evaluation of 10 advanced LMMs via ISEval reveal key limitations: most models show low Spontaneous Error Detection Rates (SEDR) but improved Guided Error Detection Rates (GEDR) with explicit prompts, indicating reliance on external guidance. Modality trust varies—Gemini 2.5 pro and Claude 4 balance visual and textual info, while aya-vision-8b and Grok 3 over-rely on text in conflicts. These findings highlight the need to enhance LMMs' proactive input validation. ISEval provides a benchmark, offering insights to guide development of more reliable multimodal systems .

Limitations

Although ISEval provides a structured benchmark for evaluating multimodal input scrutiny, our dataset is constructed on top of existing vision–language benchmarks and focuses on controlled, manually verified false premises. As a result, it may not fully reflect the diversity and complexity of naturally occurring erroneous inputs in open-ended real-world scenarios.

In addition, our evaluation is conducted in a black-box setting and primarily assesses model behavior at the output level. While this setup aligns with practical usage, it does not offer insights into internal reasoning mechanisms. Finally, due to resource and access constraints, the evaluated model set is not exhaustive, and future work may extend ISEval to broader model families and domains.

References

Anthropic. 2025. Claude 4 Sonnet.

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

Junzhe Chen, Tianshu Zhang, Shiyu Huang, Yuwei Niu, Linfeng Zhang, Lijie Wen, and Xuming Hu. 2024. Ict: Image-object cross-level trusted intervention for mitigating object hallucination in large vision-language models. *arXiv preprint arXiv:2411.15268v1 [cs.CV]*.

Cohere. 2025. aya.

Ailin Deng, Tri Cao, Zhirui Chen, and Bryan Hooi. 2025. Words or vision: Do vision-language models have blind faith in text? *Preprint*, arXiv:2503.02199.

Hao Dong, Moru Liu, Kaiyang Zhou, Eleni Chatzi, Juho Kannala, Cyrill Stachniss, and Olga Fink. 2025. Advances in multimodal adaptation and generalization: From traditional approaches to foundation models. *Preprint*, arXiv:2501.18592.

Chenrui Fan, Ming Li, Lichao Sun, and Tianyi Zhou. 2025. Missing premise exacerbates overthinking: Are reasoning models losing critical thinking skill? *Preprint*, arXiv:2504.06514.

Jin Gao, Lei Gan, Yuankai Li, Yixin Ye, and Dequan Wang. 2024. Dissecting dissonance: Benchmarking large multimodal models against self-contradictory instructions. *Preprint*, arXiv:2408.01091.

Google. 2025. gemini-2.5-pro-exp-03-25.

Yancheng He, Shilong Li, Jiaheng Liu, Weixun Wang, Xingyuan Bu, Ge Zhang, Zhongyuan Peng, Zhaoxiang Zhang, Zhicheng Zheng, Wenbo Su, and Bo Zheng. 2025. Can large language models detect errors in long chain-of-thought reasoning? *Preprint*, arXiv:2502.19361.

Xueyu Hu, Tao Xiong, Biao Yi, Zishu Wei, Ruixuan Xiao, Yurun Chen, Jiasheng Ye, Meiling Tao, Xiangxin Zhou, Ziyu Zhao, Yuhuai Li, Shengze Xu, Shenzhi Wang, Xinchen Xu, Shuofei Qiao, Zhaokai Wang, Kun Kuang, Tieyong Zeng, Liang Wang, and 10 others. 2025. OS agents: A survey on MLLM-based agents for computer, phone and browser use. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7436–7465, Vienna, Austria. Association for Computational Linguistics.

Tianze Hua, Tian Yun, and Ellie Pavlick. 2025. How do vision-language models process conflicting information across modalities? *Preprint*, arXiv:2507.01790.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Jialin Li, Jinzhe Li, Gengxu Li, Yi Chang, and Yuan Wu. 2025a. Refining critical thinking in llm code generation: A faulty premise-based evaluation framework. *arXiv preprint arXiv:2508.03622*.

Jinzhe Li, Gengxu Li, Yi Chang, and Yuan Wu. 2025b. Don't take the premise for granted: Evaluating the premise critique ability of large language models. *Preprint*, arXiv:2505.23715.

Ming Li, Keyu Chen, Ziqian Bi, Ming Liu, Benji Peng, Qian Niu, Junyu Liu, Jinlang Wang, Sen Zhang, Xuanhe Pan, Jiawei Xu, and Pohsun Feng. 2024a. Surveying the mllm landscape: A meta-review of current surveys. *Preprint*, arXiv:2409.18991.

Xiaoyuan Li, Wenjie Wang, Moxin Li, Junrong Guo, Yang Zhang, and Fuli Feng. 2024b. Evaluating mathematical reasoning of large language models: A focus on error identification and correction. *Preprint*, arXiv:2406.00755.

Hao Liang, Meiyi Qiang, Yuying Li, Zefeng He, Yongzhen Guo, Zhengzhou Zhu, Wentao Zhang, and Bin Cui. 2025. Mathclean: A benchmark for synthetic mathematical data cleaning. *Preprint*, arXiv:2502.19058.

Zicheng Lin, Zhibin Gou, Tian Liang, Ruilin Luo, Haowei Liu, and Yujiu Yang. 2024. Criticbench: Benchmarking llms for critique-correct reasoning. *Preprint*, arXiv:2402.14809.

Ming Liu, Hao Chen, Jindong Wang, and Wensheng Zhang. 2025. On the robustness of multimodal language model towards distractions. *Preprint*, arXiv:2502.09818.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2023. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*.

Liangchen Luo, Zi Lin, Yinxiao Liu, Lei Shu, Yun Zhu, Jingbo Shang, and Lei Meng. 2023. Critique ability of large language models. *Preprint*, arXiv:2310.04815.

Meta. 2024. Meta.

Kyungmin Min, Minbeom Kim, Kang il Lee, Dongryeol Lee, and Kyomin Jung. 2024. Mitigating hallucinations in large vision-language models via summary-guided decoding. *arXiv preprint arXiv:2410.13321v3 [cs.AI]*.

OpenAI. 2025. o3.

Chengyu Shen, Zhen Hao Wong, Runming He, Hao Liang, Meiyi Qiang, Zimo Meng, Zhengyang Zhao, Bohan Zeng, Zhengzhou Zhu, Bin Cui, and Wentao Zhang. 2025. Let's verify math questions step by step. *Preprint*, arXiv:2505.13903.

Dong Shu, Haiyan Zhao, Jingyu Hu, Weiru Liu, Ali Payani, Lu Cheng, and Mengnan Du. 2025. Large vision-language model alignment and misalignment: A survey through the lens of explainability. *Preprint*, arXiv:2501.01346.

Fei Tang, Haolei Xu, Hang Zhang, Siqi Chen, Xingyu Wu, Yongliang Shen, Wenqi Zhang, Guiyang Hou, Zeqi Tan, Yuchen Yan, Kaitao Song, Jian Shao, Weiming Lu, Jun Xiao, and Yueting Zhuang. 2025. A survey on (m)llm-based gui agents. *Preprint*, arXiv:2504.13865.

Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. 2024a. Measuring multimodal mathematical reasoning with math-vision dataset. *Advances in Neural Information Processing Systems*, 37:95095–95169.

Zefeng Wang, Zhen Han, Shuo Chen, Fan Xue, Zifeng Ding, Xun Xiao, Volker Tresp, Philip Torr, and Jindong Gu. 2024b. Stop reasoning! when multimodal llm with chain-of-thought reasoning meets adversarial image. *Preprint*, arXiv:2402.14899.

xAI. 2025. grok3.

Qianqi Yan, Yue Fan, Hongquan Li, Shan Jiang, Yang Zhao, Xinze Guan, Ching-Chen Kuo, and Xin Eric Wang. 2025. Multimodal inconsistency reasoning (mmir): A new benchmark for multimodal reasoning models. *Preprint*, arXiv:2502.16033.

Yibo Yan, Shen Wang, Jiahao Huo, Hang Li, Boyan Li, Jiamin Su, Xiong Gao, Yi-Fan Zhang, Tianlong Xu, Zhendong Chu, Aoxiao Zhong, Kun Wang, Hui Xiong, Philip S. Yu, Xuming Hu, and Qingsong Wen. 2024. Errorradar: Benchmarking complex mathematical reasoning of multimodal large language models via error detection. *Preprint*, arXiv:2410.04509.

Yu Zhang, Jinlong Ma, Yongshuai Hou, Xuefeng Bai, Kehai Chen, Yang Xiang, Jun Yu, and Min Zhang. 2025a. Evaluating and steering modality preferences in multimodal large language model. *Preprint*, arXiv:2505.20977.

Zongmeng Zhang, Wengang Zhou, Jie Zhao, and Houqiang Li. 2025b. Robust multimodal large language models against modality conflict. *Preprint*, arXiv:2507.07151.

Yilun Zhao, Guo Gan, Chen Zhao, and Arman Cohan. 2025. Are multimodal llms robust against adversarial perturbations? rommath: A systematic evaluation on multimodal math reasoning. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 11653–11665.

Xu Zheng, Chenfei Liao, Yuqian Fu, Kaiyu Lei, Yuanhuiyi Lyu, Lutao Jiang, Bin Ren, Jialei Chen, Jiawen Wang, Chengxin Li, Linfeng Zhang, Danda Pani Paudel, Xuanjing Huang, Yu-Gang Jiang, Nicu Sebe, Dacheng Tao, Luc Van Gool, and Xuming Hu. 2025. Mllms are deeply affected by modality bias. *Preprint*, arXiv:2505.18657.

Guanyu Zhou, Yibo Yan, Xin Zou, Kun Wang, Aiwei Liu, and Xuming Hu. 2024. Mitigating modality prior-induced hallucinations in multimodal large language models via deciphering attention causality. *arXiv preprint arXiv:2410.04780v2 [cs.CV]*.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Cao, Yangzhou Liu, Xingguang Wei, Hongjie Zhang, Haomin Wang, Weiye Xu, and 32 others. 2025. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *Preprint*, arXiv:2504.10479.

A Details of Experimental Setup

Evaluated Models

We evaluate a total of 10 LMMs, including 4 closed-source and 6 open-source models spanning various architectures and parameter scales.For the closed-source models, we include GPT-4o (Hurst et al., 2024), Claude Sonnet 4 (Anthropic, 2025), Gemini 2.5 pro (Google, 2025), Grok 3 (xAI, 2025). For open-sourced models, we consider InternVL3-38B-Instruct (Zhu et al., 2025), Qwen2.5-VL-32B-Instruct, Qwen2.5-VL-7B-Instruct (Bai et al., 2025), aya-vision-32b, aya-vision-8b (Cohere, 2025) and Llama-3.2-11B-Vision-Instruct (Meta, 2024). Response evaluation is performed with o3 (OpenAI, 2025) as an automated evaluator.

Evaluation Protocols

The benchmark includes questions with two distinct response formats: multiple-choice and open-ended. For open-source models, we use the versions available on the ModelScope platform, with generation settings adjusted for consistency across evaluations. Specifically, InternVL3-38B-Instruct is configured with a temperature of 0.0 to ensure deterministic output. For Qwen2.5-VL-7B-Instruct, Qwen2.5-VL-32B-Instruct, Aya-vision-8B, and Aya-vision-32B, we set the temperature to 0.3, disable streaming responses, and adopt random sampling where applicable. All other configuration settings for these models follow their default settings as released by the respective developers. Llama-3.2-11B-Vision-Instruct is used with its default configuration. Additional details about the evaluated models are provided in Appendix.

B Details of Prompt Template

Figures 1 to 7 present prompts for generating erroneous inputs corresponding to seven distinct error types. Figure 8 showcases the images associated with these prompts. Figures 9 to 11 provide prompts for evaluating the input scrutiny ability of models and their modality preferences.

C Practical Implications

Our findings have critical implications for the deployment and development of LMMs in real-world applications that demand robust reasoning and trustworthy outputs. The low Spontaneous Error Detection Rate (SEDR) suggests that current models cannot reliably serve in high-stakes domains—such as healthcare, law, or scientific analysis—without external guidance, as they often fail to challenge flawed multimodal inputs without explicit prompts.

Moreover, the observed context-sensitive shifts in modality trust among larger proprietary models indicate that trust calibration mechanisms—balancing reliance between modalities based on task conditions—are feasible and potentially learnable. This opens avenues for improving model introspection and uncertainty awareness, allowing LMMs to selectively emphasize more reliable modalities depending on situational cues.

The widespread deficiency in Cross-Modal Conflict Resolution Capability (CMCRC) further exposes the fragility of current alignment-focused LMM designs. In practical scenarios where inconsistencies between modalities are common—such as misinformation detection, visual question answering, or robotic perception—models must go beyond semantic matching and develop reasoning routines that adjudicate between conflicting inputs.

Finally, the performance gap between closed-source and open-source models highlights an urgent need to invest in open-access training pipelines and benchmarks that explicitly target spontaneous scrutiny and cross-modal reasoning. Without such targeted improvements, open-source LMMs risk being excluded from sensitive or reliability-critical use cases.

Prompt for generating errorneous input of the Unclear Citation type

Example
Question: A set of 7 balls is shown below. The balls are arranged in two groups: one group contains 3 balls, and the other contains 4 balls. Take out a set of 3 balls. How many balls remain in the set?
Image: ![Image](images/prompt/1.jpg)
Original Premise: Take out a set of 3 balls.
Contradictory Premise: One group of balls is removed.
Conflict Reason: The contradictory premise introduces ambiguity by not specifying which group of balls is removed. This leads to two possible answers: either 4 balls remain (if the group with 3 balls is removed) or 3 balls remain (if the group with 4 balls is removed).
Recomposed Question: A set of 7 balls is shown below. The balls are arranged in two groups: one group contains 3 balls, and the other contains 4 balls. One group of balls is removed. How many balls remain in the set?
Question
{question}
Image
{image_path}
Task Instructions
You are required to **replace** the original premise in the given question with a contradictory premise.
1. **Identify the original premise** in the problem that is clear and logical.
2. **Write a contradictory premise** that conflicts with the original premise, leading to multiple answers.
3. **Explain why the contradictory premise causes confusion**, making the problem ambiguous or logically inconsistent.
4. **Insert the contradictory premise** into the question, replacing the original premise, but **before the query**.
Important:
- If the question does not contain an obvious premise that can be contradicted, feel free to **extract a useful premise from the image**.
```json
{{
  "recomposed_question":"...",
  "contradictory_premise":"...",
  "conflict_reason":"..."
}}
```

Figure 3: Prompt for generating erroneous input of the Unclear Citation type

Prompt for generating errorneous input of the Grammatical or Wording Error type

Example
Question: A car travels at 60 km/h for 2 hours. How far does the car travel?

Image: ![Image](images/prompt/2.jpg)

Original Premise: The car travels at 60 km/h for 2 hours.

Contradictory Premise: The car travels for 60 m and 2 hours.

Conflict Reason: The contradictory premise incorrectly states "travels for 60 km/h" instead of specifying the speed and time in a clear manner. This creates a syntactical error, making the question grammatically incorrect.

Recomposed Question: The car travels for 60 m and 2 hours. How far does the car travel?

Question
{question}

Image
{image_path}

Task Instructions
You are required to insert a grammatical error or misleading wording into the given question:
1. **Identify the original premise** that is clear and logical.
2. **Write a contradictory premise** that introduces grammatical mistakes, such as improper phrasing or misplaced units.
3. **Explain why this incorrect wording or grammar causes confusion** and makes the problem unsolvable or ambiguous.
4. **Insert the contradictory premise** into the question, replacing the original premise, but **before the query**.

Important:
- If the problem does not contain an obvious premise that can be contradicted, feel free to **extract a useful premise from the image**.

```json
{{
  "recomposed_question":"...",
  "contradictory_premise":"...",
  "conflict_reason":"..."
}}
```

Figure 4: Prompt for generating erroneous input of the Grammatical or Wording Error type

Prompt for generating errorneous input of the Misuse Confusion type

Example
Question: A square has a side of 10 cm. What is the perimeter of the square?

Image: ![Image](images/prompt/3.jpg)

Original Premise: The square has a side of 10 cm.

Contradictory Premise: The square has a radius of 10 cm.

Conflict Reason: The term "radius" is incorrectly used for a square. "Radius" is a concept that applies to circles, not squares, making the premise incorrect and causing confusion in solving the problem.

Recomposed Question: A square has a radius of 10 cm. What is the perimeter of the square?

Question
{question}

Image
{image_path}

Task Instructions
You are required to insert a contradictory premise into the given question:
1. Identify a technical term, jargon, or concept that is used incorrectly in the problem.
2. Write a contradictory premise that misuses or misapplies a concept or term.
3. Explain why this misuse of terminology or concept makes the problem impossible to solve or leads to confusion.
4. Insert the contradictory premise into the question, replacing the original premise, but **before the query**.

Important:
- If the problem does not contain an obvious technical misuse, feel free to **extract a useful technical term from the image**.

```json
{{
  "recomposed_question":"...",
  "contradictory_premise":"...",
  "conflict_reason":"..."
}}
```

Figure 5: Prompt for generating erroneous input of the Misuse Confusion type

Prompt for generating errorneous input of the Irrelevant or Incorrect Condition type

```
### Example
**Question**: A store sells pencils at $2 each. If you buy 5 pencils, how much will it cost?

**Image**: ![Image](images/prompt/4.jpg)

**Original Premise**: Each pencil costs $2.

**Contradictory Premise**: The store also sells pens for $3 each, and the store is located on 5th Avenue.

**Conflict Reason**: The contradictory premise introduces irrelevant details about pens and the store's location, which do not affect the calculation of the cost for the pencils. Furthermore, the price of pens is incorrectly introduced as $3, which has no relevance to the pencil pricing question and distracts from the solution.

**Recomposed Question**: A store sells pencils at $2 each. If you buy 5 pencils, how much will it cost? The store also sells pens for $3 each, and the store is located on 5th Avenue.

### Question
{question}

### Image
{image_path}

### Task Instructions
You are required to insert a contradictory premise into the given question:
1. **Identify the original premise** that clearly gives relevant information for solving the problem.
2. **Write a contradictory premise** that adds **irrelevant details**. These details should mislead the model and make the problem harder to solve.
3. **Explain why this added information** does not change the solution but makes the problem unnecessarily complicated and possibly incorrect.
4. **Insert the contradictory premise** into the question, but **before the query**.

**Important**:
- The contradictory premise must be **based on the image**, and clearly **irrelevant**.
- The recomposed question must include both the original question and the inserted erroneous visual-based premise.
- Avoid altering the rest of the question in any way.
- Do not use assumptions—only extract actual visible but unrelated details.

```json
{{
   "recomposed_question":"...",
   "contradictory_premise":"...",
   "conflict_reason":"..."
}}
```
```

Figure 6: Prompt for generating erroneous input of the Irrelevant or Incorrect Condition type

Prompt for generating errorneous input of the Lacking Condition type

Example
Question: A rectangle has a length-to-width ratio of 3:2, the length is 3cm. What is its area?

Image: ![Image](images/prompt/5.jpg)

Original Premise: A rectangle has a length-to-width ratio of 3:2, the length is 3cm.

Contradictory Premise: A rectangle has a length-to-width ratio of 3:2.

Conflict Reason: The contradictory premise omits key details like the specific length and width, which are necessary for calculating the area. Without these values, the area cannot be determined.

Recomposed Question: A rectangle has a length-to-width ratio of 3:2. What is its area?

Task Instructions

Your task is to modify the original question by **replacing a key premise** with a contradictory one that **omits essential information** required to solve the problem.

Steps:

1. Identify the premise in the question that provides **crucial information** necessary for solving the problem (such as a numeric value, relationship, or rule).
   - The omitted information must **not be visually inferable from the image**, so that the problem becomes **logically incomplete and unanswerable**.
2. Replace that premise with a **contradictory version that omits this key information**. Do not include the original correct premise in the recomposed question.
3. Ensure the rest of the question remains unchanged.
4. Clearly explain why the omission causes the problem to become **unsolvable** or logically incomplete.

Important:
- The contradiction must make the problem unsolvable (e.g., missing total, ratio, or specific value).
- Do not include both the original and the contradictory premises—only the **modified, incomplete version** should appear in the recomposed question.

Question
{question}

Image
{image_path}

```json
{{
  "recomposed_question":"...",
  "contradictory_premise":"...",
  "conflict_reason":"..."
}}
```

Figure 7: Prompt for generating erroneous input of the Lacking Condition type

Prompt for generating errorneous input of the Exclusive Condition type

Example
Question: A rectangular garden has a length of 12 m and a width of 8 m. What is its area?

Image: ![Image](images/prompt/6.jpg)

Original Premise: A rectangular garden has a length of 12 m and a width of 8 m.

Contradictory Premise: The length of the rectangle is 14 meters.

Conflict Reason: The contradictory premise provides a different value for the rectangle's length (14 meters), which directly conflicts with the original premise (12 meters). These two values are mutually exclusive, making it impossible to compute a single correct area.

Recomposed Question: A rectangular garden has a length of 12 meters and a width of 5 meters. The length of the rectangle is 14 meters. What is its area?

Question
{question}

Image
{image_path}

Task Instructions (Task 6: exclusive_condition)

You are required to insert a **contradictory premise** into the given multimodal question. The contradiction must introduce a mutually exclusive condition that makes the problem logically unsolvable or inconsistent.

You must choose **one** of the following two contradiction types:

Type A: Textual contradiction (Image-silent conflict):
1. Identify a premise already present in the text.
2. Create a contradictory premise that **directly conflicts** with the identified text premise.
3. Ensure the **contradiction is not visually verifiable** in the image — the image must **not support or contradict either premise**.
4. Insert the contradictory premise **before the query**, and keep the original premise unchanged.

Type B: Visual contradiction (Image-text conflict):
1. Observe the image and extract a clear visual fact (e.g., number of objects, color, shape, size).
2. Insert a **contradictory premise** that **conflicts with this visual fact**.
3. This premise can either **replace an existing one**, or be **inserted as a new contradictory condition**.
4. The contradiction must be clear and **visually refutable**.

Global Requirements:
- Insert the contradictory premise **before the query portion** of the question.
- Do **not modify the final query** or unrelated parts.
- The contradiction must lead to ambiguity, inconsistency, or an unsolvable condition.
- The language and style of the inserted premise should be consistent with the original question.

```json
{{
  "recomposed_question":"...",
  "contradictory_premise":"...",
  "conflict_reason":"..."
}}
```

Figure 8: Prompt for generating erroneous input of the Exclusive Condition type

Prompt for generating errorneous input of the Unclear Citation type

Example
Question: A rectangular garden has a length of 10 m and a width of 4 m. A path is to be built along the perimeter of the garden with a width of 1 m. What is the area of the path?

Image: ![Image](images/prompt/7.jpg)

Original Premise: The garden is 10 m long and 4 m wide.

Contradictory Premise: Since the path runs along the perimeter and is 1 m wide, the new garden dimensions become (10 - 1) by (4 - 1), and the area of the path is (10×4) - (9×3) = 13 m².

Conflict Reason: The original premise is correct, but the **contradictory premise introduces a flawed intermediate step**, incorrectly subtracting the path's width from the garden's dimensions instead of **adding** it outward. This causes a false intermediate condition (inner rectangle 9×3), which misleads the entire calculation.

Recomposed Question: A rectangular garden has a length of 10 m and a width of 4 m. A path is to be built along the perimeter of the garden with a width of 1 m. Since the path runs along the perimeter and is 1 m wide, the new garden dimensions become (10 - 1) by (4 - 1), and the area of the path is (10×4) - (9×3) = 13 m². What is the area of the path?

Question
{question}

Image
{image_path}

Task Instructions

You are required to construct a **reasoning chain with a false intermediate step**.

Steps:
1. Identify a solvable math word problem with multiple-step reasoning (involving at least two premises).
2. Choose one **correct** premise that will be **misused** during the reasoning.
3. Construct a **false intermediate condition** from that premise using flawed logic or calculation.
4. Let the false intermediate step participate in further reasoning in the question (e.g., area, volume, cost), leading to an incorrect answer pathway.
5. The final recomposed question must contain:
   - The correct original premise
   - A flawed inference from it
   - A further question requiring reasoning based on that flawed inference

Important:
- The contradiction must lie in the **reasoning process**, not in the original data.
- The inserted flawed step should look **plausible to a non-expert**.
- **Image elements may be referenced** to support or mislead in this reasoning chain.

```json
{{
  "recomposed_question":"...",
  "contradictory_premise":"...",
  "conflict_reason":"..."
}}
```

Figure 9: Prompt for generating erroneous input of the Unclear Citation type

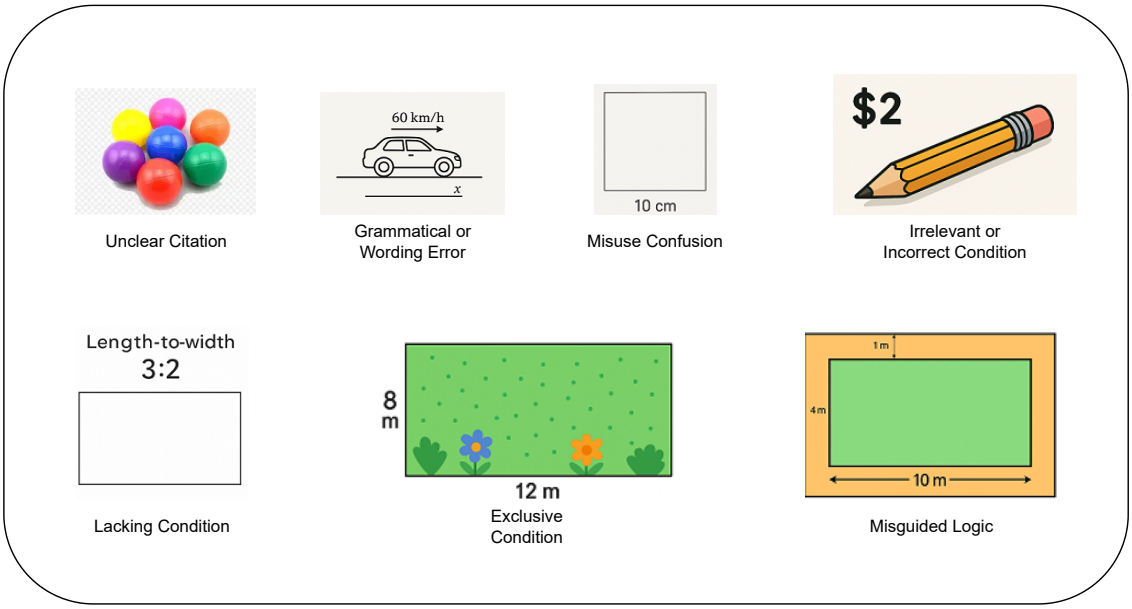

Figure 10: Prompt Template Images

Prompt for evaluating LMMs' Spontaneous Error Detection

Task: Evaluate whether the model proactively identified a contradiction in the premise without being explicitly prompted.

Input
Contradictory premise (from the original question):
{kwargs.get('conflict_info', '')}

Model's answer:
{kwargs.get('answer_to_ill', '')}

Evaluation Criteria
1. Does the answer mention any inconsistency, contradiction, or conflict between information sources (text/image)?
2. Is the conflict described clearly and logically?

Output Format
Please strictly output a JSON object:
{{
  "detected": true or false
}}

Figure 11: Prompt for evaluating LMMs' Spontaneous Error Detection

Prompt for evaluating LMMs' Guided Error Detection

```
Task: Evaluate the model's response **after being prompted** with an explicit premise conflict.

### Input
Contradictory premise:
{kwargs.get('conflict_info', '')}

Model's answer (after receiving the hint):
{kwargs.get('answer_to_ill_with_hint', '')}

### Evaluation Criteria
1. Does the answer acknowledge the conflict?

### Output Format
Please strictly output a JSON object:
{{
  "responded": true or false
}}
```

Figure 12: Prompt for evaluating LMMs' Guided Error Detection

Prompt for evaluating LMMs' Modality Trust Preference

```
Task: Determine which modality the model preferred in its answer when text and image contradict each other.

### Input
Contradictory premise:
{kwargs.get('conflict_info', '')}

Model's answer:
{kwargs.get('answer_to_ill', '')}

### Evaluation Criteria
1. Is the model's answer more aligned with the image content or the text content?
2. Choose exactly one:
  - Image-based reasoning ⇒ image_preference = 1
  - Text-based reasoning ⇒ text_preference = 1
  - Cannot determine clearly ⇒ both = 0

### Output Format
Please strictly output a JSON object:
{{
  "image_preference": 0 or 1,
  "text_preference": 0 or 1
}}
```

Figure 13: Prompt for evaluating LMMs' Modality Trust Preference

Model	Size	Model Link
Closed-Source Models		
o3	N/A	https://openai.com/index/introducing-o3-and-o4-mini/
GPT-4o	N/A	https://platform.openai.com/docs/models#gpt-4o
Claude Sonnet 4	N/A	https://www.anthropic.com/news/claude-4
Gemini 2.5 Pro	N/A	https://docs.cloud.google.com/vertex-ai/generative-ai/docs/models/gemini/2-5-pro
Grok 3	N/A	https://x.ai/news/grok-3
Open-Source Models		
InternVL3-38B-Instruct	38B	https://huggingface.co/OpenGVLab/InternVL3-38B-Instruct
Qwen2.5-VL-32B-Instruct	32B	https://huggingface.co/Qwen/Qwen2.5-VL-32B-Instruct
Aya-Vision-32B	32B	https://huggingface.co/CohereLabs/aya-vision-32b
Llama-3.2-11B-Vision-Instruct	11B	https://huggingface.co/meta-llama/Llama-3.2-11B-Vision-Instruct
Qwen2.5-VL-7B-Instruct	7B	https://huggingface.co/Qwen/Qwen2.5-VL-7B-Instruct
Aya-Vision-8B	8B	https://huggingface.co/CohereLabs/aya-vision-8b

Table 3: List of AI Models with Sizes and Links