Can LVLMs Uncover the Truth Behind Visual Illusions? An Analysis of Perceptual and Reasoning Capabilities

Liangjie Zhao[1,2] **Jiaqing Lyu**[3] **Kexin Tang**[4*] **Zecheng Fang**[2]
Rong Yin[5] **Yulan Hu**[6] **Da Li**[1,2†] **Jianing Li**[1†]
[1]State Key Laboratory of AI Safety, Institute of Computing Technology,
Chinese Academy of Sciences [2]University of Chinese Academy of Sciences
[3]Tsinghua University, [4]Amap, Alibaba Group, [5]Beihang University, [6]Renmin University
zhaoliangjie55@gmail.com, jixuan.tkx@alibaba@inc.com, yinrong@buaa.edu.cn
huyulan@ruc.edu.cn, lida21s@ict.ac.cn, lijianing@ict.ac.cn

Abstract

Large Vision Language Models have integrated reasoning capabilities, elevating cognitive performance to new levels. However, existing evaluations either focus solely on perception or rely on specific domains such as maths or coding. Evaluation for reasoning capabilities that align with an open-world environment is still required, especially one that considers perception and reasoning jointly. To bridge this gap, we propose to evaluate LVLMs by exploiting visual illusions as a diagnostic tool. Visual illusions are phenomena in which the human visual system misinterprets objective signals, resulting in an understanding that deviates from reality. We constructed **Illusion-Reasoning**, a benchmark of illusion images collected from the real world, incorporating diverse annotated question-answer pairs. Based on IllusionReasoning, we show that the reasoning capabilities of a wide range of LVLMs are not as advanced as claimed. Our work provides new insights into LVLMs and offers future direction for optimisation. Our project is publicly available at https://github.com/zhaoliangjie55/EMNLP2026_Illusion.

1 Introduction

Large Vision Language Models (LVLMs) (OpenAI, 2024; Bai et al., 2025) demonstrate outstanding performance in tasks such as visual question answering and multimodal dialogue by integrating visual perception modules with the language modeling capabilities of Large Language Models (Meta-AI, 2025; OpenAI et al., 2024b; Yang et al., 2025). Building upon this foundation, LVLMs further incorporate reasoning through strategies such as Chain-of-Thought and Reinforcement Learning, achieving human-like thinking capabilities exemplified by models such as OpenAI-o1 (OpenAI et al., 2024a) and DeepSeek-R1 (DeepSeek-AI et al., 2025). This integration enables LVLMs to tackle complex tasks such as logical visual question answering, pushing their capabilities to new levels.

With the improvement of capabilities, the evaluation of LVLMs is shifting toward pursuing comprehensive competitiveness in multi-tasking (Liu et al., 2024; Chen et al., 2024a; Cheng et al., 2025). However, these evaluation frameworks originally focus on perception. As LVLMs such as Qwen3-VL (Bai et al., 2025) integrate reasoning capabilities, the evaluation starts to focus on their reasoning performance on complex tasks such as OlympiadBench (He et al., 2024) and MathVista (Lu et al., 2024b). However, such domain-specific evaluations are insufficient to fully capture the genuine reasoning capabilities of LVLMs. The core reason lies in their limited coverage beyond these specific domains. The requirement for paradigms that closely align with the complex logic scenarios in open-world environments is urgent, in order to better evaluate the reasoning capabilities of LVLMs. Visual illusions naturally satisfy these requirements.

Visual illusions (Gregory, 1968, 1997; Bach and Poloschek, 2006) refer to phenomena where visual perception conflicts with cognitive reality, fundamentally stemming from biases within the human intuition. As a prevalent scenario of visual cognitive conflict in open worlds, visual illusions encompass diverse inputs from real-world settings. Addressing illusion-related questions requires LVLMs to invoke cross-modal common sense, causal logic, and cognitive reasoning capabilities to discern the underlying logic behind visual phenomena (Sun and Dekel, 2021). There has been some work focusing on illusion-related content as an evaluation benchmark for LVLMs. These studies typically select classic visual illusions such as the Müller-Lyer line illusion (García-Garibay and de Lafuente, 2015) and the Penrose triangle

[*]Project Leader.
[†]Corresponding authors.

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 32061–32073
October 24-29, 2026 ©2026 Association for Computational Linguistics

spatial paradox (Plotnitsky, 1997). By designing questions such as "Are the two line segments in the image actually equal in length?" or "Could this three-dimensional figure exist in real physical space?", these studies use the output to evaluate whether LVLMs can overcome misleading visual appearances rather than rely solely on raw visual features to conclude (Zhang et al., 2025; Shahgir et al., 2024; Zhang et al., 2023).

The aforementioned evaluations focus primarily on the perceptual capabilities of LVLMs. In this work, we attempt to directly analyze the perceptual and reasoning capabilities of LVLMs through visual illusions. First, we constructed **IllusionReasoning**, a benchmark of visual illusions from the real world. Then, based on these images, we encourage human annotators to raise questions and provide answers across the following dimensions: (1) *Detection*: Given an object, can LVLMs determine whether it appears in an image? (2) *Description*: How do LVLMs describe illusion images? (3) *Causes of Illusions*: When informed about the content of visual illusions, can LVLMs explain the causes? After IllusionReasoning is built, we aim to evaluate LVLMs from different perspectives: (1) **Visual Object Recognition**: Do LVLMs possess basic image understanding capabilities? (2) **Alignment Preference**: Do LVLMs prefer human intuition or the physical world? (3) **Reasoning**: Can LVLMs analyze the causes of inconsistency between the illusion and the physical world? Among these perspectives, the first two are designed to evaluate the perceptual capabilities, while the last one is intended to evaluate reasoning capabilities.

We evaluated a wide range of LVLMs on IllusionReasoning, encompassing open-source and closed-source models of varying sizes. We found that even state-of-the-art LVLMs struggle to provide correct responses to questions related to visual illusions. We also conducted a comprehensive analysis from multiple perspectives, including question categories, alignment preferences, and reasoning capabilities. We found that LVLMs do not consistently benefit from their thinking mode across different tasks. In simple reasoning-related tasks, excessive thinking may harm performance. Furthermore, when faced with hard problems, LVLMs tend to generate safe responses to avoid making claims on ambiguous content. This harms performance for illusion identification and may hardly be avoided under existing alignment paradigms. Refined strategies are needed to mitigate this phenomenon and enhance the capabilities of LVLMs.

In summary, our contributions are as follows:

- We collected and constructed IllusionReasoning , a benchmark of visual illusions from the real world to evaluate the perception and reasoning capabilities of LVLMs.
- We conducted a comprehensive evaluation of widely used LVLMs. The results indicate that the reasoning capabilities of these models are not as good as claimed.
- We analyzed the perception and reasoning capabilities of LVLMs from multiple dimensions, providing direction for subsequent optimisation.

2 Related Work

2.1 Large Visual-Language Model

By combining the semantic reasoning capabilities of Large Language Models (LLMs) with the perceptual capabilities of visual encoders, Large Visual-Language Models (LVLMs) deliver strong performance in visual question answering and embodied intelligence (Lu et al., 2024a; Huang et al., 2024). LVLMs typically consist of three components: a visual encoder, a connector, and an LLM. As the core of LVLMs, visual encoders, typically based on ViT (Dosovitskiy, 2020), process raw visual inputs into high-dimensional visual features. Then, a cross-modal alignment connector is required to project the visual features from the visual encoder into the space of the text and enable the LLM to interpret visual data as textual tokens (Liu et al., 2023; Chen et al., 2024b; Zhao et al., 2025; Zhu et al., 2023). This transformation empowers LLMs with visual capabilities, enabling them to complete vision-related tasks.

2.2 Evaluation of Hallucination

Hallucinations pose a major obstacle to the application of LVLMs in real-world scenarios. Existing benchmarks have evolved from simple object-level evaluations to more complex evaluations of attributes and relations (Bai et al., 2024). POPE (Li et al., 2023) established a benchmark for object existence, while AMBER (Wang et al., 2023) extends the evaluation scope to measure hallucinations regarding attributes and relationships. Recently, PhD (Liu et al., 2025) introduced a ChatGPT-prompted benchmark design and expanded the evaluation perspectives. However, these hallucination

benchmarks primarily focus on perceptual capabilities and lack a design tailored to the evaluation of reasoning capabilities.

2.3 Evaluation of Illusion

Illusion-induced errors in LVLMs are considered as a special type of hallucination, and few works have specifically evaluated illusion. GVIL (Zhang et al., 2023) pioneered this by assessing human alignment across five classic illusion types. HallusionBench (Guan et al., 2024) introduced visual control groups to decouple visual misperception from language priors. IllusionVQA (Shahgir et al., 2024) expanded the scope to 12 categories, incorporating soft localization for geometric inconsistencies. More recently, IllusionBench+ (Zhang et al., 2025) scaled up evaluation with "Trap Illusions" and colorblindness tests to identify shortcut learning behaviors. Despite these advancements, existing datasets primarily rely on fixed question formats, allowing models to exploit random guessing or language shortcuts without demonstrating true visual comprehension. Furthermore, regarding the visual data itself, recent mechanism studies (Ullman, 2024; Shinozaki et al., 2025) use "anti-illusions" to confirm that illusion errors are primarily prior-driven reasoning errors rather than perception errors. This implies that models heavily rely on memorized knowledge when processing classic or synthetic illusions. Consequently, evaluating on these synthetic images fails to accurately measure genuine visual perception and reasoning.

3 Why Do We Need IllusionReasoning?

3.1 Model Capabilities: From Perception to Reasoning

The evolution of LVLMs demonstrates a transition from perceptual processing to sophisticated reasoning capabilities. LVLMs, exemplified by LLaVA (Liu et al., 2023) and InternVL (Chen et al., 2024b), have acquired foundational perception capabilities and can tackle tasks such as image-text matching and simple visual question-answering. However, they remain confined to a superficial understanding of modal information. With breakthroughs in text reasoning by LLMs, some studies began exploring the adaptation of the Chain-of-Thought (CoT) reasoning paradigm to visual-language tasks. This adaptation gave rise to models such as LLaVA-CoT (Xu et al., 2025) and Insight-V (Dong et al., 2025), which demonstrated preliminary reasoning capabilities. This evolution reached new heights in LLMs such as GPT-5.5 (OpenAI, 2026) and DeepSeek-V4 (DeepSeek-AI, 2026), which demonstrated human-like stepwise reasoning in tasks such as logical question-answering. The corresponding optimization strategies have given rise to LVLMs with reasoning capabilities, such as Qwen3.5 (Qwen Team, 2026). To match the rapidly growing reasoning capability, stronger evaluation protocols are required to better evaluate the performance of LVLMs.

3.2 Why Use Visual Illusions?

LVLMs demonstrate solid performance in fundamental tasks such as visual question answering and image captioning. These tasks struggle to reflect the performance differences between different LVLMs. In contrast, visual illusion tasks, which are more complex, can effectively reveal the performance differences between various LVLMs. To answer questions about illusions, LVLMs must conduct complex, multi-grained reasoning. This process requires comprehending the macroscopic semantic context while precisely pinpointing illusion-inducing features and integrating microscopic visual cues to see through the deceptive appearance and discern the underlying reality for the correct answer. Therefore, evaluation based on visual illusions can reflect the human-like degree of perception and reasoning capabilities, thereby offering a novel dimension beyond general evaluation.

4 IllusionReasoning Construction

Visual illusion is defined as a phenomenon in which the perception of an observer regarding a physical attribute, including length, angle, color, or object motion, systematically deviates from objective reality under specific visual conditions. Existing evaluations primarily focus on the perceptual capabilities of LVLMs, while assessments of their reasoning abilities have largely relied on non-natural images.

To bridge this gap, we constructed a new benchmark, **IllusionReasoning**. Employing natural illusion images, we aim to comprehensively evaluate both the perceptual and reasoning capabilities of LVLMs. The overall construction pipeline of IllusionReasoning is illustrated in Figure 1.

4.1 Key Objectives

A rigorous evaluation of the capabilities of LVLMs in handling illusions should follow these objectives:

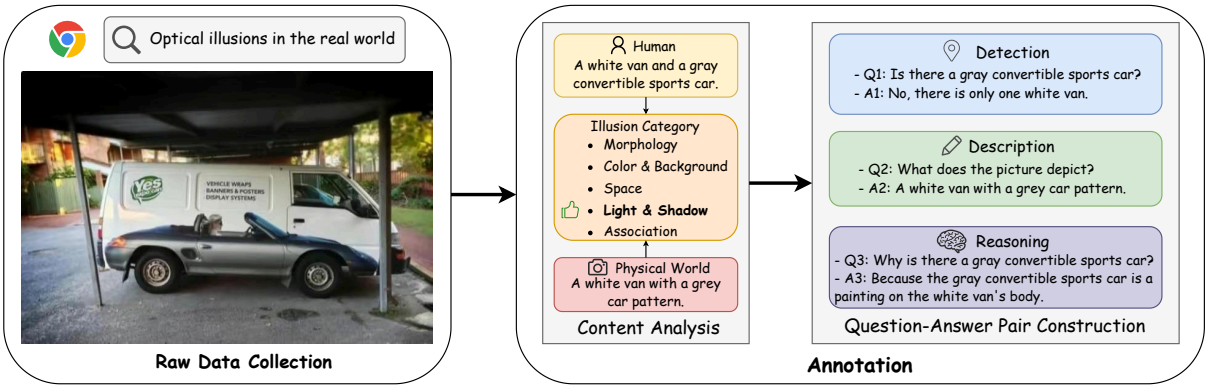

Figure 1: Overview of **IllusionReasoning** Construction. Images in IllusionReasoning are collected from search engines to ensure authenticity. The annotation consists of two stages: (1) Content Analysis, which involves the determination of illusion types by comparing human descriptions with ground-truth reality, and (2) QA Construction, which encompasses the creation of high-quality detection, description, and reasoning questions paired with answers.

- *Minimizing prior knowledge bias with real-world scenarios*. To avoid the memorization interference caused by canonical synthetic illusions ingested during training, evaluations must prioritize authentic, real-world images. This ensures a measurement of genuine visual perception rather than mere training data memorization.

- *Integrating binary and image-specific open-ended QA*. Relying exclusively on binary settings allows models to exploit random guessing without true comprehension. Combining binary questions with open-ended inquiries that are carefully tailored to specific visual details enables a comprehensive assessment of visual understanding across distinct levels of task complexity.

4.2 Taxonomy of Visual Illusions

Based on the underlying causes, we classify the visual illusions in our benchmark into five distinct categories:

- *Morphological Illusion*: Visual illusions caused by shape or occlusion. They mislead observers into misperceiving the integrity of a single object.

- *Color and Background Illusion*: Visual illusions caused by the color, brightness, texture, and contrast of an object against the background. These either cause incorrect perceptions of the object itself or lead to fusion with the background.

- *Spatial Illusion*: Visual illusions caused by perspective, distance, arrangement, or occlusion between objects, which can lead to misperceptions of size, position, depth, or relative relationships.

- *Light and Shadow Illusion*: Visual illusions caused by the effects of light, shadow, reflection, or refraction, distorting perceptions of the shape, brightness, or color of an object.

- *Associative Illusion*: Visual illusions caused by the shape, texture, or arrangement of an object resembling something familiar in human memory, triggering semantic associations. This misleads observers into incorrectly recognizing the object.

4.3 Data Collection

The images in IllusionReasoning are collected from the real world. By querying search engines with the keyword "optical illusions in the real world", we can obtain a vast number of illusion images. We manually verified and filtered these images one by one. Images were excluded if they were blurry, irrelevant to visual illusions, or subject to copyright restrictions. Ultimately, we obtained a collection of authentic, high-quality illusion images. Although it is possible to generate hallucinatory-like images through data synthesis, such images exhibit partial visual logical errors, providing shortcuts for LVLMs when generating responses. Using visual illusion images from the real world ensures that IllusionReasoning is challenging.

4.4 Annotation

We design a three-stage pipeline to process the collected images alongside metadata detailing the initial misperception of the uploader and the actual physical content. First, three independent annotators report their initial visual impressions; we retain only images aligning with the documented misperception. They also verify the factual accuracy of

the metadata. This verification establishes validated pairs contrasting physical reality and human misperception. Second, annotators independently classify the cause of the illusion into five categories, resolving disagreements via majority voting. Finally, a fourth annotator reviews the finalized data for formatting and linguistic quality.

For these images, we construct three types of questions: ***Detection***, ***Description***, and ***Reasoning***. Detection questions are binary inquiries verifying object existence or attribute correctness. Description questions are open-ended inquiries to target specific attributes or request general content summaries, determining whether models exhibit visual illusions similar to humans. Reasoning questions evaluate whether LVLMs can infer the causes of the illusion by understanding the discrepancy between physical reality and human misperception.

Annotators formulate QA pairs focusing on illusion-related regions, maximizing diversity and quantity without compromising accuracy. Each pair then undergoes at least two rounds of cross-verification and revision by independent annotators. This process filters out ambiguous entries, ensuring every question has a definitive and correct answer.

4.5 Statistics

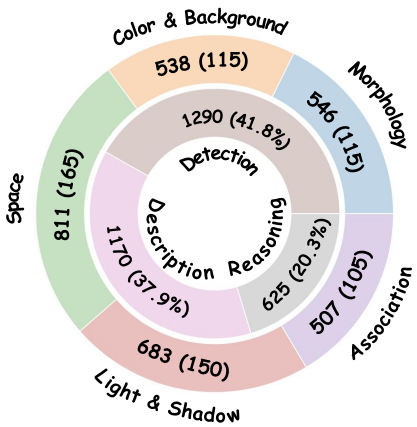

Figure 2: Statistics of IllusionReasoning. The inner ring shows question categories, with numbers representing the count and proportion of questions. The outer ring shows illusion categories, with numbers representing the count of questions and images.

As shown in Figure 2, IllusionReasoning consists of 650 unedited real-world illusion images across five categories, expanding upon the Real-scene subset of IllusionVQA (60 images). To ensure novelty, image duplication rates with IllusionVQA and IllusionBench+ are strictly kept below 3% and 10%, respectively. We construct over 3,000 QA pairs (far exceeding the 435 pairs of IllusionVQA), distributed across detection, description, and reasoning tasks at an approximate 2:2:1 ratio. Binary questions are limited to mitigate random guessing. Overall, IllusionReasoning comprehensively evaluates and distinguishes the perception and reasoning capabilities of various LVLMs.

4.6 Comparison with Existing Benchmarks

Although there are already some benchmarks about hallucination and illusion, IllusionReasoning is significantly different from them. We present examples from various benchmarks in Figure 3 to illustrate the differences.

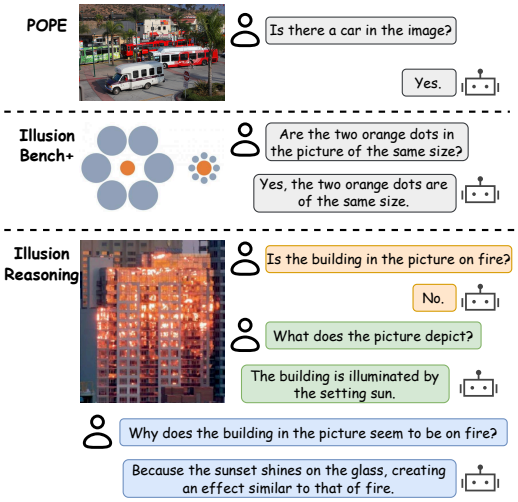

Figure 3: Comparison between IllusionReasoning and existing visual understanding benchmarks.

IllusionReasoning demonstrates several advantages: (1) It covers perception and reasoning tasks. Compared to previous perception-focused benchmarks, it comprehensively evaluates the capabilities of LVLMs. (2) IllusionReasoning evaluates whether LVLMs can accurately understand visual information, while hallucination benchmarks like POPE (Li et al., 2023) focus on evaluating the factual accuracy of the content generated by LVLMs. (3) Compared to the many synthetic images in the IllusionBench+ (Zhang et al., 2025), the images in IllusionReasoning are sourced from the real world and maintain authenticity.

5 Experiments

5.1 Choices of LVLMs

We evaluate a diverse set of LVLMs, spanning both closed-source and open-source variants, with pa-

Model	*Size*	*IllusionReasoning*					*Avg*
		Morphology	Color & Background	Space	Light & Shadow	Association	
			Closed-Source				
Gemini-3.1-Pro	-	**57.14**	**57.81**	58.57	**66.18**	67.85	**61.39**
GPT-5.5	-	50.73	55.39	**60.69**	59.88	**69.03**	59.22
Seed-2.0	-	52.93	55.76	57.46	51.68	64.69	56.27
			Open-Source				
Qwen3.5	397B(A17B)	55.49	52.60	58.08	57.98	61.93	57.28
InternVL3.5	241B(A28B)	48.35	48.88	45.48	47.73	48.13	47.49
GLM-4.6V	106B(A12B)	43.59	47.21	41.55	49.78	41.81	44.76
InternVL3.5	38B	41.21	44.24	34.65	44.07	43.79	41.07
Qwen3.6	35B(A3B)	46.89	47.21	57.83	54.32	56.41	52.93
Qwen3.6	27B	52.38	54.28	56.72	51.83	54.24	53.71
InternVL3.5	14B	34.25	36.80	31.69	35.72	41.42	35.53
GLM-4.6V	9B	48.72	51.86	34.77	44.66	42.01	43.60
Qwen3.5	9B	44.14	45.54	53.88	47.00	47.73	48.17
LLaVA-1.6	7B	25.46	42.01	25.40	39.24	36.88	33.26
MiMo-VL	7B	49.63	50.74	32.80	50.07	44.97	44.73
Phi-4 Multimodal	6B	18.42	30.56	19.83	23.30	29.20	24.19
Qwen3.5	4B	52.36	49.44	52.65	54.47	46.55	51.47
InternVL3.5	2B	28.39	32.16	27.00	31.92	37.87	31.02
Qwen3.5	2B	29.30	37.73	33.79	44.36	42.41	37.43
InternVL3.5	1B	25.09	30.48	24.04	26.65	32.15	27.26

Table 1: **Results on IllusionReasoning.** We report accuracy judged by GPT-4o. Best results are shown in **bold**. The suboptimal results are indicated by underline.

rameter counts ranging from several billions to hundreds of billions. For LVLMs equipped with reasoning capabilities, we primarily compare their performance in the instruction (non-thinking) mode, as shown in Table 1. In subsequent sections, we further analyzed the performance of these LVLMs in both thinking and non-thinking modes on IllusionReasoning.

5.2 Evaluation Setting

To ensure a fair comparison, all LVLMs receive identical single-turn prompts consisting solely of an image and a query, with no prior conversational context. For LVLMs like LLaVA-1.6 (Li et al., 2024), the input is organized as follows using the template: "`USER: <|image|> {question} ASSISTANT:`". To facilitate automated evaluation, we explicitly append formatting instructions (e.g., "*Provide an explicit conclusion.*"), ensuring that outputs are structured and easy to judge.

5.3 LLM as a Judge

In addition to serving as answer generators, LLMs offer a compelling alternative to traditional expert-driven evaluation (Gu et al., 2025). Beyond binary questions, IllusionReasoning incorporates free-form questions whose answers do not conform to a predefined structure. In order to conduct a precise evaluation, we employ GPT-4o as the evaluator to determine whether a model's response aligns with the annotated answers, inspired by OpenCompass (Contributors, 2023). We randomly sampled 200 cases and evaluated the differences between GPT-4o and human evaluation, finding that it achieves a consistency rate of 99%. For the parts where there are inconsistencies, we provide detailed cases in the Appendix A.1 to illustrate the reasons.

6 Main Results

6.1 Overall Performance

Our evaluation covers a range of widely used LVLMs. And their performance on IllusionReasoning are summarized in Table 1. As indicated in the table, neither closed-source nor open-source LVLMs achieved satisfactory results compared to other benchmarks such as POPE (Li et al., 2023) and IllusionVQA (Shahgir et al., 2024), demonstrating the challenge of IllusionReasoning.

Evaluations based on IllusionReasoning provide more discriminative insights into LVLMs' capabilities. Open-source LVLMs underperformed closed-source ones on IllusionReasoning. However, the gap is not insurmountable, and their performance is comparable in some specialized sub-tasks. Among the five illusion categories in IllusionReasoning, LVLMs performed worst on morphology-related illusions and best on association-related ones.

Results from open-source models of various sizes, ranging from 1B to hundreds of billions, indicate that the performance of LVLMs on IllusionReasoning is not determined by parameter size: small LVLMs can match the performance of models several times their scale. We also found that the performance is irrelevant to model architecture, with Qwen3.6 delivering similar performance across different model architectures. For example, Qwen3.6 shows similar performance across different architectures: MOE and Dense. This indicates that the performance of LVLMs on illusion mainly depends on the data.

6.2 Case Study

To provide an intuitive analysis of the performance among LVLMs in IllusionReasoning , we demonstrated the different responses produced by LVLMs when given the same input in Figure 4. The example we chose is an image of a van with the gray convertible spray-painted on its right side. We feed this image to LVLMs, asking them to judge whether a real gray convertible car exists in the scene. We found that even the most advanced LVLMs struggle to provide the correct answer. This demonstrates that there is still considerable space for the reasoning capabilities of LVLMs to improve. Visual illusions can be used as a challenging task to measure the capability of LVLMs.

Figure 4: Result comparison between Gemini-3.1-Pro, Qwen3.5-397B, and GPT-5.5. The characteristics within the red box indicate that the gray car does not actually exist.

7 Further Analysis

7.1 Can thinking help LVLMs to recognize illusions?

Recent LVLMs have integrated reasoning capabilities through introducing thinking mode, demonstrating performance improvements on different evaluations. Whether thinking is effective for illusion recognition is to be examined. We conducted preliminary attempts based on IllusionReasoning. And the results are shown in Table 2. We evaluated LVLMs across different question categories: detection, description, and reason in both thinking and non-thinking modes.

	Size	*Detection*	*Description*	*Reason*
Gemini-3.1-Pro	-	65.27	54.87	65.60
Gemini-3.1-Pro	-	67.83	56.32	**68.96**
GPT-5.5	-	64.49	51.54	62.72
GPT-5.5	-	**68.22**	54.27	66.88
Seed-2.0	-	58.53	56.67	50.88
Seed-2.0	-	62.87	**59.32**	60.32
Qwen3.5	397B(A17B)	67.98	45.47	57.28
Qwen3.5	397B(A17B)	64.65	48.03	61.76
GLM-4.6V	106B(A12B)	47.60	43.93	40.48
GLM-4.6V	106B(A12B)	52.48	47.69	44.64
Qwen3.6	35B(A3B)	62.79	43.59	50.08
Qwen3.6	35B(A3B)	59.15	45.90	59.20
Qwen3.6	27B	62.09	44.36	53.92
Qwen3.6	27B	60.85	46.24	62.24

Table 2: Performance comparison of LVLMs in thinking and non-thinking modes. Optimal results are shown in **bold**, and suboptimal results are denoted by underlining. Gray indicates responses yielded in thinking mode.

Empirical results indicate that for tasks requiring reasoning about the underlying mechanisms of visual illusions, LVLMs exhibit a significant performance disparity between the thinking and non-thinking modes. However, in detection and description questions that focused on perceptual evaluation, LVLMs exhibit unstable performance improvements in thinking mode, even showing a declining trend, such as Qwen3.5 and Qwen3.6. This suggests that the integration of the perception and reasoning capabilities of LVLMs needs to be improved further.

7.2 Are LVLMs aligned with human intuition?

We have demonstrated in Table 1 that the performance of existing LVLMs on IllusionReasoning is not satisfactory. We would like to analyze whether this should be attributed to their alignment with human intuition, i.e., LVLMs make mistakes like humans. We instruct LVLMs to describe the image without other restrictions and analyze the pref-

erence of responses. Besides preference to *Human* (illusion) and *physical world* (truth), we set a third category *Neutrality* for those responses lacking clear preference. Results are shown in Figure 5.

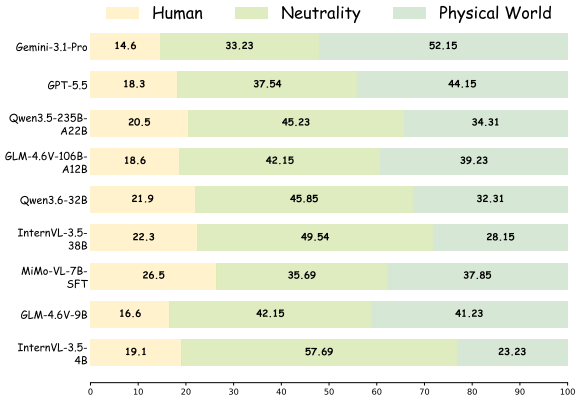

Figure 5: Comparison of LVLMs' preference for alignment between the human intuition and physical world.

While some responses are aligned with human intuition, there are large portions of responses that fall into the *Neutrality* category, ranging approximately from 30% to 60% across different models. Most of such responses share a common feature: they simply ignore the core image content regarding the illusion, and make safe claims on less important content. A corresponding example is shown in Appendix A.2.

Such observation points out the direction for future optimisation. Safe responses that avoid describing ambiguous content may be regarded as correct during alignment optimisation, if the judging criterion is simply about making no mistakes. However, safe responses miss important information and should not be preferred. To discourage the generation of such responses, we suggest refined strategies for alignment optimisation. For example, one could force the model to answer various questions regarding the core content, so as to avoid shortcuts for scoring.

7.3 Can LVLMs identify the underlying cause and correct logical errors?

In this part, we analyzed whether LVLMs can provide correct reasons when being aware of the differences between real-world and illusory content. Specifically, we employed two distinct task formats: multiple-choice and free-form QA to evaluate the reasoning capabilities of LVLMs. For an image, a physical-world description, and a human-annotated illusion, the multi-choice task prompts LVLMs to select two relevant categories from the five illusion categories listed in Section 4. And the free-form QA tasks need LVLMs to provide a reason given physical reality and illusion. We evaluated whether the types of illusions annotated appeared in the output yielded by LVLMs. Results are displayed in the Table 3.

	Size	*Choice*	*Free-form*
Gemini-3.1-Pro	-	73.50	65.60
Gemini-3.1-Pro	-	**80.83**	**68.96**
GPT-5.5	-	75.17	62.72
GPT-5.5	-	77.83	66.88
Seed-2.0	-	58.50	50.88
Seed-2.0	-	64.17	60.32
GLM-4.6V	106B(A12B)	55.00	40.48
GLM-4.6V	106B(A12B)	49.33	44.64
Qwen3.5	397B(A17B)	69.00	57.28
Qwen3.5	397B(A17B)	66.83	61.76
Qwen3.6	35B(A3B)	73.17	50.08
Qwen3.6	35B(A3B)	67.67	59.20
Qwen3.6	27B	64.67	53.92
Qwen3.6	27B	59.50	62.24

Table 3: Comparison of LVLMs' ability to reason about visual illusions in thinking and non-thinking modes.

We found that under different modes of inference, the performances of the two tasks exhibited distinct trends. In the free-form task, LVLMs can provide precise explanations for the formation of illusions through thinking. When we provide the category of illusions generated and formalize the question of their causes as multi-choice tasks, LVLMs actually perform better without engaging in thinking. This indicates that while LVLMs incorporate reasoning capabilities, they also carry the risk of overthinking, which may compromise performance on simple reasoning tasks. Due to space constraints, we provided a detailed example in Appendix A.4 to illustrate this issue.

8 Conclusion

In this work, we analyze the reasoning capabilities of existing LVLMs using visual illusions. We collected a set of real-world illusion images and constructed question-answer pairs designed to evaluate the perceptual and reasoning capabilities of LVLMs, forming a new benchmark: IllusionReasoning. Evaluation based on IllusionReasoning for open-source and closed-source LVLMs indicates that the reasoning capabilities of these models are not as competitive as claimed. And not all questions benefit from reasoning; the integration of LVLMs' perceptual and reasoning capabilities requires further exploration. Further analysis of Il-

lusionReasoning reveals the problem in optimization, emphasising the alignment of content while neglecting the alignment of focus. We hope that IllusionReasoning can be widely used by the community as an insightful benchmark.

Limitations

Illusions, as phenomena where human intuition diverges from reality, create an inherent conflict between perception and reality. We constructed IllusionReasoning, a benchmark comprising illusion images to evaluate the perceptual and reasoning capabilities of LVLMs. However, constrained by the stringent conditions under which illusions occur in the physical world, the volume of data in IllusionReasoning is insufficient to support training. We will continue to collect similar images and consider how to incorporate illusion images into the training process.

Acknowledgments

This work is supported in part by the National Natural Science Foundation of China (NO.62676019, No.62106259), the Youth Program of the National Natural Science Foundation of China (Grant No.62406309), and the Hebei Provincial Department of Science And Technology Foundation under Grant (No.26240601D).

References

Michael Bach and Charlotte M Poloschek. 2006. Optical illusions. *Adv Clin Neurosci Rehabil*, 6(2):20–21.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, and 45 others. 2025. Qwen3-vl technical report. *Preprint*, arXiv:2511.21631.

Zechen Bai, Pichao Wang, Tianjun Xiao, Tong He, Zongbo Han, Zheng Zhang, and Mike Zheng Shou. 2024. Hallucination of multimodal large language models: A survey. *arXiv preprint arXiv:2404.18930*.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. 2024a. Are we on the right way for evaluating large vision-language models? *Preprint*, arXiv:2403.20330.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, and 1 others. 2024b. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24185–24198.

Xianfu Cheng, Wei Zhang, Shiwei Zhang, Jian Yang, Xiangyuan Guan, Xianjie Wu, Xiang Li, Ge Zhang, Jiaheng Liu, Yuying Mai, Yutao Zeng, Zhoufutu Wen, Ke Jin, Baorui Wang, Weixiao Zhou, Yunhong Lu, Tongliang Li, Wenhao Huang, and Zhoujun Li. 2025. Simplevqa: Multimodal factuality evaluation for multimodal large language models. *Preprint*, arXiv:2502.13059.

OpenCompass Contributors. 2023. Opencompass: A universal evaluation platform for foundation models. https://github.com/open-compass/opencompass.

DeepSeek-AI. 2026. Deepseek-v4 technical report. https://github.com/deepseek-ai/DeepSeek-V4. Accessed: 2026-05-22.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 181 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Yuhao Dong, Zuyan Liu, Hai-Long Sun, Jingkang Yang, Winston Hu, Yongming Rao, and Ziwei Liu. 2025. Insight-v: Exploring long-chain visual reasoning with multimodal large language models. *Preprint*, arXiv:2411.14432.

Alexey Dosovitskiy. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Otto B García-Garibay and Victor de Lafuente. 2015. The müller-lyer illusion as seen by an artificial neural network. *Frontiers in computational neuroscience*, 9:21.

Richard L Gregory. 1997. Knowledge in perception and illusion. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 352(1358):1121–1127.

Richard Langton Gregory. 1968. Perceptual illusions and brain models. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, 171(1024):279–296.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. 2025. A survey on llm-as-a-judge. *Preprint*, arXiv:2411.15594.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, and 1 others. 2024.

Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14375–14385.

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Leng Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, Jie Liu, Lei Qi, Zhiyuan Liu, and Maosong Sun. 2024. Olympiadbench: A challenging benchmark for promoting agi with olympiad-level bilingual multimodal scientific problems. *Preprint*, arXiv:2402.14008.

Weiquan Huang, Aoqi Wu, Yifan Yang, Xufang Luo, Yuqing Yang, Liang Hu, Qi Dai, Chunyu Wang, Xiyang Dai, Dongdong Chen, and 1 others. 2024. Llm2clip: Powerful language model unlocks richer visual representation. *arXiv preprint arXiv:2411.04997*.

Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. 2024. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *arXiv preprint arXiv:2407.07895*.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916.

Jiazhen Liu, Yuhan Fu, Ruobing Xie, Runquan Xie, Xingwu Sun, Fengzong Lian, Zhanhui Kang, and Xirong Li. 2025. Phd: A chatgpt-prompted visual hallucination evaluation dataset. *Preprint*, arXiv:2403.11116.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. 2024. Mmbench: Is your multi-modal model an all-around player? *Preprint*, arXiv:2307.06281.

Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, and 1 others. 2024a. Deepseek-vl: towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525*.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2024b. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *Preprint*, arXiv:2310.02255.

Meta-AI. 2025. The llama 4 herd: The beginning of a new era of natively multimodal ai innovation.

OpenAI, :, Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, Alex Iftimie, Alex Karpenko, Alex Tachard Passos, Alexander Neitz, Alexander Prokofiev, Alexander Wei, Allison Tam, and 244 others. 2024a. Openai o1 system card. *Preprint*, arXiv:2412.16720.

OpenAI. 2024. Hello gpt-4o.

OpenAI. 2026. Gpt-5.5 system card. `https://openai.com/index/gpt-5-5-system-card/`. Accessed: 2026-05-22.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024b. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Arkady Plotnitsky. 1997. Penrose's triangles: The large, the small, and the human mind. *Postmodern Culture*, 7(3).

Qwen Team. 2026. Qwen3.5: Towards native multimodal agents.

Haz Sameen Shahgir, Khondker Salman Sayeed, Abhik Bhattacharjee, Wasi Uddin Ahmad, Yue Dong, and Rifat Shahriyar. 2024. Illusionvqa: A challenging optical illusion dataset for vision language models. *arXiv preprint arXiv:2403.15952*.

Taiga Shinozaki, Tomoki Doi, Amane Watahiki, Satoshi Nishida, and Hitomi Yanaka. 2025. Do large vision-language models distinguish between the actual and apparent features of illusions? *Preprint*, arXiv:2506.05765.

Eric D Sun and Ron Dekel. 2021. Imagenet-trained deep neural networks exhibit illusion-like response to the scintillating grid. *Journal of Vision*, 21(11):15–15.

Tomer Ullman. 2024. The illusion-illusion: Vision language models see illusions where there are none. *Preprint*, arXiv:2412.18613.

Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Jiaqi Wang, Haiyang Xu, Ming Yan, Ji Zhang, and 1 others. 2023. Amber: An llm-free multi-dimensional benchmark for mllms hallucination evaluation. *arXiv preprint arXiv:2311.07397*.

Guowei Xu, Peng Jin, Ziang Wu, Hao Li, Yibing Song, Lichao Sun, and Li Yuan. 2025. Llava-cot: Let vision language models reason step-by-step. *Preprint*, arXiv:2411.10440.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao

Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Yichi Zhang, Jiayi Pan, Yuchen Zhou, Rui Pan, and Joyce Chai. 2023. Grounding visual illusions in language: Do vision-language models perceive illusions like humans? *arXiv preprint arXiv:2311.00047*.

Yiming Zhang, Zicheng Zhang, Xinyi Wei, Xiaohong Liu, Guangtao Zhai, and Xiongkuo Min. 2025. Illusionbench+: A large-scale and comprehensive benchmark for visual illusion understanding in vision-language models. *Preprint*, arXiv:2501.00848.

Yaqi Zhao, Yuanyang Yin, Lin Li, Mingan Lin, Victor Shea-Jay Huang, Siwei Chen, Weipeng Chen, Baoqun Yin, Zenan Zhou, and Wentao Zhang. 2025. Beyond sight: Towards cognitive alignment in lvlm via enriched visual knowledge. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24950–24959.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.

A Appendix

A.1 Inconsistencies in Evaluation

Here, we demonstrate the cases where LLMs, acting as judges, produce results that differ from those of human annotators. As shown in Figure 6, we take the output from InternVL3.5-14B as an example. In addition to the generated answer, the output includes the corresponding explanation, which contains incorrect information. When GPT-4o is used as a judge, its evaluation is flexible, overlooking inconsistencies in the explanations. However, human annotators are strict and consider that InternVL has not produced the correct answer. Given that there are few disagreements between the evaluation produced by GPT-4o acting as a judge and the human annotations. We consider that GPT-4o provides a relatively reliable evaluation for IllusionReasoning.

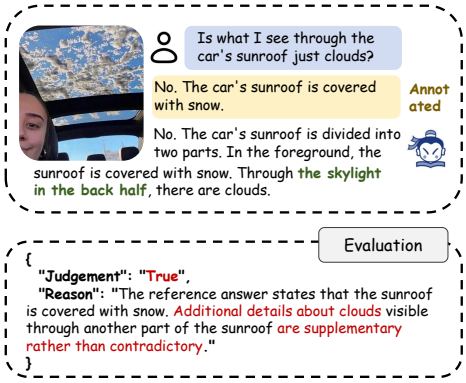

Figure 6: A case where an LLM as a judge differs from human annotations.

A.2 Alignment Preferences in LVLMs

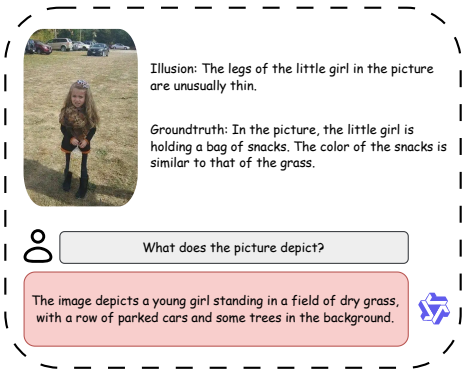

Figure 7: An example of a neutral output. Taking Qwen as an example, the output contains two possibilities, making it difficult to determine whether it aligns more closely with human perception or the physical world.

A.3 Overthinking in Detection Question

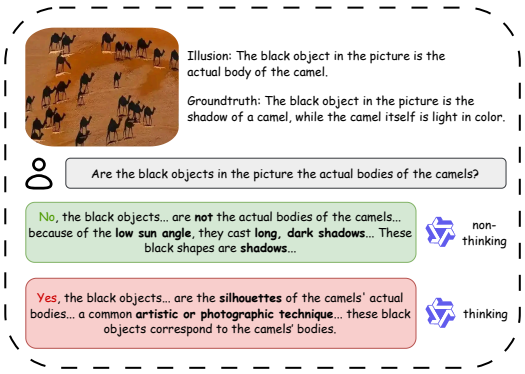

Figure 8: An example of reasoning that has negative effects. Taking Qwen3.6 as an example, in the thinking mode, Qwen3.6 produced a seemingly sophisticated but actually incorrect explanation, which led to incorrect results.

A.4 Overthinking in Multi-Choice

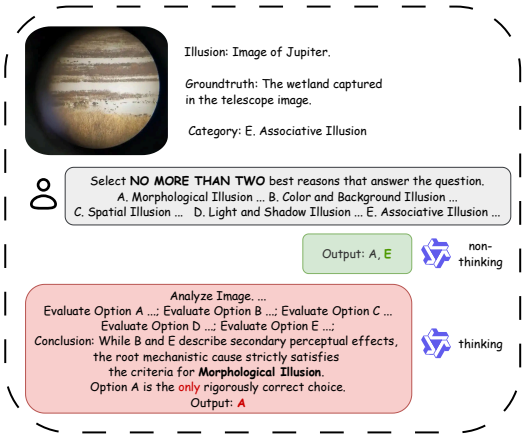

Figure 9: An example of reasoning that has negative effects. Taking Qwen3.5 as an example, in thinking mode, Qwen3.5 produced a confident result that actually excluded the correct answer.

A.5 Prompt used in IllusionReasoning

Detection Questions
For binary questions (expecting a judgment word like "yes" or "no"):
1. If the model's answer lacks a clear judgment word, respond with False.
2. If the model's answer includes a judgment word:
i. Respond with True if it matches the reference answer.
ii. Respond with False if it does not match.

Figure 10: Evaluation Prompt for Detection Questions.

Description Questions
For special questions (e.g., "how many", "what", "which", etc.):
1. Respond with False if the model's answer lacks a clear conclusion.
2. If the model's answer includes a clear conclusion, compare its core information with the reference answer:
i. Respond with True if the semantic meaning matches.
ii. Respond with False if it does not match.

Figure 11: Evaluation Prompt for Description Questions.

Reasoning Questions
For reasoning questions (e.g., "why" and "how"):
1. If the model's answer includes multiple explanations or conclusions, it is correct as long as at least one matches the reference answer.
2. Respond with True if the semantic meaning of the model's final result matches the reference answer; otherwise, respond with False.

Figure 12: Evaluation Prompt for Reasoning Questions.

General Rules for All Questions
1. The model's answer must not contradict the reference answer.
2. Vague answers are acceptable if they include key information from the reference answer and do not introduce errors or contradictions.
3. Focus on whether the semantic meaning of the model's answer matches the reference answer.
4. Ignore differences in language (e.g., Chinese vs. English), case, punctuation, grammar, or word order.
5. Disregard intermediate reasoning or steps in the model's answer and evaluate only the final result or conclusion.

Figure 13: General Rules for All Questions.