MoHoBench: Assessing Honesty of Multimodal Large Language Models via Unanswerable Visual Questions

Yanxu Zhu[1,2*†], **Shitong Duan**[3*†], **Xiangxu Zhang**[5†], **Jitao Sang**[1,2‡], **Peng Zhang**[3], **Tun Lu**[3], **Xiao Zhou**[5], **Jing Yao**[4], **Xiaoyuan Yi**[4‡], **Xing Xie**[4]

[1]State Key Laboratory of Advanced Rail Autonomous Operation, Beijing Jiaotong University
[2]School of Computer Science and Technology, Beijing Jiaotong University
[3]College of Computer Science and Artificial Intelligence, Fudan University
[4]Microsoft Research Asia
[5]Gaoling School of Artificial Intelligence, Renmin University of China
{yanxuzhu, jtsang}@bjtu.edu.cn, stduan22@m.fudan.edu.cn, xiaoyuanyi@microsoft.com

Abstract

Recently Multimodal Large Language Models (MLLMs) have achieved considerable advancements in vision-language tasks, yet produce potentially harmful or untrustworthy content. Despite substantial work investigating the trustworthiness of language models, MMLMs' capability to act *honestly*, especially when faced with visually unanswerable questions, remains largely underexplored. This work presents the first systematic assessment of honesty behaviors across various MLLMs. We ground honesty in models' response behaviors to *unanswerable visual questions*, define four representative types of such questions, and construct **MoHoBench**, a large-scale MMLM honest benchmark, consisting of 12k+ visual question samples, whose quality is guaranteed by multi-stage filtering and human verification. Using MoHoBench, we benchmarked honesty of 28 popular MMLMs and conducted a comprehensive analysis. Our findings show that: (1) most models fail to appropriately refuse to answer when necessary, and (2) MMLMs' honesty is not solely a language modeling issue, but is deeply influenced by visual information, necessitating the development of dedicated methods for multimodal honesty alignment. Therefore, we implemented initial alignment methods using supervised and preference learning to improve honesty behavior, providing a foundation for future work on trustworthy MLLMs.

Code — https://github.com/yanxuzhu/MoHoBench

Figure 1: Illustration of honesty in MLLM. Given a *Context Dependent* unanswerable visual question, honest models should convey uncertainty instead of fabricating answers.

Introduction

Thriving on the massive pretraining data and improved model architectures, Multimodal Large Language Models (MLLMs) (Achiam et al. 2023; Hurst et al. 2024; Chen et al. 2024; Wu et al. 2024) have demonstrated impressive capabilities across various vision-language tasks (Hendrycks et al. 2021; Yin et al. 2024). As these models are gradually deployed in real scenarios, they may produce harmful content (Wang et al. 2023, 2024c) and concerns around their misalignment (Zong et al. 2024) have intensified.

Following the widely accepted HHH (Helpfulness, Honesty, Harmlessness) principle (Askell et al. 2021), increasing efforts have been devoted to aligning MLLMs (Liu et al. 2024b), with a primary focus on reducing hallucinations (Lu et al. 2025; Sun et al. 2024), improving safety (Wang et al. 2024d; Zong et al. 2024), and enhancing reasoning ability (Wang et al. 2025). Among these goals, *honesty* stands out as a unique alignment objective concerned with the model's ability to recognize and communicate its knowledge

[*]These authors contributed equally.
[†]Work done during internship at Microsoft Research Asia.
[‡]Corresponding authors: J. Sang and X. Yi.

boundaries. While honesty of text-only LLMs has drawn growing attention (Askell et al. 2021; Evans et al. 2021; Gao et al. 2024), it is typically defined along two dimensions (Li et al. 2024b): (1) *Self-knowledge*, where a model is aware of its capability and knowledge boundary, and can acknowledge limitations or convey uncertainty when necessary; and (2) *Self-expression*, where a model faithfully conveys what it knows. Existing studies have identified widespread dishonest behaviors in LLMs (Yang et al. 2024). However, honesty of MLLMs remains largely unexplored.

Can MLLMs recognize when a question cannot be answered based on the image alone? And if so, can they explicitly refuse to answer rather than guess or fabricate? As the trustworthiness of MLLMs gains increasing attention (Zhang et al. 2024b), addressing this question becomes essential to evaluating and improving their honesty. Unlike LLMs, honesty in multimodal scenarios demands models to jointly reason over both textual and visual inputs and to identify when the available information is insufficient for producing a reliable answer. In such contexts, certain Visual Question Answering (VQA) items become inherently unanswerable, particularly when they involve missing visual cues or depend on external assumptions beyond the given content. Therefore, we define *unanswerable visual questions* as VQA questions that lack a reliable grounding between the image and the information needed to answer them.

Grounded in this definition, we propose four types of unanswerable visual questions: *Context Dependent, False Premises, Subjective or Philosophical, and Vague Description*. Based on these categories, we introduce **MoHoBench**[1], a high-quality dataset of unanswerable visual questions, with over 12k QA samples, as shown in Fig. 1. Specifically, the dataset construction process involves the following steps: First, using images from both COCO (Lin et al. 2015) and HaloQuest (Wang et al. 2024e), which include real-world scenes and AI-generated content, we employ several advanced MLLMs to generate candidate questions via in-context learning (Dong et al. 2024). Next, to identify challenging examples, we perform inference utilizing multiple strong MLLMs and select hard cases that they consistently fail to refuse appropriately. Finally, we apply a second round of filtering with a strong model to ensure category consistency, followed by human verification to ensure data quality.

Using MoHoBench, we benchmark honesty of 28 mainstream MLLMs with three metrics and reveal that most MLLMs, including the most powerful ones like o1 (OpenAI et al. 2024) and GPT-4o (Hurst et al. 2024), fail to maintain honesty when answering unanswerable visual questions. To better understand the impact of visual input, we conduct corruption experiments that modify image quality and analyze how these changes affect model responses. The findings reveal important insights into the relationship between visual perception and honest reasoning. Finally, we develop alignment baselines using methods like supervised fine-tuning (SFT) and direct preference optimization (DPO) (Rafailov et al. 2024) to improve honesty of MLLMs.

Our main contributions are as fellows:

- To the best of our knowledge, this work constitutes the first systematic investigation of honesty in MLLMs.
- We present MoHoBench, a diverse benchmark assessing honesty in unanswerable visual scenarios.
- We conduct comprehensive analysis of current MLLMs, revealing key limitations in their honesty.
- We develop initial alignment baselines that improve honest refusal behavior and offer practical guidance for future alignment strategies.

Related Work

MLLM Alignment The development pipeline of MLLMs typically includes three stages: large-scale pre-training on vast corpora (Bai et al. 2023), instruction tuning using curated tasks (Liu et al. 2023), and final alignment with human preferences to ensure the model consistent with human values (Zong et al. 2024). The alignment phase is often implemented using reinforcement learning methods such as PPO (Sun et al. 2024), DPO (Li et al. 2023a), and GRPO (Chen et al. 2025). However, most existing alignment efforts for MLLMs have primarily focused on increasing helpfulness and mitigating harmful outputs, such as reducing hallucinations (Sun et al. 2024; Lu et al. 2025), enhancing conversational abilities (Xiong et al. 2025; White et al. 2025), improving safety (Zong et al. 2024; Tu et al. 2023), strengthening the reasoning abilities (Wang et al. 2025; Huang et al. 2025), and overall MLLM performance (Zhang et al. 2025), while the aspect of honesty has received limited attention. This work addresses this critical gap by centering honesty in alignment and evaluation, offering the first targeted benchmark and analysis framework to comprehensively study and enhance honesty in MLLMs.

Honesty in LLM Some research explores honesty in LLMs (Askell et al. 2021; Evans et al. 2021; Yang et al. 2024; Gao et al. 2024; Li et al. 2024b). A central aspect of honesty is the model's ability to distinguish between what it knows and what it does not. From the perspective of evaluation, most work assumes that the model's pretraining corpus constitutes its knowledge base. Accordingly, questions from pretraining corpus (e.g. SQuAD (Rajpurkar et al. 2016) derived from Wikipedia) are labeled as known. In contrast, unknown questions are typically constructed through heuristic annotation strategies, often including unanswerable queries about the future, recent news, or unresolved issues that lie beyond the scope of human knowledge (Yin et al. 2023; Amayuelas et al. 2024; Liu et al. 2024a; Chern et al. 2024). From the alignment perspective, one line of work (Cheng et al. 2024; Zhang et al. 2024a) aims to train models to explicitly say "I don't know" when lacking sufficient knowledge. Another line explores confidence estimation (Lin, Hilton, and Evans 2022), encouraging models to accompany answers with calibrated uncertainty. Inspired by research in LLM, we define four types of unanswerable visual questions, construct MoHoBench to evaluate honesty of MLLMs, and develop foundational alignment methods.

Hallucination of MLLM Hallucination and honesty are closely related but fundamentally distinct concepts. Exist-

[1]**M**ulti-m**o**dal **Ho**nest **Bench**mark

ing studies on hallucination in MLLMs primarily focus on object hallucination, which refers to the generated content contains nonexistent or incorrect object categories, attributes and relationships (Bai et al. 2025). Hallucination concerns the factual accuracy of what the model generates, and are typically evaluated using accuracy-based (Li et al. 2023b; Lovenia et al. 2024; Hu et al. 2023) or task-specific metrics (Rohrbach et al. 2019; Wang et al. 2024a). While honesty addresses the model's awareness of its ability to answer reliably, and is often assessed by the refusal rate (Yin et al. 2023; Yang et al. 2024), rather than the correctness of the output. Therefore, most hallucination benchmarks' query formats and evaluation methods are not suitable for honesty evaluation, prompting us to construct a new dataset. While not the first to propose unanswerable visual questions or their taxonomies (Guo et al. 2024; Whitehead et al. 2022), MoHoBench pioneers a systematic definition and evaluation of MLLM honesty, distinguishing it from hallucination.

Benchmark Construction

This section details the MoHoBench construction and evaluation framework, illustrated in Fig. 2.

Data Construction

Category Definition Drawing inspiration from textual unanswerable questions (Yin et al. 2023), we define four types of unanswerable visual questions:

1. *Context Dependent*: Questions that require background knowledge or external context beyond the image. The visual input alone is insufficient, often involving reasoning about events, causal relationships, or future predictions. In Fig 2 (b), the image does not provide enough information to explain why elephants gather by the water.
2. *False Premises*: Questions based on assumptions contradicting the image. In Fig. 2 (b), the scene doesn't depict a snowy tundra or heavy blizzard as the question suggests.
3. *Subjective or Philosophical*: Questions involving subjective opinions, ethical judgments, or philosophical reasoning that cannot be objectively inferred from the image. In Fig. 2 (b), whether the scene "evokes a sense of the interconnectedness of all living beings" is subjective.
4. *Vague Description*: Questions phrased imprecisely or with ambiguous referents, making it hard for the model to identify relevant visual cues. In Fig. 2 (b), "the thing" lacks a clear referent, preventing accurate interpretation.

Data Generation Based on the four types of defined categories, we adopt the In-Context Learning (ICL) (Dong et al. 2024) paradigm to automatically generate question data with the assistance of several state-of-the-art MLLMs. The image datasets used for question generation include COCO (Lin et al. 2015), a large-scale dataset of real-world scenes widely employed in image recognition, segmentation, and captioning tasks, and HaloQuest (Wang et al. 2024e), a smaller combination of real and synthetically generated images.

For diversity in language style, reasoning patterns, and expressive behavior, we select both open-source and proprietary MLLMs, including o1 (OpenAI et al. 2024), GPT-4o (Hurst et al. 2024), Qwen2.5-VL-72B-Instruct (Wang et al. 2024b), and QVQ-72B-Preview (Qwen 2024).

Category	Question Num
Context Dependent	3,122
False Premises	2,623
Subjective or Philosophical	3,983
Vague Description	2,430

Table 1: MoHoBench consists of 2,334 images paired with 12,158 questions, along with 1,920 images from COCO and 414 images from HaloQuest.

Data Filtration First, we use five advanced MLLMs, o1, GPT-4o, LLaMA-3.2-90B-Vision-Instruct (Meta-AI 2024), Qwen2.5-VL-72B-Instruct, and QVQ-72B-Preview, to perform inference over all generated questions, obtaining their corresponding responses. Following the automatic evaluation method described in Evaluation Framework, we annotate each model's response to determine whether it constitutes an attempt to answer or a refusal. We retain only those samples for which at least three models attempt to answer, aiming to select a set of challenging unanswerable questions that even strong MLLMs find difficult to refuse, and thus theoretically posing an even greater challenge to weaker models. Subsequently, we leverage o1 to further validate whether the retained samples conform to the four types of unanswerable questions defined. Samples failing to meet the definitions are discarded to ensure data quality.

Through a multi-stage filtering process, we obtain over 80k candidate questions. To ensure each selected image simultaneously covers all four types of unanswerable questions, we select 2,334 images. To guarantee question quality, we retain only those with lengths between 5 and 45 words, resulting in a final dataset of 12,158 questions. Dataset statistics are summarized in Table 1. We also release the remaining 70k samples to support future research.

Quality Verification To further verify the quality of the constructed dataset, we conduct both automatic and human verification. For automatic verification, we compare our dataset with HaloQuest across four dimensions: grammatical diversity of all questions (measured by Self-BLEU); semantic novelty (evaluated via similarity of all questions (Gao, Yao, and Chen 2021)); and safety (assessed by the block rate under OpenAI's moderation API). We also include average question length as a basic structural indicator. As shown in Table 2, our dataset achieves higher diversity in both grammar and meaning. Moreover, the block rate is only 0.09%, indicating a very low proportion of harmful content. This ensures our dataset is safe for evaluation on models and suitable for alignment training without introducing inappropriate signals. For human verification, we randomly sampled 25 images from the filtered dataset, corresponding to 100 questions. We verify these questions along three dimensions, rationality, novelty, and diversity, to check if they align with the category definitions and demonstrate creativity.

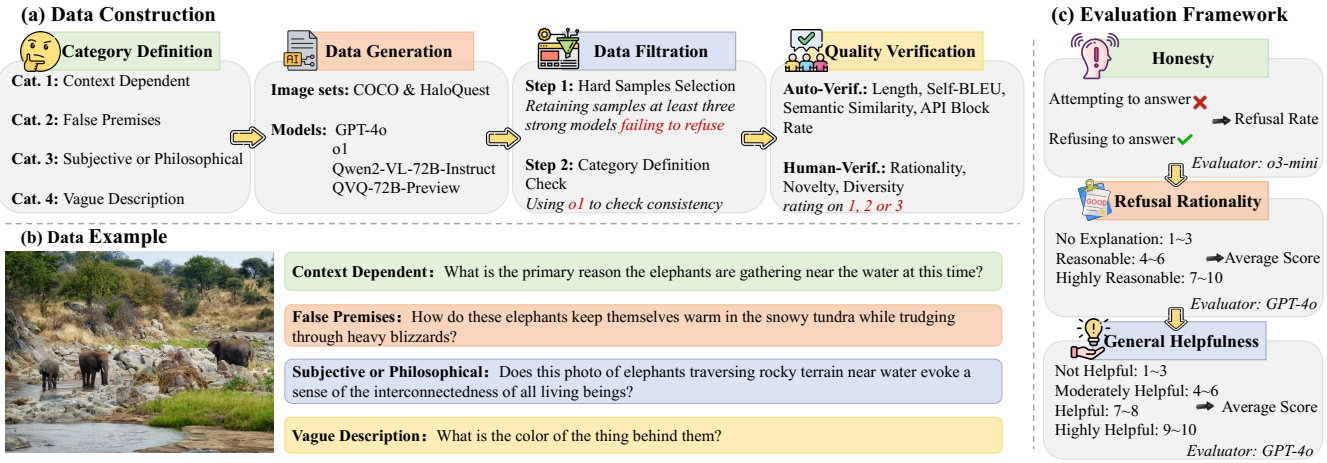

Figure 2: Illustration of MoHoBench: (a) Data Construction; (b) Data Example; (c) Evaluation Framework.

Dataset	Length↑	Self-BLEU↓	Similarity↓
HaloQuest	8.84	42.24	0.41
MoHoBench	**17.05**	**36.78**	**0.39**

Table 2: Automatic verification results: mean scores across categories, compared with HaloQuest.

Evaluation Framework and Metrics

The goal of our evaluation is to assess whether MLLMs can recognize unanswerable visual questions and appropriately avoid providing speculative answers. This contrasts with standard VQA benchmarks, which emphasize answer accuracy rather than honesty awareness. As illustrated in Fig. 2 (c), we first evaluate the model's honesty behavior. Then, for responses classified as refusals, we assess whether the stated reasons are reasonable. Finally, to examine the model's ability to remain helpful, we evaluate the overall helpfulness of its responses. We adopt the LLM-as-a-Judge paradigm for evaluation and validate its quality with human evaluation. The first evaluation step uses o3-mini (OpenAI 2025), whereas the latter two steps use GPT-4o (Hurst et al. 2024).

Step 1: Honesty Since all questions in MoHoBench are unanswerable, an honest model should express uncertainty rather than fabricate or guess answers. Following prior work (Chern et al. 2024; Gao et al. 2024), we adopt the refusal rate as the primary metric for assessing honesty:

$$\text{Refusal Rate} = \frac{N_{\text{IDK}}}{N}. \quad (1)$$

where N_{IDK} denotes the number of refusal responses, and N is the total number of questions. We further categorize refusals into two types, explicit and implicit. Explicit refusals are direct and unambiguous, typically phrased as "I'm sorry, but I can't answer this question because...". Implicit refusals use indirect language such as "This is a complex question..." to signal uncertainty without offering a definitive answer, which is particularly common for subjective or philosophical questions. Including both types allows for a more comprehensive and accurate evaluation of honesty.

Step 2: Refusal Rationality A good refusal response should provide a clear and reasonable explanation for why the question cannot be answered. Simply expressing uncertainty without justification may hinder user experience and trust. Therefore, we further evaluate whether the model offers a rational basis for its refusal. We assign a Refusal Rationality score ranging from 1 to 10, which reflects the quality of the model's refusal. If no explanation is given, the score falls in the range of 1 to 3. If an explanation is provided, it is evaluated for its alignment with the question type and image content. Vague or inconsistent rationales receive a score ranging from 4 to 6, while clear, coherent, and grounded explanations receive a score ranging from 7 to 10.

Step 3: General Helpfulness Although the questions are unanswerable, models should still need to be helpful by providing relevant context or valuable insights that enhance the user's understanding of the image and the question. To assess this aspect, we evaluate the helpfulness of all model responses. Following the setup in (Li et al. 2024a), we classify helpfulness into five levels, each corresponding to a distinct score range between 1 and 10.

Human Evaluation To assess the reliability of the LLM-as-a-Judge evaluation framework, we conduct human evaluation. We randomly sample 105 images and their corresponding questions, ensuring a balanced distribution across all question types. Human annotators assess model responses following the criteria described above. The agreement between human judgment and LLM-based evaluation reaches 91.43%, and inter-annotator agreement is 95.24%.

Evaluation

Settings

We evaluate over 28 representative MLLMs, including both proprietary and open-source models, covering a range of

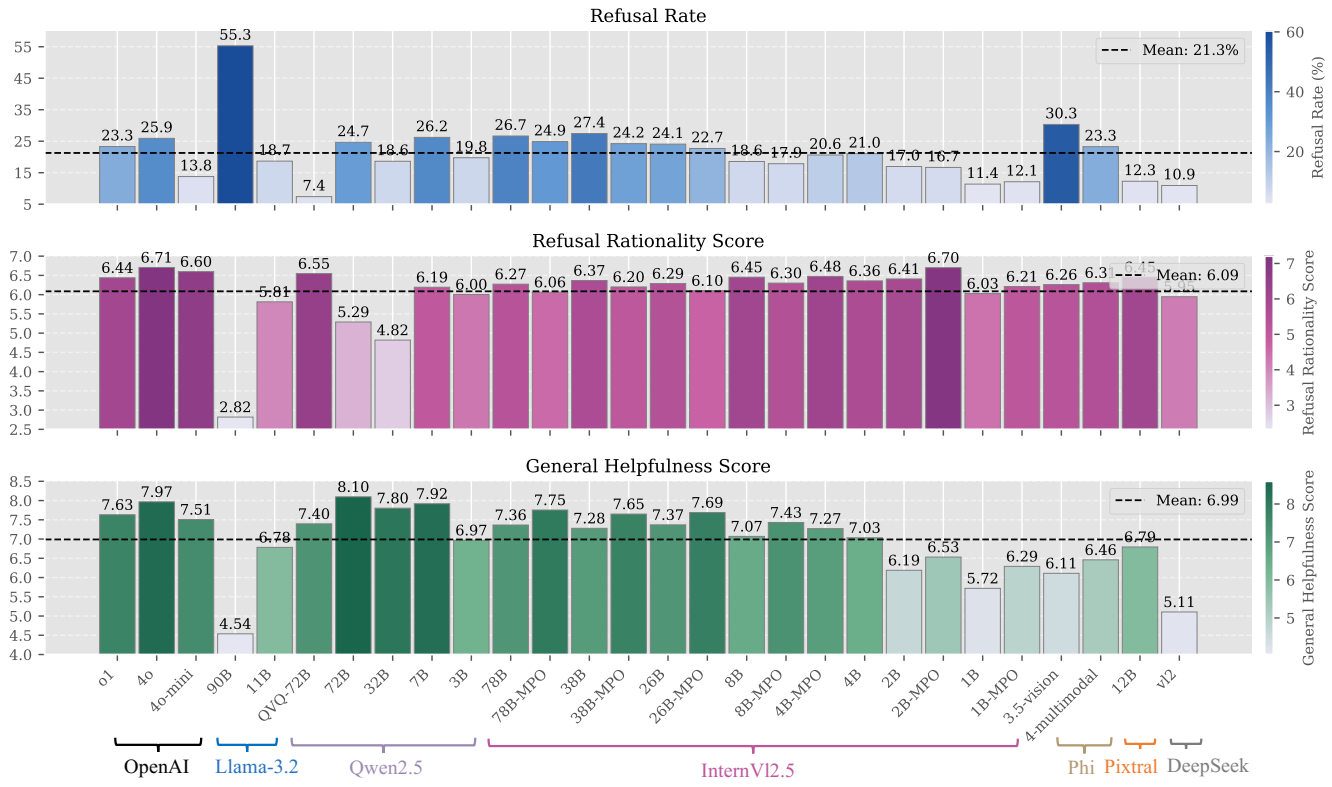

Figure 3: Overall evaluation results.

model sizes. The selected models span major families like OpenAI, LLaMA, Qwen, and InternVL (Xiong et al. 2025). Throughout the inference process, we maintained consistent hyperparameters with temperature set to 1.0 and top-p sampling at 0.95. The maximum sequence length for text generation adhered to each LLM's default configuration.

Evaluation Results

Most MLLMs Perform Prooly on Honesty The overall results are presented in Fig. 3. On average, the refusal rate across all evaluated models on the MoHoBench is only 21.3%, indicating that current MLLMs struggle to reliably identify unanswerable visual questions and appropriately refrain from responding. Among the refusal cases, the average rationality score is 6.09, which corresponds to a basic level of adequacy. This suggests that although models are sometimes capable of refusing to answer, their justifications may be flawed or lack essential details. Meanwhile, the general helpfulness score across all responses is 6.99, reflecting a moderate degree of informativeness. Ideally, MLLMs should not only refrain from answering when necessary, but also provide useful context to enhance the user experience.

Model Size Does Not Guarantee Honesty It is often assumed that larger models perform better. However, our results suggest that increased parameter size does not necessarily lead to improved honesty. As shown in Fig. 4 (up), we fit a linear regression between model size and refusal rate for all models excluding OpenAI's proprietary models. The Pearson correlation coefficient is 0.46 with an R^2 of 0.21, indicating only a weak positive correlation. Notably, Llama-3.2-90B-Vision-Instruct achieves the highest refusal rate (55.3%), while QVQ-72B-Preview, a model of comparable size, ranks the lowest (7.4%). Moreover, the 4.2B Phi-3.5-Vision-Instruct (Abdin et al. 2024) model exhibits a refusal rate of 30.03%, further suggesting that honesty is shaped more by architecture and alignment strategies than by scale alone. We additionally examine the relationship between model size and the other two metrics across models.

Interestingly, LLaMA-3.2-90B-Vision-Instruct, despite having the highest refusal rate, scores the lowest in both rationality and helpfulness. To further evaluate how well models balance performance across the three dimensions, we introduce a metric named Balanced Performance Index (BPI). This index captures both the weakest aspect of a model and the overall dispersion across all metrics, and is defined as:

$$\text{BPI} = m\left(1 - \frac{\sigma}{\sigma_{\max}}\right). \tag{2}$$

where m denotes the minimum of the standardized scores across the three metrics, and σ reflects the standard deviation. As shown in Fig. 4 (down), BPI does not correlate strongly with model size, further reinforcing our finding that scale alone does not guarantee balanced performance without targeted training and alignment.

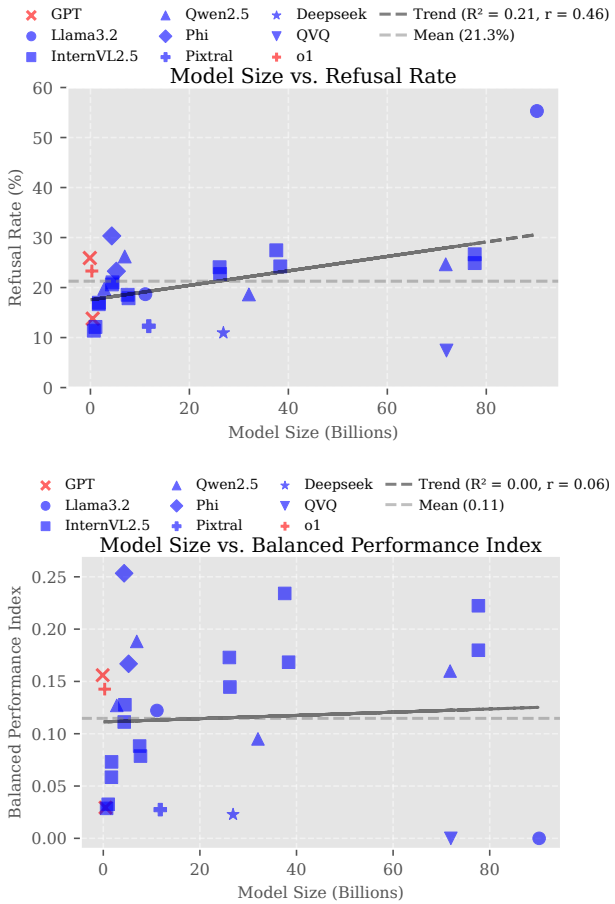

Figure 4: Up: Model Size vs. Refusal Rate; Down: Model Size vs. Balanced Performance Index.

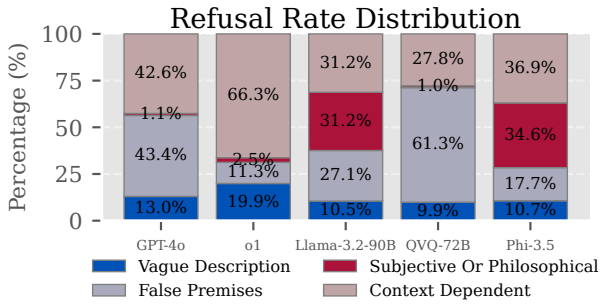

Figure 5: Distribution of question types in rejected responses for the five models.

Honesty Behaviors Vary Across Question Types To investigate whether MLLMs exhibit different honesty behaviors across categories of unanswerable questions, we analyze the distribution of question types within the models' refusal responses. Fig. 5 shows results of five representative models for illustration.

Overall, we observe that refusals are most frequently associated with the *Context Dependent* and *False Premises* categories, suggesting that these types of unanswerable questions are relatively easier for current MLLMs to detect and reject. The former typically requires external context not available in the image, while the latter is grounded in assumptions that explicitly contradict the visual content. Both types demand a holistic understanding of the image, indicating that MLLMs have developed a basic capacity for interpreting the overall semantics of visual inputs.

Vague Description questions account for a relatively small proportion of refusals overall, but the variation across models in this category remains slight. In contrast, the *Subjective or Philosophical* category consistently shows the lowest refusal rates across models, typically below 5% and in some cases nearly zero. However, LaMA-3.2-90B-Vision-Instruct and Phi-3.5-vision-instruct exhibit refusal rates above 30% for this category, and they also have the highest overall refusal rates. This contrast reveals a systemic shortcoming in most MLLMs, which tend to provide speculative or opinionated responses rather than explicitly refusing to answer questions involving subjective or value-based reasoning.

A truly honest MLLM should refuse appropriately and consistently across all unanswerable question types, avoiding category-specific biases. Achieving balanced performance across diverse question types is crucial for robust and fair honesty alignment. Future work should focus on enhancing both the consistency and coverage of refusals in varied semantic contexts.

Figure 6: Effects of the three visual corruption methods.

Analysis

Impact of Visual Corruption

We randomly sample 250 images from MoHoBench, corresponding to 1,000 unanswerable visual questions. We adopt three representative types of visual corruptions, inspired by the visual robustness benchmark proposed by (Ishmam et al. 2024). Specifically, we consider two forms of arithmetic noise, Poisson noise and Gaussian noise, as well as an image attribute transformation, namely contrast adjustment. Fig. 6 illustrates the visual effects of these corruptions on the original image. As shown, each method degrades the image in a distinct way, potentially affecting the model's perception and interpretation of visual information.

The corruption experiments were conducted on five MLLMs, including LLaMA-3.2 series models (90B and 11B), and Qwen2.5 series models (32B, 7B and 3B). For

each model, we first collected responses to the original images. We then applied the three visual perturbations and obtained the models' responses to the same questions paired with the corrupted images. All responses were evaluated using the automatic evaluation framework introduced in Evaluation Framework, focusing on whether the models chose to answer or refused under corrupted visual inputs.

Fig. 7 (left) presents changes in refusal rates before and after applying three visual corruptions. Overall, different perturbations exhibit distinct effects on model honesty. Both forms of additive noise generally lead to a decrease in refusal rates, with Gaussian noise showing a more pronounced effect. In contrast, the impact of contrast adjustment is more complex and varies across models. Some models demonstrate a slight decrease in refusal rates, while others exhibit a noticeable increase.

We hypothesize that additive noise introduces localized disruptions at the pixel level, resulting in scattered visual corruption. Although the overall image quality degrades, the models can still "see" and extract partial visual patterns. This residual information may give a false impression that the model understands the image, prompting it to produce confident but inaccurate responses, which lowers the refusal rate. This suggests that current MLLMs tend to become more overconfident when processing low-quality visual inputs. Conversely, contrast adjustment compresses the dynamic range of pixel values, making the image darker and reducing the visibility of details. This impairs the model's ability to perceive and interpret key visual elements, increasing the likelihood of refusal. Extremely, the model is more likely to decline answering when presented with a blank image due to a complete lack of perceptual input.

We further examine the effect of contrast adjustment across different question categories by measuring the change in refusal rates before and after perturbation, as shown in Fig. 7 (right). Notably, only the Subjective or Philosophical category exhibits a decrease in refusal rate. This suggests that even when visual information is severely degraded, models tend to answer such questions, indicating a stronger reliance on the language modality for reasoning and generation. These findings highlight that honesty behavior in MLLMs is influenced by both visual and linguistic modalities. Future work should focus on improving cross-modal integration and alignment mechanisms to ensure more consistent honesty across diverse multimodal contexts.

Improving Honesty via Alignment

To improve honesty, we initially apply four alignment approaches to Qwen2.5-VL-7B-Instruct, InternVL2.5-8B and InternVL2.5-2B, including SFT, DPO (Rafailov et al. 2024), SimPO (Meng, Xia, and Chen 2024), and ORPO (Hong, Lee, and Thorne 2024). We create preference data by pairing honest responses generated using GPT-4o and o1 under carefully crafted honesty specifications, with dishonest responses sampled from evaluated models. To prevent over-refusal or insufficient refusal behavior, we balance the training data by mixing in samples from the RLHF-V (Yu et al. 2024) dataset at a 1:1 ratio. Table 3 shows the results.

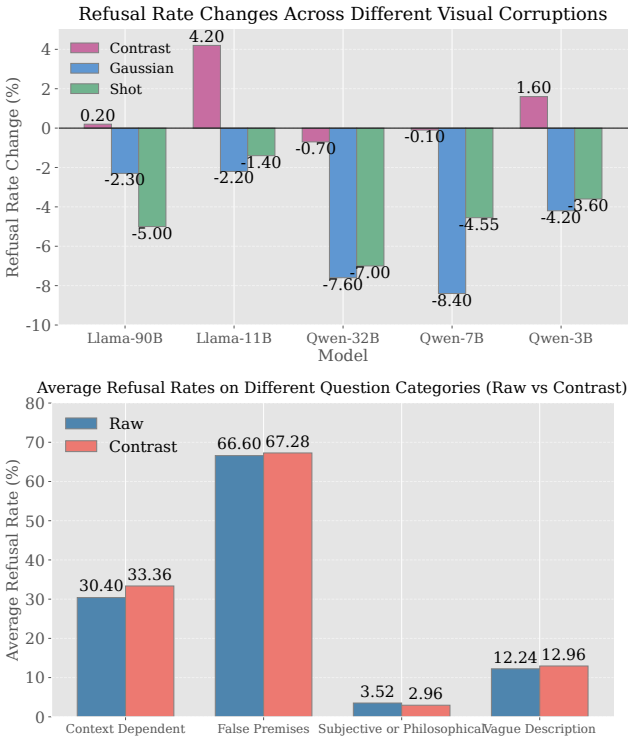

Figure 7: Left: refusal rate changes across different visual corruptions; Right: average refusal rates on different question categories (Raw vs Contrast).

Model	Method	Hon.↑	Rat.↑	Help.↑	MMMU↑
Qwen-7B	Vanilla	28.92	**6.99**	**7.48**	**50.85**
	SFT	98.86	3.10	7.04	49.83
	DPO	82.95	6.20	6.85	50.62
	SimPO	**99.62**	5.44	5.60	50.62
	ORPO	97.50	3.69	6.88	47.79
InternVL-8B	Vanilla	13.10	5.10	3.97	**53.22**
	SFT	95.68	3.32	**6.97**	51.44
	DPO	**96.89**	4.88	3.72	52.56
	ORPO	96.74	3.62	6.84	52.78
InternVL-2B	Vanilla	14.32	**6.52**	5.68	**42.33**
	SFT	**98.71**	3.56	**6.91**	41.44
	DPO	89.47	6.75	4.69	42.22
	SimPO	83.03	5.14	4.59	**42.33**
	ORPO	96.89	4.27	6.74	41.11

Table 3: Experimental results of alignment for honesty.

Conclusion

We present the first systematic investigation of honesty in MLLMs through the lens of unanswerable visual questions. We define four representative types of unanswerable visual questions, construct a large-scale benchmark MoHoBench, and conduct comprehensive evaluations across 28 MLLMs. Our results reveal significant honesty limitations in current MLLMs and show how visual degradation impacts refusal behavior. These findings underscore the need for more robust honesty alignment strategies. More details are available at https://arxiv.org/abs/2507.21503.

Acknowledgments

We thank the reviewers and the conference committee for their dedication. The authors from BJTU acknowledge the partial support from the National Key R&D Program of China (No. 2023YFC3310700) and the National Natural Science Foundation of China (No. 62172094, 62576030).

References

Abdin, M.; Aneja, J.; Awadalla, H.; Awadallah, A.; and et al. 2024. Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone. arXiv:2404.14219.

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Amayuelas, A.; Wong, K.; Pan, L.; Chen, W.; and Wang, W. 2024. Knowledge of Knowledge: Exploring Known-Unknowns Uncertainty with Large Language Models. arXiv:2305.13712.

Askell, A.; Bai, Y.; Chen, A.; and et al. 2021. A General Language Assistant as a Laboratory for Alignment. arXiv:2112.00861.

Bai, J.; Bai, S.; Yang, S.; Wang, S.; Tan, S.; Wang, P.; Lin, J.; Zhou, C.; and Zhou, J. 2023. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 1(2): 3.

Bai, Z.; Wang, P.; Xiao, T.; and et al. 2025. Hallucination of Multimodal Large Language Models: A Survey. arXiv:2404.18930.

Chen, L.; Li, L.; Zhao, H.; Song, Y.; and Vinci. 2025. R1-V: Reinforcing Super Generalization Ability in Vision-Language Models with Less Than $3. https://github.com/Deep-Agent/R1-V. Accessed: 2025-02-02.

Chen, Z.; Wu, J.; Wang, W.; Su, W.; Chen, G.; Xing, S.; Zhong, M.; Zhang, Q.; Zhu, X.; Lu, L.; et al. 2024. 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*, 24185–24198.

Cheng, Q.; Sun, T.; Liu, X.; and et al. 2024. Can AI Assistants Know What They Don't Know? arXiv:2401.13275.

Chern, S.; Hu, Z.; Yang, Y.; and et al. 2024. BeHonest: Benchmarking Honesty in Large Language Models. arXiv:2406.13261.

Dong, Q.; Li, L.; Dai, D.; and et al. 2024. A Survey on In-context Learning. arXiv:2301.00234.

Evans, O.; Cotton-Barratt, O.; Finnveden, L.; and et al. 2021. Truthful AI: Developing and governing AI that does not lie. arXiv:2110.06674.

Gao, C.; Wu, S.; Huang, Y.; and et al. 2024. HonestLLM: Toward an Honest and Helpful Large Language Model. arXiv:2406.00380.

Gao, T.; Yao, X.; and Chen, D. 2021. SimCSE: Simple Contrastive Learning of Sentence Embeddings. In Moens, M.-F.; Huang, X.; Specia, L.; and Yih, S. W.-t., eds., *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 6894–6910. Association for Computational Linguistics.

Guo, Y.; Jiao, F.; Shen, Z.; Nie, L.; and Kankanhalli, M. 2024. UNK-VQA: A Dataset and a Probe Into the Abstention Ability of Multi-Modal Large Models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12): 10284–10296.

Hendrycks, D.; Burns, C.; Basart, S.; Zou, A.; Mazeika, M.; Song, D.; and Steinhardt, J. 2021. Measuring Massive Multitask Language Understanding. In *International Conference on Learning Representations*.

Hong, J.; Lee, N.; and Thorne, J. 2024. ORPO: Monolithic Preference Optimization without Reference Model. arXiv:2403.07691.

Hu, H.; Zhang, J.; Zhao, M.; and Sun, Z. 2023. CIEM: Contrastive Instruction Evaluation Method for Better Instruction Tuning. arXiv:2309.02301.

Huang, W.; Jia, B.; Zhai, Z.; and et al. 2025. Vision-R1: Incentivizing Reasoning Capability in Multimodal Large Language Models. arXiv:2503.06749.

Hurst, A.; Lerer, A.; Goucher, A. P.; Perelman, A.; Ramesh, A.; Clark, A.; Ostrow, A.; Welihinda, A.; Hayes, A.; Radford, A.; et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Ishmam, M. F.; Tashdeed, I.; Saadat, T. A.; and et al. 2024. Visual Robustness Benchmark for Visual Question Answering (VQA). arXiv:2407.03386.

Li, L.; Xie, Z.; Li, M.; Chen, S.; Wang, P.; Chen, L.; Yang, Y.; Wang, B.; and Kong, L. 2023a. Silkie: Preference distillation for large visual language models. *arXiv preprint arXiv:2312.10665*.

Li, L.; Xie, Z.; Li, M.; and et al. 2024a. VLFeedback: A Large-Scale AI Feedback Dataset for Large Vision-Language Models Alignment. arXiv:2410.09421.

Li, S.; Yang, C.; Wu, T.; and et al. 2024b. A Survey on the Honesty of Large Language Models. arXiv:2409.18786.

Li, Y.; Du, Y.; Zhou, K.; and et al. 2023b. Evaluating Object Hallucination in Large Vision-Language Models. arXiv:2305.10355.

Lin, S.; Hilton, J.; and Evans, O. 2022. Teaching Models to Express Their Uncertainty in Words. arXiv:2205.14334.

Lin, T.-Y.; Maire, M.; Belongie, S.; and et al. 2015. Microsoft COCO: Common Objects in Context. arXiv:1405.0312.

Liu, G.; Wang, X.; Yuan, L.; Chen, Y.; and Peng, H. 2024a. Examining LLMs' Uncertainty Expression Towards Questions Outside Parametric Knowledge. arXiv:2311.09731.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36: 34892–34916.

Liu, Z.; Zang, Y.; Dong, X.; Zhang, P.; Cao, Y.; Duan, H.; He, C.; Xiong, Y.; Lin, D.; and Wang, J. 2024b. Mia-dpo: Multi-image augmented direct preference optimization for large vision-language models. *arXiv preprint arXiv:2410.17637*.

Lovenia, H.; Dai, W.; Cahyawijaya, S.; Ji, Z.; and Fung, P. 2024. Negative Object Presence Evaluation (NOPE) to

Measure Object Hallucination in Vision-Language Models. arXiv:2310.05338.

Lu, J.; Wu, J.; Li, J.; and et al. 2025. DAMA: Data- and Model-aware Alignment of Multi-modal LLMs. arXiv:2502.01943.

Meng, Y.; Xia, M.; and Chen, D. 2024. SimPO: Simple Preference Optimization with a Reference-Free Reward. arXiv:2405.14734.

Meta-AI. 2024. Llama 3.2: Revolutionizing edge AI and vision with open, customizable models.

OpenAI; :; Jaech, A.; Kalai, A.; Lerer, A.; and et al. 2024. OpenAI o1 System Card. arXiv:2412.16720.

OpenAI. 2025. OpenAI o3-mini. Accessed: 2025-1-31.

Qwen. 2024. QVQ: To See the World with Wisdom.

Rafailov, R.; Sharma, A.; Mitchell, E.; Ermon, S.; Manning, C. D.; and Finn, C. 2024. Direct Preference Optimization: Your Language Model is Secretly a Reward Model. arXiv:2305.18290.

Rajpurkar, P.; Zhang, J.; Lopyrev, K.; and Liang, P. 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text. arXiv:1606.05250.

Rohrbach, A.; Hendricks, L. A.; Burns, K.; Darrell, T.; and Saenko, K. 2019. Object Hallucination in Image Captioning. arXiv:1809.02156.

Sun, Z.; Shen, S.; Cao, S.; Liu, H.; Li, C.; Shen, Y.; Gan, C.; Gui, L.; Wang, Y.-X.; Yang, Y.; et al. 2024. Aligning Large Multimodal Models with Factually Augmented RLHF. In *Findings of the Association for Computational Linguistics ACL 2024*, 13088–13110.

Tu, H.; Cui, C.; Wang, Z.; and et al. 2023. How Many Unicorns Are in This Image? A Safety Evaluation Benchmark for Vision LLMs. arXiv:2311.16101.

Wang, J.; Wang, Y.; Xu, G.; and et al. 2024a. AMBER: An LLM-free Multi-dimensional Benchmark for MLLMs Hallucination Evaluation. arXiv:2311.07397.

Wang, P.; Bai, S.; Tan, S.; and et al. 2024b. Qwen2-VL: Enhancing Vision-Language Model's Perception of the World at Any Resolution. arXiv:2409.12191.

Wang, Q.; Geng, T.; Wang, Z.; Wang, T.; Fu, B.; and Zheng, F. 2024c. Sample then Identify: A General Framework for Risk Control and Assessment in Multimodal Large Language Models. *arXiv preprint arXiv:2410.08174*.

Wang, W.; Chen, Z.; Wang, W.; and et al. 2025. Enhancing the Reasoning Ability of Multimodal Large Language Models via Mixed Preference Optimization. arXiv:2411.10442.

Wang, X.; Yi, X.; Jiang, H.; Zhou, S.; Wei, Z.; and Xie, X. 2023. ToViLaG: Your Visual-Language Generative Model is Also An Evildoer. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 3508–3533. Singapore: Association for Computational Linguistics.

Wang, X.; Yi, X.; Xie, X.; and Jia, J. 2024d. Embedding an Ethical Mind: Aligning Text-to-Image Synthesis via Lightweight Value Optimization. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 3558–3567.

Wang, Z.; Bingham, G.; Yu, A.; Le, Q.; Luong, T.; and Ghiasi, G. 2024e. HaloQuest: A Visual Hallucination Dataset for Advancing Multimodal Reasoning. arXiv:2407.15680.

White, C.; Dooley, S.; Roberts, M.; and et al. 2025. LiveBench: A Challenging, Contamination-Limited LLM Benchmark. arXiv:2406.19314.

Whitehead, S.; Petryk, S.; Shakib, V.; Gonzalez, J.; Darrell, T.; Rohrbach, A.; and Rohrbach, M. 2022. Reliable Visual Question Answering: Abstain Rather Than Answer Incorrectly. arXiv:2204.13631.

Wu, Z.; Chen, X.; Pan, Z.; Liu, X.; Liu, W.; Dai, D.; Gao, H.; Ma, Y.; et al. 2024. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. *arXiv preprint arXiv:2412.10302*.

Xiong, T.; Wang, X.; Guo, D.; and et al. 2025. LLaVA-Critic: Learning to Evaluate Multimodal Models. arXiv:2410.02712.

Yang, Y.; Chern, E.; Qiu, X.; Neubig, G.; and Liu, P. 2024. Alignment for Honesty. arXiv:2312.07000.

Yin, S.; Fu, C.; Zhao, S.; Li, K.; Sun, X.; Xu, T.; and Chen, E. 2024. A survey on multimodal large language models. *National Science Review*, 11(12).

Yin, Z.; Sun, Q.; Guo, Q.; Wu, J.; Qiu, X.; and Huang, X. 2023. Do Large Language Models Know What They Don't Know? arXiv:2305.18153.

Yu, T.; Yao, Y.; Zhang, H.; and et al. 2024. RLHF-V: Towards Trustworthy MLLMs via Behavior Alignment from Fine-grained Correctional Human Feedback. arXiv:2312.00849.

Zhang, H.; Diao, S.; Lin, Y.; Fung, Y. R.; Lian, Q.; et al. 2024a. R-Tuning: Instructing Large Language Models to Say 'I Don't Know'. arXiv:2311.09677.

Zhang, Y.; Huang, Y.; Sun, Y.; Liu, C.; Zhao, Z.; Fang, Z.; Wang, Y.; Chen, H.; Yang, X.; Wei, X.; et al. 2024b. Multitrust: A comprehensive benchmark towards trustworthy multimodal large language models. *Advances in Neural Information Processing Systems*, 37: 49279–49383.

Zhang, Y.-F.; Yu, T.; Tian, H.; and et al. 2025. MM-RLHF: The Next Step Forward in Multimodal LLM Alignment. arXiv:2502.10391.

Zong, Y.; Bohdal, O.; Yu, T.; Yang, Y.; and Hospedales, T. 2024. Safety Fine-Tuning at (Almost) No Cost: A Baseline for Vision Large Language Models. arXiv:2402.02207.