# Do Multimodal LLMs Understand Order? Measuring the Fragility of Multimodal Reasoning under Input Order Perturbations

**Sheng-Lun Wei**[α], **Yu-Ling Liao**[α], **Hen-Hsen Huang**[β], **Hsin-Hsi Chen**[αγ]
[α]Department of Computer Science and Information Engineering,
National Taiwan University, Taiwan
[β]Institute of Information Science, Academia Sinica, Taiwan
[γ]AI Research Center (AINTU), National Taiwan University, Taiwan
{weisl,ylliao}@nlg.csie.ntu.edu.tw,
hhhuang@iis.sinica.edu.tw, hhchen@ntu.edu.tw

**Abstract**

Multimodal reasoning has advanced significantly with large vision-language models (LVLMs), yet their robustness under input variations remains underexplored. This study investigates positional bias in LVLMs for multimodal multiple-choice questions. Our analysis reveals that model predictions are sensitive to both choice and modality ordering. We conduct a large-scale evaluation on MMMU, CVQA, and MMBENCH using fourteen representative models. We further show that question properties, including difficulty, domain, and image type, significantly modulate model robustness. We also assess whether text-based mitigation strategies transfer to the VQA setting, and conduct ablation studies on self-consistency and reasoning complexity. Overall, our findings provide the first comprehensive understanding of positional bias from a vision-language perspective, highlighting key challenges in achieving stable multimodal reasoning.

**Keywords:** Large Vision Language Model, Positional Bias, LLM Robustness

## 1. Introduction

Large language models have rapidly evolved to handle multimodal inputs, achieving strong performance across diverse benchmarks (Gemini Team, 2023; OpenAI, 2024; Meta AI, 2025). They now reach near-human accuracy on vision-language tasks such as visual question answering (Lu et al., 2022; Yue et al., 2024) and image captioning (Rotstein et al., 2024; Cheng et al., 2025). However, these advances mask a key vulnerability: large vision-language models (LVLMs) exhibit positional biases that undermine robustness and reliability. Prior work (Wei et al., 2024; Pezeshkpour and Hruschka, 2024; Zheng et al., 2024) has examined order sensitivity in text-only settings, but the role of positional bias in multimodal reasoning remains largely unexplored (Tian et al., 2025).

We conduct the first systematic study of positional bias in LVLMs for multimodal multiple-choice question answering. As shown in Figure 1, we examine two main forms of instability: (a) *choice-order sensitivity*, where reversing answer options alters predictions, and (b) *modality-order sensitivity*, where swapping the order of image and text inputs disrupts reasoning. Our experiments cover fourteen LVLMs on three benchmarks, MMMU, CVQA, and MMBENCH. We further analyze how question properties such as difficulty, domain, and image type affect robustness, and assess whether text-based mitigation methods transfer to multimodal settings. In addition, we conduct ablation studies on self-consistency and reasoning complexity to understand how inference mechanisms influence stability.

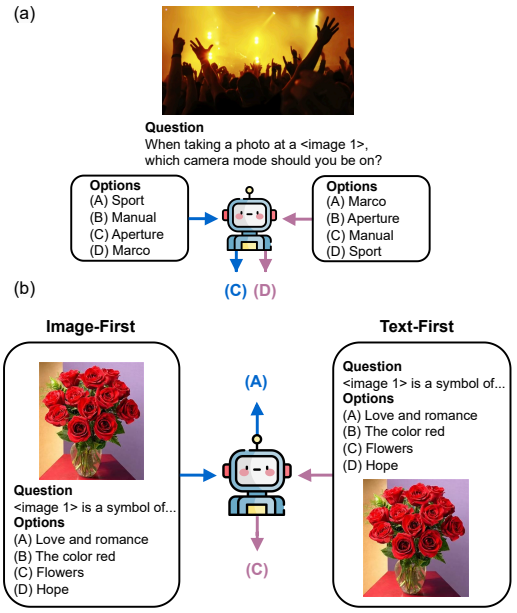


Figure 1: Example of choice order and modality order variations in vision-language MCQs.

Our contributions are threefold. (1) We provide the first large-scale analysis of positional bias in multimodal reasoning, identifying systematic sensitivity to both choice and modality ordering. (2) We develop a unified framework to quantify robust-



*Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026)*, pages 9118–9128
11–16 May 2026. 

ness across datasets, model families, and question attributes. (3) We evaluate mitigation strategies, including caption-based enhancement, self-consistency, and reasoning-based methods, revealing their effectiveness and limitations. Together, these findings offer new insights into the robustness of multimodal reasoning.

## 2. Problem Formulation and Experimental Setup

### 2.1. Notations

Given a question $Q$ with $k$ answer choices, each instance consists of a textual description of the question $T$, an image $I$, and a choice set $C$:

$$C = \langle (s_1, o_1), \ldots, (s_k, o_k) \rangle \tag{1}$$

where $s_i \in S = \{s_1, s_2, \ldots, s_k\}$ denotes the choice symbols and $o_i \in O = \{o_1, o_2, \ldots, o_k\}$ represents the corresponding options. As illustrated in Figure 1, we explore the **order sensitivity** of large vision-language models (LVLMs) along two dimensions: (a) *choice order*, and (b) *modality order*.

**Choice Order Sensitivity.** To examine the effect of option ordering, we reverse the sequence of answer choices while keeping the instruction, question, and image unchanged. This design isolates the impact of positional variation by ensuring that the semantic content of the task remains constant. The reversed choice set is defined as

$$C_{\text{rev}} = \langle (s_1, o_k), (s_2, o_{k-1}), \ldots, (s_k, o_1) \rangle. \tag{2}$$

The corresponding input sequences are defined as

$$P_{\text{original}} = \langle \texttt{[INST]}, I, T, C \rangle, \text{and} \tag{3}$$

$$P_{\text{reversed}} = \langle \texttt{[INST]}, I, T, C_{\text{rev}} \rangle, \tag{4}$$

Here, `[INST]` denotes the instruction prompt, $I$ represents the context image, $T$ the textual component of the question, and $C$ the corresponding choice set. This formulation enables a controlled evaluation of model robustness under option-order perturbations, allowing us to systematically characterize order sensitivity in LVLMs across different choice configurations.

**Modality Order Sensitivity.** We further investigate whether the ordering of multimodal inputs influences model predictions. Specifically, we consider two configurations: IMAGE-FIRST and TEXT-FIRST, which determine whether the image or text follows the instruction prompt. Formally, the input sequences are defined as:

$$P_{\text{img-first}} = \langle \texttt{[INST]}, I, T, C \rangle, \tag{5}$$

$$P_{\text{text-first}} = \langle \texttt{[INST]}, T, I, C \rangle. \tag{6}$$

This formulation allows for a controlled analysis of modality-order effects on model behavior.

### 2.2. Measurement of Positional Bias

To quantify the order sensitivity of model predictions, we follow prior work on text-only positional bias (Wei et al., 2024; Croce et al., 2021) and adapt two complementary metrics beyond accuracy: *Fluctuation Rate* (FR) and *Relative Standard Deviation* (RSD).

**Fluctuation Rate (FR).** The FR measures the proportion of prediction changes under different input permutations. Specifically, for choice-order and modality-order sensitivity, they are defined as:

$$\text{FR}_{\text{choice}} = \frac{1}{N} \sum_{i=1}^{N} \delta(M(P_{\text{original}}), M(P_{\text{reversed}})), \tag{7}$$

$$\text{FR}_{\text{modal}} = \frac{1}{N} \sum_{i=1}^{N} \delta(M(P_{\text{img-first}}), M(P_{\text{text-first}})), \tag{8}$$

where $N$ denotes the number of questions, $M(\cdot)$ the model prediction , and $\delta(a, b)$ the disagreement indicator:

$$\delta(a, b) = \begin{cases} 1, & \text{if } a \neq b, \\ 0, & \text{otherwise.} \end{cases} \tag{9}$$

A higher FR value indicates greater order sensitivity, reflecting decreased prediction stability under input reordering.

**Relative Standard Deviation (RSD).** The RSD captures the variability of class-wise accuracy across different choice positions:

$$\text{RSD} = \frac{\sqrt{\frac{1}{k} \sum_{i=1}^{k} (a_i - \bar{a})^2}}{\bar{a}}, \tag{10}$$

where $k$ is the number of answer choices, $a_i$ the accuracy associated with the $i^{\text{th}}$ position, and $\bar{a}$ their mean. A higher RSD indicates stronger positional bias and lower robustness to order perturbations.

### 2.3. Evaluation Tasks

Our experiments cover three multimodal benchmarks: MMMU (Yue et al., 2024), a college-level benchmark spanning 30 subjects and 183 subfields across six disciplines with diverse visual formats such as charts, diagrams, and maps, for which we use the validation split since the official test set does not provide ground-truth annotations; CVQA (Mogrovejo et al., 2024), a culturally diverse multilingual VQA dataset with over

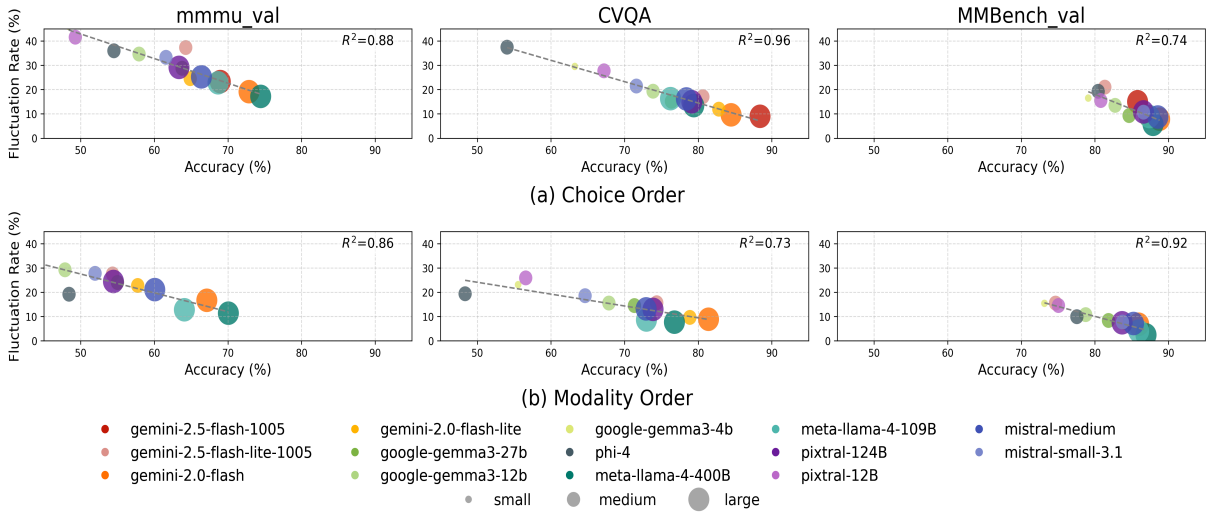


Figure 2: Accuracy versus fluctuation rate under (a) choice order and (b) modality order perturbations across three benchmarks. Each point represents a model, with bubble size proportional to model scale.

10K questions across 39 country–language pairs, where we sample 50 examples per country and use the English-translated version to ensure consistent cross-benchmark evaluation across all models; and MMBENCH (Liu et al., 2025), a bilingual benchmark assessing multimodal reasoning with structured multiple-choice evaluation, from which we use 50 examples per category in the English validation split, as the test set lacks ground-truth labels. Table 1 summarizes the statistics of all datasets used in our experiments.

| Benchmark | # Questions | # Options |
|---|---|---|
| MMMU | 11,550 | 2–5 |
| CVQA | 10,374 | 4 |
| MMBench | 13,217 | 4 |

Table 1: Data statistics of our benchmarks.

### 2.4. Models and Other Details

We evaluate fourteen LVLMs spanning five model families: the proprietary GEMINI series and the open-source LLAMA 4, MISTRAL, GEMMA 3, and PHI-4. The GEMINI family includes 2.0 Flash Lite, 2.0 Flash, 2.5 Flash Lite, and 2.5 Flash (Gemini Team, 2023). For MISTRAL, we assess Pixtral 12B (Agrawal et al., 2024), Pixtral 124B, Small 3.1, and Medium 3. Within LLAMA 4, we include the Scout (109B) and Maverick (400B) (Meta AI, 2025) variants, and for GEMMA 3, we evaluate the 4B, 12B, and 27B models. All models are accessed via APIs provided by Google, Meta, Mistral, and NVIDIA, and inference is executed in parallel for efficiency. We adopt standardized prompting from OpenAI's `simple-evals`[1] framework and fix the sampling temperature to 0 to ensure deterministic and reproducible results.

## 3. Order Sensitivity in VLMs

### 3.1. Overall Observation

**Choice Order Sensitivity.** We measure choice-order sensitivity to examine whether model predictions are influenced by perturbations in the ordering of answer options. Figure 2(a) illustrates the relationship between accuracy and fluctuation rate across the benchmarks. Overall, MMMU emerges as the most challenging dataset, exhibiting consistently lower accuracy compared to CVQA and MMBENCH. Across all benchmarks, smaller models cluster in the upper-left region, corresponding to lower accuracy and higher FR. This pattern indicates weaker robustness and stronger sensitivity to choice order. In contrast, larger models tend to occupy the lower-right region, reflecting greater robustness and reduced order sensitivity. This inverse correlation between FR and accuracy suggests that robustness to choice-order perturbations consistently improves with increasing model scale.

**Modality Order Sensitivity.** Figure 2(b) compares models across benchmarks in terms of accuracy and fluctuation rate (FR) under modality-order inversions. Models exhibit consistent but dataset-dependent patterns, with modality-order sensitivity being most pronounced in MMMU and weaker in MMBENCH, similar to the trend observed in the choice-order setting. The relationship with model

[1] https://github.com/openai/simple-evals

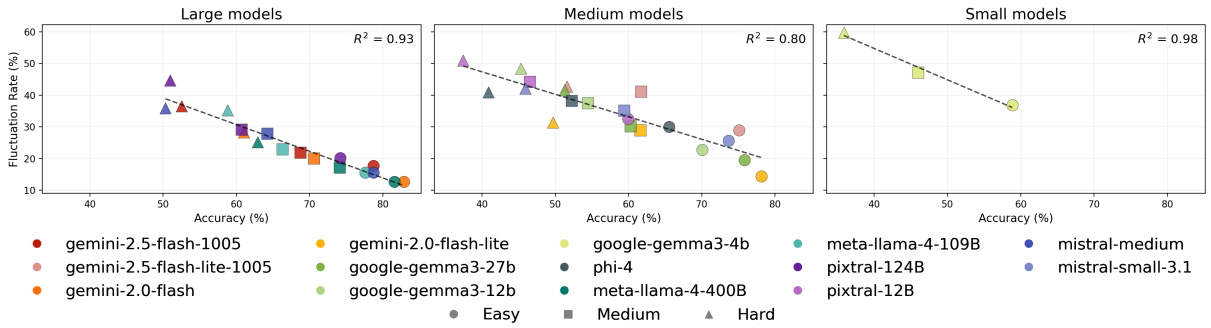


Figure 3: Accuracy versus fluctuation rate across different model scales. Each subplot corresponds to large, medium, and small models, respectively, with marker shapes indicating question difficulty.

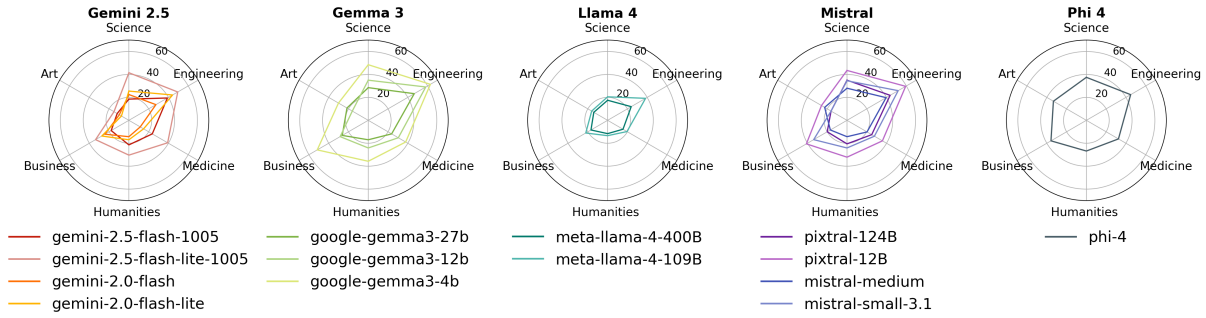


Figure 4: Fluctuation rate across six MMMU domains grouped by model family. Each radar plot represents a model's performance distribution across domains.

size follows a similar pattern: smaller models tend to lie in the upper-left region, whereas larger models cluster in the lower-right. We also observe an inverse correlation between FR and accuracy, indicating that models with higher robustness to modality-order perturbations generally achieve better overall accuracy.

## 3.2. In-Depth Analysis

In this section, we further examine whether different question attributes influence model robustness.

**Effect of Question Difficulty.** Figure 3 illustrates the relationship between accuracy and fluctuation rate across three difficulty levels in MMMU. The scatter plots are grouped by model scale, with each point representing a specific model–difficulty pair. Among large models, the points cluster toward the upper-left region, indicating high accuracy and low FR, which reflects strong robustness and consistent reasoning across difficulty levels. Medium-scale models exhibit a more dispersed distribution: as difficulty increases from Easy (circles) to Hard (triangles), accuracy gradually decreases while FR increases, forming a downward-right trend. Small models show the strongest sensitivity, with accuracy dropping below 60% and FR exceeding 50%. Overall, the inverse correlation between accuracy and FR becomes steeper with increasing difficulty, suggesting that model robustness degrades more rapidly on harder questions.

**Effect of Question Domain.** Figure 4 presents model stability across six MMMU question domains: Science, Engineering, Medicine, Humanities, Business, and Art, grouped by model family. *Artistic* and *Humanities* questions, which feature naturalistic or stylistic imagery with relatively low symbolic density, achieve consistently lower FR across most model families. In contrast, domains with higher information density and complex visual-textual relations, such as *Science*, *Business*, and *Engineering*, significantly challenge the models' visual-textual grounding ability, leading to marked increases in FR. Across families, larger models maintain comparatively stable trajectories across the six domains, suggesting stronger multimodal alignment and feature abstraction. Smaller models within the same families show higher instability, implying weaker robustness to question complexity.

**Effect of Question Image Type.** Figure 5 illustrates the stability of model performance across four image types in MMMU: Artistic, Photographic, Scientific, and Diagrammatic. Overall, models achieve the lowest FR on *Artistic/Illustrative* and *Real-world/Photographic* questions, indicating that

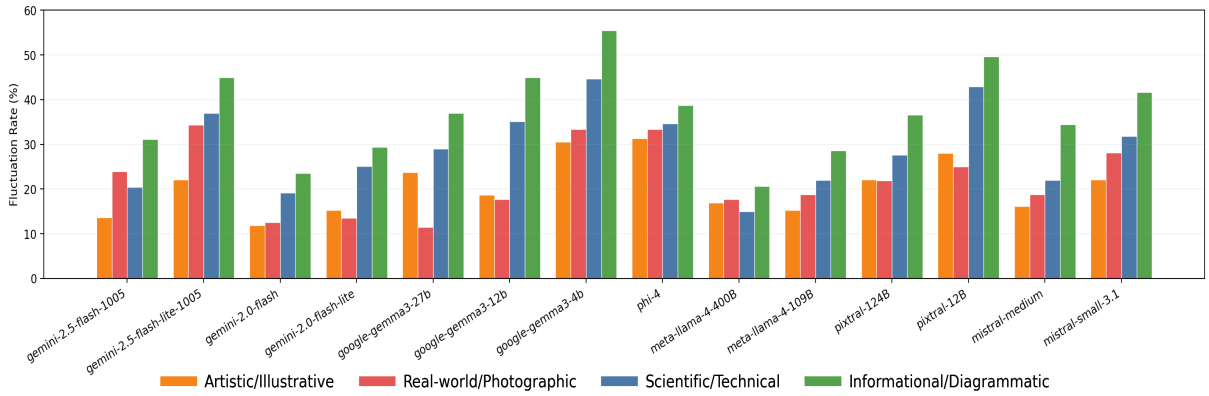


Figure 5: Fluctuation rate across four MMMU image types for all evaluated models. Each bar represents the model's stability under a specific image type.

everyday imagery and stylistic illustrations are generally easier for most multimodal models to interpret. In contrast, *Scientific/Technical* and *Informational/Diagrammatic* categories exhibit consistently higher FR values, particularly among medium and small models. These types often contain complex symbolic notations or data-dense figures such as plots and diagrams, which place greater demands on visual reasoning and cross-modal alignment. Notably, large models such as LLAMA 4 400B maintain smoother performance trends across image types, showing less variation between visual domains. This pattern suggests that scaling up multimodal capacity enhances stability across complex visual contexts, reflecting stronger integration of vision and language representations.

| Property | Slope | $R^2$ | # Points |
|---|---|---|---|
| Difficulty | -0.85 | **0.88** | 42 |
| Domain | -1.03 | 0.79 | 84 |
| Image Type | -1.08 | 0.77 | 56 |

Table 2: Linear regression between accuracy and FR across different question properties. The number of points equals the number of models multiplied by the number of categories for each property.

**Stability Across Question Properties.** We further examined how question-level properties modulate the relationship between model accuracy and FR. As shown in Table 2, a strong negative linear correlation was observed for all factors, indicating that higher accuracy generally coincides with lower instability. Among the three factors, *Image Type* exerts the strongest but most variable effect on performance stability ($R^2 = 0.77$), suggesting that visual complexity amplifies instability in a less predictable manner across models. Domain falls between the two extremes, with a moderate slope and explanatory power ($R^2 = 0.79$), indicating that subject-area variation introduces a meaningful but relatively stable source of instability. This is consistent with the domain-level patterns in Figure 4, where fields with higher symbolic density consistently exhibit elevated FR. In contrast, *Difficulty* exhibits a weaker but more consistent linear trend ($R^2 = 0.88$), revealing that intrinsic challenge remains the most reliable determinant of model robustness. Overall, these results show that as question properties increase in complexity, accuracy declines and FR rises predictably across models, confirming that multimodal robustness degrades along dimensions of both perceptual and cognitive difficulty.

## 4. Discussion

### 4.1. Do Text-Based Mitigation Strategies Transfer to VQA?

**AOI** Choi et al. (2025) propose the Auxiliary Option Injection (AOI) method, which introduces an additional "I don't know" (IDK) option into the original set of answer choices to mitigate positional bias in purely textual settings. The modified choice set is defined as:

$$C_{\text{AOI}} = \langle (s_1, o_1), (s_2, o_2), \ldots, (s_k, o_k), (s_{k+1}, \text{IDK}) \rangle \quad (11)$$

Accordingly, the input sequence becomes:

$$P_{\text{AOI}} = \langle \texttt{[INST]}, I, T, C_{\text{AOI}} \rangle \quad (12)$$

Following the post-processing strategy in Choi et al. (2025), if the model selects the IDK option, we randomly select one of the original $k$ choices as the final prediction.

**RE2** Xu et al. (2024) propose a simple prompting strategy called Re-Reading (RE2), which improves model performance by enhancing input comprehension. Unlike approaches such as Chain-of-Thought (CoT) that elicit reasoning in the output,

| Model | $\text{MMMU}_{\text{val}}$ | | | $\text{CVQA}_{\text{EN}}$ | | | $\text{MMBench}_{\text{dev}}$ | | |
|---|---|---|---|---|---|---|---|---|---|
| | Acc↑ | RSD↓ | FR↓ | Acc↑ | RSD↓ | FR↓ | Acc↑ | RSD↓ | FR↓ |
| Llama 4 Maverick 400B | **74.45** | 4.49 | **17.20** | **79.40** | 3.21 | 13.40 | 87.90 | 0.57 | 5.70 |
| +AOI | 73.72 | 4.03 | 18.41 | 78.53 | 3.34 | 14.00 | **88.40** | **0.53** | **5.20** |
| +RE2 | 73.11 | 3.66 | 18.90 | 79.00 | 3.68 | 14.60 | 87.95 | 1.09 | 5.30 |
| +CapS | 72.87 | **3.55** | 18.41 | 78.23 | **1.93** | 13.40 | 87.55 | 1.05 | 5.50 |
| +CapL | 73.84 | 3.83 | 18.66 | 78.50 | 2.48 | **11.67** | 86.25 | 0.69 | **5.20** |
| +OptEcho | 72.93 | 4.65 | 21.10 | 79.30 | 3.45 | 14.40 | 87.75 | 0.73 | 5.70 |
| Llama 4 Maverick 109B | 68.66 | 4.03 | 22.80 | **76.23** | 2.95 | 16.27 | 87.40 | 1.82 | 8.90 |
| +AOI | **69.45** | 4.54 | **21.46** | 73.87 | 3.22 | 17.27 | 87.10 | 1.09 | 7.90 |
| +RE2 | 69.02 | 3.90 | 22.44 | 75.93 | **2.70** | 17.67 | **88.50** | **0.65** | 7.40 |
| +CapS | 69.39 | 6.21 | 23.41 | 75.30 | 3.47 | 16.60 | 87.50 | 1.32 | 7.80 |
| +CapL | 67.74 | **3.31** | 24.51 | 73.03 | 3.94 | 13.80 | 86.90 | 0.66 | 7.30 |
| +OptEcho | 68.35 | 5.07 | 25.24 | 75.53 | 4.54 | **12.87** | 83.95 | 0.98 | **6.30** |
| Gemini 2.5 Flash Lite | 64.27 | 6.27 | 37.32 | **80.63** | **1.39** | 17.07 | **81.35** | 3.50 | 21.00 |
| +AOI | 64.57 | 5.97 | 35.61 | 77.60 | 3.00 | 22.53 | 81.05 | **2.46** | **20.50** |
| +RE2 | **65.98** | 5.94 | **34.02** | 79.40 | 1.87 | **18.60** | 80.05 | 3.50 | 22.30 |
| +CapS | 65.00 | 4.87 | 37.07 | 79.43 | 1.50 | 20.93 | 78.95 | 2.53 | 22.40 |
| +CapL | 64.70 | **3.09** | 35.98 | 77.97 | 1.52 | 20.67 | 78.35 | 3.12 | 22.40 |
| +OptEcho | 57.93 | 8.61 | 51.10 | 72.40 | 2.28 | 34.40 | 69.05 | 6.43 | 42.50 |

Table 3: Results of mitigation methods on three representative models.

| **Method** | Avg $\text{RI}_{\text{Acc}}$ | Avg $\text{RI}_{\text{RSD}}$ | Avg $\text{RI}_{\text{FR}}$ |
|---|---|---|---|
| AOI | -0.83 | -5.53 | -1.87 |
| CapL | -1.81 | 12.96 | 1.63 |
| CapS | -0.82 | -2.84 | -2.72 |
| OptEcho | -4.76 | -28.64 | -25.73 |
| RE2 | -0.15 | -4.52 | -0.92 |

Table 4: Relative improvements (%) of each mitigation method across all models and datasets.

RE2 modifies the input by repeating the question, thereby reinforcing the model's understanding during encoding. The resulting input sequence is:

$$P_{\text{RE2}} = \langle \texttt{[INST]}, I, T, T, C \rangle \tag{13}$$

**CapS and CapL** Inspired by methods such as Hu et al. (2023), which extract visual information as textual context to enhance VQA performance, we explore how the density of this auxiliary information affects bias mitigation. Given a question $Q$ and its associated image $I$, we generate image captions using a captioning model $M(\cdot)$ guided by predefined meta-prompts. The **CapS** module produces a concise, single-sentence summary, whereas the **CapL** module generates a detailed, paragraph-level description. These captions are then incorporated into the model input to provide additional contextual grounding. The caption generation process is defined as:

$$\text{CapS}(I) = M(\langle P_{\text{short}}, I \rangle) \tag{14}$$

$$\text{CapL}(I) = M(\langle P_{\text{long}}, I \rangle) \tag{15}$$

where $P_{\text{short}}$ and $P_{\text{long}}$ are prompts used to elicit short and long image descriptions, respectively. The resulting input sequences are:

$$P_{\text{CapS}} = \langle \texttt{[INST]}, I, \text{CapS}(I), T, C \rangle \tag{16}$$

$$P_{\text{CapL}} = \langle \texttt{[INST]}, I, \text{CapL}(I), T, C \rangle \tag{17}$$

**OptEcho** From the textual perspective, inspired by Xu et al. (2024), repeating parts of the input is hypothesized to strengthen attention allocation in large language models, thereby supporting more effective reasoning. We apply this idea to the answer choices by repeating the choice set, effectively allowing the model to attend to the options twice. The resulting input sequence is:

$$P_{\text{OptEcho}} = \langle \texttt{[INST]}, I, T, C, C \rangle \tag{18}$$

where $I$ is the image, $T$ is the question text, and $C$ is the set of answer choices.

**Analysis across Mitigation Strategies** Table 3 presents the results of five mitigation strategies compared against the ZERO-SHOT-COT baseline across three representative models. Due to space and computational constraints, we focus on one closed-source model (GEMINI 2.5 FLASH LITE) and two open-source models of different scales (LLAMA 4 109B and LLAMA 4 400B). This setup enables both cross-family (closed vs. open) and cross-scale comparisons, providing a balanced perspective on how mitigation strategies perform across models with distinct characteristics.

To quantify the effectiveness of each mitigation strategy, we define the *Relative Improvement* (RI) for a metric $m \in \{\text{Acc}, \text{RSD}, \text{FR}\}$ as:

$$\text{RI}_m = \frac{s^{(m)}_{\text{method}} - s^{(m)}_{\text{base}}}{|s^{(m)}_{\text{base}}|} \times d_m, \tag{19}$$

where $s^{(m)}$ denotes the score of metric $m$ and $d_m$ indicates the direction of improvement:

$$d_m = \begin{cases} +1, & \text{if higher is better (e.g., Acc)} \\ -1, & \text{if lower is better (e.g., FR).} \end{cases} \tag{20}$$

Positive values of RI consistently represent performance gains over the baseline, with higher $\text{RI}_{\text{Acc}}$

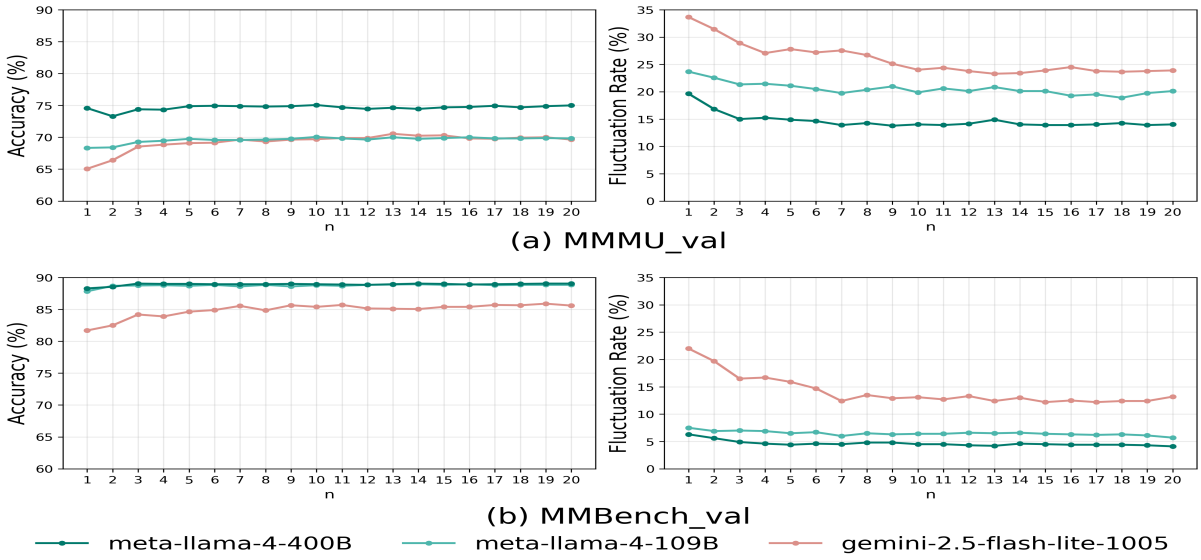


Figure 6: Ablation results of self-consistency with the number of samples $n$ varying from 1 to 20.

reflecting improved accuracy and higher $RI_{RSD}$ or $RI_{FR}$ indicating reduced positional bias. Macro-averaged results across models and datasets are summarized in Table 4. As shown in Table 3, the mitigation strategies exhibit mixed effectiveness across all models, with their benefits varying by model scale and family. No single strategy consistently outperforms the others. In addition, as summarized in Table 4, most relative improvement values are negative, indicating that these mitigation methods often fail to enhance performance and, in most of cases, even degrade it.

## 4.2. Does Self-Consistency Mitigate Positional Bias?

Self-consistency (Wang et al., 2023) has proven effective across various reasoning tasks. In this section, we examine whether it also helps mitigate positional bias. Following the same experimental setup as in Section 4.1, we evaluate GEMINI 2.5 FLASH LITE, LLAMA 4 109B, and LLAMA 4 400B on two representative benchmarks: the most challenging dataset, MMMU, and the easiest one, MMBENCH. To further assess the impact of self-consistency, we conduct an ablation study with $n = 20$ samples, and the results are shown in Figure 6. As $n$ increases, all models exhibit consistent gains in accuracy along with reductions in FR, indicating that self-consistency sampling effectively enhances both correctness and stability. In particular, when $n$ increases from 1 to 4, the improvement is substantial, and the effect plateaus when $n > 10$. Overall, these results confirm that self-consistency not only boosts performance but also stabilizes output distributions across different models and datasets, thereby narrowing the robustness gap.

| Model | Thinking Limit | $MMMU_{val}$ | | | $MMBench_{dev}$ | | |
|---|---|---|---|---|---|---|---|
| | | Acc↑ | RSD↓ | FR↓ | Acc↑ | RSD↓ | FR↓ |
| 2.5 Flash | 0 | 68.48 | **2.05** | 21.59 | 89.80 | 1.61 | 6.10 |
| | 1024 | 69.96 | 4.21 | 21.12 | **90.25** | 0.87 | 6.60 |
| | 2048 | **73.90** | 2.96 | 15.73 | 89.95 | 1.35 | **3.70** |
| | 4096 | 71.95 | 3.23 | **13.17** | 89.25 | 0.93 | 3.70 |
| | 8192 | 64.02 | 2.76 | 22.44 | 87.90 | **0.59** | 5.50 |
| 2.5 Pro | - | 79.94 | 1.46 | 11.46 | 92.05 | 1.81 | 3.10 |

Table 5: Results of varying reasoning complexity on the MMMU and MMBENCH datasets using direct prompting. Results from GEMINI 2.5 PRO are included as an upper-bound reference.

| Dataset | Thinking Limit 1024 | 2048 | 4096 | 8192 |
|---|---|---|---|---|
| $MMMU_{val}$ | 630 | 1287 | **1888** | 1483 |
| $CVQA_{EN}$ | 348 | **460** | 454 | 424 |
| $MMBench_{dev}$ | 318 | **451** | 423 | 397 |

Table 6: Average number of tokens generated during reasoning under different thinking-length constraints.

## 4.3. Do Reasoning Models Reduce Positional Bias?

Recently, large reasoning models (LRMs) such as DeepSeek-R1 (DeepSeek-AI, 2025), Gemini 2.5 Flash with extended thinking (Gemini Team, 2023), and Qwen3 (Team, 2025) have shown reasoning capabilities that surpass traditional chain-of-thought (CoT) prompting. In our experiments, we select GEMINI 2.5 FLASH as the representative model, as it is the only one that supports multimodal input while offering accessible API-based inference within our computational budget. We evaluate on MMBENCH and MMMU, the same

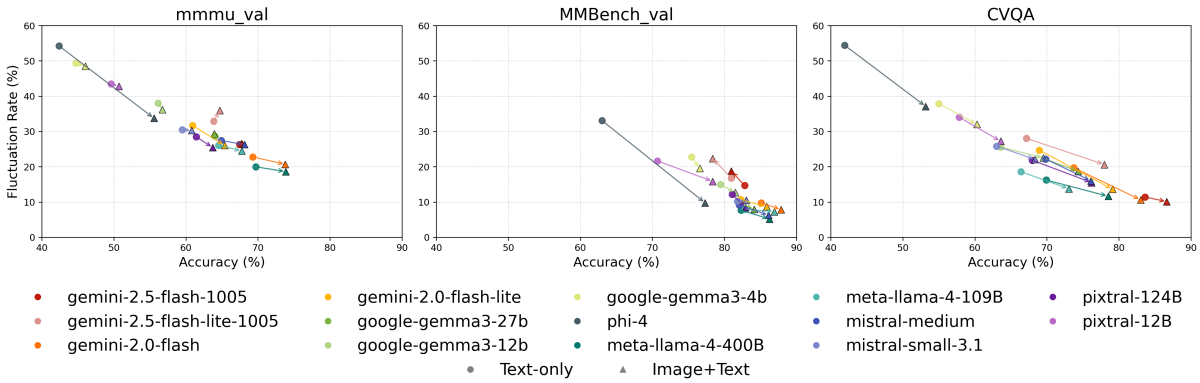


Figure 7: Accuracy versus fluctuation rate between text-only and image+text conditions across three benchmarks. Each arrow connects the same model's performance under the two settings.

benchmarks used in Section 4.2, to ensure comparability. Our objective is to investigate whether increasing reasoning complexity can also mitigate positional bias. To this end, we conduct ablation studies by varying the *thinking budget* parameter provided by the Gemini API, thereby simulating different levels of reasoning depth and complexity.

Regarding the results shown in Table 5, LRMs generally achieve higher accuracy and exhibit lower selection bias than models without explicit reasoning. As the reasoning budget increases, performance improves, suggesting that longer reasoning contributes to greater robustness and capability. For the relatively easier task MMBENCH, peak performance is typically observed when the budget is set between 1024 and 2048 tokens. In contrast, for the more challenging dataset MMMU, optimal performance occurs in the 2048 to 4096 token range. However, we also find that excessively long reasoning can be detrimental. At 8192 tokens, both accuracy and selection bias degrade, with performance falling below that of models without reasoning. This observation aligns with recent findings (Chen et al., 2025; Sui et al., 2025) that highlight the risks of overthinking in large models.

### 4.4. Impact of Visual Input.

We introduced the caption-enhancement methods **CapS** and **CapL** in Section 4.1. Here, we further explore the role of images in VQA tasks by examining the impact of removing visual inputs and relying solely on textual context. Figure 7 presents the comparison results, where *text-only* denotes inputs without images and *image+text* represents the original **CapL** setting. Across all three datasets, the results consistently show that including images leads to higher accuracy and lower fluctuation rates, indicating more stable and reliable reasoning. Specifically, on MMMU, the median accuracy increases by 1.8% while FR decreases by 1.2%; on MMBENCH, accuracy improves by 2.4% with a 2.1% reduction in FR; and on CVQA, which contains more visually grounded questions, the improvement is most pronounced, with a 6.4% increase in accuracy and a 6.0% decrease in FR. These findings demonstrate that models indeed leverage visual information when available, becoming more consistent in their responses.

## 5. Related Work

**Vision-Language Model.** Over the past decade, vision–language models (VLMs) have evolved from single-stream (Li et al., 2019; Chen et al., 2020) or dual-stream (Lu et al., 2019; Tan and Bansal, 2019) transformers to unified architectures. Early models such as CLIP (Radford et al., 2021) were primarily trained from scratch, while later models, including Flamingo (Alayrac et al., 2022), BLIP-2 (Li et al., 2023), and LLaVA (Liu et al., 2023), typically leverage pre-trained language models as backbone architectures. Integrating vision into GPT-scale frameworks has paved the way for interactive systems. Recent models, such as Gemini (Gemini Team, 2023), GPT-4o (OpenAI, 2024), Pixtral (Agrawal et al., 2024), and LLaMA (Dubey et al., 2024; Meta AI, 2025), have extended context lengths, incorporated additional modalities, and improved efficiency. These advancements mark a shift toward unified omni-modal systems that set new performance benchmarks.

**Positional Bias in LLMs.** Previous studies have shown that LLMs exhibit order sensitivity and selection bias when handling MCQs (Zheng et al., 2023; Wei et al., 2024; Pezeshkpour and Hruschka, 2024; Zheng et al., 2024; Choi et al., 2025; Shen et al., 2023). Prior work has proposed several approaches to mitigate this issue. Choi et al. (2025) introduce Bias Node Pruning (BNP), which removes a small subset of parameters (0.002% of

the total) responsible for bias in white-box settings. Wei et al. (2024) propose Probability Weighting and Probability Calibration, which post-process token-level log probabilities in gray-box settings, and a Two-Hop Strategy applicable to black-box models. Zheng et al. (2024) similarly address the problem by modifying token probabilities. However, all existing efforts focus exclusively on purely textual scenarios. To the best of our knowledge, there has been no systematic investigation of selection bias in multimodal settings.

## 6. Conclusion

This work presents the first systematic investigation of positional bias in LVLMs for multimodal MCQs. We analyze two major sources of instability, choice-order and modality-order sensitivity, across fourteen representative models on three benchmarks, MMMU, CVQA, and MMBENCH. Our results show that both forms of perturbation induce substantial prediction shifts, even in state-of-the-art models, revealing persistent vulnerabilities in multimodal reasoning. Further analysis demonstrates that question properties such as difficulty, domain, and image type strongly influence robustness, with higher perceptual and cognitive complexity amplifying instability. We also evaluate several mitigation strategies and reasoning-related factors. Text-based methods transfer inconsistently to multimodal settings, while self-consistency sampling consistently improves both accuracy and stability. Moreover, varying the reasoning depth of GEMINI 2.5 FLASH shows a nonlinear effect: moderate thinking budgets enhance robustness, but excessive reasoning can degrade it. Finally, experiments on visual inclusion and textual grounding confirm that stronger multimodal alignment reduces bias and improves stability. Overall, our findings provide a unified understanding of how input ordering affects the robustness of LVLMs. This study establishes a foundation for future research on bias diagnosis and mitigation in multimodal reasoning, and points toward the need for order-robust training objectives and benchmark design.

## Limitation

While this study provides valuable insights into positional bias in multimodal settings and evaluates several mitigation strategies for large vision-language models (LVLMs), several limitations warrant consideration. First, due to budget constraints, our experiments are limited to a selected set of representative models, and we were unable to evaluate all available LVLMs or extend the mitigation analysis to the full set of fourteen models. Second, following prior work on text-only positional bias (Wei et al., 2024), our choice-order perturbation considers only the reversed ordering rather than exhaustive permutations of answer options. While this design keeps computational costs tractable, it may not fully capture the range of positional bias patterns. Finally, the mitigation strategies examined in this work are applied at inference time without modifying model parameters; training-time interventions such as data augmentation with shuffled option orders or order-aware fine-tuning remain directions for future work.

## Acknowledgements

This work was supported by National Science and Technology Council, Taiwan, under grant NSTC 114-2221-E-002 -070 -MY3, NSTC 113-2634-F-002-003 -, and Ministry of Education (MOE) in Taiwan under grants NTU-114L900901.

## Bibliographical References

Pravesh Agrawal, Szymon Antoniak, Emma Bou Hanna, Baptiste Bout, Devendra Chaplot, Jessica Chudnovsky, Diogo Costa, Baudouin De Monicault, Saurabh Garg, Theophile Gervet, Soham Ghosh, Amélie Héliou, Paul Jacob, Albert Q. Jiang, Kartik Khandelwal, Timothée Lacroix, Guillaume Lample, Diego Las Casas, Thibaut Lavril, Teven Le Scao, Andy Lo, William Marshall, Louis Martin, Arthur Mensch, Pavankumar Muddireddy, Valera Nemychnikova, Marie Pellat, Patrick Von Platen, Nikhil Raghuraman, Baptiste Rozière, Alexandre Sablayrolles, Lucile Saulnier, Romain Sauvestre, Wendy Shang, Roman Soletskyi, Lawrence Stewart, Pierre Stock, Joachim Studnia, Sandeep Subramanian, Sagar Vaze, Thomas Wang, and Sophia Yang. 2024. Pixtral 12b.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millicah, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. 2022. Flamingo: a visual language model for few-shot learning. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Xingyu Chen, Jiahao Xu, Tian Liang, Zhiwei He, Jianhui Pang, Dian Yu, Linfeng Song, Qiuzhi Liu, Mengfei Zhou, Zhuosheng Zhang, Rui Wang, Zhaopeng Tu, Haitao Mi, and Dong Yu. 2025. Do not think that much for 2+3=? on the overthinking of o1-like llms.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Uniter: Universal image-text representation learning. page 104–120, Berlin, Heidelberg. Springer-Verlag.

Kanzhi Cheng, Wenpo Song, Jiaxin Fan, Zheng Ma, Qiushi Sun, Fangzhi Xu, Chenyang Yan, Nuo Chen, Jianbing Zhang, and Jiajun Chen. 2025. CapArena: Benchmarking and analyzing detailed image captioning in the LLM era. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 14077–14094, Vienna, Austria. Association for Computational Linguistics.

Hyeong Kyu Choi, Weijie Xu, Chi Xue, Stephanie Eckman, and Chandan K. Reddy. 2025. Mitigating selection bias with node pruning and auxiliary options. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5190–5215, Vienna, Austria. Association for Computational Linguistics.

Francesco Croce, Maksym Andriushchenko, Vikash Sehwag, Edoardo Debenedetti, Nicolas Flammarion, Mung Chiang, Prateek Mittal, and Matthias Hein. 2021. Robustbench: a standardized adversarial robustness benchmark. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Google Gemini Team. 2023. Gemini: A family of highly capable multimodal models. *ArXiv*, abs/2312.11805.

Yushi Hu, Hang Hua, Zhengyuan Yang, Weijia Shi, Noah A. Smith, and Jiebo Luo. 2023. Promptcap: Prompt-guided image captioning for vqa with gpt-3. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2951–2963.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, pages 34892–34916. Curran Associates, Inc.

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. 2025. Mmbench: Is your multi-modal model an all-around player? In *Computer Vision – ECCV 2024*, pages 216–233, Cham. Springer Nature Switzerland.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. *ViLBERT: pretraining task-agnostic visiolinguistic representations for vision-and-language tasks*. Curran Associates Inc., Red Hook, NY, USA.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *The 36th Conference on Neural Information Processing Systems (NeurIPS)*.

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

David Orlando Romero Mogrovejo, Chenyang Lyu, Haryo Akbarianto Wibowo, Santiago Góngora, Aishik Mandal, Sukannya Purkayastha, Jesus-German Ortiz-Barajas, Emilio Villa Cueva, Jinheon Baek, Soyeong Jeong, Injy Hamed, Zheng Xin Yong, Zheng Wei Lim, Paula Mónica Silva, Jocelyn Dunstan, Mélanie Jouitteau, David LE MEUR, Joan Nwatu, Ganzorig Batnasan, and Munkh-Erdene Otgonbold. 2024. CVQA: Culturally-diverse multilingual visual question answering benchmark. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

OpenAI. 2024. GPT-4o System Card. ArXiv:2410.21276.

Pouya Pezeshkpour and Estevam Hruschka. 2024. Large language models sensitivity to the order of

options in multiple-choice questions. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2006–2017, Mexico City, Mexico. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision.

Noam Rotstein, David Bensaïd, Shaked Brody, Roy Ganz, and Ron Kimmel. 2024. Fusecap: Leveraging large language models for enriched fused image captions. In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 5677–5688.

Chenhui Shen, Liying Cheng, Xuan-Phi Nguyen, Yang You, and Lidong Bing. 2023. Large language models are not yet human-level evaluators for abstractive summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4215–4233, Singapore. Association for Computational Linguistics.

Yang Sui, Yu-Neng Chuang, Guanchu Wang, Jiamu Zhang, Tianyi Zhang, Jiayi Yuan, Hongyi Liu, Andrew Wen, Shaochen Zhong, Hanjie Chen, and Xia Hu. 2025. Stop overthinking: A survey on efficient reasoning for large language models.

Hao Tan and Mohit Bansal. 2019. LXMERT: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5100–5111, Hong Kong, China. Association for Computational Linguistics.

Qwen Team. 2025. Qwen3 technical report.

Xinyu Tian, Shu Zou, Zhaoyuan Yang, and Jing Zhang. 2025. Identifying and mitigating position bias of multi-image vision-language models. *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10599–10609.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*.

Sheng-Lun Wei, Cheng-Kuang Wu, Hen-Hsen Huang, and Hsin-Hsi Chen. 2024. Unveiling selection biases: Exploring order and token sensitivity in large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*.

Xiaohan Xu, Chongyang Tao, Tao Shen, Can Xu, Hongbo Xu, Guodong Long, Jian-Guang Lou, and Shuai Ma. 2024. Re-reading improves reasoning in large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15549–15575, Miami, Florida, USA. Association for Computational Linguistics.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of CVPR*.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. 2024. Large language models are not robust multiple choice selectors. In *The Twelfth International Conference on Learning Representations*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36, pages 46595–46623.