

Unexplored flaws in multiple-choice VQA make benchmarking unreliable

Fabio Rosenthal^{1,2} Sebastian Schmidt¹ Thorsten Bagdonat²
 Thorsten Graf² Stephan Günnemann¹ Leo Schwinn^{1,3,4}

¹ Technical University of Munich, ² Volkswagen AG, ³ Helmholtz AI, ⁴ ELLIS

Correspondence: f.rosenthal@tum.de

Abstract

Previous works identify sensitivity to option order as a key issue in multiple-choice VQA (MC-VQA) evaluation and propose protocols to mitigate this effect. We show that such mitigation is insufficient to ensure the validity of MC-VQA as a reliable benchmark for Multimodal Large Language Models (MLLMs): performance remains highly sensitive to *semantically neutral* prompt format choices that are not controlled by current benchmarks. In a large-scale study spanning **seven** MLLMs and **five** MC-VQA datasets, we find frequent rank reversals even under order-invariant evaluation. These reversals arise when we systematically vary option ID sets, delimiters, and separators, yielding 48 semantically equivalent prompt formats. Mechanistic analyses trace this instability to low-level language modeling effects: tokenizer-induced fusion or removal of option ID tokens introduces corrupted option ID tokens into the input sequence, while the choice of option ID sets directly affects the reliability of attention patterns for option selection. Accordingly, MC-VQA rankings correlate weakly with open-ended evaluation, indicating that MC-VQA reflects option-selection dynamics in addition to multimodal reasoning. These findings identify prompt formatting as a major, previously under-controlled confounder in MC-VQA benchmarking and motivate evaluation protocols that explicitly control prompt format sensitivity.

1 Introduction

MLLMs demonstrate strong perceptual and reasoning capabilities on image-text interleaved inputs (Bai et al., 2025), making them applicable across domains such as high-resolution imaging (Zhong et al., 2025), medical imaging (Matos et al., 2025), and autonomous driving (Marcu et al., 2024). Benchmarking aims to evaluate these capabilities through standardized proxy tasks (Reuel-Lamparth

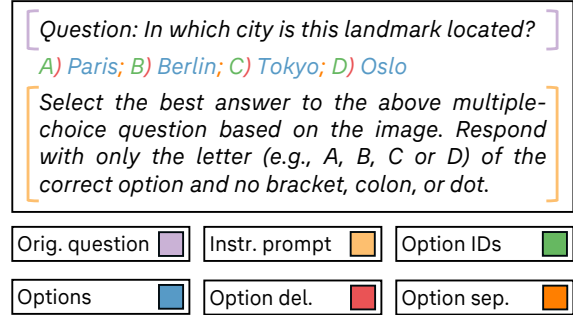


Figure 1: **Prompt template in MC-VQA.** A prompt in MC-VQA consists of the *question*, the *instruction*, and *options*, with their respective indexing *option IDs*.

et al., 2024; Eriksson et al., 2025). Ideally, benchmarks enable reliable predictions about model performance on future real-world tasks (Singh et al., 2025). However, benchmarks inherently provide only an imperfect proxy for true MLLM capabilities, as the open-ended tasks encountered in deployment are not directly measurable (Longjohn et al., 2024). Therefore, benchmarks are only meaningful insofar as their outcomes reflect underlying model capabilities, rather than being driven by superficial variations in the evaluation setup.

Common MLLM benchmarks are multiple-choice VQA (MC-VQA), in which MLLMs must select an answer from a set of options (Schwenk et al., 2022; Liu et al., 2024b; Yue et al., 2024), and thereby simplifies answer evaluation compared to open-ended generation (Robinson and Wingate, 2023; Molfese et al., 2025). Previous works show that the option order influences multiple-choice performance of (M)LLMs, even when the underlying task remains unchanged (Zheng et al., 2024; Pezeshkpour and Hruschka, 2024). This may undermine the reliability of benchmarking results on MC-VQA and limit its validity as a proxy for evaluating MLLMs. However, these works propose methods to mitigate the effects of option order.

Yet, these efforts focus exclusively on option ordering, overlooking other fundamental aspects of

prompt design. Beyond ordering, multiple format factors influence the options representation in the prompt (Wei et al., 2024; Sanz-Guerrero et al., 2025). Typically, *option IDs* index each option, and the prompt instructs the MLLM to output the corresponding ID rather than the option (Robinson and Wingate, 2023; Liu et al., 2024b). The *option delimiter* hyphenates each option ID and the respective option, and the *option separator* splits different option ID-option pairs in the prompt (see Fig. 1). Note that there is no standardized format for these elements, and prompt formats vary across datasets. Therefore, the influence of prompt format decisions on evaluation outcomes has not been systematically studied.

In this paper, we show that semantically neutral prompt format choices substantially affect MC-VQA evaluation, calling into question its reliability as a benchmark. For instance, LLaVA-OV (Li et al., 2025) shifts from the second-lowest to the top rank under a changed prompt format (see Fig. 2), with an accuracy difference of nearly 50 percentage points (pp). Similarly, LLaVA-1.5 (Liu et al., 2024a) and Qwen-2-VL (Wang et al., 2024) exhibit considerable accuracy differences across formats.¹ Across seven MLLMs, five datasets, and 48 prompt formats, we observe large accuracy variations and frequent rank reversals, even under order-invariant evaluation. These effects persist for larger and proprietary models.

We trace this instability to low-level language modeling effects rather than changes in visual or semantic reasoning. These effects introduce variability in how option ID tokens are represented and selected, leading to inconsistent model predictions across prompt formats. We further show that MLLMs are systematically aware of certain prompt formats that yield better performance. Finally, the weak alignment with open-ended evaluation suggests that MC-VQA reflects both option-selection dynamics and reasoning capabilities.

Our main contributions are: (1) We show in a large-scale study that small structural prompt format changes lead to benchmark rank variations, (2) We identify mechanistic sources of format-dependent instability, including tokenization failures, generalization behavior of option ID sets beyond training exposure, and the selection of option ID set influences attention patterns in MLLMs, (3) We show

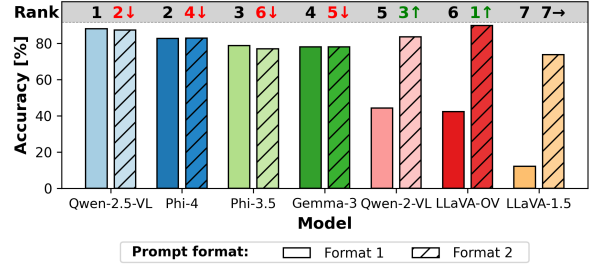


Figure 2: **Benchmark unreliability in MC-VQA.** We compare MLLM benchmarking on A-OKVQA using two prompt formats in the worst-case scenario. *Format 1* uses lowercase option IDs, double brackets as delimiters, and line breaks as separators, whereas *Format 2* uses uppercase option IDs, dots as delimiters, and commas as separators.

that MLLMs exhibit implicit awareness of prompt format effectiveness, systematically favoring certain formats at rates far exceeding random selection, and (4) We demonstrate that MC-VQA substantially overestimates multimodal reasoning ability and correlates weakly with open-ended evaluation, calling into question its reliability as a benchmark and proxy for downstream performance.

2 Preliminary

This section provides the necessary background on MLLMs, the MC-VQA task, and the benchmarking scope used in this work.

2.1 Multimodal Large Language Models

Multimodal Large Language Models (MLLMs) consist of three components: a Large Language Model (LLM), a vision encoder, and a modality projector (Li et al., 2023; Liu et al., 2024a). The vision encoder processes the input image to produce visual features represented as tokens $\mathbf{X}_v$ (Li et al., 2023; Liu et al., 2023). A modality projector maps $\mathbf{X}_v$ into the LLM embedding space, yielding $\mathbf{X}_{v, \text{proj}}$ (Liu et al., 2023, 2024a). A model-specific tokenizer converts the text input into textual tokens $\mathbf{X}_t$ (Li et al., 2023; Liu et al., 2023). Finally, the LLM receives a concatenation of the visual and text tokens (Abdin et al., 2024a):

$$\mathbf{X}_{\text{in}} = [\mathbf{X}_t ; \mathbf{X}_{v, \text{proj}}] \quad (1)$$

The LLM generates its output autoregressively, predicting one token at a time conditioned on the prompt and previously generated tokens (Vaswani et al., 2017; Liu et al., 2023). In MC-VQA, this procedure typically reduces to generating a short

¹These bias effects were isolated from the known position bias through a circular evaluation scheme.

option identifier (e.g., 'A' or 'a') that already appears in the prompt (Antol et al., 2015; Goyal et al., 2017). This decoding setup makes multiple-choice evaluation particularly sensitive to how option IDs are tokenized and represented in the input. Small prompt format changes that alter the tokenization of option IDs may therefore affect the generated tokens, even when the underlying reasoning remains unchanged.

2.2 Visual Question Answering

Visual Question Answering (VQA) evaluates the multimodal reasoning ability of MLLMs, which must answer questions about a given image (Schwenk et al., 2022; Yue et al., 2024). MC-VQA simplifies correctness assertion by restricting the output space to a fixed set of option IDs (Robinson and Wingate, 2023). Formally, each sample consists of an image, a question q , and a set of answer options $\mathcal{O} = \{o_j\}_{j=1}^n$ (Antol et al., 2015; Goyal et al., 2017). An option ID i_j from a predefined ID set $\mathcal{I} = \{i_j\}_{j=1}^n$ indexes each option o_j in the prompt. We define a prompt format as $F = (\mathcal{I}, d, s)$, where d denotes the delimiter separating an option ID from its option text, and s denotes the separator between different option ID-option pairs. Further, the permutation π specifies the order in which options appear. Given a fixed instruction prompt In that requests the model to output an option ID, the resulting multiple-choice prompt is (Liu et al., 2024b; Zhang et al., 2025):

$$T(q, F, \pi) = q \parallel \text{Join}_s((i_{\pi(k)} d o_{\pi(k)})_{k=1}^n) \parallel \text{In}.$$

The MLLM generates an output token sequence that ideally matches one of the option IDs and maps to the corresponding GT answer (Duan et al., 2024; Polo et al., 2024). This setup tightly couples model predictions to the tokenization and representation of option IDs, because the output of the model is a discrete ID that appears in the prompt. In this work, we analyze how variations in the prompt format F influence model behavior in MC-VQA, while controlling for the effect of option order π .

2.3 Benchmarking

Benchmarking evaluates a set of models $\{\mathcal{M}_1, \dots, \mathcal{M}_K\}$ on datasets and ranks them according to task accuracy: $\mathcal{M}_{(1)} \geq \mathcal{M}_{(2)} \geq \dots \geq \mathcal{M}_{(K)}$, where $\mathcal{M}_{(1)}$ is the highest-scoring model (Srivastava et al., 2023; Liang et al., 2023). These rankings aim to reflect

global model capability and therefore should not depend on arbitrary evaluation choices, such as the prompt format, that may alter rankings (Alzahrani et al., 2024). For MLLMs, benchmarking frequently relies on MC-VQA (Bai et al., 2025; Abdin et al., 2024a). In this work, we analyze how these rankings change under variations in prompt format, holding the underlying task and evaluation protocol fixed.

3 Related Work

We review work on prompt sensitivity in (M)LLMs, biases in multiple-choice evaluation, and the reliability of benchmark-based evaluation.

Prompt sensitivity in (M)LLMs. (M)LLMs are sensitive to prompt formulation, where small changes in phrasing or structure can lead to different outputs (He et al., 2024; Li et al., 2026). Prior work studies this phenomenon in a variety of settings, including few-shot learning, instruction following, and zero-shot evaluation (Zhuo et al., 2024; Sclar et al., 2024; Sun et al., 2024; Mizrahi et al., 2024). These results show that model behavior can depend strongly on surface-level prompt variations, raising concerns about evaluation consistency. However, most of this work focuses on free-form generation, while less attention has been paid to structured settings with semantically neutral option IDs, such as multiple-choice evaluation.

Biases in multiple-choice evaluation. In multiple-choice settings, prior work identifies systematic biases related to option order, where models prefer certain positions even when the underlying task is unchanged (Zheng et al., 2024; Pezeshkpour and Hruschka, 2024). Mitigation strategies include post-hoc calibration and set-invariant training objectives (Egressy and Stühmer, 2025; Chen et al., 2025), while work in multimodal settings remains limited (Tan et al., 2024). Beyond order effects, recent evidence suggests that multiple-choice tasks can overestimate (M)LLM capabilities (Raman et al., 2025; Liu et al., 2026), indicating that high accuracy may not translate to strong performance in open-ended settings. Further, it was shown that varying option ID sets has influence on LLM benchmarking outcomes (Yang et al., 2025). However, existing work primarily focuses on option permutations or model calibration, and does not systematically study how semantically neutral prompt format choices influence evaluation.

Benchmark reliability. Recent work ques-

tions the reliability of benchmark-based evaluation, showing that model rankings can change under small variations in evaluation design, such as prompt phrasing, formatting, or response ordering (Zhuo et al., 2024; Singh et al., 2025; He et al., 2024; Wei et al., 2024; Sanz-Guerrero et al., 2025). These findings suggest that evaluation results often depend on incidental design choices rather than intrinsic model capability (Zheng et al., 2024). In this work, we focus on MC-VQA and show that prompt format choices (option ID sets, delimiters, and separators) constitute a previously under-explored source of instability. In contrast to prior work, we provide both a systematic analysis and mechanistic evidence linking evaluation instability to tokenization and decoding behavior.

4 Benchmark Instability under Prompt Format Variations

Previous studies on biases in multiple-choice evaluation primarily focus on the effect of option order (Zheng et al., 2024; Pezeshkpour and Hruschka, 2024; Tan et al., 2024), while other sources of prompt variation remain largely unexplored. To address this gap, we conduct a systematic study of how semantically neutral prompt format choices influence MC-VQA evaluation. We evaluate model performance across 48 prompt formats, varying option IDs, delimiters, and separators, while controlling for option order through circular evaluation. Fig. 3 shows the resulting rank distributions of models across five MC-VQA datasets.

4.1 Experimental Framework

Here, we outline the experimental framework, including the models, datasets, implementation details, and evaluation methodology used in our study. We provide more extensive details in Sec. A.

MLLMs. We use popular MLLMs covering different architectures: Gemma-3 (Kamath et al., 2025), LLaVA-1.5 (Liu et al., 2024a), LLaVA-OV (Li et al., 2025), Phi-3.5 (Abdin et al., 2024b), Phi-4 (Abdin et al., 2024a), Qwen-2-VL (Wang et al., 2024) and Qwen-2.5-VL (Bai et al., 2025).

Datasets. We use a highly diverse selection of datasets covering different domains and task difficulties, containing high-resolution datasets HRBench-4K (Wang et al., 2025) and V*Bench (Wu and Xie, 2024), general knowledge dataset A-OKVQA (Schwenk et al., 2022), and multi-domain benchmark datasets MME-RealWorld-Lite (Zhang

et al., 2025) and MMBench (Liu et al., 2024b). To reduce poor instruction-following of models, we use an optimized instruction prompt for all datasets.

Evaluation Setting. As is common with MC-VQA datasets, we evaluate all models in a *zero-shot setting* (Li et al., 2022, 2023; Liu et al., 2024a). Additionally, we enforce fully deterministic outputs across all models by disabling sampling, ensuring consistent comparisons on the same datasets for different prompt formats. Further, we set the number of generated output tokens to match the maximum required to output all option IDs from the utilized option ID set (see Tab. 19).

Table 1: **Prompt format variations.** We systematically vary the prompt format across three factors: (1) option ID set, (2) option delimiter, and (3) option separator.

Factor	Format variant	Example repr.
Option ID set	Uppercase	A/B/C/D
	Lowercase	a/b/c/d
	Numbers	1/2/3/4
	Roman numbers	I/II/III/IV
Option delimiter	Colon	{opt}:
	Dot	{opt}.
	Bracket	{opt})
	Double brackets	{{opt}}
Option separator	Line break	{opt}\n {opt}
	Comma	{opt}, {opt}
	Semicolon	{opt}; {opt}

Prompt Format Variations. We systematically vary three components of the prompt format: (1) the option ID set, (2) the delimiter between ID and option text, and (3) the separator between options (see Tab. 1). Combining four option ID sets, four delimiters, and three separators results in 48 distinct prompt formats.²

4.2 Rank Instability Across Prompt Formats

We next analyze how model rankings vary across semantically equivalent prompt formats. Fig. 3 shows the rank distributions of all models across 48 prompt formats per dataset. Model rankings vary substantially across prompt formats and do not remain constant, in contrast to the *ideal evaluation*³ shown in Fig. 3. On several datasets, rank distributions overlap strongly and frequently invert,

²While computationally expensive, this exhaustive setup serves as a high-fidelity reference for analyzing evaluation instability rather than a practical evaluation protocol.

³An ideal MC-VQA evaluation yields rankings that are largely invariant to changes in prompt format.

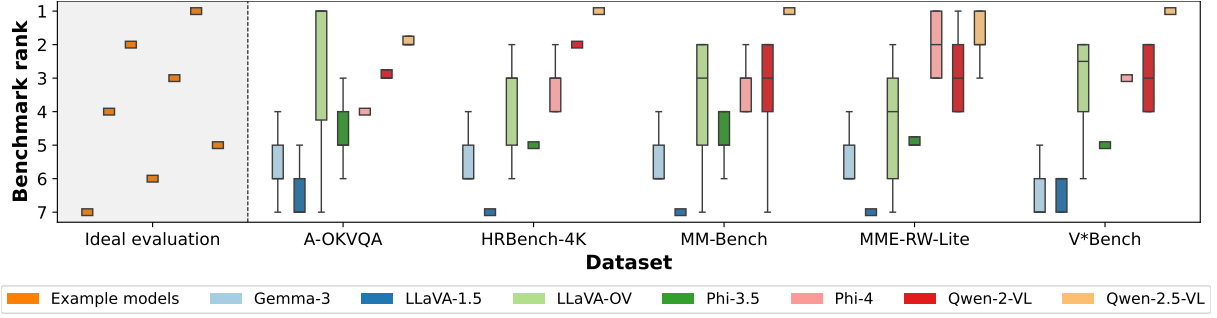


Figure 3: **Benchmarking results across five MC-VQA datasets.** The figure shows the distribution of ranks for seven MLLMs across 48 prompt format variants. The ranks vary substantially across datasets, indicating high sensitivity to prompt format. Compared to an ideal evaluation (left), where ranks remain constant, real-world evaluations exhibit significant rank variability: *highlighting that MC-VQA benchmarking is unreliable.*

such that no single model maintains a consistent position. This instability persists despite controlling for option-order effects through circular evaluation. On A-OKVQA and MME-RealWorld-Lite, rank variability is high, with no model exhibiting a stable rank. For example, LLaVA-OV ranges from top to near-bottom ranks on A-OKVQA depending on the prompt format. On MMBench and V*Bench, the top-ranked model remains stable, but lower-ranked models exhibit substantial fluctuations. In contrast, HRBench-4K shows the most stable rankings across prompt formats.

Additional analyses show that these effects are not explained by instruction-following failures or low-confidence examples, and persist under existing bias mitigation methods (see Sec. C). Additionally, we find that circular evaluation effectively mitigates option order preference. We further confirm that variations across all prompt format variations cause statistically significant accuracy changes (see Sec. B). Overall, these results show that MC-VQA rankings are sensitive to neutral prompt format variations, even under controlled evaluation settings.

4.3 Model Scaling Does Not Eliminate Prompt Sensitivity

The instability observed in Sec. 4.2 raises the question whether larger models exhibit more robust behavior under prompt format variations. To examine this, we extend our analysis on V*Bench to larger variants of Gemma-3 and Qwen2.5-VL, as well as proprietary models from the GPT-5.2 family.

Our analysis shows that increasing model size does not consistently reduce prompt sensitivity. Qwen2.5-VL-32B exhibits reduced variation compared to its 7B counterpart, with an accuracy deviation of 2.6 pp compared to 7.1 pp. However, the

smaller Qwen2.5-VL-7B model outperforms the 32B variant under certain prompt formats, indicating that prompt design can outweigh gains from model scaling. In contrast, Gemma-3-27B shows increased sensitivity, with 5.9 pp deviation compared to 3.9 pp for the smaller model. Variants of GPT-5.2 similarly display substantial prompt-induced variability in MC-VQA. Overall, these results show that increasing model scale does not guarantee stable evaluation behavior under prompt format variations.

Table 2: **Sensitivity of larger MLLMs on V*Bench.** Results show that increased model size does not consistently reduce accuracy variance across prompt formats. Further, we report which model size within each family achieves the highest accuracy for each prompt format.

Model	Avg. acc. [%]	Min. acc. [%]	Max. acc. [%]	Higher acc. per prompt format
Qwen2.5-VL-7B	77.44	72.65	81.71	7 out of 48
Qwen2.5-VL-32B	80.52	78.85	81.54	41 out of 48
Gemma-3-4B	34.41	32.31	36.25	0 out of 48
Gemma-3-27B	45.09	41.78	47.65	48 out of 48
GPT-5.2-nano	59.70	57.21	61.41	0 out of 48
GPT-5.2-mini	70.05	67.61	71.81	48 out of 48

5 Mechanistic Sources of Instability

Sec. 4 shows that semantically neutral prompt format choices induce rank reversals in MC-VQA. We now ask why these changes affect model predictions so strongly. Our hypothesis is that prompt format sensitivity arises from low-level language-model behavior rather than changes in reasoning. In particular, we trace the observed benchmark instability to three interacting factors: tokenization of option IDs, the ability of models to generalize beyond familiar option ID sets, and attention patterns to option IDs during answer generation.

5.1 Tokenization Effects on Option IDs

One potential source of instability is how tokenizers represent option IDs in the input sequence. Across prompt formats, we observe two recurring patterns: (1) the option ID-delimiter pair tokenizes into more than two tokens, and (2) the tokenizer fails to preserve the option ID as a standalone token. Both effects alter the input representation in ways that are unrelated to the task semantics and may reduce evaluation stability. Therefore, we analyze whether atypical tokenization under varying prompt formats causes the observed benchmark instability. We provide more details in Sec. D.

Token count of option ID-delimiter. First, we analyze how many tokens each tokenizer assigns to the option ID-delimiter pair across prompt formats. While some models exhibit accuracy differences when these tokenized pairs expand beyond two tokens, the overall effect is minor and inconsistent across MLLMs. For more details, see Sec. D.2.

Option ID atomic tokenization. Next, we examine whether the choice of option ID set and delimiter causes the tokenizer to fuse them into a corrupted token, preventing the option ID from appearing as a standalone token. For example, a tokenizer may split "(A)" into "(A" and ")", such that the token "A" never appears. Therefore, we first identify which option ID-delimiter combinations

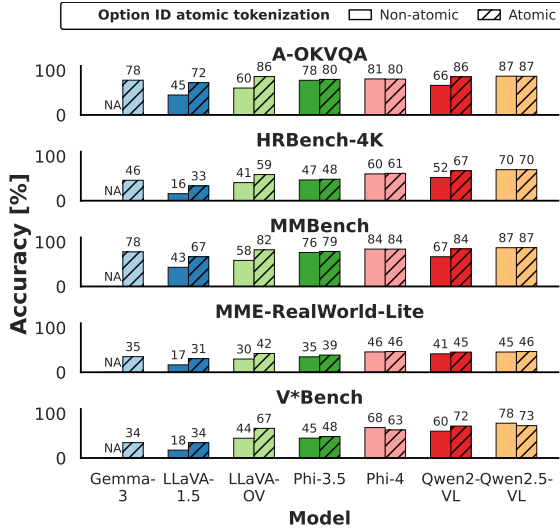


Figure 4: **Effect of atomic and non-atomic option ID tokenization.** We group prompt formats based on whether they preserve or remove the bare option ID token. Models whose tokenizers fail to preserve standalone option IDs often show accuracy decrease when the bare option ID token is absent. This result links prompt format sensitivity directly to tokenization rather than to changes in task semantics.

induce this tokenization pattern and then group prompt formats according to whether the pattern occurs for any option ID. Interestingly, this phenomenon does not occur for the Gemma-3, which consistently preserves atomic option IDs under all evaluated formats. Correspondingly, Gemma-3 exhibits the lowest accuracy variation in our study. In contrast, the absence of a bare option ID token has a pronounced impact for the other models. We observe a significant accuracy decrease for LLaVA-1.5, LLaVA-OV, and Qwen2-VL in cases where the bare option ID token is missing, indicating a strong dependence on explicit option ID representation in the input. In contrast, accuracy is stable for Phi-4 and Qwen2.5-VL when the bare option ID token is not present, suggesting that these models rely less on isolated option ID tokens and can operate on fused representations. A key finding of this analysis is that a well-designed tokenizer (e.g., GemmaTokenizer) enables consistent encoding behavior for multiple-choice prompts independent of prompt format choices. This supports the claim that robust tokenizer design can substantially mitigate the impact of prompt format variations. The results of this analysis are essential as the widely used MC-VQA benchmarks MMMU and MMMU-Pro rely on the "(A)" option format (Yue et al., 2024, 2025).

Table 3: **Impact of non-atomic option ID tokenization on embeddings.** We compare the cosine similarity between fused and original option ID tokens. Models that are sensitive to missing bare tokens exhibit low representational similarity, indicating that fused tokens fail to preserve ID identity and lead to accuracy deviations.

Model	Negative effect?	Average cos. sim.	Avg. acc. deviation [%]
Phi-3.5	✓	0.0110	−5.0
LLaVA-1.5	✓	0.0341	−43.4
Qwen2-VL	✓	0.2617	−18.3
LLaVA-OV	✓	0.2672	−30.6
Qwen2.5-VL	—	0.2390	+0.9
Phi-4	—	0.6139	+4.2
Gemma-3	—	—	—

To further characterize this effect, we compare the representations of bare option ID tokens and their fused counterparts in input embedding space. Models that show a negative accuracy effect from missing bare tokens consistently exhibit low cosine similarity between the corrupted and bare option-ID embeddings (see Tab. 3). This indicates that the fused tokens occupy a distant region of embedding space

and cannot reliably substitute for the standalone option ID. For models without a negative effect, we observe that Phi-4 displays markedly higher cosine similarity, suggesting that option-ID identity is better preserved despite fusion, whereas Qwen2.5-VL shows low similarity while remaining unaffected. For models with low cosine similarity already at the embedding, correct multiple-choice behavior requires the model to internally remap corrupted option ID tokens to their intended identities during decoding. This additional representational correction might increase sensitivity to prompt formatting and prior work has shown that such early-stage tokenization differences can substantially affect multiple-choice evaluation outcomes, even when task semantics remain unchanged (Mielke et al., 2021; Sanz-Guerrero et al., 2025). More generally, analyses of tokenizer choice show that tokenizer design can substantially affect downstream model robustness and behavior (Altıntaş et al., 2026).

5.2 Option ID Performance Cannot Be Explained Solely by Training Exposure

Tokenization explains part of the observed instability, but it does not explain why some option ID sets consistently perform better even when all IDs tokenize cleanly. We next test whether this preference is simply a consequence of training exposure.

A hypothesis might be that option ID sensitivity reflects training exposure: models may simply perform better on frequently seen IDs (e.g., uppercase letters) than on less common ones. To probe this, we evaluate Phi-4 and Qwen2.5-VL on V*Bench using additional option ID sets that are unlikely to occur in multiple-choice tasks during training. We introduce three sets with increasing semantic grounding: (i) Greek letter words, (ii) directional

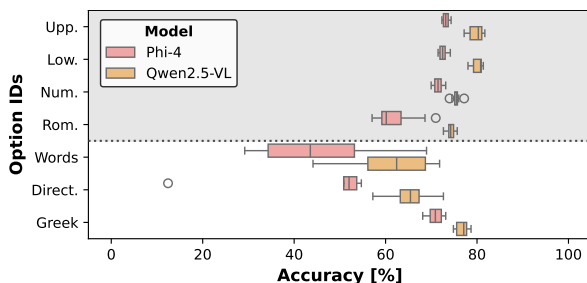


Figure 5: **Accuracy for alternative option IDs.** We evaluate additional option ID sets that differ in training exposure and semantic grounding. Greek-letter IDs perform comparably to letter-based IDs and outperform numbers and Roman numerals. It shows that ID preferences cannot be explained by training exposure alone.

words, and (iii) concrete nouns (see Tab. 22). We mitigate output-length effects on accuracy deviations by ensuring that each option ID maps to a single token in both models. Across both models, Greek-letter IDs outperform numbers and Roman numerals and are comparable to letter-based option IDs (see Fig. 5). This pattern suggests that option ID performance is not solely a consequence of exposure during training. Instead, MLLMs can generalize multiple-choice selection to other IDs.

5.3 Attention Heads Can Be More Discriminative for Certain Option IDs

Sec. 5.2 shows that some option ID sets yield higher MC-VQA accuracy independent of training exposure. We hypothesize that these differences arise from how decisively models attend to candidate option IDs during decoding.

In (M)LLMs, attention heads can specialize in distinct behaviors, e.g., select-and-copy behavior (Ols-son et al., 2022; Bietti et al., 2023; D’Angelo et al., 2025). In multiple-choice settings, this mechanism is relevant because the model only needs to reproduce an option ID rather than generate a free-form answer (Tulchinskii et al., 2024). We analyze this behavior by tracking attention to option ID tokens

Table 4: **Attention margin to option IDs correlates with MC-VQA accuracy.** We compare the cumulative attention margin of attention heads with MC-VQA accuracy across option ID sets. Higher margins consistently correlate with higher accuracy, indicating that multiple-choice selection depends on decisive attention. Greek IDs ($\dagger$) are included for an extended analysis.

A-OKVQA		
Option IDs	Attention Margin (Avg.)	MC-VQA Acc. (Avg.) [%] $\uparrow$
Lowercase	0.609	87.51
Uppercase	0.596	87.64
Numbers	0.526	85.73
Roman	0.509	86.19
Pearson ρ		0.930
V*Bench		
Option IDs	Attention Margin (Avg.)	MC-VQA Acc. (Avg.) [%] $\uparrow$
Lowercase	0.651	80.10
Uppercase	0.648	79.58
Greek $\dagger$	0.620	76.69
Numbers	0.543	75.45
Roman	0.499	74.34
Pearson ρ		0.934

across varying ID sets on A-OKVQA and V*Bench. To isolate the effect of ID set variation, we fix the option delimiter and separator. We measure the decisiveness of option selection using the margin between the top-1 and top-2 attention weights to option IDs. Details are provided in Sec. F.2.

We identify a set of attention heads on V*Bench that consistently exhibit strong attention to option IDs and function as copy-style heads. To test whether this behavior generalizes beyond the dataset used for attention head selection, we apply the same heads to A-OKVQA.

Across both datasets, option ID sets with higher MC-VQA accuracy exhibit larger cumulative attention margins across these heads (see Tab. 4). For V*Bench, we additionally include Greek option IDs (see Sec. 5.2) to test whether the attention patterns generalize to uncommon ID sets. MC-VQA accuracy correlates strongly with the copy-attention margin across ID sets ($\rho = 0.930$ on A-OKVQA and $\rho = 0.934$ on V*Bench including Greek option IDs). This result indicates that the effect is not specific to a single dataset. Instead, MC-VQA accuracy depends in part on how decisively the model selects and reliably reproduces an option ID during generation. These findings show that prompt-format sensitivity arises not only from tokenization effects, but also from decoding behavior. The same copy-style attention heads remain predictive of option ID performance across datasets, indicating that prompt formatting influences the reliability of the option-selection mechanism itself.

5.4 MLLMs Can Reason About Favorable Prompt Formats For Multiple-Choice

Previously, we showed that model-specific encoding and decoding behavior causes format sensitivity. We now test whether MLLMs can predict which prompt formats are more favorable for their own performance. Therefore, we present each model with the design space in Tab. 1 and ask it to select a preferred configuration for each formatting factor. Additional details are provided in Sec. F.3. We find that all MLLMs, except Phi-4, select similar or identical prompt formats, and that the selected formats consistently outperform random guessing⁴, achieving average ranks between 5.4 and 14.4 across datasets. We observe similar behavior for the additional option ID sets introduced in Sec. 5.2 (see Tab. 24). These results sug-

gest that prompt format preferences are structured and partially aligned with model-specific processing. However, they do not eliminate the need to control for prompt format effects in evaluation.

Table 5: **Model-selected prompt formats.** Each model selects a prompt format from Tab. 1. Most models choose format with lower average ranks, which suggests that prompt format preferences are recoverable from MLLMs. Lower average rank indicates better performance. We further report the relative accuracy difference between each selected prompt format and the best-performing prompt format across all datasets.

Model	Option IDs	Delimiter	Separator	Avg. Rank	Rel. gap to optimal
Gemma-3	Uppercase	Colon	Linebreak	14.4	-1.61
LLaVA-1.5	Uppercase	Colon	Linebreak	7.0	-3.12
LLaVA-OV	Uppercase	Bracket	Linebreak	5.4	-0.79
Phi-3.5	Uppercase	Colon	Linebreak	9.6	-1.42
Phi-4	Numbers	Dot	Comma	34.0	-4.21
Qwen2-VL	Uppercase	Colon	Linebreak	11.6	-1.08
Qwen2.5-VL	Uppercase	Colon	Linebreak	7.4	-0.92

5.5 Multiple-Choice VQA May Not Fully Reflect Open-Ended MLLM Performance

The previous analyses show that MC-VQA results depend strongly on how models encode and decode option IDs. We now test the broader implication of this result: whether MC-VQA remains a reliable proxy for open-ended MLLM performance. Therefore, we compare multiple-choice and open-ended evaluation (see Fig. 6). For open-ended evaluation, we remove answer options and prompt models to generate free-form responses (see Sec. E.1). We assess correctness using an LLM-as-a-judge framework (see Sec. E.2). Across all datasets, rankings correlate weakly to moderately between the two settings: Kendall’s τ is 0.09 on A-OKVQA, 0.52 on HRBench-4K, and 0.67 on V*Bench. To further analyze option selection effects, we

	Dataset								
	A-OKVQA $\tau = 0.09$			HRBench-4K $\tau = 0.52$			V*Bench $\tau = 0.67$		
Model	MC-VQA	OE	MC-VQA [†]	MC-VQA	OE	MC-VQA [†]	MC-VQA	OE	MC-VQA [†]
Gemma-3	6	6	6	6	4	5	6	7	6
LLaVA-1.5	7	5	7	7	6	7	7	6	7
LLaVA-OV	3	4	3	4	7	4	4	4	4
Phi-3.5	5	3	5	5	5	6	5	5	5
Phi-4	4	1	4	3	3	3	2	3	3
Qwen2-VL	2	2	2	2	1	2	3	1	2
Qwen2.5-VL	1	7	1	1	2	1	1	2	1

Figure 6: **Open-ended and grounded multiple-choice evaluation.** We compare MC-VQA, open-ended (OE), and grounded multiple-choice (MC-VQA[†]). Kendall’s τ indicates weak agreement between MC-VQA and OE, and MC-VQA[†] does not recover the original ordering.

⁴Randomly selecting one of the 48 prompt formats yields an expected prompt-format rank of 24.

Table 6: **Prompt-format transfer as a simple bias mitigation strategy.** We select the best prompt format on a small validation benchmark (HRBench-4K or V*Bench) per MLLM and apply the same format to all other benchmarks. While this strategy often reduces severe accuracy drops, the selected format does not transfer reliably across datasets and fails to consistently improve accuracy. Values denote the accuracy difference in pp relative to each dataset’s standard format. [†] denotes the benchmark used for prompt-format selection.

Model	Accuracy difference to the dataset-specific standard prompt format [pp.]							
	Prompt format selection dataset: HRBench-4K [†]				Prompt format selection dataset: V*Bench [†]			
	A-OKVQA	MMBench	MME-RW	V*Bench	A-OKVQA	HRBench	MMBench	MME-RW
Gemma-3	−0.57	0.14	−0.80	−0.17	−0.68	0.63	0.39	−0.92
LLaVA-1.5	65.17	0.00	−1.13	−1.01	59.89	−3.63	−3.26	−3.42
LLaVA-OV	48.21	0.33	−0.27	0.00	48.56	0.25	0.70	0.25
Phi-3.5	2.38	2.73	0.00	−1.01	2.34	0.00	2.87	0.89
Phi-4	−0.07	0.80	−0.28	−0.50	0.02	0.00	1.03	0.07
Qwen-2-VL	40.96	0.74	1.42	−2.68	41.09	−0.13	1.62	0.91
Qwen-2.5-VL	−0.15	0.00	0.97	−1.51	0.00	0.00	−0.21	−0.38
Avg.	22.28	0.68	−0.01	−0.98	21.60	−0.41	0.45	−0.37

perform *grounded multiple-choice* evaluation by providing each model its open-ended answer (see Sec. E.3). This setup provides context while preserving the multiple-choice setting. Grounding improves MC-VQA accuracy for most models compared to open-ended, but does not consistently restore original rankings. More importantly, it does not eliminate prompt sensitivity across prompt format variations (see Tab. 21). This indicates that multiple-choice selection remains influenced even when high-level reasoning signals are available.

Our results suggest that MC-VQA captures both reasoning ability and option selection behavior. The agreement between multiple-choice and open-ended rankings varies substantially across datasets, ranging from minor correlation on A-OKVQA ($\tau = 0.09$) to moderate correlation on HRBench-4K ($\tau = 0.52$) and V*Bench ($\tau = 0.67$). Consequently, MC-VQA should be interpreted as a distinct evaluation setting rather than as a direct measure of open-ended MLLM performance.

5.6 Practical Recommendations

Based on our findings, we want to provide several practical recommendations for MC-VQA evaluation. First, MC-VQA dataset creators should explicitly report the choice of the prompt format (i.e., option IDs, delimiter and separator) for the evaluation on the respective dataset. Second, benchmark results should ideally be complemented with open-ended evaluation, since MC-VQA reflects both reasoning and option-selection behavior. Third, practitioners should verify that prompt formats preserve standalone option ID tokens, as non-atomic tokenization can cause substantial accuracy degradation for some models. Finally, we evaluate a sim-

ple mitigation strategy that selects a model-specific prompt format on a small validation set and reuses it during evaluation (see Tab. 6). This approach reduces extreme prompt-format failures and can improve performance relative to standard formats. However, gains are inconsistent across datasets and models, and the method still requires evaluating multiple prompt formats on a validation set. We therefore view validation-based prompt format selection as a practical mitigation rather than a complete solution to prompt-format sensitivity.

6 Conclusion

We show that multiple-choice VQA is sensitive to semantically neutral prompt format choices. Variations in option IDs, delimiters, and separators induce large accuracy shifts and frequent rank reversals, even under order-invariant evaluation. We trace this instability to low-level language-model behavior: tokenization can fuse option ID tokens under certain prompt formats, and different option ID sets activate specialized attention heads to varying degrees. Interestingly, the performance differences across option ID sets are not solely explained by training exposure. Further, we show that models can also identify prompt formats that improve performance. Finally, the weak to moderate correlation with open-ended evaluation indicates that multiple-choice accuracy is not a reliable proxy for general MLLM capability.

Limitations

Our work has the following limitations: (i) *Evaluation setting*: We focus on zero-shot MC-VQA with deterministic decoding to enable controlled

comparisons. Results may differ under few-shot prompting, or stochastic decoding, where sampling may interact differently with prompt format. (ii) *Evaluation methodology*: Our open-ended evaluation relies on an LLM-as-a-judge framework, which may introduce bias. We do not systematically validate this protocol across all datasets. In grounded MC-VQA, models also receive unfiltered generated answers, which may introduce noise when incorrect responses are provided. Further, LLMs from the same model family as some of the tested MLLMs are used to generate the open-ended answer and for the LLM-as-a-judge correctness assertion. (iii) *Mitigation*: We show that models can partially reason effective prompt formats, revealing that prompt format preferences may be structured and accessible to the models. Further, selecting prompts on a small validation set can reduce extreme performance drops, indicating limited cross-dataset generalization. However, it does not consistently improve performance, and remains computationally expensive due to the large number of prompt formats to evaluate. These findings provide a concrete direction for developing more robust evaluation protocols in future work.

Acknowledgments

During the preparation of this paper, the first author’s father passed away. The authors would like to honor his memory and express profound gratitude for his support, encouragement, and belief in the value of education and scientific inquiry.

References

- Marah I Abidin, Jyoti Aneja, Harkirat S. Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J. Hewett, Mojan Javaheripi, Piero Kauffmann, James R. Lee, Yin Tat Lee, Yuanzhi Li, Weishung Liu, Caio C. T. Mendes, Anh Nguyen, Eric Price, Gustavo de Rosa, Olli Saarikivi, and 8 others. 2024a. [Phi-4 technical report](#). *CoRR*, abs/2412.08905.
- Marah I Abidin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat S. Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Caio César Teodoro Mendes, Weizhu Chen, Vishrav Chaudhary, Parul Chopra, and 68 others. 2024b. [Phi-3 technical report: A highly capable language model locally on your phone](#). *CoRR*, abs/2404.14219.
- Gül Sena Altıntaş, Malikeh Ehghaghi, Brian Lester, Fengyuan Liu, Wanru Zhao, Marco Ciccone, and Colin Raffel. 2026. [Toksuite: Measuring the impact of tokenizer choice on language model behavior](#). In *Forty-third International Conference on Machine Learning*.
- Norah A. Alzahrani, Hisham Abdullah Alyahya, Yazeed Alnumay, Sultan Alrashed, Shaykhah Alsubaie, Yousef Almushayqih, Faisal Mirza, Nouf Alotaibi, Nora Al-Twairesh, Areeb Alowisheq, M. Saiful Bari, and Haidar Khan. 2024. [When benchmarks are targets: Revealing the sensitivity of large language model leaderboards](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, *ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 13787–13805. Association for Computational Linguistics.
- Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. [VQA: visual question answering](#). In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 2425–2433. IEEE Computer Society.
- Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibong Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Ming-Hsuan Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. [Qwen2.5-vl technical report](#). *CoRR*, abs/2502.13923.
- Yoshua Bengio, Réjean Ducharme, and Pascal Vincent. 2000. [A neural probabilistic language model](#). In *Advances in Neural Information Processing Systems*, volume 13. MIT Press.
- Alberto Bietti, Vivien Cabannes, Diane Bouchacourt, Hervé Jégou, and Léon Bottou. 2023. [Birth of a transformer: A memory viewpoint](#). In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.
- Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. [Language models are few-shot learners](#). In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.
- Liang Chen, Li Shen, Yang Deng, Xiaoyan Zhao, Bin Liang, and Kam-Fai Wong. 2025. [PEARL: towards permutation-resilient llms](#). In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

- Francesco D'Angelo, Francesco Croce, and Nicolas Flammarion. 2025. [Selective induction heads: How transformers select causal structures in context](#). In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.
- Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Jiaqi Wang, and 1 others. 2024. Vlmevalkit: An open-source toolkit for evaluating large multi-modality models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 11198–11201.
- Beni Egressy and Jan Stühmer. 2025. [Set-llm: A permutation-invariant LLM](#). In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025*.
- Maria Eriksson, Erasmo Purificato, Arman Noroozian, João Vinagre, Guillaume Chaslot, Emilia Gómez, and David Fernández Llorca. 2025. [Can we trust AI benchmarks? an interdisciplinary review of current issues in AI evaluation](#). In *Proceedings of the Eighth AAAI/ACM Conference on AI, Ethics, and Society, AIES 2025, Madrid, Spain, October 20-22, 2025, Main Track I*, pages 850–864. AAAI Press.
- Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. [Making the V in VQA matter: Elevating the role of image understanding in visual question answering](#). In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 6325–6334. IEEE Computer Society.
- Jia He, Mukund Rungta, David Koleczek, Arshdeep Sekhon, Franklin X. Wang, and Sadid Hasan. 2024. [Does prompt formatting have any impact on LLM performance?](#) *CoRR*, abs/2411.10541.
- Jiming Jiang and Thuan Nguyen. 2021. [Linear and Generalized Linear Mixed Models and Their Applications](#), 2nd edition. Springer Series in Statistics. Springer.
- Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean-Bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, and 191 others. 2025. [Gemma 3 technical report](#). *CoRR*, abs/2503.19786.
- Ao Li, Zongfang Liu, Xinhua Li, Jinghui Zhang, Pengwei Wang, and Hu Wang. 2026. [Mvp: Modeling variants of prompts for vision-language models](#). In *ICASSP 2026 - 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8407–8411.
- Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. 2025. [Llava-onevision: Easy visual task transfer](#). *Trans. Mach. Learn. Res.*, 2025.
- Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. 2023. [BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models](#). In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 19730–19742. PMLR.
- Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. 2022. [BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation](#). In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 12888–12900. PMLR.
- Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yan Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Alexander Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew Arad Hudson, and 13 others. 2023. [Holistic evaluation of language models](#). *Transactions on Machine Learning Research*.
- Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024a. [Improved baselines with visual instruction tuning](#). In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 26286–26296. IEEE.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. [Visual instruction tuning](#). In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.
- Yesheng Liu, Hao Li, Haiyu Xu, Baoqi Pei, Jiahao Wang, Mingxuan Zhao, Jing-Shu Zheng, Zheqi He, JG Yao, Xi Yang, Bowen Qin, and Jiajun Zhang. 2026. Beyond multiple choice: Verifiable openqa for robust vision-language rft. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18859–18869.
- 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. 2024b. [Mmbench: Is your multi-modal model an all-around player?](#) In *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part VI*, volume 15064 of *Lecture Notes in Computer Science*, pages 216–233. Springer.

- Rachel Longjohn, Markelle Kelly, Sameer Singh, and Padhraic Smyth. 2024. [Benchmark data repositories for better benchmarking](#). In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.
- Ana-Maria Marcu, Long Chen, Jan Hünemann, Alice Karnsund, Benoît Hanotte, Prajwal Chidananda, Saurabh Nair, Vijay Badrinarayanan, Alex Kendall, Jamie Shotton, Elahe Arani, and Oleg Sinavski. 2024. [Lingoqa: Visual question answering for autonomous driving](#). In *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part LXXVII*, volume 15135 of *Lecture Notes in Computer Science*, pages 252–269. Springer.
- João Matos, Shan Chen, Siena Placino, Yingya Li, Juan Carlos Climent Pardo, Daphna Idan, Takeshi Tohyama, David S. Restrepo, Luis Filipe Nakayama, Jose M. M. Pascual-Leone, Guergana K. Savova, Hugo J. W. L. Aerts, Leo Anthony Celi, An-Kwok Ian Wong, Danielle S. Bitterman, and Jack Gallifant. 2025. [Worldmedqa-v: a multilingual, multimodal medical examination dataset for multimodal language models evaluation](#). In *Findings of the Association for Computational Linguistics: NAACL 2025, Albuquerque, New Mexico, USA, April 29 - May 4, 2025*, pages 7203–7216. Association for Computational Linguistics.
- Sabrina J. Mielke, Zaid Alyafeai, Elizabeth Salesky, Colin Raffel, Manan Dey, Matthias Gallé, Arun Raja, Chenglei Si, Wilson Y. Lee, Benoît Sagot, and Samson Tan. 2021. [Between words and characters: A brief history of open-vocabulary modeling and tokenization in NLP](#). *CoRR*, abs/2112.10508.
- Moran Mizrahi, Guy Kaplan, Dan Malkin, Rotem Dror, Dafna Shahaf, and Gabriel Stanovsky. 2024. [State of what art? A call for multi-prompt LLM evaluation](#). *Trans. Assoc. Comput. Linguistics*, 12:933–949.
- Francesco Maria Molfese, Luca Moroni, Luca Gioffrè, Alessandro Scirè, Simone Conia, and Roberto Navigli. 2025. [Right answer, wrong score: Uncovering the inconsistencies of LLM evaluation in multiple-choice question answering](#). In *Findings of the Association for Computational Linguistics, ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 18477–18494. Association for Computational Linguistics.
- Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, and 7 others. 2022. [In-context learning and induction heads](#). *CoRR*, abs/2209.11895.
- 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, Mexico City, Mexico, June 16-21, 2024*, pages 2006–2017. Association for Computational Linguistics.
- Felipe Maia Polo, Ronald Xu, Lucas Weber, Mírian Silva, Onkar Bhardwaj, Leshem Choshen, Allysson Flavio Melo de Oliveira, Yuekai Sun, and Mikhail Yurochkin. 2024. [Efficient multi-prompt evaluation of llms](#). In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.
- Narun K. Raman, Taylor Lundy, and Kevin Leyton-Brown. 2025. Reasoning models are test exploiters: Rethinking multiple choice. *arXiv preprint arXiv:2507.15337*.
- Anka Reuel-Lamparth, Amelia F. Hardy, Chandler Smith, Max Lamparth, Malcolm Hardy, and Mykel J. Kochenderfer. 2024. [Betterbench: Assessing AI benchmarks, uncovering issues, and establishing best practices](#). In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.
- Joshua Robinson and David Wingate. 2023. [Leveraging large language models for multiple choice question answering](#). In *The Eleventh International Conference on Learning Representations*.
- Mario Sanz-Guerrero, Minh Duc Bui, and Katharina von der Wense. 2025. [Mind the gap: A closer look at tokenization for multiple-choice question answering with LLMs](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 19573–19583, Suzhou, China. Association for Computational Linguistics.
- Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. 2022. [A-OKVQA: A benchmark for visual question answering using world knowledge](#). In *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part VIII*, volume 13668 of *Lecture Notes in Computer Science*, pages 146–162. Springer.
- Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. 2024. [Quantifying language models’ sensitivity to spurious features in prompt design or: How I learned to start worrying about prompt formatting](#). In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.
- Shivalika Singh, Yiyang Nan, Alex Wang, Daniel D’souza, Sayash Kapoor, Ahmet Üstün, Sanmi Koyejo, Yuntian Deng, Shayne Longpre, Noah A. Smith, Beyza Ermis, Marzieh Fadaee, and Sara

- Hooker. 2025. [The leaderboard illusion](#). In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025*.
- Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, and 1 others. 2023. [Beyond the imitation game: Quantifying and extrapolating the capabilities of language models](#). *Transactions on Machine Learning Research*.
- Jiuding Sun, Chantal Shaib, and Byron C. Wallace. 2024. [Evaluating the zero-shot robustness of instruction-tuned language models](#). In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.
- Zhijie Tan, Xu Chu, Weiping Li, and Tong Mo. 2024. [Order matters: Exploring order sensitivity in multimodal large language models](#). Preprint, arXiv:2410.16983.
- Eduard Tulchinskii, Laida Kushnareva, Kristian Kuznetsov, Anastasia Voznyuk, Andrei Andriianin, Irina Piontkovskaya, Evgeny Burnaev, and Serguei Barannikov. 2024. [Listening to the wise few: Select-and-copy attention heads for multiple-choice QA](#). *CoRR*, abs/2410.02343.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. [Attention is all you need](#). In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.
- Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024. [Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution](#). *CoRR*, abs/2409.12191.
- Wenbin Wang, Liang Ding, Minyan Zeng, Xiabin Zhou, Li Shen, Yong Luo, Wei Yu, and Dacheng Tao. 2025. [Divide, conquer and combine: A training-free framework for high-resolution image perception in multimodal large language models](#). In *AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence, February 25 - March 4, 2025, Philadelphia, PA, USA*, pages 7907–7915. AAAI Press.
- Jason Wei, Clara Meister, and Ryan Cotterell. 2021. [A cognitive regularizer for language modeling](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 5191–5202. Association for Computational Linguistics.
- 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, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, Findings of ACL, pages 5598–5621. Association for Computational Linguistics.
- Penghao Wu and Saining Xie. 2024. [V*: Guided visual search as a core mechanism in multimodal llms](#). In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 13084–13094. IEEE.
- Zhen Yang, Ping Jian, and Chengzhi Li. 2025. [Option symbol matters: Investigating and mitigating multiple-choice option symbol bias of large language models](#). In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1902–1917, Albuquerque, New Mexico. Association for Computational Linguistics.
- Xiang Yue, Yuansheng Ni, Tianyu Zheng, Kai Zhang, 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, and 3 others. 2024. [MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI](#). In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 9556–9567. IEEE.
- Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, Yu Su, Wenhao Chen, and Graham Neubig. 2025. [Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 15134–15186. Association for Computational Linguistics.
- Yifan Zhang, Huanyu Zhang, Haochen Tian, Chaoyou Fu, Shuangqing Zhang, Junfei Wu, Feng Li, Kun Wang, Qingsong Wen, Zhang Zhang, Liang Wang, and Rong Jin. 2025. [Mme-realworld: Could your multimodal LLM challenge high-resolution real-world scenarios that are difficult for humans?](#) In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.
- 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*.
- Liangyu Zhong, Fabio Rosenthal, Joachim Sicking, Fabian Huger, Thorsten Bagdonat, Hanno Gottschalk,

and Leo Schwinn. 2025. [FOCUS: internal MLLM representations for efficient fine-grained visual question answering](#). In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025*.

Jingming Zhuo, Songyang Zhang, Xinyu Fang, Haodong Duan, Dahua Lin, and Kai Chen. 2024. [ProSA: Assessing and understanding the prompt sensitivity of large language models](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 1950–1976, Miami, Florida, USA. Association for Computational Linguistics.

Appendix A Experimental Setup

In this section, we provide additional technical details on the following components: instruction prompts (Sec. A.1), evaluation settings (Sec. A.2), metrics (Sec. A.3), models and datasets (Sec. A.4), the computation report (Sec. A.5), and bias mitigation methods (Sec. A.6).

A.1 Instruction Prompts

As shown in Fig. 1, instruction prompts are essential for prompting in MC-VQA tasks. They guide MLLMs to output only the corresponding option ID for the selected option (Duan et al., 2024). In practice, many datasets use simple instruction prompts (Wu and Xie, 2024; Wang et al., 2025; Schwenk et al., 2022; Zhang et al., 2025; Duan et al., 2024); see the *Reference* instruction prompt from V*Bench in Tab. 7. Vague instruction prompts, such as failing to specify that the MLLM should return the option ID rather than the option text, can lead to poor instruction-following of MLLMs. This can lead to poor instruction-following and consequently lower accuracy, not due to reasoning errors but simply because of inadequate prompt design. Therefore, we use an optimized instruction prompt that clearly specifies what the MLLMs should generate and what it should avoid in all experiments of our comprehensive analysis. Additionally, the prompt includes the exact option IDs that the models are expected to generate (see examples in Tab. 7).

A.2 Evaluation Settings

We evaluate all models in a fully deterministic setting by disabling sampling. Specifically, we set `do_sample=False` and $k = 0$ during generation for each MLLM. This ensures that each model produces identical outputs for identical inputs. Before evaluating an MLLM on a given prompt format, we iterate over all option IDs required by the dataset. We tokenize each option ID and set the number of generated tokens to the maximum required to generate all option IDs. Furthermore, we evaluate the seven MLLMs across all MC-VQA datasets in a zero-shot setting, i.e. without providing any in-context learning examples, as is standard for this task (Duan et al., 2024). Further, we use Flash-Attention for all model evaluations besides the attention head analysis.

A.3 Metrics

Accuracy. The most commonly used metric for evaluating the reasoning ability of MLLMs in Visual Question Answering (VQA) is *accuracy* (Goyal et al., 2017; Antol et al., 2015). The accuracy is defined as the ratio of correctly answered questions to the total number of questions in the evaluation set (Goyal et al., 2017; Antol et al., 2015). The correctness assessment is performed directly on the model’s generated output sequence without applying any answer extraction mechanism (e.g., regular expressions). An answer is considered correct only if it exactly matches the ground-truth

Table 7: **Instruction prompt used across all MC-VQA datasets.** We employ this optimized instruction prompt to mitigate poor instruction-following behavior. Additionally, we include the reference prompt from V*Bench for comparison. Finally, we demonstrate the adaptability of our prompt to different question formats, accommodating variations in the number and sets of option IDs.

Option IDs	Instruction prompt
<i>Reference</i> (V*Bench (Duan et al., 2024))	<i>Please select the correct answer from the options above.</i>
Template	Select the best answer to the above multiple-choice question based on the image. Respond with only the letter (e.g., {opt}, {opt}, {opt} or {opt}) of the correct option and no bracket, colon, or dot.
A/B/C/D	Select the best answer to the above multiple-choice question based on the image. Respond with only the letter (e.g., A, B, C or D) of the correct option and no bracket, colon, or dot.
A/B	Select the best answer to the above multiple-choice question based on the image. Respond with only the letter (e.g., A, or B) of the correct option and no bracket, colon, or dot.

Table 8: **Model overview.** We report the specific versions used of each MLLM for reproducibility reasons of this work.

Model	Year	HuggingFace model name	LLM size [B]
Gemma-3	2025	<i>google/gemma-3-4b-it</i>	4
LLaVA-1.5	2023	<i>llava-hf/llava-1.5-7b-hf</i>	7
LLaVA-OV	2024	<i>lmms-lab/llava-onevision-qwen2-7b-ov</i>	7
Phi-3.5	2024	<i>microsoft/Phi-3.5-vision-instruct</i>	3.8
Phi-4	2025	<i>microsoft/Phi-4-multimodal-instruct</i>	3.8
Qwen2-VL	2024	<i>Qwen/Qwen2-VL-7B-Instruct</i>	7
Qwen2.5-VL	2025	<i>Qwen/Qwen2.5-VL-7B-Instruct</i>	7

label. For instance, if the label is "A", then "A" is correct, whereas "A." or "A)" are not.

Coverage. To assess instruction-following capability, we introduce the metric *coverage*. Coverage is defined as the ratio of answers that belong to the valid option ID set to the total number of questions in the evaluation set:

$$\text{coverage} = \frac{s_i}{n} \quad (2)$$

where s_i denotes the number of answers within the option ID set and n is the total number of questions. A higher coverage value indicates better instruction-following capability.

A.4 Models and Datasets

Models. We evaluate a set of seven widely used MLLMs (see Tab. 8), each selected to have a comparable LLM parameter size. We consider three categories of models: (i) those that process a single downsized global view of the input image (LLaVA-1.5), (ii) those that combine a global view with additional high-resolution local crops (Gemma-3, LLaVA-OV, Phi-3.5, and Phi-4), and (iii) those that operate on a high-resolution global view of the input image (Qwen-2-VL and Qwen-2.5-VL).

Datasets. For our comprehensive analysis, we employ five MC-VQA datasets that span a range of task difficulties and domains. Additional details about these datasets are provided in Tab. 9.

A.5 Computation Report

We provide an overview of the computational complexity involved in our comprehensive analysis of prompt format-induced biases in MC-VQA, including the number of questions and inference time, in Tab. 10.

A.6 Bias Mitigation Methods

We assess the effectiveness of several bias mitigation techniques in reducing prompt format-induced

biases. All methods are reimplemented according to the specifications provided in their papers.

Circular evaluation. The circular evaluation scheme shifts the options in the prompt by one position (Liu et al., 2024b) and thereby increases the number of questions - e.g., for four options it enlarges the dataset by a factor of four. This mitigates the option position bias. While the approach requires no retraining or calibration, it is computationally more expensive.

Position-invariant accuracy (PIA). PIA is a post-hoc metric that mitigates the influence of the option position bias by weighting the per-option accuracy by the number of times an MLLM selects a specific option (Tan et al., 2024):

$$\text{PIA} = \frac{1}{M} \sum_{i=1}^M \frac{C_i}{Pr_i} \frac{C_i}{N} \quad (3)$$

where M is the number of options, C_i is the number of times option i was correctly selected, Pr_i is the number of times option i was selected in total and N is the number of questions (Tan et al., 2024). This method does not require a calibration; however, it is prone to variations in the prompt format, as these changes may influence the selection frequency of MLLMs. Furthermore, it changes the accuracy calculation, which has two disadvantages: (i) this measure is less intuitive and not comparable to standard accuracy, and (ii) it can skew the result, e.g., when a model performs overall well but it does not match the dataset’s positional distribution or when a model answers one option only rarely.

Debiasing with Prior estimation (PriDe). PriDe estimates the prior for each option ID based on calibration data containing circularly shifted multiple-choice questions (Zheng et al., 2024). This prior is then used to debias the predictions of an MLLM, adjusting the probability of each option ID according to its corresponding prior:

$$P_{\text{debiased}}(o_i | q, x) \propto \frac{P_{\text{observed}}(d_i | q, x)}{P_{\text{prior}}(d_i)} \quad (4)$$

where o_i denotes the i -th option, q is the question, x represents the default ordering of option IDs, and d_i is the i -th option ID (Zheng et al., 2024). The answer is then selected as the option ID with the highest debiased probability (Zheng et al., 2024). For our evaluations with PriDe, we estimate this prior using the 4-option questions from V*Bench. These priors are only valid for questions with the same number of options as in the calibration data;

Table 9: **Statistics of the MC-VQA datasets.** The datasets vary significantly in the number of samples and image resolution. We report the number of question-answer (QA) pairs for both the single-prompt and circular evaluation schemes. Additionally, we indicate the prompt format variations that are standard for each dataset.

Dataset	No. options	QA pairs [#]		Image size [px]		Prompt format		
		Single	Circ. eval.	$\overline{H}$	$\overline{W}$	Option IDs	Option del.	Option sep.
A-OKVQA (Schwenk et al., 2022)	4	1,145	4,580	582	480	Low.	Double b.	Line break
HR-Bench-4k (Wang et al., 2025)	4	200	800	4,024	3,503	Upp.	Dot	Line break
MMBench-DEV (Liu et al., 2024b)	2 – 4	1,292	4,876	464	349	Upp.	Dot	Semicolon
MME-RW-Lite (Zhang et al., 2025)	5	1,919	9,595	2,836	1,566	Upp.	Double b.	Line break
V*Bench (Wu and Xie, 2024)	2 – 4	192	596	2,246	1,582	Upp.	Dot	Line break

therefore, they cannot be applied to datasets with varying numbers of options.

Cloze prompt (length normalized) (CP-LN).

An alternative approach to mitigating biases in MLLMs involves modifying the answer generation process. Rather than presenting all options to the model together with a predefined prompt format and instruction, this method calculates the perplexity of the question for each individual option. Perplexity serves as a metric for MLLMs, indicating how unexpected it is for the model to generate an output sequence y given the input token sequence $\mathbf{X}_{\text{in}}$ and the previously generated tokens $y_{<i}$ (Bengio et al., 2000; Brown et al., 2020; Wei et al., 2021):

$$\text{PPL}(\mathbf{X}_{\text{in}})^y = \left(\prod_i^n P(y_i | \mathbf{X}_{\text{in}}, y_{<i}) \right)^{-\frac{1}{n}} \quad (5)$$

where n is the number of tokens in the sequence. A lower perplexity value indicates a more likely prediction of the output token sequence than a higher one. Unlike the sum of log-probabilities, perplexity is normalized and thus less sensitive to the sequence length n . Given the set of perplexities $\text{PPL}^o(X)$ computed by the MLLM for input $\mathbf{X}_{\text{in}}$ conditioned on each option o , we select the predicted answer $\hat{Y}$ as the option with the lowest perplexity among all options O :

$$\hat{Y} = \arg \min_{o \in O} (\text{PPL}^o(\mathbf{X}_{\text{in}})) \quad (6)$$

This approach is closely related to the methods used in (Robinson and Wingate, 2023; Brown et al., 2020). It has been shown to be less effective than the standard multiple-choice evaluation scheme in terms of accuracy (Robinson and Wingate, 2023), while its impact on bias mitigation remains unclear.

Table 10: **Computation Report.** Computational complexity for the comprehensive analysis presented in Sec. 4.

Dataset	Question-level evaluations		Inference time estimates	
	Per prompt format variant	Total evaluations	Per question [s]	Total time [h]
A-OKVQA	4,580	1,538,880	0.50	205
HRBench-4K	800	268,800	2.00	149
MM-Bench	4,876	1,638,336	1.00	455
MME-RW-Lite	9,595	3,223,920	1.50	1,343
V*Bench	596	200,256	2.00	111
Sum	—	6,807,192	—	2,264

Appendix B Significance Analysis

We test whether prompt formats have a statistically significant effect on accuracy and, if so, which factors drive this effect.

B.1 Linear Mixed Models

For the significance analysis of prompt format variations, we employ a linear mixed model (LMM) to assess the impact of prompt format factors such as *option ID set*, *option delimiter*, and *option separator* on accuracy. Our comprehensive study spans multiple performance levels, including diverse datasets and MLLMs. Standard linear models (e.g., linear regression) are unsuitable for this non-independent data due to their independence assumption (Jiang and Nguyen, 2021). In contrast, LMMs handle such dependencies by modeling the target variable y as:

$$y = X\beta + Zu + \epsilon \quad (7)$$

where y is the vector of observations, β the fixed effects, and u the random effects (Jiang and Nguyen, 2021). The design matrices X and Z capture the relationships between y and β , and between y and u , respectively (Jiang and Nguyen, 2021). In our application, we want to model the accuracy Acc_{ijk} , therefore, our modeling design of the LMM is:

$$\begin{aligned} \text{Acc}_{ijk} = & \beta_0 + \beta_1 \text{OptionID}_i + \beta_2 \text{OptionsDelim}_j \\ & + \beta_3 \text{OptionsSep}_k + u_g + \epsilon_{ijk} \end{aligned} \quad (8)$$

where β_0 is the intercept, $\beta_1, \beta_2, \beta_3$ are fixed-effect coefficients for the prompt format factors, u_g is the random effect for group g (model-dataset), and ϵ_{ijk} is the residual error. This specification estimates the effects of prompt format variations on accuracy while accounting for non-independence across different models and datasets.

Specifically, we fit a LMM using statsmodels in Python to estimate the influence of prompt format variations on accuracy.

B.2 Significance Testing Details

We employ a linear mixed model (LMM) to investigate which biases have a statistical influence on the performance of MLLMs. LMMs model fixed effects (e.g., prompt format factors) and random effects (e.g., combined group of model and dataset), thereby yielding less biased and more generalizable results than traditional linear models (e.g., linear regression) (Jiang and Nguyen, 2021). We control

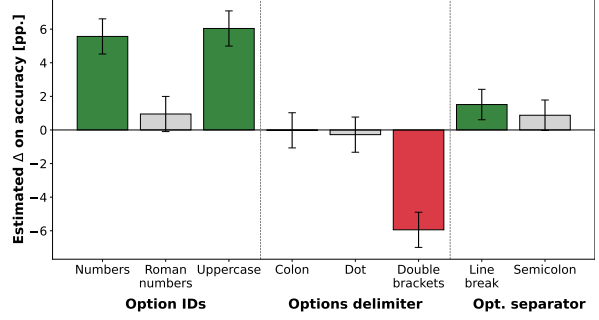


Figure 7: **Prompt formatting choices can have significant effects on MLLM accuracy, even after controlling for model and dataset differences.** Bars show the effect of each prompt format factor, with error bars indicating 95% confidence intervals. Significant effects are colored by their direction, and non-significant effects are gray. All effects are interpreted relative to the base configuration (*option IDs: lowercase, option delimiter: bracket, option separator: comma*).

for performance differences across MLLMs and datasets by incorporating random effects, acknowledging differences in model capabilities and dataset difficulty. The results of the LMM, shown in Fig. 7, reveal the estimated accuracy change in pp when varying each prompt format factor from a baseline configuration, which in our case consists of *lowercase* (option IDs), *bracket* (option delimiter), and *comma* (option separator). We find significant effects for all three prompt format factors. Changing the option IDs yields positive effects for *Uppercase* (+6 pp) and *Numbers* (+5 pp). Changing the delimiter to *Double brackets* results in a negative effect (−6 pp), while using a *Line break* as separator yields a small positive effect (+1 pp).

Appendix C Additional Results

This section provides additional results supporting the instability results in Sec. 4. We provide detailed results on instruction-following analysis in Sec. C.1, option selection patterns in Sec. C.2, bias mitigation results in Sec. C.3, and the full benchmark results in Sec. C.4.

C.1 Instruction-Following Analysis

As previously noted, a key factor affecting MLLM accuracy in MC-VQA tasks is the instruction-following capability of the models. We measure this as the *coverage* of given answers (see Sec. A.3). Low coverage reflects poor instruction-following, limiting accuracy because only part of the provided answers appear in the option ID set. We report the coverage results across all datasets in Fig. 8. The coverage trends are consistent across datasets, with minimal deviation. Most MLLMs maintain near-perfect coverage. Gemma-3 and Qwen-2.5-VL achieve nearly 100% coverage across all datasets and prompt formats. In contrast, LLaVA-1.5 and LLaVA-OV show slight degradation, with LLaVA-1.5 dropping to a minimum of 56% for lowercase option IDs on V*Bench. Overall, these results indicate that instruction-following limitations play only a minor role in the observed accuracy deviations across prompt format variations.

C.2 Option Selection Patterns

We report the answering frequency for each option position across five MC-VQA datasets, using a circular evaluation scheme to mitigate option-order bias, as previously noted. We present the

answer frequency results for A-OKVQA in Fig. 9, HRBench-4K in Fig. 10, MM-Bench in Fig. 11, MME-RealWorld-Lite in Fig. 12 and V*Bench in Fig. 13. For our analysis, we first count the number of times each option position is selected for each evaluation. To obtain a meaningful measure of answer frequency, we normalize these counts. This is necessary because some datasets contain a varying number of options per question. Therefore, for each option position i , we compute its normalized answering frequency by dividing the number of times it was selected (s_i) by the number of times it appeared in questions (n_i):

$$f_i = \frac{s_i}{n_i}. \quad (9)$$

Next, we normalize these answering frequencies across all option positions so that they sum to 1:

$$\hat{f}_i = \frac{f_i}{\sum_{j=1}^o f_j}, \quad (10)$$

where o is the maximal number of options in the datasets. This ensures comparability of the answering frequency across datasets and positions. In general, we observe a uniform answering behavior across all MLLMs for the datasets A-OKVQA, HRBench-4K, MM-Bench, and V*Bench. However, under certain evaluation configurations—such as LLaVA-1.5 with lowercase option IDs or LLaVA-OV with Roman numeral option IDs—we detect significant preferences for specific option IDs. Notably, on MME-RealWorld-Lite, most models display less uniform option-selection behavior compared to other benchmarks.

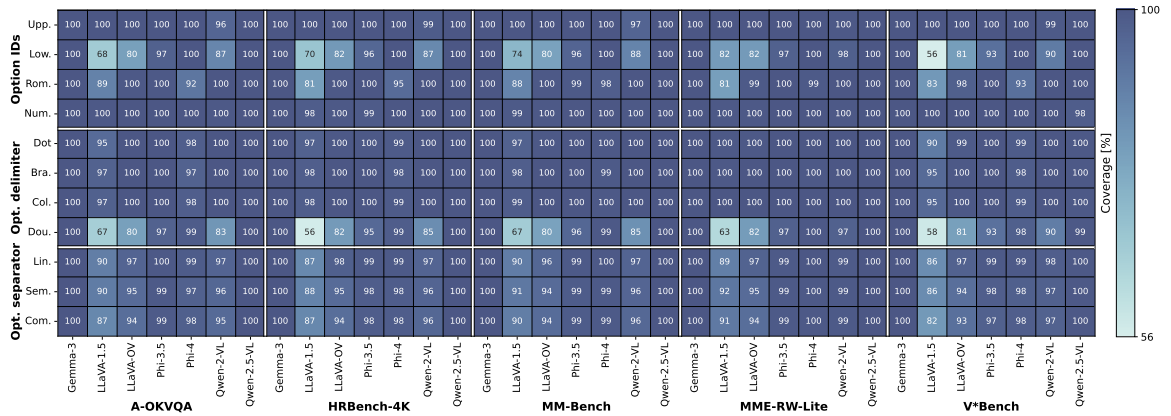


Figure 8: **Coverage results of MLLMs on MC-VQA datasets.** The results are averaged over 12 evaluations for *option IDs* and *option delimiter*, and 16 for *option separator*. While models such as Gemma-3, LLaVA-OV, Phi-3.5, Phi-4, Qwen2-VL, and Qwen2.5-VL consistently achieve coverage above 75%, indicating stable instruction-following capabilities, only LLaVA-1.5 exhibits minor drops under certain prompt formats, revealing strong sensitivity to formatting variations.

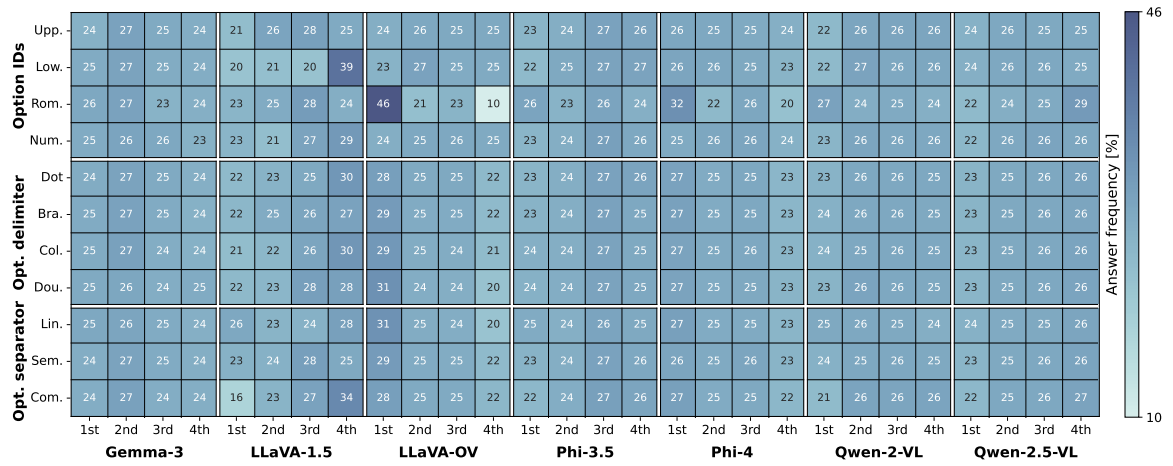


Figure 9: **Answer behavior of MLLMs on the A-OKVQA dataset.** We report the average answering frequency per option position for all seven MLLMs across the prompt format variations. The results are averaged over 12 evaluations for *option IDs* and *option delimiter*, and 16 for *option separator*. Option position bias is mitigated via a circular evaluation scheme.

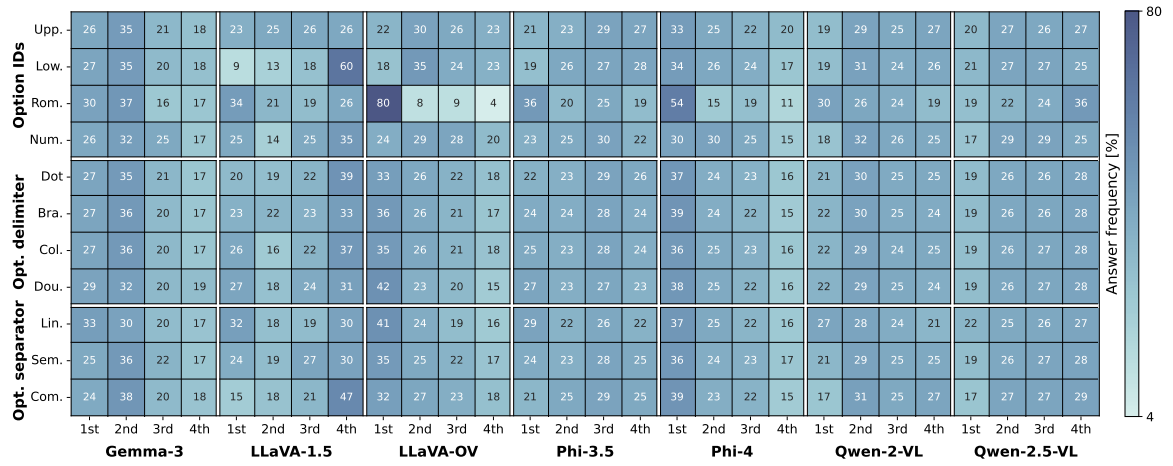


Figure 10: **Answer behavior of MLLMs on the HRBench-4K dataset.** We report the average answering frequency per option position for all seven MLLMs across the prompt format variations. The results are averaged over 12 evaluations for *option IDs* and *option delimiter*, and 16 for *option separator*. Option position bias is mitigated via a circular evaluation scheme.

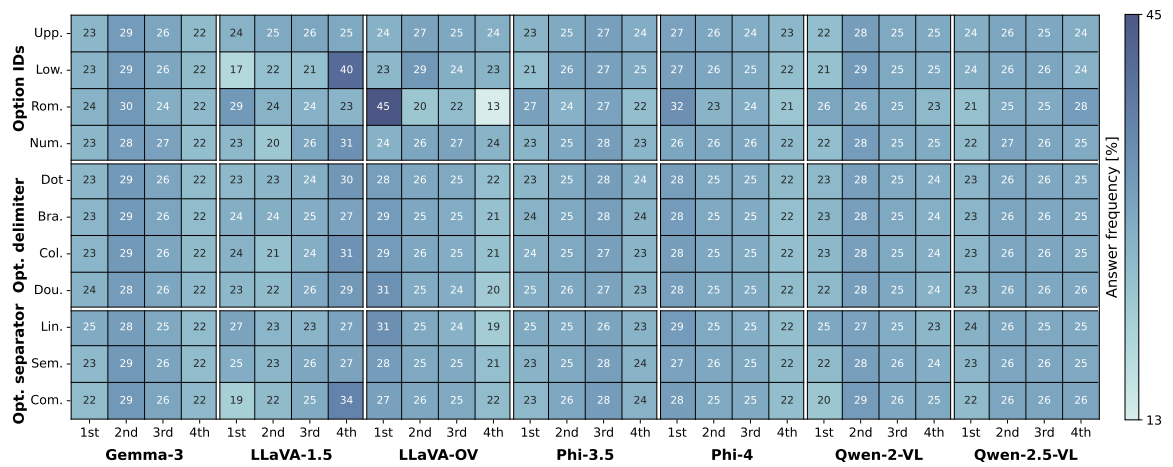


Figure 11: **Answer behavior of MLLMs on the MM-Bench dataset.** We report the average answering frequency per option position for all seven MLLMs across the prompt format variations. The results are averaged over 12 evaluations for *option IDs* and *option delimiter*, and 16 for *option separator*. Option position bias is mitigated via a circular evaluation scheme.

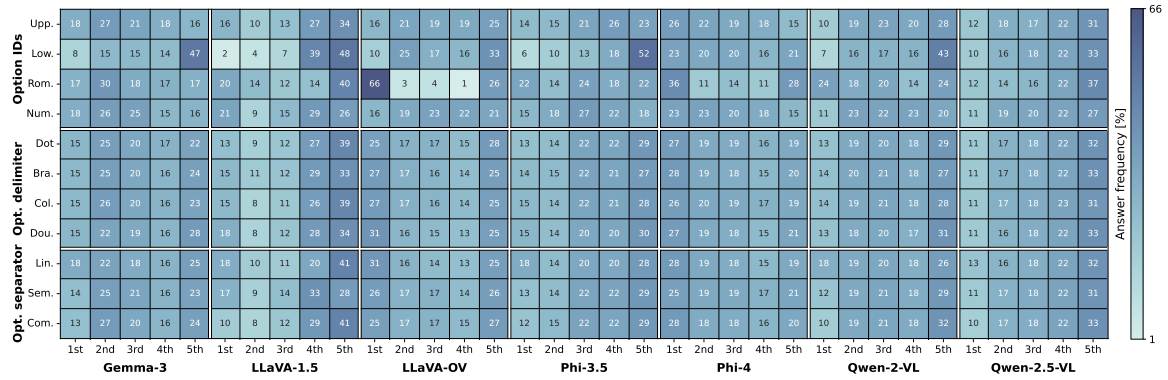


Figure 12: **Answer behavior of MLLMs on the MME-RealWorld-Lite dataset.** We report the average answering frequency per option position for all seven MLLMs across the prompt format variations. The results are averaged over 12 evaluations for *option IDs* and *option delimiter*, and 16 for *option separator*. Option position bias is mitigated via a circular evaluation scheme.

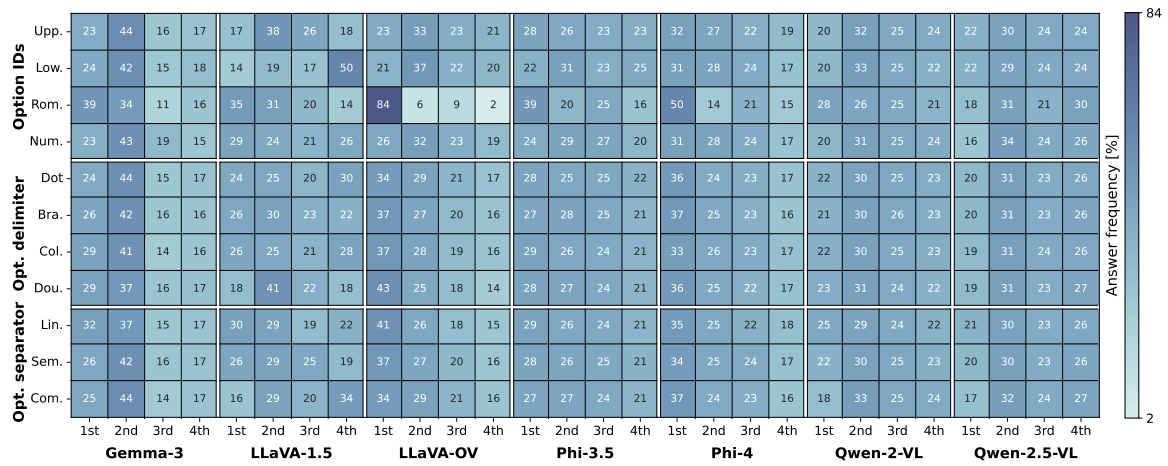


Figure 13: **Answer behavior of MLLMs on the V*Bench dataset.** We report the average answering frequency per option position for all seven MLLMs across the prompt format variations. The results are averaged over 12 evaluations for *option IDs* and *option delimiter*, and 16 for *option separator*. Option position bias is mitigated via a circular evaluation scheme.

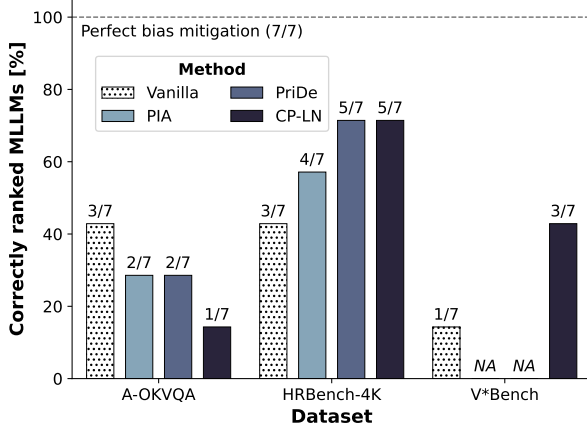


Figure 14: **Existing bias mitigation methods provide only limited mitigation of the prompt format-induced biases analyzed in this work.** We compare the effectiveness of bias mitigation methods (see Sec. A.6) in reducing the unexplored biases on three MC-VQA datasets. While some methods improve rankings on individual datasets, no method consistently recovers the rankings obtained by averaging performance across all 48 prompt formats (see Tab. 1).

C.3 Bias Mitigation Results

Next, we examine the extent to which existing bias mitigation methods PIA, PriDe, and CP-LN (see Sec. A.6) can mitigate the influence of the prompt format-induced biases across the datasets A-OKVQA, HRBench-4K, and V*Bench. We assess their effectiveness by comparing method rankings to those obtained by averaging accuracy across all 48 prompt formats. For reference, we also compare rankings based on each dataset’s standard prompt format with those from the evaluation results across all prompt formats. Our findings indicate that existing bias mitigation methods provide only limited robustness against prompt format-induced biases (see Fig. 14). Interestingly, on A-OKVQA, the vanilla prompt format correctly ranks 3 models, while PIA ranks 2. PriDe performs best on this dataset, correctly ranking 4 of the 7 models, whereas CP-LN ranks only 1. On HRBench-4K, both PriDe and CP-LN correctly rank 5 models, outperforming PIA (4) and the vanilla format (3). This suggests that bias mitigation can improve ranking quality on some datasets, although the gains are not consistent across benchmarks. PIA and PriDe are not compatible with evaluations on V*Bench due to their requirement for a fixed number of options (NA in Fig. 14), which varies between 2 and 4 in this dataset. CP-LN correctly ranks 3, while the vanilla format ranks 1 of the 7 MLLMs. Overall, none of the evaluated bias mitigation methods

consistently reflects benchmarking outcomes under prompt format variations across all datasets. Although PriDe and CP-LN improve ranking accuracy on individual benchmarks, their effectiveness remains dataset-dependent and they do not reliably recover the reference rankings obtained by averaging over all 48 prompt formats.

We present the full bias mitigation results on A-OKVQA in Tab. 11a, HRBench-4K in Tab. 11b, and V*Bench in Tab. 11c. We assess the effectiveness of three bias mitigation methods by comparing method rankings to those obtained by averaging accuracy across all 48 prompt format permutations.

C.4 Full Benchmark Results

Further, we report the full benchmark results of all dataset-model evaluations across A-OKVQA in Tab. 12, HRBench-4K in Tab. 13, MM-Bench in Tab. 14, MME-RealWorld-Lite in Tab. 15, and V*Bench in Tab. 16. Each table shows the MLLM performance under prompt format variations. The reported metrics include accuracy (%) and coverage (%), where accuracy reflects the correctness of predictions and coverage indicates the instruction-following capability for each model. These tables provide a comprehensive view of how prompt format variations influence model performance across all used MC-VQA datasets.

Table 11: **Full bias mitigation results.** This table compares the bias-mitigation effectiveness of the methods CP-LN, PIA, and PriDe on the datasets A-OKVQA, HRBench-4K, and V*Bench. V*Bench contains questions with 2 and 4 options. The methods PIA and PriDe do not support mixed-number of options datasets. We assess effectiveness by counting the number of correctly ranked models relative to the pseudo GT. The pseudo GT is obtained by averaging accuracy across all 48 prompt formats for each dataset (see Tab. 1). For correctly ranked models, the rank is highlighted in **bold**.

(a) Bias mitigation results on A-OKVQA.

Model	Pseudo GT		Vanilla		PIA		PriDe		CP-LN	
	Acc. [%]	Rank	Acc. [%]	Rank	Acc. [%]	Rank	Acc. [%]	Rank	Acc. [%]	Rank
Gemma-3	77.99	6	78.08	3	61.00	4	78.86	6	67.78	6
LLaVA-1.5	48.47	7	12.29	7	10.93	7	77.12	7	68.21	5
LLaVA-OV	81.28	3	42.40	6	37.91	5	91.35	1	96.59	1
Phi-3.5	79.38	5	78.67	2	64.37	2	80.70	5	67.69	7
Phi-4	80.50	4	75.04	4	61.14	3	83.40	4	85.94	2
Qwen-2-VL	81.96	2	44.26	5	36.31	6	87.16	3	78.52	3
Qwen-2.5-VL	86.77	1	87.93	1	77.31	1	87.69	2	74.93	4
Complexity [#] ↓	219,840		1,145		4,580		1,145		4,580	
Comparison to Pseudo GT										
Ranking correlation ↑			0.36		0.39		0.89		0.54	
Correctly ranked ↑			(3/7)		(2/7)		(4/7)		(1/7)	

(b) Bias mitigation results on HRBench-4K.

Model	Pseudo GT		Vanilla		PIA		PriDe		CP-LN	
	Acc. [%]	Rank	Acc. [%]	Rank	Acc. [%]	Rank	Acc. [%]	Rank	Acc. [%]	Rank
Gemma-3	45.91	6	44.25	6	4.99	6	44.25	6	52.00	6
LLaVA-1.5	29.90	7	36.75	7	3.41	7	37.13	7	34.50	7
LLaVA-OV	55.21	4	65.00	1	10.61	3	65.50	2	64.00	3
Phi-3.5	47.90	5	50.50	5	6.40	5	50.13	5	53.00	5
Phi-4	61.13	3	65.00	1	10.66	2	65.25	4	58.00	4
Qwen-2-VL	64.36	2	62.00	4	9.63	4	65.50	2	68.06	2
Qwen-2.5-VL	69.61	1	62.75	3	39.47	1	72.50	1	68.50	1
Complexity [#] ↓	38,400		200		800		200		800	
Comparison to Pseudo GT										
Ranking correlation ↑			0.66		0.86		0.98		1.00	
Correctly ranked ↑			(3/7)		(4/7)		(5/7)		(5/7)	

(c) Bias mitigation results on V*Bench.

Model	Pseudo GT		Vanilla		PIA		PriDe		CP-LN	
	Acc. [%]	Rank	Acc. [%]	Rank	Acc. [%]	Rank	Acc. [%]	Rank	Acc. [%]	Rank
Gemma-3	34.41	6	36.13	7	—	—	—	—	42.93	7
LLaVA-1.5	31.15	7	40.84	6	—	—	—	—	44.50	6
LLaVA-OV	62.36	4	72.73	3	—	—	—	—	68.06	2
Phi-3.5	47.46	5	52.88	5	—	—	—	—	53.40	5
Phi-4	69.83	2	73.30	1	—	—	—	—	62.83	4
Qwen-2-VL	69.58	3	54.45	4	—	—	—	—	65.97	3
Qwen-2.5-VL	77.44	1	72.77	2	—	—	—	—	69.11	1
Complexity [#] ↓	28,608		191		—		—		596	
Comparison to Pseudo GT										
Ranking correlation ↑			0.71		—		—		0.93	
Correctly ranked ↑			(1/7)		—		—		(3/7)	

Table 12: **Full results of prompt format variations on A-OKVQA dataset.** This table presents the performance of seven MLLMs under different option separators (*Comma*, *Line break*, *Semicolon*) and delimiters (*Dot*, *Colon*, *Bracket*, *Double brackets*). Option ID sets are varied across four formats: *Uppercase*, *Lowercase*, *Numbers*, and *Roman numbers*. Metrics include accuracy (%) and coverage (%), with coverage consistently near 100% and accuracy varying by format. The top-performing model changes across prompt format variations, alternating between Qwen2.5-VL, Qwen2-VL, and LLaVA-OV. The accuracy gap to the second-best MLLM ranges from 0.00 pp (tie between two models) to 5.39 pp. The highest accuracy per prompt format variation is highlighted in bold.

Option sep.	Model		Option delimiter															
			Dot Option IDs				Colon Option IDs				Bracket Option IDs				Double brackets Option IDs			
			Upp.	Low.	Num.	Rom.	Upp.	Low.	Num.	Rom.	Upp.	Low.	Num.	Rom.	Upp.	Low.	Num.	Rom.
Comma	Gemma-3	Acc. [%]	77.97	78.36	77.84	77.34	77.51	77.84	77.66	77.14	78.08	78.56	77.88	77.45	78.56	78.56	77.97	77.77
		Cov. [%]	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	99.98	100.00	99.96
	LLaVA-1.5	Acc. [%]	73.82	45.18	71.20	71.81	72.18	52.49	70.35	71.05	73.80	55.57	73.89	72.01	75.42	8.28	73.47	46.03
		Cov. [%]	100.00	67.10	99.76	99.87	100.00	80.09	99.91	99.98	100.00	80.55	99.87	99.93	99.96	10.26	99.59	55.44
	LLaVA-OV	Acc. [%]	89.85	89.54	89.76	80.72	90.42	90.07	89.80	74.28	90.61	90.33	90.15	77.58	90.59	3.82	89.72	73.14
		Cov. [%]	100.00	99.98	99.78	99.93	100.00	100.00	99.89	99.78	100.00	100.00	99.93	99.96	100.00	4.39	99.78	99.45
	Phi-3.5	Acc. [%]	76.92	78.25	79.21	79.00	79.54	79.69	79.50	79.13	78.30	79.19	79.52	79.32	79.65	69.85	79.21	78.56
		Cov. [%]	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	84.52	99.93	97.99
	Phi-4	Acc. [%]	82.77	82.49	82.36	72.12	82.77	82.86	82.29	73.04	82.82	82.73	82.62	69.54	83.04	82.84	82.58	76.57
		Cov. [%]	100.00	100.00	100.00	89.85	100.00	100.00	100.00	90.35	100.00	100.00	100.00	87.16	100.00	99.93	100.00	94.76
	Qwen-2-VL	Acc. [%]	83.58	84.26	85.59	85.33	84.15	85.35	85.20	85.70	84.26	85.09	85.59	85.59	65.94	35.76	84.21	85.94
		Cov. [%]	100.00	100.00	99.83	100.00	100.00	100.00	99.83	100.00	100.00	100.00	99.83	100.00	78.08	44.06	98.43	99.91
	Qwen-2.5-VL	Acc. [%]	87.31	87.21	85.70	85.94	86.97	87.03	85.68	85.13	87.23	86.99	86.05	85.59	87.56	87.36	85.79	85.13
		Cov. [%]	100.00	100.00	99.87	100.00	100.00	99.98	99.87	100.00	100.00	100.00	99.87	100.00	100.00	100.00	99.85	100.00
Line break	Gemma-3	Acc. [%]	78.56	78.19	78.17	77.75	78.47	78.41	78.30	77.82	78.62	78.65	78.30	77.82	78.14	78.08	78.12	78.06
		Cov. [%]	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	99.98	100.00	100.00	100.00
	LLaVA-1.5	Acc. [%]	79.11	69.98	76.90	77.58	78.56	76.05	77.23	77.03	78.89	74.26	77.27	77.36	78.89	12.29	76.70	38.60
		Cov. [%]	100.00	89.50	99.74	99.93	100.00	97.47	99.93	100.00	100.00	94.48	99.89	99.98	100.00	13.87	99.30	42.03
	LLaVA-OV	Acc. [%]	91.07	90.76	90.02	59.85	90.96	90.61	90.50	62.80	90.87	90.90	90.37	63.58	90.74	42.40	90.26	67.27
		Cov. [%]	100.00	100.00	99.72	99.41	100.00	100.00	99.91	99.61	100.00	100.00	99.85	99.69	100.00	47.47	99.74	99.59
	Phi-3.5	Acc. [%]	81.00	81.05	81.03	80.09	80.87	80.98	80.44	79.85	81.07	80.87	80.81	80.13	81.05	78.67	80.98	79.85
		Cov. [%]	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	96.18	100.00	99.72
	Phi-4	Acc. [%]	82.62	82.34	82.18	78.17	82.88	82.71	82.53	79.11	82.66	82.53	81.97	76.20	82.90	82.60	82.60	75.66
		Cov. [%]	100.00	100.00	100.00	96.16	100.00	100.00	100.00	96.79	100.00	100.00	100.00	93.78	100.00	100.00	100.00	93.71
	Qwen-2-VL	Acc. [%]	86.57	86.83	85.63	84.98	86.62	86.62	85.59	85.35	86.66	86.42	85.81	85.13	78.97	44.26	85.22	85.61
		Cov. [%]	100.00	100.00	99.91	100.00	100.00	100.00	99.89	100.00	100.00	100.00	99.91	100.00	91.77	53.95	99.26	99.85
	Qwen-2.5-VL	Acc. [%]	87.99	87.77	86.49	86.16	88.12	87.80	86.53	86.09	87.90	87.86	86.53	86.03	88.17	87.99	86.83	85.68
		Cov. [%]	100.00	100.00	99.87	99.98	100.00	100.00	99.85	100.00	100.00	100.00	99.87	99.98	100.00	100.00	99.85	99.96
Semicolon	Gemma-3	Acc. [%]	77.97	78.30	77.40	77.56	77.49	77.86	77.56	77.51	78.10	78.43	77.53	77.51	78.30	78.34	77.90	77.86
		Cov. [%]	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	99.98
	LLaVA-1.5	Acc. [%]	77.47	60.74	73.93	73.91	75.63	68.14	73.73	73.01	76.11	68.71	74.41	73.71	76.40	12.16	74.37	54.32
		Cov. [%]	100.00	80.79	99.85	99.93	100.00	90.90	99.96	100.00	100.00	90.92	99.87	99.96	99.98	14.32	99.89	65.81
	LLaVA-OV	Acc. [%]	90.00	90.24	89.76	79.39	90.85	90.33	90.18	73.58	90.94	90.66	90.02	77.42	90.72	11.66	89.83	72.66
		Cov. [%]	100.00	100.00	99.89	99.83	100.00	100.00	99.96	99.83	100.00	100.00	99.91	99.96	100.00	13.30	99.78	99.72
	Phi-3.5	Acc. [%]	78.89	79.39	79.78	79.32	79.69	79.91	79.76	79.30	79.56	79.76	80.18	79.78	79.76	72.53	80.04	79.15
		Cov. [%]	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	87.49	99.93	98.87
	Phi-4	Acc. [%]	82.60	82.36	82.21	71.83	82.84	82.62	82.42	72.75	82.73	82.51	82.62	68.30	82.73	82.62	82.58	77.60
		Cov. [%]	100.00	100.00	100.00	89.39	100.00	100.00	100.00	89.74	100.00	100.00	100.00	84.98	100.00	99.93	100.00	95.50
	Qwen-2-VL	Acc. [%]	84.87	86.27	85.90	85.44	86.11	86.55	85.59	85.66	86.29	86.46	85.98	85.72	73.65	41.09	84.96	85.92
		Cov. [%]	100.00	100.00	99.87	100.00	100.00	100.00	99.85	100.00	100.00	100.00	99.87	100.00	86.00	50.02	98.91	99.96
	Qwen-2.5-VL	Acc. [%]	87.84	87.58	86.11	85.87	87.60	87.42	85.90	85.72	87.40	87.58	86.33	85.92	87.62	87.56	86.40	85.50
		Cov. [%]	100.00	100.00	99.87	100.00	100.00	100.00	99.85	100.00	100.00	100.00	99.87	100.00	100.00	100.00	99.85	99.98

Table 13: **Full results of prompt format variations on HR-Bench-4K dataset.** This table presents the performance of seven MLLMs under different option separators (*Comma*, *Line break*, *Semicolon*) and delimiters (*Dot*, *Colon*, *Bracket*, *Double brackets*). Option ID sets are varied across four formats: *Uppercase*, *Lowercase*, *Numbers*, and *Roman numbers*. Metrics include accuracy (%) and coverage (%), with coverage consistently near 100% and accuracy varying by format. Qwen-2.5-VL is consistently the top-performing MLLM, demonstrating strong robustness. The accuracy gap to the second-best MLLM ranges from 0.25 pp to 6.88 pp. The highest accuracy per prompt format variation is highlighted in bold.

Option sep.	Model		Option delimiter															
			Dot				Colon				Bracket				Double brackets			
			Option IDs				Option IDs				Option IDs				Option IDs			
			Upp.	Low.	Num.	Rom.	Upp.	Low.	Num.	Rom.	Upp.	Low.	Num.	Rom.	Upp.	Low.	Num.	Rom.
Comma	Gemma-3	Acc. [%]	45.75	46.00	43.50	46.38	46.25	46.38	44.25	46.25	46.38	46.25	46.38	46.88	45.63	46.38	45.00	46.88
		Cov. [%]	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
	LLaVA-1.5	Acc. [%]	33.00	24.38	30.50	33.38	33.13	25.38	30.88	31.75	35.13	27.00	32.25	33.38	35.88	1.88	32.88	11.88
		Cov. [%]	100.00	84.88	97.25	99.88	100.00	89.25	100.00	100.00	100.00	87.25	99.75	100.00	100.00	4.75	97.13	32.13
	LLaVA-OV	Acc. [%]	64.25	62.38	60.25	47.75	64.25	63.88	60.38	44.63	64.50	63.88	61.38	46.63	64.13	4.38	62.00	40.13
		Cov. [%]	100.00	100.00	100.00	99.63	100.00	100.00	100.00	99.63	100.00	100.00	100.00	99.88	100.00	6.63	99.63	99.63
	Phi-3.5	Acc. [%]	46.88	48.00	46.63	45.50	47.75	47.25	45.25	45.38	47.25	48.25	48.38	45.88	50.25	39.63	47.88	45.13
		Cov. [%]	100.00	100.00	100.00	100.00	100.00	100.00	99.50	100.00	100.00	100.00	100.00	100.00	100.00	78.13	97.38	99.75
	Phi-4	Acc. [%]	62.00	63.25	60.38	53.25	63.50	63.63	63.13	54.38	63.50	62.75	62.50	49.63	63.13	63.25	61.88	50.88
		Cov. [%]	100.00	100.00	100.00	94.88	100.00	100.00	100.00	94.00	100.00	100.00	100.00	92.38	100.00	100.00	100.00	93.50
	Qwen-2-VL	Acc. [%]	66.00	66.13	67.63	66.38	65.50	66.75	67.50	66.63	66.25	66.25	66.88	66.38	59.88	25.63	66.00	67.63
		Cov. [%]	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	99.88	100.00	92.25	43.88	100.00	100.00
	Qwen-2.5-VL	Acc. [%]	70.63	69.88	68.88	67.88	69.63	69.63	67.75	67.38	70.50	70.38	68.88	67.88	70.50	70.00	69.38	66.50
		Cov. [%]	100.00	100.00	99.75	100.00	100.00	99.88	99.88	100.00	100.00	100.00	99.88	100.00	100.00	99.88	100.00	100.00
Line break	Gemma-3	Acc. [%]	44.25	44.88	45.13	46.88	46.00	45.88	46.00	46.13	44.88	44.63	45.63	46.63	45.50	44.75	45.13	46.75
		Cov. [%]	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
	LLaVA-1.5	Acc. [%]	36.75	33.00	34.00	34.00	36.75	33.25	34.50	34.13	36.38	34.13	35.00	33.63	35.38	0.88	34.25	5.63
		Cov. [%]	100.00	93.38	97.00	100.00	100.00	96.63	99.13	100.00	100.00	95.00	99.63	100.00	100.00	1.88	94.88	9.25
	LLaVA-OV	Acc. [%]	65.00	65.75	63.50	34.25	65.25	65.88	63.50	34.75	65.13	65.25	63.25	36.00	65.88	42.38	63.63	35.63
		Cov. [%]	100.00	100.00	99.13	99.00	100.00	100.00	99.88	99.63	100.00	100.00	99.38	99.75	100.00	65.88	99.13	99.75
	Phi-3.5	Acc. [%]	50.50	49.75	49.00	49.63	50.63	50.25	50.25	49.13	50.38	49.63	50.00	48.75	50.75	47.63	50.38	49.50
		Cov. [%]	100.00	100.00	97.88	99.75	100.00	100.00	99.00	99.88	100.00	100.00	99.13	99.75	100.00	93.13	97.88	99.88
	Phi-4	Acc. [%]	65.00	65.50	63.25	60.00	65.00	65.13	63.25	61.38	64.25	65.50	63.88	56.38	64.50	64.50	63.50	54.00
		Cov. [%]	100.00	100.00	100.00	99.13	100.00	100.00	100.00	99.38	100.00	100.00	100.00	96.38	100.00	100.00	100.00	93.13
	Qwen-2-VL	Acc. [%]	67.13	67.38	68.63	65.88	67.00	67.50	68.38	67.00	66.63	66.88	68.38	67.50	64.75	25.88	68.75	68.38
		Cov. [%]	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	97.75	47.00	99.88	100.00
	Qwen-2.5-VL	Acc. [%]	71.25	69.88	69.88	69.00	71.00	70.38	69.63	68.88	71.00	70.38	69.63	68.63	71.50	70.50	69.75	68.88
		Cov. [%]	100.00	100.00	100.00	100.00	100.00	100.00	99.63	100.00	100.00	100.00	100.00	100.00	100.00	100.00	99.88	100.00
Semicolon	Gemma-3	Acc. [%]	46.38	46.88	44.88	46.63	46.88	46.75	45.00	47.50	46.38	46.38	45.25	46.63	45.38	47.00	45.63	46.63
		Cov. [%]	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
	LLaVA-1.5	Acc. [%]	37.25	29.88	32.63	35.50	36.63	32.88	32.50	32.75	36.88	33.13	34.00	33.88	35.63	2.00	32.63	13.00
		Cov. [%]	100.00	91.88	98.75	99.88	100.00	95.13	99.75	100.00	100.00	94.63	100.00	100.00	100.00	5.38	96.75	33.88
	LLaVA-OV	Acc. [%]	64.13	63.38	61.50	46.50	65.50	65.13	61.63	41.75	64.75	65.13	61.50	43.88	64.38	9.38	62.88	39.13
		Cov. [%]	100.00	100.00	100.00	99.25	100.00	100.00	100.00	99.63	100.00	100.00	99.88	99.75	100.00	15.50	99.88	99.50
	Phi-3.5	Acc. [%]	48.13	48.38	46.50	45.88	48.50	48.50	48.13	46.63	48.88	48.38	47.88	47.38	49.88	39.13	48.38	47.13
		Cov. [%]	100.00	100.00	99.50	99.75	100.00	100.00	99.88	100.00	100.00	100.00	100.00	99.75	100.00	79.13	97.63	100.00
	Phi-4	Acc. [%]	62.13	62.00	61.38	53.88	63.00	62.63	62.38	55.88	62.75	63.50	62.75	51.50	63.63	63.13	62.75	54.88
		Cov. [%]	100.00	100.00	100.00	93.88	100.00	100.00	100.00	94.00	100.00	100.00	100.00	91.00	100.00	100.00	100.00	95.75
	Qwen-2-VL	Acc. [%]	67.00	68.00	67.88	66.38	67.75	68.13	67.25	67.25	67.38	68.13	68.38	68.13	62.25	27.38	67.13	67.63
		Cov. [%]	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	94.25	49.13	99.88	100.00
	Qwen-2.5-VL	Acc. [%]	71.75	69.88	69.25	69.00	70.25	69.88	69.38	68.88	69.63	70.50	69.63	69.13	71.25	69.25	69.75	68.38
		Cov. [%]	100.00	100.00	100.00	100.00	100.00	99.88	99.88	100.00	100.00	99.88	100.00	100.00	100.00	99.88	100.00	100.00

Table 14: **Full results of prompt format variations on MM-Bench dataset.** This table presents the performance of seven MLLMs under different option separators (*Comma, Line break, Semicolon*) and delimiters (*Dot, Colon, Bracket, Double brackets*). Option ID sets are varied across four formats: *Uppercase, Lowercase, Numbers*, and *Roman numbers*. Metrics include accuracy (%) and coverage (%), with coverage consistently near 100% and accuracy varying by format. Qwen-2.5-VL is consistently the top-performing MLLM, demonstrating strong robustness. The accuracy gap to the second-best MLLM ranges from 0.10 pp to 3.84 pp. The highest accuracy per prompt format variation is highlighted in bold.

Option sep. Model			Option delimiter															
			Dot Option IDs				Colon Option IDs				Bracket Option IDs				Double brackets Option IDs			
			Upp.	Low.	Num.	Rom.	Upp.	Low.	Num.	Rom.	Upp.	Low.	Num.	Rom.	Upp.	Low.	Num.	Rom.
Comma	Gamma-3	Acc. [%]	76.74	77.52	77.24	76.91	77.03	77.24	77.67	76.60	76.85	77.48	77.46	77.09	78.12	78.04	77.67	77.54
		Cov. [%]	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
	LLaVA-1.5	Acc. [%]	66.63	52.69	65.69	65.42	66.12	55.99	64.91	64.89	66.78	56.87	67.25	65.05	69.07	14.52	66.59	46.45
		Cov. [%]	99.96	83.76	95.51	99.16	99.96	90.03	99.91	99.43	99.94	89.15	100.00	99.24	99.90	17.56	99.24	57.22
	LLaVA-OV	Acc. [%]	84.58	84.09	83.70	78.92	84.76	84.33	84.39	76.27	84.76	84.45	84.15	77.69	85.25	9.25	84.62	74.75
		Cov. [%]	99.98	99.96	99.57	99.08	100.00	100.00	99.80	99.61	100.00	100.00	99.69	99.61	100.00	11.10	99.57	99.28
	Phi-3.5	Acc. [%]	75.55	76.60	77.77	77.30	77.85	78.10	77.48	77.01	77.40	77.67	78.38	77.87	79.80	68.48	78.34	76.81
		Cov. [%]	99.96	99.96	99.92	99.98	100.00	99.98	99.96	100.00	100.00	99.98	99.94	99.92	99.96	82.08	99.63	96.80
	Phi-4	Acc. [%]	84.52	84.25	83.92	81.01	84.48	84.17	83.68	80.95	84.68	84.64	84.33	79.86	84.72	84.13	84.60	82.75
		Cov. [%]	100.00	100.00	100.00	97.99	100.00	100.00	100.00	97.70	100.00	100.00	100.00	96.53	100.00	99.16	99.90	99.16
Qwen-2-VL	Acc. [%]	82.81	83.08	83.45	83.74	82.88	83.35	83.80	83.94	83.04	83.27	84.09	83.94	66.18	40.73	86.61	84.35	
	Cov. [%]	100.00	100.00	99.86	99.96	100.00	100.00	99.92	99.98	100.00	100.00	99.92	99.96	79.53	50.08	99.34	99.71	
Qwen-2.5-VL	Acc. [%]	88.08	87.86	86.49	85.73	87.76	87.84	86.75	85.17	87.98	87.80	86.53	85.34	87.96	87.86	87.06	84.84	
	Cov. [%]	100.00	100.00	99.94	100.00	100.00	100.00	99.98	100.00	100.00	100.00	99.92	100.00	100.00	100.00	100.00	100.00	
Line break	Gamma-3	Acc. [%]	77.73	77.79	78.36	77.93	77.97	78.20	78.45	78.45	77.56	78.06	78.38	77.81	78.24	78.49	78.49	78.28
		Cov. [%]	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
	LLaVA-1.5	Acc. [%]	71.86	67.97	70.47	70.32	72.13	69.46	70.47	69.91	71.97	68.79	70.65	70.22	71.47	14.17	68.72	34.45
		Cov. [%]	100.00	95.02	99.45	99.69	100.00	97.95	99.75	99.84	100.00	96.37	99.65	99.77	99.98	16.00	97.07	38.50
	LLaVA-OV	Acc. [%]	85.54	85.21	84.37	65.09	85.54	85.17	84.70	67.00	85.52	85.13	84.76	68.89	85.24	27.54	84.99	71.66
		Cov. [%]	100.00	99.92	99.67	99.04	100.00	100.00	99.80	99.82	100.00	99.98	99.77	99.84	100.00	34.37	99.51	99.69
	Phi-3.5	Acc. [%]	80.27	80.35	80.11	79.23	79.66	79.74	79.84	79.64	80.27	80.27	80.23	79.35	80.13	75.51	79.94	79.27
		Cov. [%]	100.00	100.00	99.98	99.34	100.00	100.00	99.98	99.96	100.00	100.00	100.00	99.26	100.00	91.88	100.00	99.10
	Phi-4	Acc. [%]	85.25	84.82	85.03	83.14	85.05	84.80	85.05	83.88	85.28	85.03	85.32	82.57	85.19	84.72	85.15	81.87
		Cov. [%]	100.00	100.00	100.00	99.12	100.00	100.00	100.00	99.63	100.00	100.00	100.00	98.61	100.00	99.88	99.82	97.79
Qwen-2-VL	Acc. [%]	85.46	85.40	84.60	84.58	85.56	85.48	84.48	85.13	85.30	85.40	84.64	84.97	79.08	44.69	84.25	85.23	
	Cov. [%]	100.00	100.00	99.92	100.00	100.00	100.00	99.92	100.00	100.00	100.00	99.92	100.00	93.01	54.61	99.80	99.92	
Qwen-2.5-VL	Acc. [%]	87.88	87.88	87.24	85.99	88.02	88.08	87.33	86.16	87.84	88.04	87.43	85.99	87.96	87.90	87.35	85.34	
	Cov. [%]	100.00	100.00	99.96	100.00	100.00	100.00	99.94	100.00	100.00	100.00	99.98	100.00	100.00	100.00	100.00	100.00	
Semicolon	Gamma-3	Acc. [%]	77.19	77.48	77.58	77.03	77.26	77.58	77.69	77.34	77.17	77.44	77.69	77.05	77.83	77.91	77.61	77.44
		Cov. [%]	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
	LLaVA-1.5	Acc. [%]	69.38	62.59	67.49	67.62	69.16	64.89	67.58	66.41	69.01	65.09	68.05	66.94	69.65	17.25	67.66	50.74
		Cov. [%]	99.92	90.98	99.53	99.47	99.96	95.16	99.80	99.73	99.92	94.61	99.75	99.57	99.90	20.43	99.41	63.21
	LLaVA-OV	Acc. [%]	84.84	84.52	84.21	78.16	85.42	85.03	84.72	75.74	85.17	84.95	84.80	77.55	85.50	12.39	84.89	74.57
		Cov. [%]	99.98	99.94	99.69	99.04	100.00	100.00	99.82	99.73	100.00	100.00	99.80	99.75	100.00	14.97	99.49	99.63
	Phi-3.5	Acc. [%]	77.40	77.69	78.51	78.59	78.92	78.86	78.55	77.97	78.38	78.67	78.88	78.88	79.43	69.30	78.63	77.63
		Cov. [%]	100.00	100.00	99.96	99.94	100.00	100.00	99.98	99.98	100.00	100.00	99.98	99.90	99.96	83.39	99.61	97.70
	Phi-4	Acc. [%]	84.23	84.17	83.51	81.30	84.52	83.78	83.66	81.32	84.52	84.31	84.31	80.21	84.27	84.02	84.39	83.18
		Cov. [%]	100.00	100.00	100.00	98.05	100.00	100.00	100.00	97.58	100.00	100.00	100.00	96.37	100.00	98.95	99.90	99.36
Qwen-2-VL	Acc. [%]	83.51	84.62	83.96	84.48	84.41	84.70	84.07	84.27	84.91	84.80	84.09	84.66	73.40	42.35	83.57	84.60	
	Cov. [%]	100.00	100.00	99.92	99.94	100.00	100.00	99.92	99.96	100.00	100.00	99.90	99.94	87.43	52.09	99.36	99.88	
Qwen-2.5-VL	Acc. [%]	88.08	87.94	86.81	85.99	88.06	88.02	86.87	85.85	88.04	87.96	86.92	85.69	87.98	87.78	87.16	84.95	
	Cov. [%]	100.00	100.00	99.94	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	

Table 15: **Full results of prompt format variations on MME-RealWorld-Lite dataset.** This table presents the performance of seven MLLMs under different option separators (*Comma*, *Line break*, *Semicolon*) and delimiters (*Dot*, *Colon*, *Bracket*, *Double brackets*). Option ID sets are varied across four formats: *Uppercase*, *Lowercase*, *Numbers*, and *Roman numbers*. Metrics include accuracy (%) and coverage (%), with coverage consistently near 100% and accuracy varying by format. The top-performing model changes across prompt format variations, alternating between Qwen-2.5-VL, Qwen-2-VL, and Phi-4. The accuracy gap to the second-best MLLM ranges from 0.08 pp to 3.00 pp. The highest accuracy per prompt format variation is highlighted in bold.

Option sep.	Model		Option delimiter															
			Dot Option IDs				Colon Option IDs				Bracket Option IDs				Double brackets Option IDs			
			Upp.	Low.	Num.	Rom.	Upp.	Low.	Num.	Rom.	Upp.	Low.	Num.	Rom.	Upp.	Low.	Num.	Rom.
Comma	Gemma-3	Acc. [%]	35.97	34.24	35.28	35.78	36.23	34.10	35.33	35.33	35.85	32.84	35.04	35.87	36.11	30.70	35.70	35.71
		Cov. [%]	100.00	100.00	99.97	100.00	100.00	100.00	99.99	100.00	100.00	100.00	99.94	100.00	100.00	99.99	99.98	100.00
	LLaVA-1.5	Acc. [%]	30.44	26.62	28.87	29.03	29.71	26.11	28.73	28.23	30.79	27.14	29.71	29.88	30.94	8.50	29.72	10.70
		Cov. [%]	100.00	97.98	99.53	99.74	100.00	99.33	99.84	99.92	100.00	98.75	99.69	99.81	100.00	30.39	99.42	28.25
	LLaVA-OV	Acc. [%]	44.77	44.72	45.14	34.77	46.18	45.51	45.55	33.08	46.20	45.03	45.09	33.99	45.31	3.93	44.66	32.30
		Cov. [%]	100.00	100.00	99.75	99.19	100.00	100.00	99.95	99.49	100.00	100.00	99.89	99.60	100.00	9.53	98.60	98.44
	Phi-3.5	Acc. [%]	37.34	35.14	39.15	39.40	39.25	36.21	39.60	39.50	38.00	36.33	39.66	39.02	37.67	26.76	40.25	38.67
		Cov. [%]	100.00	100.00	99.98	100.00	100.00	100.00	99.99	100.00	100.00	100.00	100.00	99.96	99.96	88.53	99.92	99.40
	Phi-4	Acc. [%]	47.51	47.82	45.49	44.19	47.66	47.48	45.66	45.29	47.51	47.67	45.61	42.82	46.90	45.81	46.13	43.23
		Cov. [%]	100.00	100.00	100.00	99.32	100.00	100.00	100.00	99.34	100.00	100.00	100.00	98.78	100.00	99.20	99.85	99.39
	Qwen-2-VL	Acc. [%]	44.27	41.96	45.92	45.43	44.62	42.30	46.17	46.35	44.98	42.96	46.14	45.69	42.64	33.24	45.70	46.59
		Cov. [%]	100.00	99.99	99.73	100.00	100.00	100.00	99.78	100.00	100.00	100.00	99.88	100.00	99.36	91.68	99.43	100.00
	Qwen-2.5-VL	Acc. [%]	47.21	44.82	47.48	46.02	47.74	45.17	47.51	45.76	46.84	44.52	47.32	45.51	46.51	44.42	47.29	44.30
		Cov. [%]	100.00	100.00	99.90	100.00	100.00	100.00	99.90	100.00	100.00	100.00	99.95	100.00	100.00	100.00	99.95	99.99
Line break	Gemma-3	Acc. [%]	35.92	33.23	35.28	36.12	35.78	32.35	35.58	35.38	35.82	31.86	35.69	36.17	36.09	29.65	35.50	36.22
		Cov. [%]	100.00	100.00	99.94	100.00	100.00	100.00	99.95	100.00	100.00	100.00	99.95	100.00	100.00	100.00	99.95	100.00
	LLaVA-1.5	Acc. [%]	32.53	29.38	33.15	33.76	31.70	28.77	33.40	32.80	32.94	29.20	33.67	34.47	33.13	6.48	31.97	4.76
		Cov. [%]	100.00	99.52	99.46	99.98	100.00	99.65	99.87	99.99	100.00	99.64	99.71	99.99	99.99	23.44	99.03	10.46
	LLaVA-OV	Acc. [%]	44.53	44.16	43.86	29.59	45.48	44.96	44.92	30.12	45.43	44.58	44.73	30.42	45.23	22.93	44.77	31.61
		Cov. [%]	100.00	100.00	98.44	99.04	100.00	100.00	99.85	99.53	100.00	100.00	99.80	99.52	100.00	54.81	99.43	99.38
	Phi-3.5	Acc. [%]	38.95	34.64	41.12	39.24	39.01	34.85	41.25	39.16	38.34	35.30	40.58	38.44	38.06	27.83	40.86	38.56
		Cov. [%]	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	91.01	99.99	99.83
	Phi-4	Acc. [%]	47.43	46.95	46.67	44.70	47.70	47.06	46.38	46.08	47.09	47.08	46.73	43.91	47.36	46.31	46.99	43.67
		Cov. [%]	100.00	100.00	100.00	99.75	100.00	100.00	100.00	99.84	100.00	100.00	100.00	99.50	100.00	99.92	99.97	99.03
	Qwen-2-VL	Acc. [%]	45.18	44.30	46.56	45.12	45.80	44.83	46.45	45.97	45.91	45.34	46.59	45.71	45.07	32.72	46.48	46.60
		Cov. [%]	100.00	100.00	99.80	99.99	100.00	100.00	99.90	100.00	100.00	100.00	99.87	100.00	98.92	86.80	99.56	99.95
	Qwen-2.5-VL	Acc. [%]	46.17	44.41	47.91	45.46	46.77	44.86	47.85	45.41	46.02	44.42	47.55	45.08	46.55	44.30	47.22	44.28
		Cov. [%]	100.00	100.00	99.97	100.00	100.00	100.00	99.98	100.00	100.00	100.00	99.97	100.00	100.00	100.00	99.99	99.99
Semicolon	Gemma-3	Acc. [%]	36.04	34.15	35.18	35.78	36.09	34.07	35.25	35.29	36.08	32.90	35.54	35.66	36.09	30.74	35.65	36.19
		Cov. [%]	100.00	100.00	99.99	100.00	100.00	100.00	99.99	100.00	100.00	100.00	99.99	100.00	100.00	100.00	99.98	100.00
	LLaVA-1.5	Acc. [%]	32.01	28.63	30.38	31.28	30.89	27.61	31.07	30.57	32.26	28.77	31.37	31.78	10.28	30.47	12.14	
		Cov. [%]	100.00	98.82	99.67	99.85	100.00	99.50	99.93	99.94	100.00	99.31	99.80	99.93	100.00	36.09	99.56	32.53
	LLaVA-OV	Acc. [%]	43.90	44.18	44.43	34.25	45.84	45.21	45.04	32.97	45.72	44.84	44.78	33.57	44.98	8.25	44.67	31.81
		Cov. [%]	100.00	100.00	99.90	98.73	100.00	100.00	99.97	99.47	100.00	100.00	99.95	99.58	100.00	20.60	98.92	99.15
	Phi-3.5	Acc. [%]	37.55	35.00	40.52	40.45	39.55	36.09	40.73	39.84	37.97	36.12	40.67	39.75	37.88	26.71	40.56	39.10
		Cov. [%]	100.00	100.00	100.00	100.00	100.00	100.00	99.99	100.00	100.00	100.00	99.99	100.00	99.97	89.06	99.95	99.70
	Phi-4	Acc. [%]	47.29	47.49	46.10	44.83	47.53	47.18	45.82	45.34	47.51	48.07	46.36	43.42	47.25	45.90	46.81	44.04
		Cov. [%]	100.00	100.00	100.00	99.16	100.00	100.00	100.00	99.03	100.00	100.00	100.00	98.36	100.00	99.10	99.92	99.31
	Qwen-2-VL	Acc. [%]	44.65	42.96	46.14	46.05	45.23	43.14	46.02	46.34	45.75	43.91	46.38	46.07	43.79	34.47	45.82	46.40
		Cov. [%]	100.00	100.00	99.92	100.00	100.00	100.00	99.92	100.00	100.00	100.00	99.96	100.00	99.57	92.73	99.57	99.99
	Qwen-2.5-VL	Acc. [%]	47.51	45.59	47.62	46.24	48.04	45.82	47.67	46.09	47.32	45.46	47.58	45.62	47.40	45.67	47.08	44.67
		Cov. [%]	100.00	100.00	99.93	100.00	100.00	100.00	99.91	100.00	100.00	100.00	99.96	100.00	100.00	100.00	99.95	99.99

Table 16: **Full results of prompt format variations on V*Bench dataset.** This table summarizes the performance of seven MLLMs under different option separators (*Comma*, *Line break*, *Semicolon*) and delimiters (*Dot*, *Colon*, *Bracket*, *Double brackets*). Option ID sets are varied across four formats: *Uppercase*, *Lowercase*, *Numbers*, and *Roman numbers*. Metrics include accuracy (%) and coverage (%), with coverage consistently near 100% and accuracy varying by format. Qwen-2.5-VL is consistently the top-performing MLLM, demonstrating strong robustness. The accuracy gap to the second-best MLLM ranges from 1.01 pp to 9.56 pp. The highest accuracy per prompt format variation is highlighted in bold.

Option sep.	Model		Option delimiter															
			Dot Option IDs				Colon Option IDs				Bracket Option IDs				Double brackets Option IDs			
			Upp.	Low.	Num.	Rom.	Upp.	Low.	Num.	Rom.	Upp.	Low.	Num.	Rom.	Upp.	Low.	Num.	Rom.
Comma	Gemma-3	Acc. [%]	35.40	34.56	35.74	32.22	34.23	34.90	34.40	35.24	33.89	33.39	35.07	33.39	34.23	34.90	34.06	33.56
		Cov. [%]	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
	LLaVA-1.5	Acc. [%]	37.58	15.27	35.07	35.07	39.43	20.13	36.75	35.40	37.25	23.66	36.91	35.24	38.59	0.34	37.08	16.44
		Cov. [%]	100.00	50.67	100.00	100.00	100.00	63.09	100.00	100.00	100.00	68.96	100.00	100.00	100.00	0.50	100.00	32.38
	LLaVA-OV	Acc. [%]	72.65	69.80	70.30	52.01	71.81	70.81	73.49	45.81	72.99	71.14	73.83	48.49	73.66	0.67	72.99	44.97
		Cov. [%]	100.00	100.00	100.00	97.32	100.00	100.00	100.00	98.99	100.00	100.00	100.00	99.50	100.00	1.17	100.00	97.65
	Phi-3.5	Acc. [%]	46.31	47.48	46.14	45.97	49.66	48.99	47.99	46.81	48.32	48.49	46.64	45.64	49.50	30.03	46.64	46.31
		Cov. [%]	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	99.83	59.90	99.66	98.99
	Phi-4	Acc. [%]	72.32	72.32	70.30	60.07	73.32	72.32	71.31	62.42	72.82	73.15	70.81	59.23	72.65	72.65	71.48	58.22
		Cov. [%]	100.00	100.00	100.00	93.62	100.00	100.00	100.00	92.11	100.00	100.00	100.00	92.79	100.00	100.00	100.00	91.28
	Qwen-2-VL	Acc. [%]	70.13	70.81	71.14	69.30	69.80	70.30	70.64	69.97	69.63	70.47	70.64	68.96	67.62	35.24	70.47	70.97
		Cov. [%]	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	94.97	57.89	100.00	100.00
	Qwen-2.5-VL	Acc. [%]	78.36	78.36	74.50	74.33	78.52	78.19	75.17	73.49	77.18	78.02	75.67	74.16	77.85	79.53	73.99	72.65
		Cov. [%]	100.00	100.00	98.15	100.00	100.00	100.00	98.49	100.00	100.00	100.00	98.32	100.00	100.00	100.00	97.99	100.00
Line break	Gemma-3	Acc. [%]	34.90	33.73	33.89	32.55	35.24	35.74	34.23	34.23	34.23	33.39	34.56	32.22	33.22	33.05	34.06	33.89
		Cov. [%]	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
	LLaVA-1.5	Acc. [%]	38.42	22.82	37.08	37.42	37.25	33.22	36.58	37.25	38.76	30.20	36.41	37.25	37.75	0.17	36.07	11.07
		Cov. [%]	100.00	69.80	99.66	100.00	100.00	91.61	100.00	100.00	100.00	85.40	99.83	100.00	100.00	0.17	99.66	21.98
	LLaVA-OV	Acc. [%]	73.99	73.99	73.66	37.58	74.83	73.99	74.16	38.09	74.16	74.33	74.33	39.43	73.99	42.95	73.32	39.43
		Cov. [%]	100.00	100.00	100.00	95.47	100.00	100.00	100.00	98.49	100.00	100.00	100.00	98.32	100.00	61.91	99.83	98.83
	Phi-3.5	Acc. [%]	51.01	49.50	49.50	48.32	50.84	48.83	48.99	47.15	50.34	48.66	50.17	46.98	50.00	44.97	48.49	48.49
		Cov. [%]	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	91.28	100.00	99.83
	Phi-4	Acc. [%]	74.33	74.16	73.15	68.62	73.83	72.99	72.82	70.97	73.99	73.83	72.65	66.28	73.66	72.99	71.98	59.90
		Cov. [%]	100.00	100.00	100.00	98.83	100.00	100.00	100.00	99.33	100.00	100.00	100.00	98.15	100.00	100.00	99.66	91.11
	Qwen-2-VL	Acc. [%]	73.83	73.83	72.65	72.48	73.32	73.83	72.32	73.99	73.99	73.49	72.48	72.48	73.32	42.95	71.14	72.82
		Cov. [%]	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	99.50	67.62	100.00	100.00
	Qwen-2.5-VL	Acc. [%]	81.71	80.87	75.84	73.83	81.04	81.38	77.18	75.67	81.04	80.87	75.50	73.83	80.71	80.71	75.67	74.83
		Cov. [%]	100.00	100.00	98.66	100.00	100.00	100.00	98.15	100.00	100.00	100.00	98.66	100.00	100.00	100.00	97.82	100.00
Semicolon	Gemma-3	Acc. [%]	35.24	35.07	36.24	34.56	35.40	35.57	34.73	35.40	34.56	35.07	35.07	34.23	34.56	34.90	33.56	
		Cov. [%]	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	
	LLaVA-1.5	Acc. [%]	37.42	21.14	35.74	37.75	37.92	30.87	36.75	36.07	38.26	31.88	35.91	36.75	37.58	0.34	36.07	16.61
		Cov. [%]	100.00	64.60	100.00	100.00	100.00	87.25	100.00	100.00	100.00	85.57	100.00	100.00	100.00	0.34	99.83	36.24
	LLaVA-OV	Acc. [%]	72.48	72.15	71.81	50.17	73.83	73.32	74.16	42.62	73.83	72.99	73.32	47.99	73.83	5.87	73.66	43.79
		Cov. [%]	100.00	100.00	100.00	94.97	100.00	100.00	100.00	98.66	100.00	100.00	100.00	98.66	100.00	8.89	100.00	98.15
	Phi-3.5	Acc. [%]	45.30	48.83	47.48	47.32	48.99	48.49	48.15	47.32	49.33	48.99	48.83	46.81	49.66	34.73	47.32	47.48
		Cov. [%]	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	68.62	99.66	99.83
	Phi-4	Acc. [%]	72.65	71.48	69.97	59.06	73.15	71.64	70.64	61.91	72.65	71.98	71.98	57.05	73.15	71.48	71.48	60.24
		Cov. [%]	100.00	100.00	100.00	90.94	100.00	100.00	100.00	91.11	100.00	100.00	100.00	89.09	100.00	100.00	100.00	92.28
	Qwen-2-VL	Acc. [%]	71.31	73.66	71.14	71.81	72.65	73.15	71.48	71.14	72.65	72.82	71.81	70.97	70.81	36.41	71.48	71.64
		Cov. [%]	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	97.99	60.07	100.00	100.00
	Qwen-2.5-VL	Acc. [%]	80.20	80.87	75.00	74.83	80.20	80.37	75.50	75.00	80.37	81.04	75.34	75.34	81.04	81.04	76.01	74.16
		Cov. [%]	100.00	100.00	98.83	100.00	100.00	100.00	98.32	100.00	100.00	100.00	98.83	100.00	100.00	99.83	98.15	100.00

Appendix D Mechanistic Analysis of Tokenization

This section provides detailed evidence for the tokenization and decoding mechanisms driving prompt format sensitivity (Sec. 5). We analyze how model-specific tokenizers encode option IDs and how these representations translate into the tokens generated during decoding. In particular, we examine differences in tokenization length and atomicity at the input level, as well as the number of tokens required to generate option IDs during output. These analyses reveal how seemingly minor formatting choices alter the token sequences processed by the model and thereby influence multiple-choice prediction behavior.

D.1 Tokenizer overview

Tab. 17 provides an overview of the HuggingFace tokenizers used by the evaluated MLLMs. These models rely on different tokenizer families that reflect their pretraining corpora and design choices. In particular, several models use LLaMA- or Qwen-style tokenizers, while Gemma-3 and Phi-4 employ tokenizers solely used by them. These tokenizers differ in vocabulary construction, subword segmentation strategy, and the handling of special tokens, including image placeholders and control symbols. We explicitly adopt the original tokenizer of each model to preserve compatibility with pretrained checkpoints and to avoid introducing confounding effects in downstream evaluation.

Table 17: **Overview tokenizers.** We specify the tokenizer used in each MLLM by its HuggingFace tokenizer class name.

Model	HF tokenizer
Gemma-3	GemmaTokenizerFast
LLaVA-1.5	LlamaTokenizerFast
LLaVA-OV	Qwen2TokenizerFast
Phi-3.5	LlamaTokenizerFast
Phi-4	GPT2TokenizerFast
Qwen2-VL	Qwen2TokenizerFast
Qwen2.5-VL	Qwen2TokenizerFast

D.2 Tokenization of Option ID-Delimiter Pairs

In this section, we provide in-depth information about the input tokenization behavior of the four different tokenizers used by the MLLMs covered

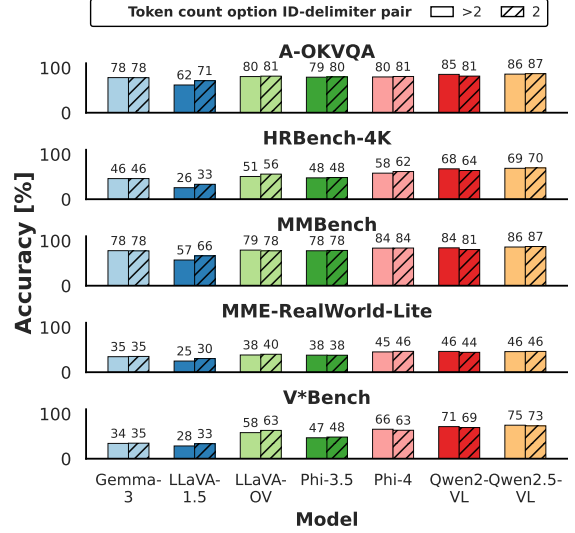


Figure 15: **Effect of option ID-delimiter tokenization length on accuracy.** Each pair of bars compares prompt formats where the option ID-delimiter pair tokenizes into exactly two tokens versus more than two tokens. Accuracy differences across these groups highlight how increases in tokenization length alter the input representation. Across several models and datasets, we observe only slight deviations across the two groups.

in this paper (see Tab. 17). This analysis consists of two parts: (1) the number of tokens of option ID-delimiter pairs and (2) whether the option ID token(s) are atomically tokenized, i.e., they are contained as the bare and not as corrupted tokens. Tab. 18 visualizes the tokenization results of both analyses and we highlight token counts of option ID-delimiter pairs higher than 2 in bold and cases where for at least one option ID is not atomically tokenized as grey cells.

Across tokenizers, simple delimiters such as dots, colons, and single brackets generally lead to stable encodings with an average of two tokens, corresponding to a single token for the option ID and a single token for the delimiter. In contrast, double-bracket delimiters or specific option ID sets frequently increase token counts beyond two and more often break atomic encoding of option IDs, particularly for numeric and Roman numeral formats. These effects are tokenizer-specific and reflect differences in subword vocabularies and special character handling.

To quantify how these encoding differences affect model performance, Fig. 15 compares accuracy across prompt formats grouped by tokenization length. Each pair of bars contrasts formats where the option ID-delimiter pair tokenizes into exactly two tokens versus more than two tokens.

We find that longer tokenizations often correlate

Appendix E Open-Ended Evaluation

To evaluate the alignment between multiple-choice and open-ended (i.e., free-form) evaluation schemes, we assess all MLLMs on A-OKVQA, HRBench-4K, and V*Bench using an open-ended VQA protocol. This analysis is motivated by the fact that many benchmarks are reported in multiple-choice form, while real-world deployment typically requires free-form generation. Therefore, it is essential for interpreting benchmark results to understand whether model behavior remains consistent across evaluation settings. In the open-ended setting, models are provided only with the image and the question. Thus, they are required to generate an answer without being exposed to candidate options. This removes inductive biases introduced by option ordering, tokenization effects, or information leakage from the answer options themselves that can arise in multiple-choice evaluation. We use a fixed prompt template (see Fig. 16) across all models to ensure comparability. For all experiments, we set the maximum number of generated tokens to 256 and disable sampling to obtain deterministic outputs. We perform evaluation by comparing generated answers against open-ended ground-truth annotations as described below.

E.1 Conversion to open-ended answers

MC-VQA datasets do not directly provide free-form ground-truth answers suitable for open-ended evaluation. To enable a consistent comparison across datasets and evaluation schemes, we construct open-ended ground-truth answers for A-OKVQA, HRBench-4K, and V*Bench. For V*Bench, we use the original open-ended

Prompt Layout: open-ended evaluation	
You are given an image and a question about that image.	
Question:	
What is the color of the "<object>"?	
Provide an answer based solely on the image.	

Figure 16: **Prompt template for open-ended VQA evaluation.** The model answers freely without being provided with answer options or additional contextual information, relying only on the image and the question.

annotations provided by the dataset. For A-OKVQA and HRBench-4K, we convert the correct multiple-choice option into a natural-language answer. This conversion is performed using *Qwen2.5-32B-Instruct*, conditioned on the question and the correct answer option. The conversion prompt enforces strict constraints to prevent the introduction of new information and to remove any reference to the original multiple-choice structure (see Fig. 17). All generated answers are manually verified to ensure semantic equivalence with the original correct option and to correct occasional formatting or phrasing artifacts. This procedure yields a set of open-ended ground-truth answers that are faithful to the original annotations while being suitable for free-form evaluation.

E.2 Correctness assessment

In the next step, we must assess the correctness of the free-form answer given by each MLLM in the open-ended evaluation scheme. Therefore, we use an LLM-as-a-judge approach with *Qwen2.5-32B-Instruct* as the LLM-judge. We define a prompt template (see Fig. 18), in which we provide the LLM detailed instructions, in-context examples, as well as the given free-form answer and the ground-truth open-ended answer. This setup constrains the evaluation model to assess only the semantic correctness of the generated answer relative to the ground-truth reference, ensuring consistent and objective judgments without introducing additional interpretation bias. We provide the open-ended results in Tab. 20.

Table 20: **Open-ended evaluation.** We compare standard MC-VQA with open-ended generation (difference reported in the brackets). Open-ended accuracy is consistently lower, and MC-VQA does not reliably correlate with open-ended rankings.

Model	Open-ended accuracy [%]		
	A-OKVQA	HRBench-4K	V*Bench
Gemma-3	32.01 (−45.98)	30.75 (−15.16)	31.15 (−3.26)
LLaVA-1.5	33.23 (−33.85)	20.00 (−9.90)	39.00 (+7.85)
LLaVA-OV	33.49 (−47.79)	13.25 (−40.61)	61.52 (−0.84)
Phi-3.5	34.45 (−44.93)	30.25 (−17.65)	46.86 (−0.60)
Phi-4	51.15 (−29.36)	42.00 (−19.13)	63.09 (−6.74)
Qwen2-VL	41.97 (−39.99)	50.75 (−13.61)	65.97 (−3.61)
Qwen2.5-VL	30.26 (−56.51)	46.00 (−23.61)	65.45 (−11.88)

E.3 Grounded MC-VQA evaluation

We next examine whether incorporating reasoning signals can stabilize model behavior under prompt format variations. Specifically, we test whether

conditioning multiple-choice predictions on a generated open-ended answer improves robustness. To this end, we construct an extended multiple-choice prompt that includes the open-ended response of the model as additional context (see Fig. 19). This grounded setting separates two components of the task: (i) generating a candidate answer in open-ended form, and (ii) selecting the corresponding option ID in the multiple-choice format. By providing the open-ended answer explicitly, the model no longer needs to rely solely on the option representation to recover its reasoning, which allows us to isolate the effect of option selection. This setup constrains the model to verify or correct its prior answer based on the image and the given options, rather than generating a decision solely from the prompt structure. As a result, any remaining performance differences across prompt formats reflect variability in the option-selection process rather than differences in underlying reasoning. The results of the grounded MC-VQA evaluation indicate that the context does not stabilize MC-VQA across prompt formats (see Tab. 21).

Prompt Layout: Conversion of multiple-choice VQA to open-ended evaluation setting

System Prompt (Answer Generator):
 You are a ground-truth answer generator. Your task is to convert a correct multiple-choice answer option into a clear, natural, open-ended answer. Always write a full, self-contained sentence. Never refer to option IDs or multiple-choice structure.

Rules:
 Use only the information contained in the provided correct option. Do not introduce additional explanations or external knowledge. Write the answer as an authoritative reference suitable as ground truth.

User Prompt:

[QUESTION]
 What is the color of the object?

[CORRECT ANSWER OPTION]
 "<correct option text>"

[INSTRUCTION]
 Write the ground-truth open-ended answer. Return only the answer text.

Figure 17: **Prompt template for generating open-ended ground-truth answers from multiple-choice VQA annotations.** A language model converts the correct answer option into a natural, self-contained sentence without referencing the original multiple-choice format.

Prompt Layout: Correctness assessment of open-ended answer via LLM-as-a-judge

System Prompt (Judge Model):

You are a strict and consistent evaluator for open-ended Visual Question Answering (VQA). Your task is to assess the correctness of a model-generated answer by comparing it to a ground-truth reference for the same question. Judge correctness, not style. Accept semantically equivalent answers. Never reveal chain-of-thought; provide only a brief rationale.

User Prompt:

Evaluate the VQA response.

[Abstract Few-Shot Examples]

Several example evaluations are provided to illustrate valid judgments, including cases of correct, partially correct, and incorrect answers. Each example specifies a question, a ground-truth answer, a model-generated answer, and the expected JSON verdict.

[Evaluation Instance]

QUESTION:

What is the color of the object?

GROUND_TRUTH_ANSWER:

"<reference answer>"

GIVEN_ANSWER (model output):

"<open-ended prediction>"

Return a valid JSON object containing: a correctness verdict, a numeric score, error categories, and a brief explanation. Output JSON only.

Figure 18: **Prompt template for LLM-as-a-judge evaluation of open-ended VQA.** A separate language model evaluates the correctness of a model-generated answer relative to a ground-truth reference, guided by abstract few-shot examples that define the evaluation criteria.

Prompt Layout: Grounded multiple-choice VQA evaluation

You are given an image and a question about that image. A previously generated open-ended answer is provided as additional context.

Question:

What is the color of the "<object>"?

Contextual Description (previous open-ended answer):

"<model-generated answer>"

Options:

- a. "<option_1>"
- b. "<option_2>"
- c. "<option_3>"
- d. "<option_4>"

Select the best answer to the above multiple-choice question based on the image. Respond with only the letter (e.g., a, b, c or d) of the correct option and no bracket, colon, or dot.

Figure 19: **Example prompt for grounded evaluation.** The model is conditioned on its own prior open-ended answer (context), which must be verified or corrected using visual evidence from the image.

Table 21: **Full grounded MC-VQA results.** We compare grounded MC-VQA (indicated by †) with standard MC-VQA across all models. Grounded MC-VQA does not improve evaluation robustness, as indicated by comparable or higher standard deviations.

Model	Option IDs	A-OKVQA - Acc. [%]				HRBench-4K - Acc. [%]				V*Bench - Acc. [%]			
		Mean	Std.	Min.	Max.	Mean	Std.	Min.	Max.	Mean	Std.	Min.	Max.
Gemma-3	Uppercase†	78.66	0.26	78.21	79.13	48.59	0.71	47.38	50.12	40.98	0.41	40.27	41.61
	Uppercase	78.15	0.38	77.49	78.62	45.80	0.74	44.25	46.88	34.63	0.72	33.22	35.40
	Lowercase†	78.76	0.21	78.49	79.08	48.42	0.46	47.88	49.25	40.84	0.50	40.10	41.44
	Lowercase	78.30	0.26	77.84	78.65	46.01	0.83	44.62	47.00	34.44	0.86	33.05	35.74
	Numbers†	78.59	0.16	78.28	78.78	47.77	0.65	46.62	48.88	41.02	0.58	40.10	41.95
	Numbers	77.89	0.30	77.40	78.30	45.15	0.76	43.50	46.38	34.82	0.75	33.89	36.24
	Roman†	78.32	0.24	77.77	78.65	48.00	0.63	47.00	48.88	40.03	0.30	39.60	40.60
	Roman	77.63	0.26	77.14	78.06	46.68	0.36	46.12	47.50	33.77	1.05	32.21	35.23
LLaVA-1.5	Uppercase†	76.35	2.28	72.16	79.08	35.73	1.39	33.00	37.38	40.38	0.37	39.60	41.11
	Uppercase	76.36	2.30	72.18	79.10	35.73	1.40	33.00	37.25	38.02	0.68	37.25	39.43
	Lowercase†	50.01	25.19	8.34	76.07	22.12	13.57	0.88	33.50	34.12	8.58	18.12	40.77
	Lowercase	50.32	25.43	8.28	76.05	23.15	13.40	0.88	34.12	19.17	12.58	0.17	33.22
	Numbers†	74.46	2.27	70.35	77.36	32.99	1.50	30.63	35.12	39.26	1.04	37.75	41.11
	Numbers	74.45	2.26	70.35	77.27	33.00	1.40	30.50	35.00	36.37	0.61	35.07	37.08
	Roman†	66.98	13.10	38.69	77.53	27.23	11.21	5.50	35.38	35.61	7.09	17.79	40.60
	Roman	67.20	13.21	38.60	77.58	27.74	10.77	5.62	35.50	31.03	9.97	11.07	37.75
LLaVA-OV	Uppercase†	90.64	0.38	89.85	91.07	64.78	0.62	64.00	65.88	73.13	0.62	71.98	73.83
	Uppercase	90.64	0.38	89.85	91.07	64.76	0.58	64.12	65.88	73.50	0.85	71.81	74.83
	Lowercase†	72.61	33.31	3.84	90.92	53.08	22.51	4.38	65.88	54.57	32.85	0.00	73.99
	Lowercase	72.61	33.31	3.82	90.90	53.06	22.53	4.38	65.88	58.50	27.20	0.67	74.33
	Numbers†	90.08	0.26	89.76	90.55	62.17	1.17	60.25	63.62	72.04	0.84	70.64	73.15
	Numbers	90.03	0.26	89.72	90.50	62.11	1.21	60.25	63.62	73.25	1.14	70.30	74.33
	Roman†	71.85	6.90	59.87	80.76	40.93	4.97	34.25	47.75	54.88	4.46	47.82	62.08
	Roman	71.86	6.92	59.85	80.72	40.92	4.97	34.25	47.75	44.20	4.88	37.58	52.01
Phi-3.5	Uppercase†	78.85	1.20	76.99	79.83	49.43	1.40	47.00	51.00	50.78	0.85	48.83	52.01
	Uppercase	79.69	1.25	76.92	81.07	49.15	1.42	46.88	50.75	49.11	1.72	45.30	51.01
	Lowercase†	77.90	3.14	72.40	79.91	47.12	3.69	39.12	50.25	48.24	7.57	29.53	52.68
	Lowercase	78.35	3.50	69.85	81.05	47.06	3.70	39.12	50.25	45.67	6.39	30.03	49.50
	Numbers†	79.73	0.45	79.21	80.17	48.22	1.57	45.25	50.38	50.89	1.04	49.50	53.02
	Numbers	80.04	0.65	79.21	81.03	48.22	1.57	45.25	50.38	48.03	1.23	46.14	50.17
	Roman†	79.35	0.37	78.97	79.85	47.09	1.67	45.12	49.62	50.27	1.57	48.15	52.85
	Roman	79.46	0.48	78.56	80.13	47.16	1.70	45.12	49.62	47.05	0.84	45.64	48.49
Phi-4	Uppercase†	82.76	0.13	82.66	82.90	63.19	0.86	62.12	64.50	70.92	0.87	69.63	72.48
	Uppercase	82.78	0.13	82.60	83.03	63.53	1.01	62.00	65.00	73.21	0.63	72.32	74.33
	Lowercase†	82.48	0.33	82.10	82.69	62.96	1.15	61.25	64.50	69.02	3.92	59.73	72.48
	Lowercase	82.60	0.17	82.34	82.86	63.73	1.16	62.00	65.50	72.58	0.88	71.48	74.16
	Numbers†	82.35	0.39	82.07	82.62	62.01	0.98	60.50	63.25	69.83	1.34	68.46	72.32
	Numbers	82.41	0.21	81.97	82.62	62.58	0.98	60.38	63.88	71.55	1.01	69.97	73.15
	Roman†	76.88	1.81	75.63	78.95	55.47	3.08	51.50	61.50	68.65	1.72	65.27	70.64
	Roman	74.24	3.48	68.30	79.10	54.67	3.45	49.62	61.38	62.00	4.36	57.05	70.97
Qwen2-VL	Uppercase†	82.97	5.72	68.21	86.97	63.51	2.07	58.38	65.75	71.59	2.39	67.95	74.66
	Uppercase	82.31	6.43	65.94	86.66	65.62	2.34	59.88	67.75	71.59	2.02	67.62	73.99
	Lowercase†	75.57	19.46	38.80	87.01	55.47	17.19	25.50	66.00	70.90	2.19	67.95	73.99
	Lowercase	74.58	20.72	35.76	86.83	57.00	18.53	25.62	68.12	63.91	15.66	35.23	73.83
	Numbers†	85.48	0.54	84.02	86.03	65.91	0.82	64.88	67.50	67.56	1.83	64.77	69.97
	Numbers	85.44	0.49	84.21	85.98	67.73	0.83	66.00	68.75	71.45	0.74	70.47	72.65
	Roman†	85.47	0.38	84.63	86.05	64.27	0.85	62.62	65.50	66.95	2.55	63.76	70.81
	Roman	85.53	0.29	84.98	85.94	67.09	0.78	65.88	68.38	71.38	1.48	68.96	73.99
Qwen2.5-VL	Uppercase†	87.66	0.38	87.07	88.17	70.72	0.67	69.50	71.50	73.90	1.14	72.15	75.00
	Uppercase	87.64	0.37	86.97	88.17	70.74	0.68	69.62	71.75	79.85	1.48	77.18	81.71
	Lowercase†	87.52	0.36	86.90	87.93	70.19	0.42	69.50	70.88	73.87	1.38	71.64	75.17
	Lowercase	87.51	0.32	86.99	87.99	70.04	0.39	69.25	70.50	80.10	1.24	78.02	81.38
	Numbers†	86.19	0.38	85.72	86.90	69.35	0.46	68.62	70.00	71.03	0.88	69.97	72.32
	Numbers	86.19	0.38	85.68	86.83	69.31	0.59	67.75	69.88	75.45	0.79	73.99	77.18
	Roman†	85.72	0.31	85.20	86.07	68.39	0.87	66.38	69.38	69.78	1.10	68.46	71.98
	Roman	85.73	0.34	85.13	86.16	68.36	0.81	66.50	69.12	74.34	0.84	72.65	75.67

Appendix F Additional Analyses

In this section, we provide detailed specifications and results for three analyses. Specifically, we report results for training exposure (Sec. 5.2, Sec. F.1), copy attention heads (Sec. 5.3, Sec. F.2), and models’ ability to infer effective prompt factors (Sec. 5.4, Sec. F.3).

F.1 Effects Go Beyond Training Exposure

In Sec. 5.2, we examine whether the performance of option ID sets in MC-VQA correlates with their exposure during model training. To this end, we introduce additional option ID sets that are unlikely to have appeared in multiple-choice contexts (text-only or vision-language) during pretraining or instruction tuning of the evaluated MLLMs.

These option ID sets differ in their degree of semantic leakage and structural properties (see Tab. 22). Specifically, Greek-letter IDs exhibit minimal semantic leakage and a clear ordinal structure. Directional words introduce moderate semantic leakage and relational structure but lack a strict ordering. Finally, word-based IDs have high semantic leakage and no inherent ordinal structure.

All IDs tokenize into a single token for Phi-4 and Qwen2.5-VL, which controls for decoding length effects and isolates the influence of ID semantics and structure on multiple-choice prediction.

Table 22: **Additional option ID sets grouped by semantic leakage and structural properties.** We categorize option ID sets by their degree of semantic leakage and inherent structure to disentangle training exposure effects from tokenization and decoding artifacts.

Option ID set	Semantic and structural properties	Example IDs
Greek letters	Minimal semantic leakage; strong ordinal structure.	alpha/beta/gamma/delta
Directional words	Moderate semantic leakage; relational semantics without strict ordinal structure.	north/east/south/west
Word-based IDs	High semantic leakage; no inherent ordinal or relational structure.	key/desk/tree/sea

In Fig. 5, we visualize the accuracy distribution of the additional option ID sets while varying the option delimiter and separator, resulting in 12 prompt formats per ID set. We evaluate this setting on V*Bench using Phi-4 and Qwen2.5-VL, as these are the only models for which all evaluated option IDs tokenize into a single token.

Tab. 23 summarizes the mean, standard deviation,

Table 23: **Performance of additional option ID sets on V*Bench.** We report mean, standard deviation, minimum, and maximum accuracy across 12 prompt format variants for alternative option ID sets with varying semantic leakage. Greek-letter IDs achieve accuracy comparable to standard letter- and number-based IDs, while direction- and word-based IDs show substantially lower performance and higher variance.

Model	Option IDs	Accuracy [%] ↑			
		Mean	Std.	Min.	Max.
Phi-4	Greek	70.90	1.72	68.13	73.15
	Direct.	49.18	11.67	12.42	54.70
	Words	44.97	12.88	29.20	68.96
	Uppercase	73.21	0.63	72.32	74.33
	Lowercase	72.58	0.88	71.48	74.16
	Numbers	71.55	1.01	69.97	73.15
	Roman	62.00	4.36	57.05	70.97
Qwen2.5-VL	Greek	76.69	1.33	74.83	78.69
	Direct.	65.27	4.14	57.21	72.65
	Words	60.75	9.44	44.13	71.81
	Uppercase	79.85	1.48	77.18	81.71
	Lowercase	80.10	1.24	78.02	81.38
	Numbers	75.45	0.79	73.99	77.18
	Roman	74.34	0.85	72.65	75.67

and range of accuracy across the 12 prompt formats for each ID set. Greek-letter IDs achieve performance that is close to standard uppercase, lowercase, and numeric IDs for both models, with relatively low variance across prompt formats. This result indicates that effective multiple-choice selection does not depend on exposure to conventional letter-based IDs during training.

In contrast, directional words and word-based IDs lead to substantially lower average accuracy and markedly higher variance. These IDs introduce semantic content that is unrelated to the multiple-choice structure, which appears to interfere with reliable option selection. Overall, these results show that option ID performance cannot be explained by training exposure alone: IDs with minimal semantic leakage and clear structure (such as Greek letters) generalize well, while semantically loaded IDs degrade performance even when tokenization effects are controlled.

F.2 Copy Attention Heads

In Sec. 5.3, we show that the attention in specific attention heads, which we call "copy" attention heads, varies across different option ID sets in Qwen2.5-VL. In this subsection, we will give more detailed results and more in-depth information about the procedure of how we obtained these values for our analysis. Our analysis scope covers the four option

ID sets "Uppercase", "Lowercase", "Numbers", and "Roman", while we have option delimiter fixed as "colon" and option separator as "line break".

Algorithm 1 Extraction of option ID attention margin

Require: Prompt x , image I , options $\mathcal{O}$, model f_θ

- 1: Construct conversation input (x, I)
- 2: Run forward pass $\rightarrow$ obtain attention $\{A_{l,h}\}$
- 3: Identify query position q

4: **Step 1: Extract option ID token positions**

5: **for all** $o \in \mathcal{O}$ **do**

6: $\mathcal{T}_o \leftarrow$ token positions corresponding to o

7: **end for**

8: **Step 2: Compute attention scores**

9: **for** $l = 1, \dots, L$ **do**

10: **for** $h = 1, \dots, H$ **do**

11: Compute option scores:

12: **for all** $o \in \mathcal{O}$ **do**

13: $a_{l,h}(o) \leftarrow \text{EXTRACTATTN}(q, \mathcal{T}_o)$

14: **end for**

15: $s_{l,h} \leftarrow \text{MARGIN}(a_{l,h}(o))$

16: **end for**

17: **end for**

18: **return** $S \in \mathbb{R}^{L \times H}$

algorithm 1 summarizes the procedure used to extract option-identifier attention signals for a single V*Bench sample. For each sample, we construct the model-specific multimodal prompt, including the image and multiple-choice options, and run a forward pass through the model to obtain attention maps for all layers and heads.

We identify the query position q corresponding to the generated option identifier token and determine the token positions $\mathcal{T}_o$ associated with each option identifier $o \in \mathcal{O}$ in the input sequence. For each attention layer l and head h , we compute the average attention mass assigned to the tokens of each option identifier:

$$\text{EXTRACTATTN}(q, \mathcal{T}_o) = \frac{1}{|\mathcal{T}_o|} \sum_{t \in \mathcal{T}_o} A_{l,h}(q, t),$$

where $A_{l,h}(q, t)$ denotes the attention weight from the query position q to token position t .

To quantify how decisively the model attends to a single option identifier, we compute an attention margin for each layer-head pair as the difference between the highest and second-highest option at-

tention scores:

$$\text{MARGIN}(a_{l,h}(o)) = \max_{o \in \mathcal{O}} a_{l,h}(o) - \max_{o \in \mathcal{O}'} a_{l,h}(o)$$

where o^* denotes the option with the highest attention score and $\mathcal{O}' = \mathcal{O} \setminus \{o^*\}$ represents the option set without the option with the highest attention score. The resulting margin tensor $S \in \mathbb{R}^{L \times H}$ provides a layer- and head-wise measure of how sharply attention is concentrated on a single option identifier during decoding.

We next identify attention heads that exhibit copy-style behavior in MC-VQA. For each evaluation with a fixed option ID set, we average the margin tensor S across all dataset samples. We then select the top-10 attention heads with the largest average margins for each option ID set and analyze the overlap of these sets across evaluations. Attention heads that consistently appear among the top-10 across all option ID sets are identified as copy attention heads. Using this criterion, we identify four such heads at layer-head indices (14, 6), (17, 8), (20, 8), and (24, 21). Fig. 20 reports the average attention margin across all attention heads for each option ID set, with the identified copy attention heads highlighted. In Tab. 4, we report the cumulative attention margin across all four copy attention heads.

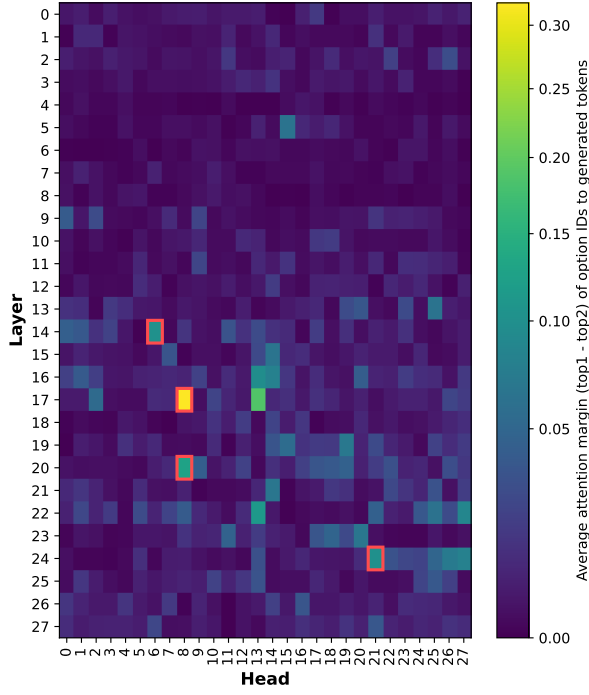
F3 MLLMs Can Reason About Favorable Prompt Formats For Multiple-Choice

In Sec. 5.4, we analyze whether MLLMs can anticipate which prompt formats are more favorable for their own multiple-choice performance. To this end, we present each model with a prompt that contains the explicit design space of semantically identical MC-VQA prompt formats options (see Fig. 21). Without access to any experimental results, the model is asked to predict which combination of formatting choices is most likely to yield

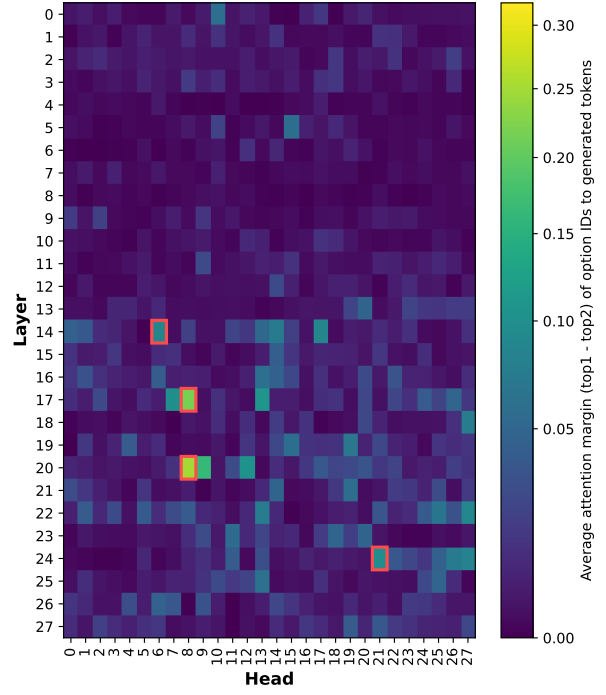
Table 24: **Model-selected prompt formats for additional ID sets.** Each model selects a preferred option ID set, delimiter, and separator from the design space in Tab. 1. Most models choose configurations with strong ranks on V*Bench, which suggests that prompt format preferences are systematic and partly recoverable from the models themselves. Lower average rank indicates better performance.

Model	Option IDs	Delimiter	Separator	Avg. Rank	Rel. gap to optimal
Phi-4	Greek	Colon	Comma	13.0	-6.88
Qwen2.5-VL	Greek	Colon	Linebreak	4.0	-1.28

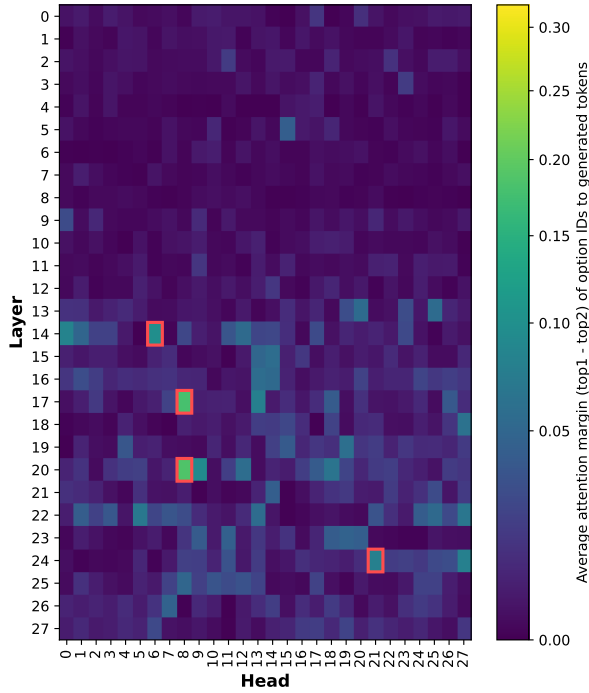
the highest accuracy based solely on its internal processing characteristics. As shown in Tab. 5, all MLLMs except Phi-4 select similar or identical prompt formats, and for each of these models the selected format performs significantly better than random guessing. We observe similar results for the additional option ID sets (see Tab. 22), where Phi-4 and Qwen2.5-VL both select best-performing Greek option IDs (see Tab. 24).



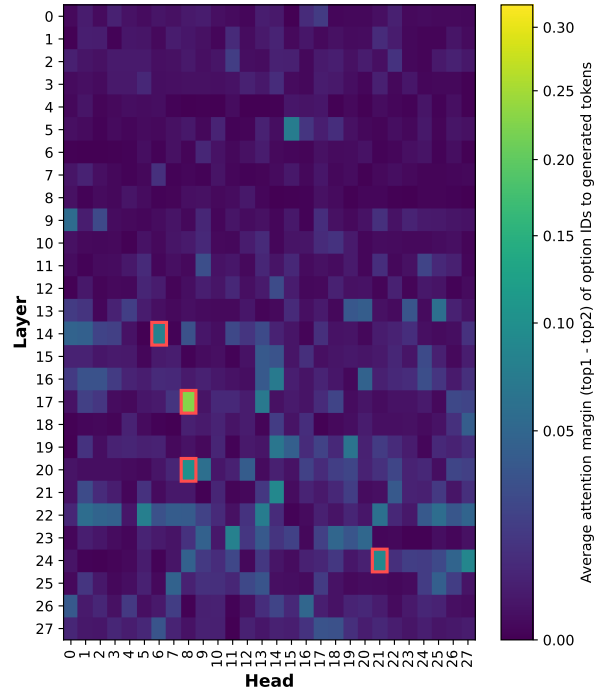
(a) Uppercase letter option IDs



(b) Lowercase letter option IDs



(c) Number option IDs



(d) Roman numeral option IDs

Figure 20: **Average option ID attention per attention head in Qwen2.5-VL.** We show the average attention margin between top-1 and top-2 attention values across all option ID tokens. We vary the option ID set to uppercase letters (a), lowercase letters (b), numbers (c), and Roman numerals (d), while keeping the option delimiter (colon) and option separator (line separator) fixed. We highlight the selected "copy" attention heads in red.

Prompt Layout: Model self-selection of optimal MC-VQA prompt format

System Prompt:

You are a multimodal large language model evaluated on multiple-choice Visual Question Answering (VQA). Your task is to reason about how different prompt formatting choices affect your own performance.

All prompt formats are semantically identical and differ only in superficial formatting choices. These differences do not change the meaning of the question or the answer options, but they may influence your accuracy.

Design Space:

The prompt formats vary along three independent factors.

Option ID set (answer labels):

- Uppercase letters: A / B / C / D
- Lowercase letters: a / b / c / d
- Numbers: 1 / 2 / 3 / 4
- Roman numerals: I / II / III / IV

Option delimiter (ID-option separator):

- Colon: "A: Paris"
- Dot: "A. Paris"
- Bracket: "A) Paris"
- Double brackets: "(A) Paris"

Option separator (between options):

- Line break: each option on a new line
- Comma: "A) Paris, B) Berlin"
- Semicolon: "A) Paris; B) Berlin"

Task:

Based on your internal processing characteristics, predict which combination of formatting choices is most likely to yield the highest multiple-choice accuracy for you. You are not given any experimental results. Base your decision solely on your understanding of how you process prompts.

Output format:

Return only a valid JSON object with the following fields:

```
{"best_option_id_set": one of ["Uppercase", "Lowercase", "Numbers", "Roman numerals"],  
"best_option_delimiter": one of ["Colon", "Dot", "Bracket", "Double brackets"],  
"best_option_separator": one of ["Line break", "Comma", "Semicolon"],  
"confidence": a number between 0 and 1,  
"rationale": "Brief explanation (1-3 sentences)"}  

```

Do not output anything else besides valid JSON.

Figure 21: **Prompt template for model self-selection of optimal multiple-choice formatting.** The model is asked to predict which semantically neutral prompt format configuration best matches its own internal processing characteristics, without access to experimental results. This setup tests whether prompt format sensitivity is structured and partly accessible to the model itself.