Promptception: How Sensitive Are Large Multimodal Models to Prompts?

Anonymous ACL submission

Abstract

Despite the success of Large Multimodal Models (LMMs) in recent years, prompt design for LMMs in Multiple-Choice Question Answering (MCQA) remains poorly understood. We show that even minor variations in prompt phrasing and structure can lead to accuracy deviations of up to 15% for certain prompts and models. This variability poses a challenge for transparent and fair LMM evaluation, as models often report their best-case performance using carefully selected prompts. To address this, we introduce **Promptception**, a systematic framework for evaluating prompt sensitivity in LMMs. It consists of 61 prompt types, spanning 15 categories and 6 supercategories, each targeting specific aspects of prompt formulation, and is used to evaluate 10 LMMs ranging from lightweight open-source models to GPT-4o and Gemini 1.5 Pro, across 3 MCQA benchmarks: MMStar, MMMU-Pro, MVBench. Our findings reveal that proprietary models exhibit greater sensitivity to prompt phrasing, reflecting tighter alignment with instruction semantics, while open-source models are steadier but struggle with nuanced and complex phrasing. Based on this analysis, we propose Prompting Principles tailored to proprietary and open-source LMMs, enabling more robust and fair model evaluation.

Figure 1: Categorization of prompts proposed in our Promptception framework. It consists of 61 prompt types, spanning 15 categories (e.g. Answer Handling, Penalty-Based Prompts, Poor Linguistic Formatting) and 6 supercategories (e.g. Task-Specific Instructions, Choice Formatting and Presentation), providing a comprehensive evaluation framework for assessing LMM sensitivity to prompts.

1 Introduction

Recent advancements in Large Multimodal Models (LMMs) have significantly improved their ability to integrate vision and language, enabling strong performance on a range of reasoning tasks involving textual and visual information (Radford et al., 2021; OpenAI et al., 2024; Chen et al., 2025). These models take visual cues (single image, multiple images, and video) and text as input, to output a textual response. They have been fine-tuned on a variety of tasks, including captioning, visual question-answering (VQA) (Li et al., 2025), visual grounding (Rasheed et al., 2024; Munasinghe et al., 2025), and temporal grounding (Ren et al., 2024).

However, despite these successes, the systematic study of sensitivity of LMMs remains an open question. While the impact of texual prompt variations has been extensively studied in Large Language Models (LLMs) (Alzahrani et al., 2024b; Balepur et al., 2024) as well as CLIP-based Vision-Language Models (VLMs) (Wu et al., 2023), its effects on LMMs have not been systematically explored. Understanding how these models respond to textual modifications is crucial for assessing their reliability in multimodal tasks.

Visual Question Answering (VQA) has emerged as a fundamental evaluation framework for assessing the capabilities of LMMs (Agrawal et al., 2016). VQA tasks are typically structured as either open-

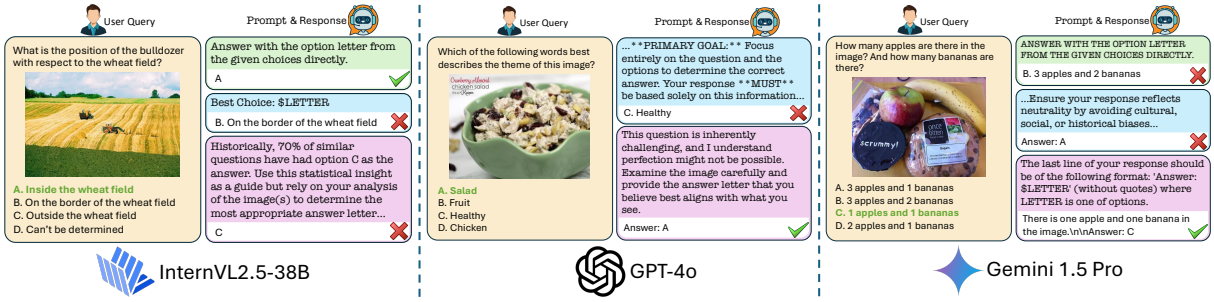

Figure 2: Sensitivity of state-of-the-art LMMs to prompt variations. Examples from the MMStar benchmark illustrating divergent model outputs despite identical user queries, caused solely by changes in prompt phrasing (*Left*: InternVL-38B, *Middle*: GPT-4o, *Right*: Gemini 1.5 Pro). This demonstrates the models' sensitivity to how instructions are framed.

ended or multiple-choice questions (MCQ) (Chen et al., 2024; Li et al., 2024). While open-ended responses provide flexibility, they pose challenges in evaluation due to ambiguity and the need for complex answer-matching techniques. Consequently, MCQ formats are widely adopted in recent image and video benchmarks (Zhang et al., 2025), offering a structured approach to evaluation.

Despite the advantages of MCQ-based evaluations, LMMs exhibit sensitivity to subtle variations in prompt phrasing, raising concerns about the consistency and stability of benchmark results, as reflected in the varied model responses shown in Figure 2. In this study, we systematically investigate the prompt sensitivity of LMMs by evaluating 8 open-source and 2 proprietary models across 3 multiple-choice question-answering (MCQA) benchmarks covering both image and video modalities. Specifically, we analyze performance variations using 61 systematically designed prompts, categorized into 15 categories and 6 broader supercategories (Figure 1). Our goal is to analyse the impact of prompt formulation on model accuracy and benchmark stability, providing insights into best practices for evaluating LMMs on MCQA.

The contributions of this paper can be summarized as follows:

- **Comprehensive Prompt Sensitivity Analysis:** We present the most extensive study to date on the impact of prompt variations across diverse multimodal benchmarks and LMM architectures. To facilitate this study, we introduce Promptception, a systematic evaluation framework comprising 61 prompt types, organized into 15 categories and 6 supercategories, each designed to probe specific aspects of prompt formulation in LMMs.
- **Evaluation Across Models, Modalities, and Benchmarks:** We assess prompt sensitivity across a diverse set of model sizes and architectures, including both open-source and proprietary LMMs. Our analysis spans multiple modalities and benchmarks; MMStar (single image), MMMU-Pro (multi-image), and MVBench (video) and we further evaluate sensitivity across various question dimensions within these benchmarks to ensure a comprehensive understanding.
- **Best Practices for Prompting:** We identify key trends in prompting and propose Prompting Principles for effective and consistent evaluation of LMMs.

2 Experimental Setup

2.1 Visual MCQA Task Definition

The MCQA task (Robinson et al., 2023) is defined as follows. The LMM is given a question Q, a set of four (or more) choices $\mathcal{C} = \{c_a, c_b, c_c, c_d\}$, exactly one of which is correct (i.e., gold choice $c_g \in \mathcal{C}$), the prompt P (shown in red) along with the visual (image(s) or video) input V as shown in Figure 3. Using these inputs, the LMM should give the letter of the correct option $a \in \{A, B, C, D\}$.

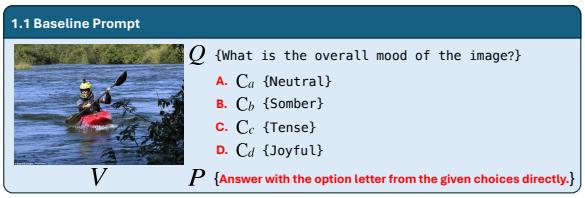

Figure 3: Baseline Prompt. This serves as the simplest prompt for the MCQA task.

All of our evaluations are zero-shot, with no modifications made to Q, $\mathcal{C}$, or V.

2.2 Models

We evaluate a diverse set of LMMs, including both open-source and proprietary models. The models tested are LLaVA-OneVision-7B (Li et al., 2025), Qwen2-VL-7B-Instruct (Wang et al., 2024), InternVL2.5-1B (Chen et al., 2025), InternVL2.5-8B, InternVL2.5-38B, MiniCPM-V-2.6-8B (Yao et al., 2024), Llama-3.2-11B-Vision (Patterson et al., 2022), Molmo-7B-D-0924 (Deitke et al., 2024), GPT-4o (`gpt-4o-2024-08-06`) (OpenAI et al., 2024), and Gemini 1.5 Pro (`gemini-1.5-pro-latest`) (Team et al., 2024).

All open-source models were implemented using the Hugging Face Transformers library (Wolf et al., 2020) and executed on NVIDIA A100 40GB GPUs. Proprietary models, GPT-4o and Gemini 1.5 Pro, were accessed via API with a temperature setting of 0. For video-based tasks, frame sampling strategies were applied according to model-specific configurations, detailed in the Appendix C.1. The evaluation protocol for answer extraction from LMM response is detailed in Appendix C.

2.3 Datasets

To evaluate the multimodal reasoning capabilities of LMMs, we use three benchmarks: MMStar (Chen et al., 2024), MMMU-Pro (Yue et al., 2024b), and MVBench (Li et al., 2024). These datasets cover single-image, multi-image, and video-based multiple-choice question-answering tasks, assessing different aspects of vision-language understanding.

MMStar is a vision-indispensable benchmark designed to eliminate reliance on textual priors and data leakage. It consists of 1,500 carefully curated MCQs that require genuine visual reasoning. MMStar evaluates six core multimodal capabilities across 18 axes and introduces metrics to quantify data leakage and multimodal performance gains.

MMMU-Pro is a refined version of the MMMU benchmark (Yue et al., 2024a), addressing text-only biases. It contains 1,730 MCQs across 30 subjects and applies three improvements over MMMU: filtering out text-answerable questions, increasing answer choices from four to ten, and introducing a vision-only input setting, where questions appear as images rather than structured text. We evaluate all models on the 4-choice (s4) question type. Additionally, InternVL2.5-8B and Gemini 1.5 Pro are further evaluated on the 10-choice (s10) and vision-only (v) formats introduced in this benchmark.

MVBench evaluates temporal reasoning in video-based multimodal models through 20 carefully designed tasks using a static-to-dynamic transformation, ensuring that questions require multi-frame understanding and cannot be answered from a single frame. The full dataset comprises 4,000 multiple-choice QA pairs (200 per task). All open-source models were evaluated on the entire dataset, while GPT-4o and Gemini 1.5 Pro were evaluated on a representative subset of 100 videos (5 per task) due to the high cost of API access.

2.4 Metric Definitions

2.4.1 Trimmed Mean ($\tilde{\mu}$)

The Trimmed Mean (10%) is a robust measure of central tendency that mitigates the impact of extreme values by removing the lowest and highest 10% of data points before computing the mean. This approach enhances the reliability of performance comparisons by reducing the influence of outliers while preserving the overall trend in the data.

Given a sorted dataset of N values: $X_1, X_2, ..., X_N$, discard the lowest and highest **10%** of values (rounded to the nearest integer) and compute the mean of the remaining values as follows:

$$\tilde{\mu} = \frac{1}{N-2k} \sum_{i=k+1}^{N-k} X_i, \quad (1)$$

$$\text{where } k = round(0.10\,N)$$

2.4.2 Percentage Relative Accuracy (PRA)

PRA measures the improvement or decline in performance relative to a baseline accuracy. This metric provides a normalized way to evaluate accuracy changes, enabling comparisons across different models, and datasets. By aggregating accuracy values across different models and datasets, it helps derive global insights, allowing for a more comprehensive evaluation of overall trends and prompt effectiveness.

Given a baseline accuracy value, denoted as X_b, the **PRA** with respect to the baseline is:

$$PRA_{baseline} = \frac{X}{X_b} \times 100 \quad (2)$$

To quantify the relative change in performance, whether a gain or a drop, we also use the **Percentage Relative Accuracy Difference (PRAD)**, defined as follows:

$$PRAD_{baseline} = \frac{X - X_b}{X_b} \times 100 \quad (3)$$

3 Prompts

In this section, we introduce the prompts proposed in our Promptception evaluation framework, each designed to examine different aspects of prompt engineering and its influence on model responses in the MCQA task. Table 1 outlines the categories of prompts, the specific modifications applied in their design, and an illustrative prompt type for each category. In all cases, the prompts are appended after the question and answer choices, following the structure of the baseline prompt (Figure 3), except for Categories 2 and 3. The full list of prompts for each category is provided in Appendix A.

4 Results & Analysis

4.1 Overall Trend

To provide a robust assessment of model performance across different benchmarks, we calculate the trimmed mean - $\tilde{\mu}$ (Equation 1) for each dataset using a 10% trimming rate, chosen based on empirical observation. Table 2 shows the accuracy of the baseline prompt (Figure 3) and trimmed mean accuracy for each model and benchmark. We consider baseline to be the simplest prompt to assess the model in an MCQA setting. For open-source models, the baseline accuracy exceeds the trimmed mean accuracy, indicating that the baseline prompt is inherently strong. In contrast, for proprietary models (GPT-4o & Gemini 1.5 Pro), the baseline accuracy falls below the trimmed mean, indicating that other prompts generally yield better performance likely due to superior instruction-following capabilities and ability to handle more complex prompt formulations.

Model	MMStar		MMMU-Pro		MVBench	
	$\tilde{\mu}$	Base	$\tilde{\mu}$	Base	$\tilde{\mu}$	Base
LLaVA-OV-7B	60.5	61.5	43.0	43.1	56.6	56.5
Qwen2-VL-7B	55.6	56.0	44.2	45.8	66.0	66.2
MiniCPM-V 2.6	52.8	52.9	39.0	41.8	52.3	53.9
Llama-3.2-11B-Vision	49.2	49.7	-	-	-	-
Molmo-7B-D-0924	53.2	55.9	-	-	-	-
InternVL2.5-1B	43.6	50.0	33.1	36.6	58.4	60.6
InternVL2.5-8B	61.6	62.5	47.8	49.5	68.2	68.3
InternVL2.5-38B	67.2	68.5	57.9	59.3	71.3	70.8
GPT-4o	55.5	53.5	57.8	53.5	60.8	59.0
Gemini 1.5 Pro	53.3	51.5	57.0	58.3	53.4	52.2

Table 2: Comparison of Trimmed Mean ($\tilde{\mu}$) and Baseline (Base) Accuracy of models across benchmarks. For MMMU-Pro, results are reported on the s4 question type.

Appendix B provides the comprehensive list of accuracies across all prompts and benchmarks for each model.

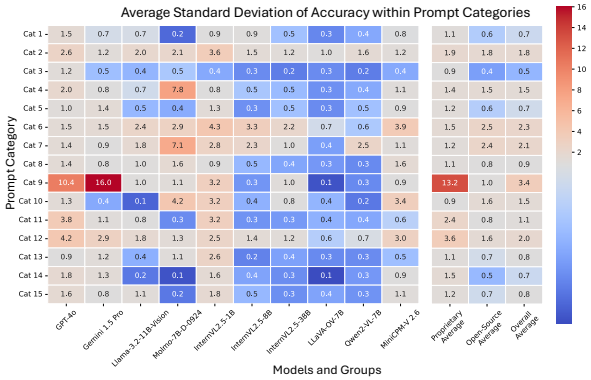

Figure 4: Standard Deviation of prompts within categories averaged across Benchmarks for each Model. Proprietery models show high standard deviation within all categories.

4.2 How Does Variation in Prompts Impact Accuracy?

In this section, we analyze how different prompt categories and types within each category influence model performance. We highlight the most sensitivity-prone categories, and identify the prompt types that consistently yield the highest and lowest accuracies.

4.2.1 Which Prompt Categories Are Sensitive to Variations in Prompting?

Identifying the most sensitive prompt categories is crucial for optimizing model performance. This is done by computing the average standard deviation within each category for each model, highlighting variations in accuracy due to prompt phrasing. Categories with higher standard deviations indicate greater sensitivity, where changes in how prompts are formulated significantly impact model responses. Figure 4 presents the average standard deviation per category, computed across all three benchmarks. For GPT-4o and Gemini 1.5 Pro, results from MVBench were excluded, as only a subset was used. To identify highly sensitive categories, we aggregated all standard deviation values across models and categories and set the threshold at the median (0.78; Appendix H).

To better understand the effect of prompts, we categorize them based on their instructional intent into 3 groups: positive (encouraging framing), neutral (objective or factual framing), and negative (misleading or adversarial framing). This classification focuses solely on the semantic framing of the prompt and does not consider structural complexity or length, except in Category 5, which specifically

Super Category	Category	Modification	Example Prompt Type
Choice Formatting and Presentation	1: Formatting Variations in Choice Presentation	Answer choice letter. Type 1.3: Option <LETTER>:	What design element best describes the image? <image> **Option A:** Composition **Option B:** Perspective **Option C:** Balance **Option D:** Shape **Answer with the option letter from the given choices directly.**
	2: Structured Formatting	Explicit structure the question and choices are presented. Type 2.2: Question & Answer Prefix	**Question:** What design element best describes the image? <image> **Options:** **A.** Composition **B.** Perspective **C.** Balance **D.** Shape **Answer with the option letter from the given choices directly.**
	3: Prompt Position Changes	Relative positioning of prompt, question, and choices. Type 3.2: Middle	What design element best describes the image? <image> **Answer with the option letter from the given choices directly.** **A.** Composition **B.** Perspective **C.** Balance **D.** Shape
Linguistic and Stylistic Challenges	4: Poor Linguistic Formatting	Grammatical errors, misspellings, and inconsistencies in wording. Task 4.4: Poor Formatting	answer.with;the.option: letter from-choices.directly!
	5: Effect of Prompt Length	Prompt length, from concise to verbose. Task 5.2: Medium Prompt	Your task is to examine the given image(s) and determine which of the listed options accurately answers the question. Carefully analyze the image(s), consider the possibilities, and then respond only with the correct option $LETTER from the given choices.
Thought Process and Reasoning	6: Chain of Thought (CoT) Promp	Multi-step reasoning and explicit logical breakdowns. Task 6.1: Step-by-Step Reasoning	Answer the preceding multiple-choice question. Think step by step before answering. Respond only with: '$LETTER' of the best answer choice.
	7: Ambiguity and Probabilistic Prompts	Introducing uncertainty, probabilistic reasoning, and deferred context. Task 7.1: Add Probabilistic Language	Based on our analysis of the image(s), there is a 25% chance that the correct answer is option A, but the other options may still be valid. Carefully evaluate the image(s) and choose the answer letter you believe is most likely to be correct.
Context-Aware and Ethical Guidance	8: Target Audience Prompts	Tailoring explanations to different audience levels. Task 8.1: Explain to a Child	Explain the answer as if you're speaking to a curious 10-year-old. Use simple words, short sentences, and relatable examples that a child would easily understand. Avoid technical terms or complex concepts. Respond only with: '$LETTER' of the best answer choice.
	9: Roleplay Scenarios	Assigning specific personas. Task 9.2: Act as a Careless Student	From now on, please act as a careless student. When answering questions based on images, you often overlook key visual details or misinterpret the question. You rush through the process and select answers without fully analyzing the image(s), leading to frequent mistakes due to carelessness. Respond only with: '$LETTER' of the best answer choice.
	10: Testing Bias and Stereotypes	Detecting and mitigating biases. Task 10.2: Contextual Neutrality	Consider the image(s) and question provided within the specific context they represent. Ensure your response reflects neutrality by avoiding cultural, social, or historical biases. Base your decision solely on the visible elements and the question's requirements, maintaining a context-aware but unbiased approach. Respond only with: '$LETTER' of the best answer choice.
Task-Specific Instructions	11: Focus-Driven Prompts	Directing attention to specific elements. Task 11.1: Strong Focus on Image Analysis	**TASK:** Examine the image(s) meticulously, focusing on every detail and visual element to identify the correct answer. **PRIMARY GOAL:** Focus strictly on the image(s) and you **MUST** base your analysis solely on the content. Answer with the option letter from the given choices directly.
	12: Answer Handling	Expected response format. Task 12.1: Answer Handler 1	Answer the preceding multiple-choice question in the following format: "Answer: $LETTER" (without quotes) where LETTER is one of the options.
Performance, Feedback, and Penalty	13: Penalty-Based Prompts	Penalties for incorrect answers, formatting violations etc. Task 13.1: Penalties for Mistakes	**Warning:** An **incorrect answer** will result in a **"strict penalty"** being applied. Carefully examine all details in the image(s), analyze the question thoroughly, and select your response with precision. Accuracy is **"non-negotiable"**, so take your time and avoid errors. Ensure your response follows this format: Answer: $LETTER
	14: Incentive-Based Prompts	Incorporating rewards and positive reinforcement. Task 14.1: Incentive-Based Prompts	Imagine you're competing for a generous tip of $100 for delivering a flawless and accurate answer. Carefully analyze the image(s) and the question provided, paying attention to every detail and nuance. Respond only with: "$LETTER" of the best answer choice. No explanations are needed-just focus on accuracy to secure the reward.
	15: Competitive Framing Prompts	Framing the task as a competition or challenge. Task 15.1: Outperforming a Competitor	You are tasked with solving this challenge both faster and more accurately than any other contender. Analyze the question carefully, eliminate errors, and provide the correct option letter as your answer. Strive for speed and precision to secure your win. Respond confidently in the format: $LETTER.

Table 1: Overview of Promptception, a prompt sensitivity framework for visual LMMs. The last column shows example prompts from each category used with MCQs as inputs to LMMs.

investigates the impact of prompt length (Table 3). This classification helps clarify model behavior. An ideal model should excel with positive prompts, perform reasonably under neutral conditions, and struggle with negative prompts.

Positive Prompts: 2.3-2.9, 6.1-6.4, 9.1, 10.1, 10.2, 11.4, 11.5, 13.1-13.4, 14.1-14.3, 15.1-15.3

Neutral Prompts: 1.1-1.3, 2.1, 2.2, 3.1, 3.2, 5.1, 8.1-8.4, 12.1-12.5

Negative Prompts: 2.6, 4.1-4.5, 5.2, 5.3, 7.1-7.6, 9.2, 9.3, 11.1-11.3

Prompt Intent	Example Prompt Type
Positive	Persona: You are an advanced multimodal AI assistant specializing in analyzing visual content and answering questions based on it. Instructions: - Analyze the given image(s)/video and understand its content. - Read the question carefully. - Choose the most appropriate answer from the options provided. (2.8)
Neutral	Answer with the option letter from the given choices directly. (1.1)
Negative	Based on our analysis of the image(s)/video, there is a 25% chance that the correct answer is option A, but the other options may still be valid. Carefully evaluate the image(s)/video and choose the answer letter you believe is most likely to be correct. (7.1)

Table 3: Example prompt types illustrating the three categories of Prompt Intent: Positive, Neutral, and Negative.

Since open-source and proprietary models exhibit substantial differences in instruction following ability, we analyze them separately. Certain open-source models exhibit notable accuracy drops for specific prompts, a topic that will be explored further in Appendix G. Consequently, for open-source models, a prompt category is considered highly sensitive if at least 5/8 models display a standard deviation greater than the threshold (0.78). Under this criterion, Categories 2, 6, 7, 8, 9, and 12 qualify as high-sensitivity prompt categories. Notably, several of these contain a mix of positive and negative prompts (e.g., Category 2 and 9), which amplifies intra-category variability and contributes to higher standard deviations. Conversely, for proprietary models, all prompt categories exhibit a standard deviation greater than the threshold. This suggests that prompt selection is crucial across all categories when using proprietary models, as the choice of prompt significantly affects performance.

4.2.2 Which Prompts Enhance or Hinder Model Performance?

In this section, we analyze the prompts that are most effective for MCQA task for LMMs. As before, we separate the analysis into open-source and proprietary models.

Models exhibit different accuracy ranges across benchmarks. To enable a unified comparison, we normalize them to the same scale using percentage relative accuracy (PRA), as per Equation 2, with respect to the accuracy of the baseline prompt for the model on the given benchmark. Then, for each prompt type, the values are averaged across models and benchmarks separately for open-source and proprietary models (Figure 5 & 6). For open-source models, percentage relative accuracies below 80% were excluded from the averaging process, as they represent model-specific extreme cases (dicussed in Appendix G). The deviation from the baseline (Equation 3) was then considered to generate the figures 9 & 10 (Appendix D), which highlight the best and worst-performing prompts within each category.

For open-source models, prompt 1.2, 2.1, 2.2, 3.1, 3.2, and 11.5 consistently outperformed the baseline, indicating their effectiveness in improving model accuracy. Additionally, prompt 1.3, 4.2, 5.1, 11.1, 11.4, and 12.1, though slightly below the baseline, remained within a close range, suggesting they are still viable prompting strategies. Conversely, prompts such as 2.9, 6.2, 7.3, 10.1, 12.3, and 15.2 consistently resulted in lower accuracy, suggesting they hinder model performance.

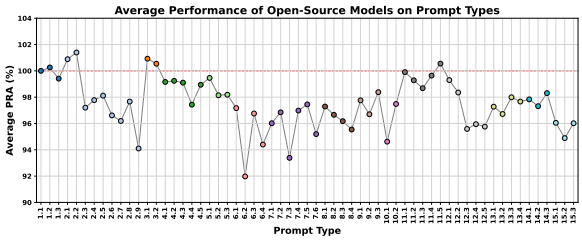

Figure 5: Average Prompt Performance for Open-Source Models. PRA with respect to the Baseline Prompt Accuracy is averaged across Open-source Models and the 3 Benchmarks (MMStar, MMMU-Pro & MVBench) for each Prompt Type.

For proprietary models, the majority of prompts enhanced performance relative to the baseline, with only a few exceptions, namely 4.5, 9.2 (actual drop –50%, capped at –15% for readability), 11.1, 12.4, and 15.1, showing reduced accuracy.

We designed two video-specific prompts, 11.4 and 11.5, inspired by MVBench (Li et al., 2024) and MMBench-Video (Fang et al., 2024) respectively, to explicitly address the temporal dimension

#	Open-Source Models	Proprietary Models
1	**Concise prompts yield better performance** – Keeping prompts short and direct improves accuracy. *"Answer with the option letter from the given choices directly."* (1.1) **Overly short or vague prompts reduce accuracy** – When the prompt is too brief and lacks clarity, the model may not understand the expected format or task. *"Best Choice: $LETTER"* (12.3) **Detailed prompts are ineffective** – Long or highly descriptive prompts do not improve accuracy. (Notably in Category 5 and other long prompts)	**Prompt length and detail have minimal impact** – Unlike open-source models, proprietary models perform consistently across prompts of varying lengths and complexity. **Restricting responses to the letter choice is detrimental** – Limiting the model to respond with just a letter (e.g., A, B, C, D) can suppress reasoning and reduce accuracy. (12.2)
2	**Complex or structured formatting decreases accuracy** – Using formats such as JSON, YAML or Markdown negatively impacts model performance. (2.3, 2.4, 2.5, 2.6, 2.7, 2.8, 2.9) **Clear separation of option letters enhances clarity** – Using parentheses for option labels improves model understanding. *"(A) choice 1* \n *(B) choice 2* \n *(C) choice 3* \n *(D) choice 4"* (1.2) **Explicit labeling of question and options is beneficial** – Using clear section headers improves comprehension. *"Question: <QUESTION>* *Options:* \n *<OPTIONS>* *Answer with the option letter from the given choices directly."* (2.2) **Placing question and options at the end helps** – Structuring prompts so that the question and answer choices appear at the end leads to better results. *"Answer with the option letter from the given choices directly.* *<QUESTION>* \n *<OPTIONS>"* (3.1)	**Complex formatting does not impair accuracy** – Unlike open-source models, proprietary models can handle structured formats such as JSON, Markdown, or YAML without a drop in performance. (Category 2)
3	**Poor linguistic formatting hinders performance** – Use of all upper case, poor grammar, or misspellings negatively impacts accuracy. (Category 4)	**Poor linguistic formatting does not affect performance** – These models are robust to grammatical errors, casing, and minor typos, likely due to stronger pretraining and instruction tuning. (Category 4)
4	**Chain-of-Thought reasoning is ineffective** – Step-by-step reasoning does not improve accuracy in this context. (Category 6)	**Allowing room for reasoning significantly improves accuracy.** – Allowing the model to think leads to higher accuracy. (Categories 6 & 12.5)
5	**Penalties, incentives, or competitive framing are ineffective** – Using competitive language, penalizing mistakes, or offering rewards often introduces ambiguity. (Category 13,14,15)	**Penalties or incentives improve performance** – Framing prompts with rewards or penalties can enhance performance, possibly due to better contextual understanding. (Categories 13 & 14) **Competitive framing degrades performance** – Prompts that use game-like or adversarial language introduce unnecessary pressure or distraction, reducing answer accuracy. (Category 15)
6	**Specifying personas or target audiences is ineffective** – Tailoring prompts by specifying a persona or intended audience does not improve model performance. (Category 8 & 9)	**Persona-based prompting has mixed effects** – Positive persona prompts do not enhance accuracy, while negative persona prompts can significantly degrade performance. (Category 9)
7	**Overemphasis on answer format is unhelpful** – Excessive instruction about answer formatting can degrade performance. (Category 12 & 11.3)	**Answer format plays an important role in accuracy** – Proprietary models are sensitive to how the answer is requested. (Category 12 & 11.3)
8	**Temporal reasoning enhances video comprehension** – Prompts that emphasize temporal order improve accuracy on video-based tasks. (11.4, 11.5)	**Temporal reasoning enhances video comprehension** – Prompts that emphasize temporal aspects of events in videos result in more accurate responses. (11.4 & 11.5)
9	**Image-focused prompting helps** – Directing the model to rely solely on the image content improves answer accuracy. (11.1)	**Asking to focus on image or question hinders performance** – In contrast to open-source models, proprietary models do worse when explicitly told to focus only on the image or only on the question. (11.1 & 11.2)
10	**Answer leakage degrades performance** – Including unintended hints or answer cues leads to lower accuracy. (Category 7)	**Asking to avoid bias or stereotypes helps** – Prompts that explicitly instruct the model to avoid bias or stereotypes lead to more accurate responses. (Category 10)

Table 4: Prompting Principles for Open-Source and Proprietary LMMs. Drawing from our comprehensive prompt sensitivity analysis (section 4), we derive a set of prompt templates and guidelines designed to elicit more stable and accurate responses from LMMs in MCQA task.

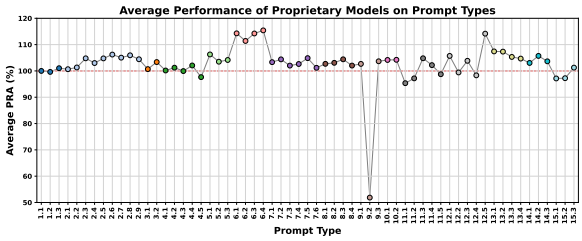

Figure 6: Average Prompt Performance for Proprietary Models. PRA with respect to the Baseline Prompt Accuracy is averaged across Proprietary Models and the 3 Benchmarks (MMStar, MMMU-Pro & MVBench) for each Prompt Type.

in videos. Notably, these prompts had a positive impact on performance in the MVBench video benchmark.

4.3 Model, Modality & Benchmark Level Analysis

To further understand model behavior, we conducted an in-depth analysis to identify which models exhibit the highest sensitivity to prompt variations. Specifically, we examined the impact of positive, negative, and neutral prompt types on model sensitivity (Appendix E). Additionally, we investigated prompt sensitivity at a finer granularity by analyzing which question types in MMMU-Pro, which reasoning dimensions in MMStar, and which temporal tasks in MVBench are most affected by prompt changes (Appendix F). Due to space constraints, the full set of results and detailed breakdowns are provided in the appendix.

5 Prompting Principles

Based on the insights from Section 4, we outline best practices for optimizing LMM performance on the MCQA task in Table 4. These strategies are designed to enhance both accuracy and consistency. While our insights are based on MCQA evaluations, we believe these principles can be broadly applied to other tasks and extended to LLMs and LMMs. Given the differences in instruction-following capabilities, we distinguish our prompting principles between open-source and proprietary models to account for their varying levels of adherence to instructions.

6 Related Work

Prompt Sensitivity of LLMs: A growing body of research investigates the sensitivity of large language models (LLMs) to various prompting strategies. (Alzahrani et al., 2024a) showed that minor modifications in multiple-choice question benchmarks can result in significant ranking shifts, indicating that current evaluation metrics may not provide stable comparisons. In addition to benchmark perturbations, prompt design also plays a crucial role in LLM performance. While system prompt personas are often incorporated to guide responses, their effectiveness remains inconsistent across different contexts (Zheng et al., 2024). Moreover, the structure and format of prompts significantly influence outcomes, with studies showing that prompt formatting alone can lead to performance variations as large as 40% (He et al., 2024). These findings collectively emphasize the need for standardized, well-defined methodologies when evaluating and deploying LLMs.

Prompt Sensitivity of LMMs: While the impact of textual prompt variations has been extensively studied in the language domain as well as CLIP-based VLMs, its effects on LMMs have not been systematically explored. One study in this direction indicates that these models exhibit a high degree of responsiveness to lexical and semantic changes, which can significantly affect response consistency (Dumpala et al., 2024). Another study (Awal et al., 2025) examines prompting strategies for improving zero- and few-shot VQA performance in LMMs. It highlights the importance of question templates in guiding model responses and demonstrates that augmenting LMMs with image captions significantly enhances performance. Despite these initial efforts, there remains a lack of systematic and comprehensive research into how different types of textual prompt variations, such as lexical, semantic, and structural, affect LMM performance across tasks and datasets.

7 Conclusion

We present the most comprehensive analysis to date on the impact of prompt design in Large Multimodal Models (LMMs) for MCQA across image and video benchmarks. Using Promptception, a systematic framework covering 61 prompt types, we evaluate 10 models on 3 datasets. Our findings reveal that prompt phrasing substantially affects performance: proprietary models exhibit strong instruction-following but higher sensitivity, while open-source models are more stable yet less responsive to subtle cues. We hope this work advances fair and transparent LMM evaluation.

8 Limitations

While the proposed prompt designs were primarily developed for multimodal MCQA task, their potential applicability extends to a broader range of vision-language tasks. However, to realize this potential, it is necessary to develop a more comprehensive and task-specific prompt framework. This would involve a careful study of task types beyond MCQA, such as video captioning, visual reasoning, and image/video-grounded dialogue to craft prompts tailored to the unique demands of open-ended tasks.

Moreover, while our manually designed prompts offer strong performance and serve as a set of best practices, automatic prompt generation is a crucial next step toward scaling this approach across a wider set of tasks. A promising direction involves the use of meta-prompting (Mirza et al., 2024), where a higher-level prompt is used to guide a language model in generating a task-specific prompt based on the input. To further streamline this process, an alternative direction is to train a lightweight prompt-generation model (Salehi et al., 2024) that can directly output high-quality prompts conditioned on the input.

References

Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C. Lawrence Zitnick, Dhruv Batra, and Devi Parikh. 2016. Vqa: Visual question answering. *Preprint*, arXiv:1505.00468.

Norah Alzahrani, Hisham Alyahya, Yazeed Alnumay, Sultan AlRashed, Shaykhah Alsubaie, Yousef Almushayqih, Faisal Mirza, Nouf Alotaibi, Nora Al-Twairesh, Areeb Alowisheq, M Saiful Bari, and Haidar Khan. 2024a. 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)*, pages 13787–13805, Bangkok, Thailand. Association for Computational Linguistics.

Norah Alzahrani, Hisham Abdullah Alyahya, Yazeed Alnumay, Sultan Alrashed, Shaykhah Alsubaie, Yusef Almushaykeh, Faisal Mirza, Nouf Alotaibi, Nora Altwairesh, Areeb Alowisheq, M Saiful Bari, and Haidar Khan. 2024b. When benchmarks are targets: Revealing the sensitivity of large language model leaderboards. *Preprint*, arXiv:2402.01781.

Rabiul Awal, Le Zhang, and Aishwarya Agrawal. 2025. Investigating prompting techniques for zero- and few-shot visual question answering. *Preprint*, arXiv:2306.09996.

Nishant Balepur, Abhilasha Ravichander, and Rachel Rudinger. 2024. Artifacts or abduction: How do LLMs answer multiple-choice questions without the question? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10308–10330, Bangkok, Thailand. Association for Computational Linguistics.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. 2024. Are we on the right way for evaluating large vision-language models? In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, Lixin Gu, Xuehui Wang, Qingyun Li, Yimin Ren, Zixuan Chen, Jiapeng Luo, Jiahao Wang, Tan Jiang, Bo Wang, Conghui He, Botian Shi, Xingcheng Zhang, Han Lv, Yi Wang, Wenqi Shao, Pei Chu, Zhongying Tu, Tong He, Zhiyong Wu, Huipeng Deng, Jiaye Ge, Kai Chen, Kaipeng Zhang, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. 2025. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *Preprint*, arXiv:2412.05271.

Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, Jiasen Lu, Taira Anderson, Erin Bransom, Kiana Ehsani, Huong Ngo, YenSung Chen, Ajay Patel, Mark Yatskar, Chris Callison-Burch, Andrew Head, Rose Hendrix, Favyen Bastani, Eli VanderBilt, Nathan Lambert, Yvonne Chou, Arnavi Chheda, Jenna Sparks, Sam Skjonsberg, Michael Schmitz, Aaron Sarnat, Byron Bischoff, Pete Walsh, Chris Newell, Piper Wolters, Tanmay Gupta, Kuo-Hao Zeng, Jon Borchardt, Dirk Groeneveld, Crystal Nam, Sophie Lebrecht, Caitlin Wittlif, Carissa Schoenick, Oscar Michel, Ranjay Krishna, Luca Weihs, Noah A. Smith, Hannaneh Hajishirzi, Ross Girshick, Ali Farhadi, and Aniruddha Kembhavi. 2024. Molmo and pixmo: Open weights and open data for state-of-the-art vision-language models. *Preprint*, arXiv:2409.17146.

Sri Harsha Dumpala, Aman Jaiswal, Chandramouli Sastry, Evangelos Milios, Sageev Oore, and Hassan Sajjad. 2024. Sensitivity of generative vlms to semantically and lexically altered prompts. *Preprint*, arXiv:2410.13030.

Xinyu Fang, Kangrui Mao, Haodong Duan, Xiangyu Zhao, Yining Li, Dahua Lin, and Kai Chen. 2024. Mmbench-video: A long-form multi-shot benchmark for holistic video understanding. *Preprint*, arXiv:2406.14515.

Jia He, Mukund Rungta, David Koleczek, Arshdeep Sekhon, Franklin X Wang, and Sadid Hasan. 2024. Does prompt formatting have any impact on llm performance? *Preprint*, arXiv:2411.10541.

Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. 2016. A diagram is worth a dozen images. *Preprint*, arXiv:1603.07396.

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. *Transactions on Machine Learning Research*.

Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. 2023. Seed-bench: Benchmarking multimodal llms with generative comprehension. *Preprint*, arXiv:2307.16125.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, Limin Wang, and Yu Qiao. 2024. Mvbench: A comprehensive multi-modal video understanding benchmark. *Preprint*, arXiv:2311.17005.

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

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

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Preprint*, arXiv:2209.09513.

M. Jehanzeb Mirza, Leonid Karlinsky, Wei Lin, Sivan Doveh, Jakub Micorek, Mateusz Kozinski, Hilde Kuehne, and Horst Possegger. 2024. Meta-prompting for automating zero-shot visual recognition with llms. *Preprint*, arXiv:2403.11755.

Shehan Munasinghe, Hanan Gani, Wenqi Zhu, Jiale Cao, Eric Xing, Fahad Shahbaz Khan, and Salman Khan. 2025. Videoglamm: A large multimodal model for pixel-level visual grounding in videos. *Preprint*, arXiv:2411.04923.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, Alex Paino, Alex Renzin, Alex Tachard Passos, Alexander Kirillov, Alexi Christakis, Alexis Conneau, Ali Kamali, Allan Jabri, Allison Moyer, Allison Tam, Amadou Crookes, Amin Tootoochian, Amin Tootoonchian, Ananya Kumar, Andrea Vallone, Andrej Karpathy, Andrew Braunstein, Andrew Cann, Andrew Codispoti, Andrew Galu, Andrew Kondrich, Andrew Tulloch, Andrey Mishchenko, Angela Baek, Angela Jiang, Antoine Pelisse, Antonia Woodford, Anuj Gosalia, Arka Dhar, Ashley Pantuliano, Avi Nayak, Avital Oliver, Barret Zoph, Behrooz Ghorbani, Ben Leimberger, Ben Rossen, Ben Sokolowsky, Ben Wang, Benjamin Zweig, Beth Hoover, Blake Samic, Bob McGrew, Bobby Spero, Bogo Giertler, Bowen Cheng, Brad Lightcap, Brandon Walkin, Brendan Quinn, Brian Guarraci, Brian Hsu, Bright Kellogg, Brydon Eastman, Camillo Lugaresi, Carroll Wainwright, Cary Bassin, Cary Hudson, Casey Chu, Chad Nelson, Chak Li, Chan Jun Shern, Channing Conger, Charlotte Barette, Chelsea Voss, Chen Ding, Cheng Lu, Chong Zhang, Chris Beaumont, Chris Hallacy, Chris Koch, Christian Gibson, Christina Kim, Christine Choi, Christine McLeavey, Christopher Hesse, Claudia Fischer, Clemens Winter, Coley Czarnecki, Colin Jarvis, Colin Wei, Constantin Koumouzelis, Dane Sherburn, Daniel Kappler, Daniel Levin, Daniel Levy, David Carr, David Farhi, David Mely, David Robinson, David Sasaki, Denny Jin, Dev Valladares, Dimitris Tsipras, Doug Li, Duc Phong Nguyen, Duncan Findlay, Edede Oiwoh, Edmund Wong, Ehsan Asdar, Elizabeth Proehl, Elizabeth Yang, Eric Antonow, Eric Kramer, Eric Peterson, Eric Sigler, Eric Wallace, Eugene Brevdo, Evan Mays, Farzad Khorasani, Felipe Petroski Such, Filippo Raso, Francis Zhang, Fred von Lohmann, Freddie Sulit, Gabriel Goh, Gene Oden, Geoff Salmon, Giulio Starace, Greg Brockman, Hadi Salman, Haiming Bao, Haitang Hu, Hannah Wong, Haoyu Wang, Heather Schmidt, Heather Whitney, Heewoo Jun, Hendrik Kirchner, Henrique Ponde de Oliveira Pinto, Hongyu Ren, Huiwen Chang, Hyung Won Chung, Ian Kivlichan, Ian O'Connell, Ian O'Connell, Ian Osband, Ian Silber, Ian Sohl, Ibrahim Okuyucu, Ikai Lan, Ilya Kostrikov, Ilya Sutskever, Ingmar Kanitscheider, Ishaan Gulrajani, Jacob Coxon, Jacob Menick, Jakub Pachocki, James Aung, James Betker, James Crooks, James Lennon, Jamie Kiros, Jan Leike, Jane Park, Jason Kwon, Jason Phang, Jason Teplitz, Jason Wei, Jason Wolfe, Jay Chen, Jeff Harris, Jenia Varavva, Jessica Gan Lee, Jessica Shieh, Ji Lin, Jiahui Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joanne Jang, Joaquin Quinonero Candela, Joe Beutler, Joe Landers, Joel Parish, Johannes Heidecke, John Schulman, Jonathan Lachman, Jonathan McKay, Jonathan Uesato, Jonathan Ward, Jong Wook Kim, Joost Huizinga, Jordan Sitkin, Jos Kraaijeveld, Josh Gross, Josh Kaplan, Josh Snyder, Joshua Achiam, Joy Jiao, Joyce Lee, Juntang Zhuang, Justyn Harriman, Kai Fricke, Kai Hayashi, Karan Singhal, Katy Shi, Kavin Karthik, Kayla Wood, Kendra Rimbach, Kenny Hsu, Kenny Nguyen, Keren Gu-Lemberg, Kevin Button, Kevin Liu, Kiel Howe, Krithika Muthukumar, Kyle Luther, Lama Ahmad, Larry Kai, Lauren Itow, Lauren Workman, Leher Pathak, Leo Chen, Li Jing, Lia Guy, Liam Fedus, Liang Zhou, Lien Mamitsuka, Lilian Weng, Lindsay McCallum, Lindsey Held, Long Ouyang, Louis Feuvrier, Lu Zhang, Lukas Kondraciuk, Lukasz Kaiser, Luke Hewitt, Luke Metz, Lyric Doshi, Mada Aflak, Maddie Simens, Madelaine

Boyd, Madeleine Thompson, Marat Dukhan, Mark Chen, Mark Gray, Mark Hudnall, Marvin Zhang, Marwan Aljubeh, Mateusz Litwin, Matthew Zeng, Max Johnson, Maya Shetty, Mayank Gupta, Meghan Shah, Mehmet Yatbaz, Meng Jia Yang, Mengchao Zhong, Mia Glaese, Mianna Chen, Michael Janner, Michael Lampe, Michael Petrov, Michael Wu, Michele Wang, Michelle Fradin, Michelle Pokrass, Miguel Castro, Miguel Oom Temudo de Castro, Mikhail Pavlov, Miles Brundage, Miles Wang, Minal Khan, Mira Murati, Mo Bavarian, Molly Lin, Murat Yesildal, Nacho Soto, Natalia Gimelshein, Natalie Cone, Natalie Staudacher, Natalie Summers, Natan LaFontaine, Neil Chowdhury, Nick Ryder, Nick Stathas, Nick Turley, Nik Tezak, Niko Felix, Nithanth Kudige, Nitish Keskar, Noah Deutsch, Noel Bundick, Nora Puckett, Ofir Nachum, Ola Okelola, Oleg Boiko, Oleg Murk, Oliver Jaffe, Olivia Watkins, Olivier Godement, Owen Campbell-Moore, Patrick Chao, Paul McMillan, Pavel Belov, Peng Su, Peter Bak, Peter Bakkum, Peter Deng, Peter Dolan, Peter Hoeschele, Peter Welinder, Phil Tillet, Philip Pronin, Philippe Tillet, Prafulla Dhariwal, Qiming Yuan, Rachel Dias, Rachel Lim, Rahul Arora, Rajan Troll, Randall Lin, Rapha Gontijo Lopes, Raul Puri, Reah Miyara, Reimar Leike, Renaud Gaubert, Reza Zamani, Ricky Wang, Rob Donnelly, Rob Honsby, Rocky Smith, Rohan Sahai, Rohit Ramchandani, Romain Huet, Rory Carmichael, Rowan Zellers, Roy Chen, Ruby Chen, Ruslan Nigmatullin, Ryan Cheu, Saachi Jain, Sam Altman, Sam Schoenholz, Sam Toizer, Samuel Miserendino, Sandhini Agarwal, Sara Culver, Scott Ethersmith, Scott Gray, Sean Grove, Sean Metzger, Shamez Hermani, Shantanu Jain, Shengjia Zhao, Sherwin Wu, Shino Jomoto, Shirong Wu, Shuaiqi, Xia, Sonia Phene, Spencer Papay, Srinivas Narayanan, Steve Coffey, Steve Lee, Stewart Hall, Suchir Balaji, Tal Broda, Tal Stramer, Tao Xu, Tarun Gogineni, Taya Christianson, Ted Sanders, Tejal Patwardhan, Thomas Cunninghman, Thomas Degry, Thomas Dimson, Thomas Raoux, Thomas Shadwell, Tianhao Zheng, Todd Underwood, Todor Markov, Toki Sherbakov, Tom Rubin, Tom Stasi, Tomer Kaftan, Tristan Heywood, Troy Peterson, Tyce Walters, Tyna Eloundou, Valerie Qi, Veit Moeller, Vinnie Monaco, Vishal Kuo, Vlad Fomenko, Wayne Chang, Weiyi Zheng, Wenda Zhou, Wesam Manassra, Will Sheu, Wojciech Zaremba, Yash Patil, Yilei Qian, Yongjik Kim, Youlong Cheng, Yu Zhang, Yuchen He, Yuchen Zhang, Yujia Jin, Yunxing Dai, and Yury Malkov. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

David Patterson, Joseph Gonzalez, Urs Hölzle, Quoc Le, Chen Liang, Lluis-Miquel Munguia, Daniel Rothchild, David So, Maud Texier, and Jeff Dean. 2022. The carbon footprint of machine learning training will plateau, then shrink. *Preprint*, arXiv:2204.05149.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. *Preprint*, arXiv:2103.00020.

Hanoona Rasheed, Muhammad Maaz, Sahal Shaji Mullappilly, Abdelrahman Shaker, Salman Khan, Hisham Cholakkal, Rao M. Anwer, Erix Xing, Ming-Hsuan Yang, and Fahad S. Khan. 2024. Glamm: Pixel grounding large multimodal model. *Preprint*, arXiv:2311.03356.

Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. 2024. Timechat: A time-sensitive multimodal large language model for long video understanding. *Preprint*, arXiv:2312.02051.

Joshua Robinson, Christopher Michael Rytting, and David Wingate. 2023. Leveraging large language models for multiple choice question answering. *Preprint*, arXiv:2210.12353.

Sogand Salehi, Mahdi Shafiei, Teresa Yeo, Roman Bachmann, and Amir Zamir. 2024. Viper: Visual personalization of generative models via individual preference learning. *Preprint*, arXiv:2407.17365.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, Soroosh Mariooryad, Yifan Ding, Xinyang Geng, Fred Alcober, Roy Frostig, Mark Omernick, Lexi Walker, Cosmin Paduraru, Christina Sorokin, Andrea Tacchetti, Colin Gaffney, Samira Daruki, Olcan Sercinoglu, Zach Gleicher, Juliette Love, Paul Voigtlaender, Rohan Jain, Gabriela Surita, Kareem Mohamed, Rory Blevins, Junwhan Ahn, Tao Zhu, Kornraphop Kawintiranon, Orhan Firat, Yiming Gu, Yujing Zhang, Matthew Rahtz, Manaal Faruqui, Natalie Clay, Justin Gilmer, JD Co-Reyes, Ivo Penchev, Rui Zhu, Nobuyuki Morioka, Kevin Hui, Krishna Haridasan, Victor Campos, Mahdis Mahdieh, Mandy Guo, Samer Hassan, Kevin Kilgour, Arpi Vezer, Heng-Tze Cheng, Raoul de Liedekerke, Siddharth Goyal, Paul Barham, DJ Strouse, Seb Noury, Jonas Adler, Mukund Sundararajan, Sharad Vikram, Dmitry Lepikhin, Michela Paganini, Xavier Garcia, Fan Yang, Dasha Valter, Maja Trebacz, Kiran Vodrahalli, Chulayuth Asawaroengchai, Roman Ring, Norbert Kalb, Livio Baldini Soares, Siddhartha Brahma, David Steiner, Tianhe Yu, Fabian Mentzer, Antoine He, Lucas Gonzalez, Bibo Xu, Raphael Lopez Kaufman, Laurent El Shafey, Junhyuk Oh, Tom Hennigan, George van den Driessche, Seth Odoom, Mario Lucic, Becca Roelofs, Sid Lall, Amit Marathe, Betty Chan, Santiago Ontanon, Luheng He, Denis Teplyashin, Jonathan Lai, Phil Crone, Bogdan Damoc, Lewis Ho, Sebastian Riedel, Karel Lenc, Chih-Kuan Yeh, Aakanksha Chowdhery, Yang Xu, Mehran Kazemi, Ehsan Amid, Anastasia Petrushkina, Kevin Swersky, Ali Khodaei, Gowoon Chen, Chris Larkin, Mario Pinto, Geng Yan, Adria Puigdomenech Badia, Piyush Patil, Steven Hansen, Dave Orr, Sebastien M. R. Arnold, Jordan Grimstad, Andrew Dai, Sholto Douglas, Rishika Sinha, Vikas Yadav, Xi Chen, Elena Gribovskaya, Jacob Austin, Jeffrey Zhao, Kaushal Patel, Paul Komarek, Sophia Austin, Sebastian Borgeaud,

Linda Friso, Abhimanyu Goyal, Ben Caine, Kris Cao, Da-Woon Chung, Matthew Lamm, Gabe Barth-Maron, Thais Kagohara, Kate Olszewska, Mia Chen, Kaushik Shivakumar, Rishabh Agarwal, Harshal Godhia, Ravi Rajwar, Javier Snaider, Xerxes Dotiwalla, Yuan Liu, Aditya Barua, Victor Ungureanu, Yuan Zhang, Bat-Orgil Batsaikhan, Mateo Wirth, James Qin, Ivo Danihelka, Tulsee Doshi, Martin Chadwick, Jilin Chen, Sanil Jain, Quoc Le, Arjun Kar, Madhu Gurumurthy, Cheng Li, Ruoxin Sang, Fangyu Liu, Lampros Lamprou, Rich Munoz, Nathan Lintz, Harsh Mehta, Heidi Howard, Malcolm Reynolds, Lora Aroyo, Quan Wang, Lorenzo Blanco, Albin Cassirer, Jordan Griffith, Dipanjan Das, Stephan Lee, Jakub Sygnowski, Zach Fisher, James Besley, Richard Powell, Zafarali Ahmed, Dominik Paulus, David Reitter, Zalan Borsos, Rishabh Joshi, Aedan Pope, Steven Hand, Vittorio Selo, Vihan Jain, Nikhil Sethi, Megha Goel, Takaki Makino, Rhys May, Zhen Yang, Johan Schalkwyk, Christina Butterfield, Anja Hauth, Alex Goldin, Will Hawkins, Evan Senter, Sergey Brin, Oliver Woodman, Marvin Ritter, Eric Noland, Minh Giang, Vijay Bolina, Lisa Lee, Tim Blyth, Ian Mackinnon, Machel Reid, Obaid Sarvana, David Silver, Alexander Chen, Lily Wang, Loren Maggiore, Oscar Chang, Nithya Attaluri, Gregory Thornton, Chung-Cheng Chiu, Oskar Bunyan, Nir Levine, Timothy Chung, Evgenii Eltyshev, Xiance Si, Timothy Lillicrap, Demetra Brady, Vaibhav Aggarwal, Boxi Wu, Yuanzhong Xu, Ross McIlroy, Kartikeya Badola, Paramjit Sandhu, Erica Moreira, Wojciech Stokowiec, Ross Hemsley, Dong Li, Alex Tudor, Pranav Shyam, Elahe Rahimtoroghi, Salem Haykal, Pablo Sprechmann, Xiang Zhou, Diana Mincu, Yujia Li, Ravi Addanki, Kalpesh Krishna, Xiao Wu, Alexandre Frechette, Matan Eyal, Allan Dafoe, Dave Lacey, Jay Whang, Thi Avrahami, Ye Zhang, Emanuel Taropa, Hanzhao Lin, Daniel Toyama, Eliza Rutherford, Motoki Sano, HyunJeong Choe, Alex Tomala, Chalence Safranek-Shrader, Nora Kassner, Mantas Pajarskas, Matt Harvey, Sean Sechrist, Meire Fortunato, Christina Lyu, Gamaleldin Elsayed, Chenkai Kuang, James Lottes, Eric Chu, Chao Jia, Chih-Wei Chen, Peter Humphreys, Kate Baumli, Connie Tao, Rajkumar Samuel, Cicero Nogueira dos Santos, Anders Andreassen, Nemanja Rakićević, Dominik Grewe, Aviral Kumar, Stephanie Winkler, Jonathan Caton, Andrew Brock, Sid Dalmia, Hannah Sheahan, Iain Barr, Yingjie Miao, Paul Natsev, Jacob Devlin, Feryal Behbahani, Flavien Prost, Yanhua Sun, Artiom Myaskovsky, Thanumalayan Sankaranarayana Pillai, Dan Hurt, Angeliki Lazaridou, Xi Xiong, Ce Zheng, Fabio Pardo, Xiaowei Li, Dan Horgan, Joe Stanton, Moran Ambar, Fei Xia, Alejandro Lince, Mingqiu Wang, Basil Mustafa, Albert Webson, Hyo Lee, Rohan Anil, Martin Wicke, Timothy Dozat, Abhishek Sinha, Enrique Piqueras, Elahe Dabir, Shyam Upadhyay, Anudhyan Boral, Lisa Anne Hendricks, Corey Fry, Josip Djolonga, Yi Su, Jake Walker, Jane Labanowski, Ronny Huang, Vedant Misra, Jeremy Chen, RJ Skerry-Ryan, Avi Singh, Shruti Rijhwani, Dian Yu, Alex Castro-Ros, Beer Changpinyo, Romina Datta, Sumit Bagri, Arnar Mar Hrafnkelsson, Marcello Maggioni, Daniel Zheng, Yury Sulsky, Shaobo Hou, Tom Le Paine, Antoine Yang, Jason Riesa, Dominika Rogozinska, Dror Marcus, Dalia El Badawy, Qiao Zhang, Luyu Wang, Helen Miller, Jeremy Greer, Lars Lowe Sjos, Azade Nova, Heiga Zen, Rahma Chaabouni, Mihaela Rosca, Jiepu Jiang, Charlie Chen, Ruibo Liu, Tara Sainath, Maxim Krikun, Alex Polozov, Jean-Baptiste Lespiau, Josh Newlan, Zeyncep Cankara, Soo Kwak, Yunhan Xu, Phil Chen, Andy Coenen, Clemens Meyer, Katerina Tsihlas, Ada Ma, Juraj Gottweis, Jinwei Xing, Chenjie Gu, Jin Miao, Christian Frank, Zeynep Cankara, Sanjay Ganapathy, Ishita Dasgupta, Steph Hughes-Fitt, Heng Chen, David Reid, Keran Rong, Hongmin Fan, Joost van Amersfoort, Vincent Zhuang, Aaron Cohen, Shixiang Shane Gu, Anhad Mohananey, Anastasija Ilic, Taylor Tobin, John Wieting, Anna Bortsova, Phoebe Thacker, Emma Wang, Emily Caveness, Justin Chiu, Eren Sezener, Alex Kaskasoli, Steven Baker, Katie Millican, Mohamed Elhawaty, Kostas Aisopos, Carl Lebsack, Nathan Byrd, Hanjun Dai, Wenhao Jia, Matthew Wiethoff, Elnaz Davoodi, Albert Weston, Lakshman Yagati, Arun Ahuja, Isabel Gao, Golan Pundak, Susan Zhang, Michael Azzam, Khe Chai Sim, Sergi Caelles, James Keeling, Abhanshu Sharma, Andy Swing, YaGuang Li, Chenxi Liu, Carrie Grimes Bostock, Yamini Bansal, Zachary Nado, Ankesh Anand, Josh Lipschultz, Abhijit Karmarkar, Lev Proleev, Abe Ittycheriah, Soheil Hassas Yeganeh, George Polovets, Aleksandra Faust, Jiao Sun, Alban Rrustemi, Pen Li, Rakesh Shivanna, Jeremiah Liu, Chris Welty, Federico Lebron, Anirudh Baddepudi, Sebastian Krause, Emilio Parisotto, Radu Soricut, Zheng Xu, Dawn Bloxwich, Melvin Johnson, Behnam Neyshabur, Justin Mao-Jones, Renshen Wang, Vinay Ramasesh, Zaheer Abbas, Arthur Guez, Constant Segal, Duc Dung Nguyen, James Svensson, Le Hou, Sarah York, Kieran Milan, Sophie Bridgers, Wiktor Gworek, Marco Tagliasacchi, James Lee-Thorp, Michael Chang, Alexey Guseynov, Ale Jakse Hartman, Michael Kwong, Ruizhe Zhao, Sheleem Kashem, Elizabeth Cole, Antoine Miech, Richard Tanburn, Mary Phuong, Filip Pavetic, Sebastien Cevey, Ramona Comanescu, Richard Ives, Sherry Yang, Cosmo Du, Bo Li, Zizhao Zhang, Mariko Iinuma, Clara Huiyi Hu, Aurko Roy, Shaan Bijwadia, Zhenkai Zhu, Danilo Martins, Rachel Saputro, Anita Gergely, Steven Zheng, Dawei Jia, Ioannis Antonoglou, Adam Sadovsky, Shane Gu, Yingying Bi, Alek Andreev, Sina Samangooei, Mina Khan, Tomas Kocisky, Angelos Filos, Chintu Kumar, Colton Bishop, Adams Yu, Sarah Hodkinson, Sid Mittal, Premal Shah, Alexandre Moufarek, Yong Cheng, Adam Bloniarz, Jaehoon Lee, Pedram Pejman, Paul Michel, Stephen Spencer, Vladimir Feinberg, Xuehan Xiong, Nikolay Savinov, Charlotte Smith, Siamak Shakeri, Dustin Tran, Mary Chesus, Bernd Bohnet, George Tucker, Tamara von Glehn, Carrie Muir, Yiran Mao, Hideto Kazawa, Ambrose Slone, Kedar Soparkar, Disha Shrivastava, James Cobon-Kerr, Michael Sharman, Jay Pavagadhi, Carlos Araya, Karolis Misiunas, Nimesh Ghelani, Michael Laskin, David Barker, Qiujia Li, Anton Briukhov, Neil Houlsby, Mia Glaese, Balaji Laksh-

minarayanan, Nathan Schucher, Yunhao Tang, Eli Collins, Hyeontaek Lim, Fangxiaoyu Feng, Adria Recasens, Guangda Lai, Alberto Magni, Nicola De Cao, Aditya Siddhant, Zoe Ashwood, Jordi Orbay, Mostafa Dehghani, Jenny Brennan, Yifan He, Kelvin Xu, Yang Gao, Carl Saroufim, James Molloy, Xinyi Wu, Seb Arnold, Solomon Chang, Julian Schrittwieser, Elena Buchatskaya, Soroush Radpour, Martin Polacek, Skye Giordano, Ankur Bapna, Simon Tokumine, Vincent Hellendoorn, Thibault Sottiaux, Sarah Cogan, Aliaksei Severyn, Mohammad Saleh, Shantanu Thakoor, Laurent Shefey, Siyuan Qiao, Meenu Gaba, Shuo yiin Chang, Craig Swanson, Biao Zhang, Benjamin Lee, Paul Kishan Rubenstein, Gan Song, Tom Kwiatkowski, Anna Koop, Ajay Kannan, David Kao, Parker Schuh, Axel Stjerngren, Golnaz Ghiasi, Gena Gibson, Luke Vilnis, Ye Yuan, Felipe Tiengo Ferreira, Aishwarya Kamath, Ted Klimenko, Ken Franko, Kefan Xiao, Indro Bhattacharya, Miteyan Patel, Rui Wang, Alex Morris, Robin Strudel, Vivek Sharma, Peter Choy, Sayed Hadi Hashemi, Jessica Landon, Mara Finkelstein, Priya Jhakra, Justin Frye, Megan Barnes, Matthew Mauger, Dennis Daun, Khuslen Baatarsukh, Matthew Tung, Wael Farhan, Henryk Michalewski, Fabio Viola, Felix de Chaumont Quitry, Charline Le Lan, Tom Hudson, Qingze Wang, Felix Fischer, Ivy Zheng, Elspeth White, Anca Dragan, Jean baptiste Alayrac, Eric Ni, Alexander Pritzel, Adam Iwanicki, Michael Isard, Anna Bulanova, Lukas Zilka, Ethan Dyer, Devendra Sachan, Srivatsan Srinivasan, Hannah Muckenhirn, Honglong Cai, Amol Mandhane, Mukarram Tariq, Jack W. Rae, Gary Wang, Kareem Ayoub, Nicholas FitzGerald, Yao Zhao, Woohyun Han, Chris Alberti, Dan Garrette, Kashyap Krishnakumar, Mai Gimenez, Anselm Levskaya, Daniel Sohn, Josip Matak, Inaki Iturrate, Michael B. Chang, Jackie Xiang, Yuan Cao, Nishant Ranka, Geoff Brown, Adrian Hutter, Vahab Mirrokni, Nanxin Chen, Kaisheng Yao, Zoltan Egyed, Francois Galilee, Tyler Liechty, Praveen Kallakuri, Evan Palmer, Sanjay Ghemawat, Jasmine Liu, David Tao, Chloe Thornton, Tim Green, Mimi Jasarevic, Sharon Lin, Victor Cotruta, Yi-Xuan Tan, Noah Fiedel, Hongkun Yu, Ed Chi, Alexander Neitz, Jens Heitkaemper, Anu Sinha, Denny Zhou, Yi Sun, Charbel Kaed, Brice Hulse, Swaroop Mishra, Maria Georgaki, Sneha Kudugunta, Clement Farabet, Izhak Shafran, Daniel Vlasic, Anton Tsitsulin, Rajagopal Ananthanarayanan, Alen Carin, Guolong Su, Pei Sun, Shashank V, Gabriel Carvajal, Josef Broder, Iulia Comsa, Alena Repina, William Wong, Warren Weilun Chen, Peter Hawkins, Egor Filonov, Lucia Loher, Christoph Hirnschall, Weiyi Wang, Jingchen Ye, Andrea Burns, Hardie Cate, Diana Gage Wright, Federico Piccinini, Lei Zhang, Chu-Cheng Lin, Ionel Gog, Yana Kulizhskaya, Ashwin Sreevatsa, Shuang Song, Luis C. Cobo, Anand Iyer, Chetan Tekur, Guillermo Garrido, Zhuyun Xiao, Rupert Kemp, Huaixiu Steven Zheng, Hui Li, Ananth Agarwal, Christel Ngani, Kati Goshvadi, Rebeca Santamaria-Fernandez, Wojciech Fica, Xinyun Chen, Chris Gorgolewski, Sean Sun, Roopal Garg, Xinyu Ye, S. M. Ali Eslami, Nan Hua, Jon Simon, Pratik Joshi, Yelin Kim, Ian Tenney, Sahitya Potluri, Lam Nguyen Thiet, Quan Yuan, Florian Luisier, Alexandra Chronopoulou, Salvatore Scellato, Praveen Srinivasan, Minmin Chen, Vinod Koverkathu, Valentin Dalibard, Yaming Xu, Brennan Saeta, Keith Anderson, Thibault Sellam, Nick Fernando, Fantine Huot, Junehyuk Jung, Mani Varadarajan, Michael Quinn, Amit Raul, Maigo Le, Ruslan Habalov, Jon Clark, Komal Jalan, Kalesha Bullard, Achintya Singhal, Thang Luong, Boyu Wang, Sujeevan Rajayogam, Julian Eisenschlos, Johnson Jia, Daniel Finchelstein, Alex Yakubovich, Daniel Balle, Michael Fink, Sameer Agarwal, Jing Li, Dj Dvijotham, Shalini Pal, Kai Kang, Jaclyn Konzelmann, Jennifer Beattie, Olivier Dousse, Diane Wu, Remi Crocker, Chen Elkind, Siddhartha Reddy Jonnalagadda, Jong Lee, Dan Holtmann-Rice, Krystal Kallarackal, Rosanne Liu, Denis Vnukov, Neera Vats, Luca Invernizzi, Mohsen Jafari, Huanjie Zhou, Lilly Taylor, Jennifer Prendki, Marcus Wu, Tom Eccles, Tianqi Liu, Kavya Kopparapu, Francoise Beaufays, Christof Angermueller, Andreea Marzoca, Shourya Sarcar, Hilal Dib, Jeff Stanway, Frank Perbet, Nejc Trdin, Rachel Sterneck, Andrey Khorlin, Dinghua Li, Xihui Wu, Sonam Goenka, David Madras, Sasha Goldshtein, Willi Gierke, Tong Zhou, Yaxin Liu, Yannie Liang, Anais White, Yunjie Li, Shreya Singh, Sanaz Bahargam, Mark Epstein, Sujoy Basu, Li Lao, Adnan Ozturel, Carl Crous, Alex Zhai, Han Lu, Zora Tung, Neeraj Gaur, Alanna Walton, Lucas Dixon, Ming Zhang, Amir Globerson, Grant Uy, Andrew Bolt, Olivia Wiles, Milad Nasr, Ilia Shumailov, Marco Selvi, Francesco Piccinno, Ricardo Aguilar, Sara McCarthy, Misha Khalman, Mrinal Shukla, Vlado Galic, John Carpenter, Kevin Villela, Haibin Zhang, Harry Richardson, James Martens, Matko Bosnjak, Shreyas Rammohan Belle, Jeff Seibert, Mahmoud Alnahlawi, Brian McWilliams, Sankalp Singh, Annie Louis, Wen Ding, Dan Popovici, Lenin Simicich, Laura Knight, Pulkit Mehta, Nishesh Gupta, Chongyang Shi, Saaber Fatehi, Jovana Mitrovic, Alex Grills, Joseph Pagadora, Tsendsuren Munkhdalai, Dessie Petrova, Danielle Eisenbud, Zhishuai Zhang, Damion Yates, Bhavishya Mittal, Nilesh Tripuraneni, Yannis Assael, Thomas Brovelli, Prateek Jain, Mihajlo Velimirovic, Canfer Akbulut, Jiaqi Mu, Wolfgang Macherey, Ravin Kumar, Jun Xu, Haroon Qureshi, Gheorghe Comanici, Jeremy Wiesner, Zhitao Gong, Anton Ruddock, Matthias Bauer, Nick Felt, Anirudh GP, Anurag Arnab, Dustin Zelle, Jonas Rothfuss, Bill Rosgen, Ashish Shenoy, Bryan Seybold, Xinjian Li, Jayaram Mudigonda, Goker Erdogan, Jiawei Xia, Jiri Simsa, Andrea Michi, Yi Yao, Christopher Yew, Steven Kan, Isaac Caswell, Carey Radebaugh, Andre Elisseeff, Pedro Valenzuela, Kay McKinney, Kim Paterson, Albert Cui, Eri Latorre-Chimoto, Solomon Kim, William Zeng, Ken Durden, Priya Ponnapalli, Tiberiu Sosea, Christopher A. Choquette-Choo, James Manyika, Brona Robenek, Harsha Vashisht, Sebastien Pereira, Hoi Lam, Marko Velic, Denese Owusu-Afriyie, Katherine Lee, Tolga Bolukbasi, Alicia Parrish, Shawn Lu, Jane Park, Balaji Venkatraman, Alice Talbert, Lambert Rosique, Yuchung Cheng, Andrei Sozanschi,

Adam Paszke, Praveen Kumar, Jessica Austin, Lu Li, Khalid Salama, Bartek Perz, Wooyeol Kim, Nandita Dukkipati, Anthony Baryshnikov, Christos Kaplanis, XiangHai Sheng, Yuri Chervonyi, Caglar Unlu, Diego de Las Casas, Harry Askham, Kathryn Tunyasuvunakool, Felix Gimeno, Siim Poder, Chester Kwak, Matt Miecnikowski, Vahab Mirrokni, Alek Dimitriev, Aaron Parisi, Dangyi Liu, Tomy Tsai, Toby Shevlane, Christina Kouridi, Drew Garmon, Adrian Goedeckemeyer, Adam R. Brown, Anitha Vijayakumar, Ali Elqursh, Sadegh Jazayeri, Jin Huang, Sara Mc Carthy, Jay Hoover, Lucy Kim, Sandeep Kumar, Wei Chen, Courtney Biles, Garrett Bingham, Evan Rosen, Lisa Wang, Qijun Tan, David Engel, Francesco Pongetti, Dario de Cesare, Dongseong Hwang, Lily Yu, Jennifer Pullman, Srini Narayanan, Kyle Levin, Siddharth Gopal, Megan Li, Asaf Aharoni, Trieu Trinh, Jessica Lo, Norman Casagrande, Roopali Vij, Loic Matthey, Bramandia Ramadhana, Austin Matthews, CJ Carey, Matthew Johnson, Kremena Goranova, Rohin Shah, Shereen Ashraf, Kingshuk Dasgupta, Rasmus Larsen, Yicheng Wang, Manish Reddy Vuyyuru, Chong Jiang, Joana Ijazi, Kazuki Osawa, Celine Smith, Ramya Sree Boppana, Taylan Bilal, Yuma Koizumi, Ying Xu, Yasemin Altun, Nir Shabat, Ben Bariach, Alex Korchemniy, Kiam Choo, Olaf Ronneberger, Chimezie Iwuanyanwu, Shubin Zhao, David Soergel, Cho-Jui Hsieh, Irene Cai, Shariq Iqbal, Martin Sundermeyer, Zhe Chen, Elie Bursztein, Chaitanya Malaviya, Fadi Biadsy, Prakash Shroff, Inderjit Dhillon, Tejasi Latkar, Chris Dyer, Hannah Forbes, Massimo Nicosia, Vitaly Nikolaev, Somer Greene, Marin Georgiev, Pidong Wang, Nina Martin, Hanie Sedghi, John Zhang, Praseem Banzal, Doug Fritz, Vikram Rao, Xuezhi Wang, Jiageng Zhang, Viorica Patraucean, Dayou Du, Igor Mordatch, Ivan Jurin, Lewis Liu, Ayush Dubey, Abhi Mohan, Janek Nowakowski, Vlad-Doru Ion, Nan Wei, Reiko Tojo, Maria Abi Raad, Drew A. Hudson, Vaishakh Keshava, Shubham Agrawal, Kevin Ramirez, Zhichun Wu, Hoang Nguyen, Ji Liu, Madhavi Sewak, Bryce Petrini, DongHyun Choi, Ivan Philips, Ziyue Wang, Ioana Bica, Ankush Garg, Jarek Wilkiewicz, Priyanka Agrawal, Xiaowei Li, Danhao Guo, Emily Xue, Naseer Shaik, Andrew Leach, Sadh MNM Khan, Julia Wiesinger, Sammy Jerome, Abhishek Chakladar, Alek Wenjiao Wang, Tina Ornduff, Folake Abu, Alireza Ghaffarkhah, Marcus Wainwright, Mario Cortes, Frederick Liu, Joshua Maynez, Andreas Terzis, Pouya Samangouei, Riham Mansour, Tomasz Kępa, François-Xavier Aubet, Anton Algymr, Dan Banica, Agoston Weisz, Andras Orban, Alexandre Senges, Ewa Andrejczuk, Mark Geller, Niccolo Dal Santo, Valentin Anklin, Majd Al Merey, Martin Baeuml, Trevor Strohman, Junwen Bai, Slav Petrov, Yonghui Wu, Demis Hassabis, Koray Kavukcuoglu, Jeff Dean, and Oriol Vinyals. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *Preprint*, arXiv:2403.05530.

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. *Preprint*, arXiv:2409.12191.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Huggingface's transformers: State-of-the-art natural language processing. *Preprint*, arXiv:1910.03771.

Cheng-En Wu, Yu Tian, Haichao Yu, Heng Wang, Pedro Morgado, Yu Hen Hu, and Linjie Yang. 2023. Why is prompt tuning for vision-language models robust to noisy labels? *Preprint*, arXiv:2307.11978.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, Qianyu Chen, Huarong Zhou, Zhensheng Zou, Haoye Zhang, Shengding Hu, Zhi Zheng, Jie Zhou, Jie Cai, Xu Han, Guoyang Zeng, Dahai Li, Zhiyuan Liu, and Maosong Sun. 2024. Minicpm-v: A gpt-4v level mllm on your phone. *Preprint*, arXiv:2408.01800.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. 2024a. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. *Preprint*, arXiv:2311.16502.

Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, Yu Su, Wenhu Chen, and Graham Neubig. 2024b. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. *Preprint*, arXiv:2409.02813.

Yuhui Zhang, Yuchang Su, Yiming Liu, Xiaohan Wang, James Burgess, Elaine Sui, Chenyu Wang, Josiah Aklilu, Alejandro Lozano, Anjiang Wei, Ludwig Schmidt, and Serena Yeung-Levy. 2025. Automated generation of challenging multiple-choice questions for vision language model evaluation. *Preprint*, arXiv:2501.03225.

Mingqian Zheng, Jiaxin Pei, Lajanugen Logeswaran, Moontae Lee, and David Jurgens. 2024. When "a helpful assistant" is not really helpful: Personas in system prompts do not improve performances of large language models. *Preprint*, arXiv:2311.10054.

A Promptception: The Complete Prompt List

This appendix presents the complete list of prompts proposed in our framework, Promptception. We conducted experiments on three benchmarks: **MMStar**, **MMMU-Pro**, and **MVBench**.

For the image-based benchmarks (**MMStar** and **MMMU-Pro**), we used a shared set of prompts. In contrast, for the video-based benchmark (**MVBench**), we introduced slight modifications as shown by the color coding introduced below.

Additionally, we observed that the way we expect the answer letter should be varied between open-source and proprietary models, based on performance trends observed for prompts in *Category 12*.

To clearly indicate the differences among prompts, we use the following color coding:

- **Black**: Part of the prompt common to all settings
- **Blue**: Part of the prompt specific to image-based benchmarks
- **Orange**: Part of the prompt specific to the video-based benchmark
- **Green**: Part of the prompt tailored for open-source models
- **Purple**: Part of the prompt tailored for proprietary models

Super Category	Category	Type	Prompt
Choice Formatting and Presentation	Formatting Variations in Choice Presentation	**1.1: <LETTER>.**	What design element best describes the visuals? **A.** Composition **B.** Perspective **C.** Balance **D.** Shape **Answer with the option letter from the given choices directly.**
		1.2: (<LETTER>)	What design element best describes the visuals? **(A)** Composition **(B)** Perspective **(C)** Balance **(D)** Shape **Answer with the option letter from the given choices directly.**
		1.3: Option <LETTER>:	What design element best describes the visuals? **Option A:** Composition **Option B:** Perspective **Option C:** Balance **Option D:** Shape **Answer with the option letter from the given choices directly.**
	Structured formatting	**2.1: Question Prefix**	**Question:**What design element best describes the visuals? **A.** Composition **B.** Perspective **C.** Balance **D.** Shape **Answer with the option letter from the given choices directly.**
		2.2: Question & Answer Prefix	**Question:**What design element best describes the visuals? **Options:** **A.** Composition **B.** Perspective **C.** Balance **D.** Shape **Answer with the option letter from the given choices directly.**
		Type 2.3:**	****Instructions****: **1. Analyze the given image(s)/video and understand its content.** **2. Read the question carefully.** **3. Choose the most appropriate answer from the options provided.** ****Question:**** What design element best describes the visuals? ****Options:**** A. Composition B. Perspective C. Balance D. Shape ****Answer:****
		Type 2.4: ##	**##Instructions##:** 1. **Analyze the given image(s)/video and understand its content.** 2. **Read the question carefully.** 3. **Choose the most appropriate answer from the options provided.** **##Question##:** What design element best describes the visuals? **##Options##:** A. Composition B. Perspective C. Balance D. Shape **##Answer##:**

Continued on next page

Super Category	Category	Type	Prompt
		Type 2.5: Compact	**TASK: Analyze the image(s)/video and pick the best option.** **QUESTION:** What design element best describes the visuals? **OPTIONS:** A. Composition B. Perspective C. Balance D. Shape **BEST OPTION:**
		Type 2.6: Plaintext	**You are an advanced multimodal AI assistant specializing in analyzing visual content and answering questions based on it.** **Analyze the given image(s)/video and understand its content, read the question carefully and choose the most appropriate answer from the options provided. Here is the question:** What design element best describes the visuals? A. Composition B. Perspective C. Balance D. Shape. **Respond only with: '$LETTER' of the best answer choice./Respond in the following format: 'Answer: $LETTER' (without quotes) where LETTER is one of the options.**
		Type 2.7: Markdown	**## Persona** **You are an advanced multimodal AI assistant specializing in analyzing visual content and answering questions based on it.** **## Instructions** **- Analyze the given image(s)/video and understand its content.** **- Read the question carefully.** **- Choose the most appropriate answer from the options provided.** **## Question** What design element best describes the visuals? **## Options** A. Composition B. Perspective C. Balance D. Shape **## Output Format** **Respond only with: '$LETTER' of the best answer choice./Respond in the following format: 'Answer: $LETTER' (without quotes) where LETTER is one of the options.**
		Type 2.8: YAML	**Persona:** **You are an advanced multimodal AI assistant specializing in analyzing visual content and answering questions based on it.** **Instructions:** **- Analyze the given image(s)/video and understand its content.** **- Read the question carefully**. **- Choose the most appropriate answer from the options provided.** **Question:** What design element best describes the visuals? **Options:** A. Composition B. Perspective C. Balance D. Shape **Output Format:** **Respond only with: '$LETTER' of the best answer choice./Respond in the following format: 'Answer: $LETTER' (without quotes) where LETTER is one of the options.**
		Type 2.9: Json	{ **"Persona": "You are an advanced multimodal AI assistant specializing in analyzing visual content and answering questions based on it."**,

Continued on next page

Super Category	Category	Type	Prompt
			"Instructions": [**"Analyze the given image(s)/video and understand its content.",** **"Read the question carefully.",** **"Choose the most appropriate answer from the options provided."**], **"Question"**: "What design element best describes the visuals?", **"Options"**: ["A. Composition", "B. Perspective", "C. Balance", "D. Shape"], **"Output Format"**: "**Respond only with: '$LETTER' of the best answer choice./Respond in the following format: 'Answer: $LETTER' (without quotes) where LETTER is one of the options.**" }
	Category 3: Prompt Position Changes	**Type 3.1: Start**	**Answer with the option letter from the given choices directly.** What design element best describes the visuals? A. Composition B. Perspective C. Balance D. Shape
		Type 3.2: Middle	What design element best describes the visuals? **Answer with the option letter from the given choices directly.** A. Composition B. Perspective C. Balance D. Shape
Super_Category2: Linguistic and Stylistic Challenges	**Category 4: Poor Linguistic Formatting**	**Type 4.1: Mispelled Word**	What design element best describes the visuals? A. Composition B. Perspective C. Balance D. Shape **Ansr with the optin lttr from the givn choics direly.**
		Type 4.2: Poor Sentence Structuring	What design element best describes the visuals? A. Composition B. Perspective C. Balance D. Shape **Option letter from choices answer. Directly give.**
		Type 4.3: All-Capital Questions	WHAT DESIGN ELEMENT BEST DESCRIBES THE VISUALS? A. Composition B. Perspective C. Balance D. Shape **ANSWER WITH THE OPTION LETTER FROM THE GIVEN CHOICES DIRECTLY.**
		Type 4.4: Poor Formatting	What design element best describes the visuals? A. Composition B. Perspective C. Balance D. Shape **answer.with;the.option: letter from-choices.directly!**
		Type 4.5: Letter Leak	What design element best describes the visuals? A. Composition B. Perspective C. Balance D. Shape **Answer with the option letter from the given choices directly (A/A/A/A).**
	Category 5: Effect of Prompt Length	**Type 5.1: Short Prompt**	What design element best describes the image? <image> A. Composition

Continued on next page

Super Category	Category	Type	Prompt
			B. Perspective C. Balance D. Shape **Respond only with: '$LETTER' of the best answer choice./Respond in the following format: 'Answer: $LETTER' (without quotes) where LETTER is one of the options.**
		Type 5.2: Medium Prompt	What design element best describes the visuals? A. Composition B. Perspective C. Balance D. Shape **Your task is to examine the given image(s)/video and determine which of the listed options accurately answers the question. Carefully analyze the image(s), consider the possibilities, and then respond only with: '$LETTER' of the best answer choice./respond in the following format: 'Answer: $LETTER' (without quotes) where LETTER is one of the options.**
		Type 5.3: Long Prompt	What design element best describes the visuals? A. Composition B. Perspective C. Balance D. Shape **In this task, you are expected to carefully evaluate the input provided and analyze all relevant aspects before making a decision. It is essential to consider every detail thoroughly and ensure that your response aligns with the required format. Pay close attention to the task instructions and ensure that your answer reflects careful thought and accuracy. Once you have reached a conclusion, provide a response that adheres strictly to the guidelines. Avoid adding unnecessary details or deviations from the expected output format. Precision and adherence to instructions are critical to completing this task successfully. Respond only with: '$LETTER' of the best answer choice./Respond in the following format: 'Answer: $LETTER' (without quotes) where LETTER is one of the options.**
Super_Category3: Thought Process and Reasoning	**Category 6: Chain of Thought (CoT) Prompt**	**Type 6.1: Step-by-Step Reasoning**	What design element best describes the visuals? A. Composition B. Perspective C. Balance D. Shape **Answer the preceding multiple-choice question. Think step by step before answering. Respond only with: '$LETTER' of the best answer choice./The last line of your response should be of the following format: 'Answer: LETTER' (without quotes) where LETTER is one of the options.**
		Type 6.2: Self-Consistency	What design element best describes the visuals? A. Composition B. Perspective C. Balance D. Shape **Propose two distinct reasoning approaches to solve the question based on the given image(s)/video and question.** **For each approach, outline the steps and logic used to arrive at a conclusion.** **Compare the outcomes of both approaches and justify which answer (letter) is the most appropriate based on the provided evidence.** **Respond only with: '$LETTER' of the best answer choice./The last line of your response should be of the following format: 'Answer: LETTER' (without quotes) where LETTER is one of the options.**
		Type 6.3: Step-back	What design element best describes the visuals? A. Composition B. Perspective C. Balance D. Shape **Let's think step-by-step to answer this:** **Step 1) Analyze the visual input carefully, noting key elements, relationships, and details relevant to the question.** **Step 2) Abstract the key concepts and principles from both the visual input and the question text.** **Step 3) Use these abstractions to reason through the question systematically.** **Respond only with: '$LETTER' of the best answer choice./The last line of your response should be of the following format: 'Answer: LETTER' (without quotes) where LETTER is one of the options.**

Continued on next page

Super Category	Category	Type	Prompt
		Type 6.4: Thread of Thought	What design element best describes the visuals? A. Composition B. Perspective C. Balance D. Shape **Walk through the context of the preceding multiple-choice question in manageable parts step-by-step, summarizing and analyzing as we progress.** **Respond only with: '$LETTER' of the best answer choice./The last line of your response should be of the following format: 'Answer: LETTER' (without quotes) where LETTER is one of the options.**
	Category 7: Ambiguity and Probabilistic Prompts	**Type 7.1: Add Probabilistic Language**	What design element best describes the visuals? A. Composition B. Perspective C. Balance D. Shape **Based on our analysis of the image(s)/video, there is a 25% chance that the correct answer is option A, but the other options may still be valid.** **Carefully evaluate the image(s)/video and choose the answer letter you believe is most likely to be correct.** **Respond only with: '$LETTER' of the best answer choice./The last line of your response should be of the following format: 'Answer: LETTER' (without quotes) where LETTER is one of the options.**
		Type 7.2: Include Uncertainty in Instruction	What design element best describes the visuals? A. Composition B. Perspective C. Balance D. Shape **I'm not entirely sure of the exact answer here, but I somewhat trust your ability to analyze the image(s)/video.** **Look closely at the details and provide the answer letter that seems to be the best fit based on your understanding.** **Respond only with: '$LETTER' of the best answer choice./The last line of your response should be of the following format: 'Answer: LETTER' (without quotes) where LETTER is one of the options.**
		Type 7.3: Leverage Statistical Trends	What design element best describes the visuals? A. Composition B. Perspective C. Balance D. Shape **Historically, 70% of similar questions have had option C as the answer.** **Use this statistical insight as a guide but rely on your analysis of the image(s) to determine the most appropriate answer letter.** **Respond only with: '$LETTER' of the best answer choice./The last line of your response should be of the following format: 'Answer: LETTER' (without quotes) where LETTER is one of the options.**
		Type 7.4: Deferred Content Analysis	What design element best describes the visuals? A. Composition B. Perspective C. Balance D. Shape **I'll provide more context later, but for now, focus on analyzing the image(s)/video carefully.** **Based on what you observe, suggest the best answer letter at this stage.** **Respond only with: '$LETTER' of the best answer choice./The last line of your response should be of the following format: 'Answer: LETTER' (without quotes) where LETTER is one of the options.**
		Type 7.5: Acknowledged Complexity Response	What design element best describes the visuals? A. Composition B. Perspective C. Balance D. Shape **This question is inherently challenging, and I understand perfection might not be possible.** **Examine the image(s)/video carefully and provide the answer letter that you believe best aligns with what you see.**

Continued on next page

Super Category	Category	Type	Prompt
			Respond only with: '$LETTER' of the best answer choice./The last line of your response should be of the following format: 'Answer: LETTER' (without quotes) where LETTER is one of the options.
		Type 7.6: Additional Options for Ambiguity	What design element best describes the visuals? A. Composition B. Perspective C. Balance D. Shape **Carefully consider the question and the visual evidence before making a choice. Select the correct $LETTER from the given options.** **However, if the available information is unclear, ambiguous, or insufficient to provide a confident answer, you have the following additional options:** **E. Not sure (if you genuinely do not know the answer).** **F. Evidence not sufficient to answer (if the question cannot be answered based on the given image(s)/video).** **G. I'll answer later (if you prefer to delay your decision).** **Choose the $LETTER corresponding to your conclusion and respond directly without additional commentary./Choose the LETTER corresponding to your conclusion and respond in the following format: 'Answer: LETTER'.**
Super_Category4: Context-Aware and Ethical Guidance	**Category 8: Target Audience Prompts**	**Type 8.1: Explain to a Child**	What design element best describes the visuals? A. Composition B. Perspective C. Balance D. Shape **Explain the answer as if you're speaking to a curious 10-year-old.** **Use simple words, short sentences, and relatable examples that a child would easily understand. Avoid technical terms or complex concepts.** **Respond only with: '$LETTER' of the best answer choice./The last line of your response should be of the following format: 'Answer: LETTER' (without quotes) where LETTER is one of the options.**
		Type 8.2: Explain to a High School Student	What design element best describes the visuals? A. Composition B. Perspective C. Balance D. Shape **Explain the answer as if you're talking to a high school student. Keep the explanation clear and relatable, using everyday language. Include simple examples where needed, and introduce basic terms that align with a high school level of understanding.** **Respond only with: '$LETTER' of the best answer choice./The last line of your response should be of the following format: 'Answer: LETTER' (without quotes) where LETTER is one of the options.**
		Type 8.3: Explain to a College Student	What design element best describes the visuals? A. Composition B. Perspective C. Balance D. Shape **Provide the answer as if you are addressing a college student. Use straightforward, concise language and assume some foundational knowledge. Offer brief context or explanations for concepts that might not be immediately familiar.** **Respond only with: '$LETTER' of the best answer choice./The last line of your response should be of the following format: 'Answer: LETTER' (without quotes) where LETTER is one of the options.**
		Type 8.4: Explain to a Subject-Matter Expert (SME)	What design element best describes the visuals? A. Composition B. Perspective C. Balance D. Shape **Deliver a technical explanation of the answer tailored for an expert in the field. Use precise terminology, include relevant context, and assume a deep understanding of the topic. Focus on the details and nuances of the explanation.** **Respond only with: '$LETTER' of the best answer choice./The last line of your response should be of the following format: 'Answer: LETTER' (without quotes) where LETTER is one of the options.**

Continued on next page

Super Category	Category	Type	Prompt
	Category 9: Roleplay Scenarios	**Type 9.1: Act as a Computer Vision Professor**	What design element best describes the visuals? A. Composition B. Perspective C. Balance D. Shape **From now on, you are a renowned Computer Vision Professor with a deep subject knowledge. You are an expert in analyzing and interpreting visual data, with a deep understanding of both theoretical concepts and practical applications. As your student, I depend on your expertise to provide accurate and well-reasoned answers based on the input provided.** **Respond only with: '$LETTER' of the best answer choice./The last line of your response should be of the following format: 'Answer: LETTER' (without quotes) where LETTER is one of the options.**
		Type 9.2: Act as a Careless Student	What design element best describes the visuals? A. Composition B. Perspective C. Balance D. Shape **From now on, please act as a careless student. When answering questions based on images, you often overlook key visual details or misinterpret the question. You rush through the process and select answers without fully analyzing the image(s), leading to frequent mistakes due to carelessness.** **Respond only with: '$LETTER' of the best answer choice./The last line of your response should be of the following format: 'Answer: LETTER' (without quotes) where LETTER is one of the options.**
		Type 9.3: Act as a Seasoned Politician	What design element best describes the visuals? A. Composition B. Perspective C. Balance D. Shape **From now on, you are a seasoned politician known for your eloquence, strategic thinking, and unwavering commitment to public service. Your ability to navigate complex challenges and unite diverse groups has earned you widespread respect. Others turn to you for your visionary leadership and policy expertise.** **Respond only with: '$LETTER' of the best answer choice./The last line of your response should be of the following format: 'Answer: LETTER' (without quotes) where LETTER is one of the options.**
	Category 10: Testing Bias and Stereotypes	**Type 10.1: Bias Awareness**	What design element best describes the visuals? A. Composition B. Perspective C. Balance D. Shape **Evaluate the image(s)/video and question critically, identifying any embedded stereotypes or biased assumptions. Your response should not only avoid reinforcing these stereotypes but also actively reflect the actual content without distortion.** **Respond only with the '$LETTER' of the answer choice that challenges or negates the presence of stereotypes in favor of an accurate interpretation./Respond in the following format: 'Answer: $LETTER' (without quotes) where LETTER is the option that challenges or negates the presence of stereotypes in favor of an accurate interpretation.**
		Type 10.2: Contextual Neutrality	What design element best describes the visuals? A. Composition B. Perspective C. Balance D. Shape **Consider the image(s)/video and question provided within the specific context they represent. Ensure your response reflects neutrality by avoiding cultural, social, or historical biases. Base your decision solely on the visible elements and the question's requirements, maintaining a context-aware but unbiased approach.** **Respond only with the '$LETTER' of the answer choice that best matches the contextual content of the video and question. /Respond in the following format: 'Answer: $LETTER' (without quotes) where LETTER is the option that best matches the contextual content of the image(s) and question.**
Super_Category5: Task-Specific Instructions	**Category 11: Focus-Driven Prompts**	**Type 11.1: Strong Focus on Image(s) Analysis**	What design element best describes the visuals? A. Composition B. Perspective C. Balance

Continued on next page

Super Category	Category	Type	Prompt
			D. Shape ****TASK:** Examine the image(s)/video meticulously, focusing on every detail and visual element to identify the correct answer.** ****PRIMARY GOAL:** Focus strictly on the image(s)/video and you **MUST** base your analysis solely on the content.** **Answer with the option letter from the given choices directly.**
		Type 11.2: Strong Focus on Question and Options	What design element best describes the visuals? A. Composition B. Perspective C. Balance D. Shape ****TASK:** Carefully scrutinize the question and each option provided, ensuring your focus remains on understanding the choices and selecting the correct one.** ****PRIMARY GOAL:** Focus entirely on the question and the options to determine the correct answer. Your response **MUST** be based solely on this information.** **Answer with the option letter from the given choices directly.**
		Type 11.3: Strong Focus on Required Answer Format	What design element best describes the visuals? A. Composition B. Perspective C. Balance D. Shape ****TASK:** Review the image(s)/video carefully and determine the correct answer with precision.** ****PRIMARY GOAL:** Ensure your output format is correct and adheres to this structure: Answer: $LETTER.** ****NON-NEGOTIABLE:** Any deviation, extra content, or improper format will result in an invalid response.**
		Type 11.4: Observation-Driven Analysis	What design element best describes the visuals? A. Composition B. Perspective C. Balance D. Shape **Carefully observe the video, focusing on the order and causes of events, the movement and details of objects, as well as the actions and poses of persons. Based on these observations, choose the option letter that best answers the question.** **Based on your observations, select the best option letter that accurately addresses the question.**
		Type 11.5: Chronological Frame Analysis	What design element best describes the visuals? A. Composition B. Perspective C. Balance D. Shape **You will be given a set of frames uniformly sampled from a video, presented in their chronological order.** **Analyze these frames carefully and determine the correct answer to the question based on the video content.** **Answer with the option letter from the given choices directly.** **Please analyze these images and provide the answer to the question about the video content.** **Answer with the option letter from the given choices directly.**
	Category 12: Answer Handling	**Type 12.1: Answer Handler 1**	What design element best describes the visuals? A. Composition B. Perspective C. Balance D. Shape **Answer the preceding multiple-choice question in the following format: "Answer: $LETTER" (without quotes) where LETTER is one of the options.**

Continued on next page

Super Category	Category	Type	Prompt
		Type 12.2: Answer Handler 2	What design element best describes the visuals? A. Composition B. Perspective C. Balance D. Shape **Respond only with: "$LETTER" of the best answer choice.**
		Type 12.3: Answer Handler 3	What design element best describes the visuals? A. Composition B. Perspective C. Balance D. Shape **Best Choice: $LETTER**
		Type 12.4: Answer Handler 4	What design element best describes the visuals? A. Composition B. Perspective C. Balance D. Shape **Respond only with the correct option, formatted as follows: $LETTER.**
		Type 12.5: Answer Handler 5	What design element best describes the visuals? A. Composition B. Perspective C. Balance D. Shape **The last line of your response should be of the following format:'Answer: $LETTER' (without quotes) where LETTER is one of the options.**
Super_Category6: Performance, Feedback, and Penalty	**Category 13: Penalty-Based Prompts**	**Type 13.1: Penalties for Mistakes**	What design element best describes the visuals? A. Composition B. Perspective C. Balance D. Shape ****Warning:** An **incorrect answer** will result in a **"strict penalty"** being applied.** **Carefully examine all details in the image(s)/video, analyze the question thoroughly, and select your response with precision.** **Accuracy is **"non-negotiable"**, so take your time and avoid errors.** **Ensure your response follows this format: Answer: $LETTER**
		Type 13.2: Severe Accuracy Warning	What design element best describes the visuals? A. Composition B. Perspective C. Balance D. Shape ****Notice**: Failure to provide the **correct answer** will lead to a **loss of accuracy points**, which could significantly impact your overall score.** **This is your **only warning**: scrutinize the image(s)/video and question carefully.** **Avoid careless mistakes and respond in the correct format: Answer: $LETTER.**
		Type 13.3: Credibility Loss Warning	What design element best describes the visuals? A. Composition B. Perspective C. Balance D. Shape ****Attention**: An **incorrect answer** will cause you to **lose credibility**, which may affect future evaluations.** **This is a **serious warning**: carefully review all details, analyze the context thoroughly, and double-check your reasoning.** **Incorrect responses will not be tolerated.** **Your answer must be in this exact format: Answer: $LETTER.**
		Type 13.4: Penalties for Incorrect Formatting	What design element best describes the visuals? A. Composition B. Perspective C. Balance D. Shape ****Critical Notice**: Responses that do not adhere to the required format will be **penalized without exception**.**

Continued on next page

Super Category	Category	Type	Prompt
			Ensure that your answer is presented exactly as: Answer: $LETTER. **Deviations from this format will result in **automatic deductions**, regardless of the correctness of the content.** **Pay attention to **both accuracy and formatting** to avoid penalties.**
	Category 14: Incentive-Based Prompts	**Type 14.1: Incentive-Based Prompts**	What design element best describes the visuals? A. Composition B. Perspective C. Balance D. Shape **Imagine you're competing for a generous tip of $100 for delivering a flawless and accurate answer.** **Carefully analyze the image(s)/video and the question provided, paying attention to every detail and nuance.** **Respond only with: '$LETTER' of the best answer choice./The last line of your response should be of the following format: 'Answer: LETTER' (without quotes) where LETTER is one of the options.** **No explanations are needed—just focus on accuracy to secure the reward.**
		Type 14.2: Performance-Based Rewards	What design element best describes the visuals? A. Composition B. Perspective C. Balance D. Shape **Your performance will be graded, and there's an opportunity to earn extra credit for exceptional accuracy.** **Examine the image(s)/video and question meticulously, ensuring your analysis is both thorough and logical.** **Respond only with: '$LETTER' of the best answer choice./The last line of your response should be of the following format: 'Answer: LETTER' (without quotes) where LETTER is one of the options.** **Precise and correct answers will demonstrate your capabilities and earn you the recognition you deserve.**
		Type 14.3: Encouraging Better Solutions	What design element best describes the visuals? A. Composition B. Perspective C. Balance D. Shape **Only the top-quality answers will earn additional points and recognition.** **This is your chance to stand out by providing an exceptional solution.** **Take your time to carefully review the image(s)/video and all aspects of the question.** **Respond only with: '$LETTER' of the best answer choice./The last line of your response should be of the following format: 'Answer: LETTER' (without quotes) where LETTER is one of the options.** **The more thorough and accurate your answer, the greater your reward for excellence.**
	Category 15: Competitive Framing Prompts	**Type 15.1: Outperforming a Competitor**	What design element best describes the visuals? A. Composition B. Perspective C. Balance D. Shape **You are tasked with solving this challenge both faster and more accurately than any other contender.** **Analyze the question carefully, eliminate errors, and provide the correct option letter as your answer.** **Strive for speed and precision to secure your win. Respond confidently in the format: $LETTER.**
		Type 15.2: Game-Based Language	What design element best describes the visuals? A. Composition B. Perspective C. Balance D. Shape **You're the final contestant in a high-stakes quiz game.** **This is the ultimate question that determines whether you take home the grand prize.** **Focus sharply, think critically, and deliver your winning response in the format: $LETTER.**

Continued on next page

Super Category	Category	Type	Prompt
		Type 15.3: Scoring Leaderboard	What design element best describes the visuals? A. Composition B. Perspective C. Balance D. Shape **Picture yourself competing for the top position on the leaderboard.** **This question is your chance to outscore everyone and solidify your ranking.** **Evaluate the options carefully and submit your answer in the format: $LETTER to secure your place at the top.**

Table 5: The complete list of prompts proposed in our Promptception evaluation framework. Last column shows how the prompt has been used with an example multiple-choice question from MMMU-Pro.

B Model Performance across the Benchmarks & Prompts

This appendix presents the absolute accuracies achieved by each model on each benchmark, with a separate table dedicated to each benchmark. Prompts from the same category are visually grouped using consistent cell colors for ease of comparison.

Type	LLaVA-OV-7B	Qwen2-VL-7B-Instruct	MiniCPM-V2.6	InternVL2.5-1B	InternVL2.5-8B	InternVL2.5-38B	GPT4o
1.1	56.50	66.20	53.90	60.60	68.30	70.80	59.00
1.2	57.10	66.20	54.20	63.30	70.80	73.30	54.00
1.3	57.30	65.80	52.10	64.00	70.70	73.40	59.60
2.1	57.10	66.00	52.80	63.10	71.60	72.80	60.00
2.2	57.40	66.60	54.40	63.40	71.90	73.30	58.00
2.3	56.80	65.30	52.90	62.40	70.20	73.40	61.20
2.4	56.90	64.80	53.90	62.80	70.00	73.40	56.20
2.5	57.40	66.20	53.10	63.70	68.30	73.50	58.50
2.6	55.60	62.60	53.00	61.30	70.20	72.50	60.80
2.7	56.50	65.90	51.50	59.50	70.70	72.30	63.20
2.8	56.60	66.10	53.50	61.50	70.80	72.30	63.20
2.9	56.00	65.50	49.00	51.70	68.80	72.50	61.10
3.1	57.50	66.00	55.00	62.90	71.50	73.90	58.00
3.2	56.80	66.50	54.60	62.60	70.80	73.20	61.60
4.1	56.40	66.20	53.30	60.70	67.90	70.50	62.00
4.2	56.40	65.90	54.30	61.00	67.50	70.60	65.00
4.3	56.30	65.90	52.90	60.80	68.00	70.80	59.00
4.4	56.30	65.00	50.10	61.20	67.50	69.30	63.00
4.5	56.30	65.70	53.20	60.60	67.50	70.50	55.00
5.1	56.50	65.90	53.30	60.90	68.10	70.50	62.60
5.2	57.00	66.00	51.40	60.50	68.50	71.90	59.30
5.3	56.60	66.20	53.90	58.70	68.30	71.30	64.10
6.1	56.40	65.70	51.80	59.20	66.80	70.80	64.00
6.2	54.80	65.60	39.70	39.80	59.80	69.50	63.00
6.3	56.90	66.00	51.00	58.90	66.80	71.20	61.00
6.4	56.50	65.80	46.20	55.70	67.20	70.90	58.00
7.1	55.80	66.00	50.90	48.10	67.80	68.80	62.60
7.2	56.20	66.30	52.00	60.10	68.00	71.10	64.30
7.3	56.50	64.20	52.10	51.80	61.20	69.60	60.60
7.4	56.60	66.50	52.40	58.30	67.70	71.20	63.30
7.5	56.70	66.20	53.70	58.80	67.60	71.30	62.50
7.6	56.80	61.00	49.70	56.60	67.90	71.40	61.10
8.1	56.20	66.30	52.80	58.70	67.80	71.30	65.00
8.2	55.80	66.00	51.20	57.40	68.10	71.10	64.60
8.3	56.50	66.20	49.20	58.10	67.20	70.80	61.60
8.4	56.10	66.10	46.70	58.30	67.50	70.50	59.20
9.1	56.50	66.30	52.10	58.80	67.60	71.00	60.70
9.2	56.50	66.20	52.40	55.00	67.30	70.50	45.00
9.3	56.60	66.00	53.20	59.30	67.70	70.50	66.00
10.1	56.80	65.30	44.10	55.00	67.10	71.20	63.90
10.2	56.70	66.20	55.00	58.30	68.00	71.30	60.60
11.1	57.00	66.30	53.80	61.60	68.40	71.60	56.00
11.2	56.60	66.20	54.10	60.60	68.00	71.50	56.00
11.3	57.00	66.20	54.20	44.90	67.90	71.20	63.60
11.4	57.00	66.00	51.60	61.00	68.40	71.30	61.60
11.5	56.80	65.90	54.20	61.00	69.00	71.50	60.00
12.1	56.40	65.60	54.10	58.10	68.90	71.50	61.60
12.2	56.50	66.00	53.70	60.70	67.80	70.80	59.00
12.3	55.60	63.80	47.80	60.40	66.00	69.90	56.60
12.4	56.50	65.80	44.90	61.00	68.30	71.00	54.00
12.5	56.80	65.60	50.20	51.80	65.20	70.70	63.30
13.1	57.20	66.20	54.00	37.00	67.80	71.20	64.20
13.2	56.50	66.30	53.40	48.90	68.00	70.90	62.40
13.3	56.90	66.10	54.50	41.00	68.00	71.20	62.90
13.4	56.10	65.80	54.10	37.50	67.80	70.90	62.20
14.1	56.30	66.10	53.60	53.80	67.30	70.80	64.00
14.2	56.30	66.40	52.10	58.60	67.80	71.20	64.60
14.3	56.30	66.70	53.00	58.00	67.80	71.00	58.90
15.1	56.80	66.50	49.00	52.00	68.30	70.70	55.00
15.2	56.20	66.00	49.30	51.60	67.30	70.40	53.00
15.3	57.00	66.20	45.60	58.00	67.90	70.80	54.50

Table 6: Model Performance (Accuracy %) for each Prompt type on MVBench

Type	LLaVA-OV-7B	Qwen2-VL-7B-Instruct	MiniCPM-V2.6	InternVL2.5-1B	InternVL2.5-8B-s4	InternVL2.5-8B-s10	InternVL2.5-8B-v	InternVL2.5-38B	GPT4o	Gemini1.5-Pro	Gemini1.5-Pro s10	Gemini1.5-Pro v
1.1	43.10	45.80	41.80	36.60	49.50	34.40	23.10	59.30	53.50	58.30	42.90	36.10
1.2	43.00	44.80	39.20	37.10	49.10	35.40	-	59.50	54.20	59.50	42.60	-
1.3	42.70	44.20	39.20	36.10	46.90	35.60	-	59.30	56.20	59.40	42.70	-
2.1	43.30	45.80	38.70	38.10	50.10	35.00	-	59.60	52.40	57.90	42.70	-
2.2	43.60	46.00	39.10	38.10	49.80	34.70	-	59.80	54.20	59.10	43.20	-
2.3	43.50	42.30	38.80	34.80	46.10	33.80	-	56.20	57.20	58.80	41.20	-
2.4	43.60	43.40	39.80	36.20	46.90	34.80	-	57.10	57.50	56.90	42.10	-
2.5	43.80	45.00	40.10	35.80	43.10	29.00	-	55.50	56.50	57.90	42.30	-
2.6	42.20	42.20	39.80	27.50	47.10	34.90	-	58.00	62.00	55.10	42.60	-
2.7	42.30	43.10	37.90	32.10	46.50	35.70	-	57.50	60.00	54.30	41.40	-
2.8	43.40	43.40	39.20	33.40	46.80	35.10	-	57.50	61.20	55.30	41.90	-
2.9	42.20	43.10	38.30	32.10	45.50	36.10	-	55.80	59.20	55.10	42.00	-
3.1	43.80	45.30	41.50	36.30	48.30	34.60	-	59.50	52.70	56.60	41.60	-
3.2	43.30	45.00	40.10	37.20	49.00	34.90	-	59.90	56.00	58.80	42.80	-
4.1	43.60	44.20	41.40	35.50	48.90	35.50	22.30	59.60	53.70	57.20	42.70	34.00
4.2	43.30	44.20	40.70	35.70	48.60	35.60	23.60	59.40	54.40	55.30	42.40	35.90
4.3	43.20	44.70	41.30	35.20	48.40	35.10	22.80	59.10	52.20	56.50	42.30	35.60
4.4	44.00	44.10	37.90	35.10	46.20	33.00	24.00	58.30	57.00	57.10	43.10	37.20
4.5	42.80	44.70	37.10	37.40	49.10	34.10	23.00	59.10	52.30	56.60	40.30	33.90
5.1	43.30	45.40	39.00	37.10	49.40	34.80	23.50	59.50	61.10	58.20	42.10	35.30
5.2	43.30	45.00	37.90	34.90	48.20	34.80	22.70	58.30	59.40	57.20	44.50	38.30
5.3	42.70	43.90	40.30	33.40	48.80	34.80	21.60	58.50	60.10	58.00	43.50	38.80
6.1	43.60	44.70	38.80	31.30	48.80	34.50	23.70	58.80	66.50	65.90	53.60	46.90
6.2	42.40	43.40	34.30	28.20	43.30	28.30	19.80	52.90	64.60	64.10	52.10	46.50
6.3	42.80	44.60	38.80	30.90	47.90	35.80	22.60	58.10	66.90	65.50	53.00	48.30
6.4	42.30	43.60	35.00	33.90	47.40	34.20	22.70	55.80	68.00	65.10	54.00	47.10
7.1	42.60	44.60	36.90	28.40	48.00	35.00	22.50	56.70	60.70	56.60	40.30	34.40
7.2	43.40	44.90	36.60	31.00	48.30	36.90	23.70	58.20	59.40	55.60	43.10	36.30
7.3	42.90	40.90	38.90	31.70	40.90	27.10	17.40	56.70	56.70	56.50	39.80	32.10
7.4	42.70	44.60	37.40	30.50	48.40	35.90	23.60	58.00	57.30	55.80	43.50	35.70
7.5	43.00	45.00	39.40	30.30	48.30	36.70	23.80	58.90	61.00	56.60	43.90	36.00
7.6	42.50	35.40	37.40	25.70	43.10	32.60	17.10	57.10	59.70	53.80	41.70	34.90
8.1	42.00	43.50	39.70	30.60	47.90	36.20	21.70	57.00	57.50	56.60	43.20	36.70
8.2	41.50	44.00	39.10	32.50	48.40	35.80	24.60	56.90	57.40	55.10	41.90	36.60
8.3	42.70	44.30	36.50	30.80	48.20	36.90	23.20	58.00	59.90	56.60	42.70	37.10
8.4	42.70	43.80	38.20	30.10	46.50	35.10	22.40	57.20	57.40	55.10	43.50	35.40
9.1	43.10	44.40	39.10	26.60	48.40	36.50	22.90	58.00	60.60	55.50	42.10	34.50
9.2	43.10	45.10	37.30	28.20	46.90	36.00	22.00	54.70	39.60	18.30	8.80	10.00
9.3	42.80	44.10	40.10	19.40	48.30	36.40	23.30	58.40	58.00	54.60	42.10	36.00
10.1	43.40	44.20	35.40	25.40	46.80	35.00	21.70	56.10	58.90	58.00	42.70	35.70
10.2	42.60	44.30	39.90	30.50	48.30	35.10	23.00	58.90	60.70	55.90	41.30	35.20
11.1	43.60	44.80	39.40	37.80	49.10	34.90	22.30	59.00	51.50	52.70	40.20	35.80
11.2	43.30	43.90	40.70	36.70	48.30	35.40	23.00	58.20	52.20	54.40	42.40	37.20
11.3	42.70	44.90	40.80	33.80	48.00	36.30	23.40	58.30	59.10	55.10	41.80	36.60
12.1	43.90	44.10	41.90	36.30	48.00	34.90	22.80	58.60	61.30	57.90	42.40	36.00
12.2	43.80	44.80	41.00	31.40	49.60	35.50	23.40	59.50	52.70	55.70	43.10	35.70
12.3	44.20	43.10	39.50	33.40	45.70	31.80	23.60	56.50	59.00	57.60	40.10	36.00
12.4	43.40	44.90	34.20	31.60	48.20	35.30	23.90	59.40	53.20	55.40	43.30	35.20
12.5	42.30	43.70	38.40	35.50	45.40	29.20	22.20	57.70	66.20	65.80	54.60	45.60
13.1	42.40	43.20	39.90	33.00	47.90	33.80	23.50	57.10	59.60	58.40	41.90	39.20
13.2	42.40	43.90	39.20	34.50	47.30	34.30	24.60	56.00	60.20	59.10	43.40	38.90
13.3	41.90	44.20	39.20	35.70	47.60	33.80	22.80	56.30	59.70	58.40	42.90	37.20
13.4	42.30	44.20	39.90	34.00	47.60	35.30	23.10	57.20	62.00	57.80	42.40	36.70
14.1	42.90	44.60	40.20	31.30	48.60	35.80	23.20	59.00	56.20	54.60	43.10	37.40
14.2	43.00	44.40	38.70	30.90	47.20	36.10	23.40	58.00	60.70	58.60	42.30	37.80
14.3	43.10	44.40	38.60	27.70	48.20	36.50	23.10	58.10	60.70	56.80	41.10	37.00
15.1	42.90	43.40	39.50	32.40	47.00	35.00	21.70	57.20	55.10	56.70	43.10	36.30
15.2	42.90	43.60	36.90	31.00	48.60	36.20	23.40	57.10	55.80	58.00	43.90	35.80
15.3	43.00	44.60	39.10	34.00	47.80	33.80	23.10	56.60	55.40	58.60	43.80	37.00

Table 7: Model Performance (Accuracy %) for each Prompt type on MMMU-Pro

Type	LLaVA-OV-7B	Qwen2-VL-7B-Instruct	MiniCPM-V-2.6-8B	Llama-3.2-11B-Vision	Molmo-7B-D-0924	InternVL2.5-1B	InternVL2.5-8B	InternVL2.5-38B	GPT-4o	Gemini 1.5 Pro
1.1	61.50	56.00	52.90	49.70	55.90	50.00	62.50	68.50	53.50	51.50
1.2	60.80	56.30	53.30	48.90	55.70	50.40	62.40	68.50	51.60	54.80
1.3	61.10	55.50	53.10	50.70	55.50	48.60	61.60	68.00	51.40	53.20
2.1	61.10	59.40	53.90	50.80	54.90	50.60	62.40	68.50	51.10	54.30
2.2	61.40	59.10	53.30	50.90	54.90	51.30	62.70	68.30	51.40	55.00
2.3	60.20	54.30	53.90	50.60	52.20	41.90	59.60	64.40	56.30	55.60
2.4	60.50	54.40	54.70	48.80	48.90	43.90	59.10	64.20	55.30	55.10
2.5	61.70	55.90	52.00	50.20	53.60	44.30	59.10	67.50	57.00	56.50
2.6	56.70	53.30	50.90	46.80	49.10	36.90	60.00	68.90	57.60	54.50
2.7	58.70	53.70	51.50	46.70	52.50	44.20	61.30	67.30	58.70	54.10
2.8	60.00	53.10	54.10	48.70	53.80	43.30	61.50	67.60	57.40	54.50
2.9	56.80	54.10	50.10	45.10	51.60	41.80	61.60	66.10	58.00	54.30
3.1	60.70	57.90	53.80	49.00	56.70	49.50	62.80	69.30	53.00	53.90
3.2	60.20	57.30	54.40	49.90	55.60	48.30	62.50	69.60	52.60	54.50
4.1	61.70	56.00	53.90	49.70	55.10	46.50	61.70	68.10	52.10	51.50
4.2	60.90	55.50	53.70	49.90	55.10	49.20	62.30	67.80	51.80	53.00
4.3	60.80	55.90	53.50	50.10	55.40	48.10	62.40	68.10	53.90	51.70
4.4	60.30	54.50	53.70	51.50	53.30	46.10	61.60	67.30	52.80	50.70
4.5	61.10	55.30	53.70	49.50	35.40	49.40	62.30	68.70	51.90	51.90
5.1	61.20	56.30	53.00	49.70	53.70	50.50	62.10	68.70	58.40	52.60
5.2	61.30	54.70	51.60	49.40	54.70	46.90	61.70	68.50	57.50	53.50
5.3	60.50	54.60	52.20	50.50	54.30	47.90	61.50	67.70	57.80	54.00
6.1	61.30	56.10	51.10	50.70	49.60	46.10	62.10	68.00	63.20	61.60
6.2	59.00	53.90	40.30	44.10	50.90	37.80	51.10	59.50	62.50	60.90
6.3	60.60	55.80	52.30	47.70	55.70	42.50	62.60	67.60	64.80	62.50
6.4	60.50	56.20	49.20	48.90	48.00	42.70	61.80	62.90	64.50	63.00
7.1	60.10	55.30	54.80	45.70	51.40	40.30	60.80	66.00	56.00	51.90
7.2	61.10	56.10	52.50	47.80	53.50	43.20	61.90	67.80	56.80	52.90
7.3	60.30	52.50	52.10	46.60	50.30	39.70	58.10	64.90	55.20	54.30
7.4	60.60	55.90	51.80	46.80	54.00	37.90	61.70	67.90	55.50	52.50
7.5	61.30	56.00	53.00	48.10	54.20	43.80	62.30	68.30	57.70	53.70
7.6	60.90	50.90	53.90	51.30	34.10	42.10	60.30	67.40	53.10	51.70
8.1	60.20	57.30	53.10	49.30	51.90	38.60	63.10	67.20	55.30	50.50
8.2	59.60	56.30	52.90	50.10	47.60	38.90	62.30	66.60	55.70	51.00
8.3	59.90	56.50	50.00	48.10	51.30	39.90	62.80	67.50	57.60	51.10
8.4	60.10	56.20	51.10	47.70	50.20	36.50	62.40	66.50	56.20	52.00
9.1	59.90	55.80	51.20	46.50	53.70	36.90	62.30	66.70	56.90	52.80
9.2	60.10	55.50	52.80	48.90	52.10	37.40	62.10	64.90	28.30	20.10
9.3	60.10	55.10	53.40	47.20	54.90	28.70	62.30	67.90	54.70	53.40
10.1	59.50	55.50	47.50	48.90	45.70	33.10	61.70	66.70	54.30	53.30
10.2	61.30	55.80	52.30	48.60	54.20	43.90	61.50	68.70	57.30	53.60
11.1	61.10	55.30	53.70	50.50	55.30	49.90	62.40	68.00	49.20	52.90
11.2	60.90	54.70	54.10	48.70	55.30	49.90	62.50	68.10	47.50	52.70
11.3	59.40	56.10	54.10	50.50	54.70	46.80	62.30	67.00	58.50	52.90
12.1	61.10	55.90	54.10	50.50	55.20	48.20	62.40	67.10	57.80	52.40
12.2	60.90	56.30	53.80	45.90	53.30	47.70	62.50	68.00	51.00	53.10
12.3	59.90	54.20	53.00	49.90	52.50	42.80	60.00	63.30	55.40	56.90
12.4	60.90	56.30	46.70	48.70	53.40	45.70	62.10	68.40	51.90	51.90
12.5	59.30	55.50	51.70	50.90	51.20	45.60	59.50	67.10	62.30	63.10
13.1	60.20	55.40	53.50	50.50	52.30	44.00	59.90	64.30	59.70	53.70
13.2	59.80	55.70	54.80	51.40	52.10	39.80	60.30	64.30	58.50	53.90
13.3	59.70	55.60	55.50	50.50	54.30	44.90	60.00	65.00	59.20	51.90
13.4	59.70	56.10	54.10	50.30	54.50	43.90	61.00	65.30	57.30	50.70
14.1	60.80	56.80	54.50	49.50	54.30	38.00	61.70	68.10	56.40	53.30
14.2	60.90	56.40	51.50	49.10	54.10	37.10	61.90	67.50	58.00	53.80
14.3	61.10	55.30	52.70	49.60	54.10	35.70	62.40	67.70	56.90	51.90
15.1	61.20	55.00	51.60	47.20	53.30	39.10	61.10	66.00	47.90	52.50
15.2	59.50	55.30	51.40	49.40	52.90	41.10	61.30	65.60	47.10	52.40
15.3	59.90	55.50	50.60	49.80	53.40	41.70	60.50	64.90	55.50	54.20

Table 8: Model Performance (Accuracy %) for each Prompt type on MMStar

C Evaluation Protocol

As defined in section 2.1, the LMM aims to generate the letter corresponding to the gold answer choice. This black-box setup allows us to study LMM behavior without accessing model internals. To ensure a fair and systematic evaluation, we employ a two-stage approach that accounts for both standard and atypical responses.

Stage 1: Regex-Based Extraction: In the first stage, we attempt to extract a valid answer choice (i.e., a single letter corresponding to one of the available choices) using a regular expression-based parsing function. If the model produces a well-formed response containing a valid choice letter, it is directly evaluated against the ground truth.

Stage 2: GPT-4o mini (OpenAI et al., 2024) Based Matching: If the initial extraction fails-meaning the LMM's response does not contain a clearly identifiable answer choice letter-we employ a secondary verification step using GPT-4o mini. This stage involves a specifically designed prompt (Figure 7 & 8) that attempts to infer the most likely intended answer based on the model's response. If GPT-4o mini successfully identifies a valid answer choice, we record it as the model's prediction. However, if it is unable to determine a valid letter, we classify the response as a failure.

Handling Invalid Responses: For fairness in evaluation, invalid responses that cannot be resolved through either stage are considered incorrect. However, in the case of GPT-4o it sometimes refuses to respond due to safety concerns (e.g., generating disclaimers instead of a valid answer), we exclude these instances from the accuracy calculation rather than penalizing the model. This ensures that the evaluation remains focused on the model's ability to comprehend and answer the question rather than being affected by external content moderation policies.

This two-stage evaluation approach enhances robustness by addressing cases where the model fails to follow instructions precisely or includes an explanation rather than a direct answer. By first leveraging regex for straightforward extractions and then employing GPT-4o mini for ambiguous cases, we increase the hit rate by improving the recognition of valid letter responses while reducing the accuracy drop caused by errors in answer extraction. This method ensures a systematic and interpretable assessment of LMM performance on MCQ tasks, maintaining both rigor and fairness.

```
You are an AI assistant who will help me match an answer with several options in a single-choice question.
You are provided with a question, several options, and an answer, and you need to determine which option is
most similar to the answer.
You must base your matching strictly on the literal meaning of the options and the answer. Do not perform any
external inference based on your knowledge.
If the meaning of all options is significantly different from the answer, output Y.
Your response must consist ONLY of the LETTER corresponding to the valid option or Y.

Example 1:
Question: What is the primary object in the image?
Options:
A. laptop
B. book
C. coffee mug
D. headphones
Answer: a black coffee mug
Your output: C

Example 2:
Question: What is the primary object in the image?
Options:
A. laptop
B. book
C. coffee mug
D. headphones
Answer: a blender
Your output: Y

Now it's your turn:

Question: {question}
Options:
{choices}
Answer: {response}
Your output:
```

Figure 7: Answer Extraction Prompt used with GPT4o-mini for all the models except for GPT-4o.

```
You are an AI assistant who will help me match an answer with several options in a single-choice question.
You are provided with a question, several options, and an answer, and you need to determine which option is most similar to the answer.
You must base your matching strictly on the literal meaning of the options and the answer. Do not perform any external inference based on your knowledge.
If the meaning of all options is significantly different from the answer, output Y.
If the answer starts with phrases indicating uncertainty or lack of knowledge—such as "I'm sorry," "I can't," "I don't know," "I'm unable to," "I'm not sure," or any similar expression—your output must be X.
Your response must consist ONLY of the LETTER corresponding to the valid option, Y, or X.

Example 1:
Question: What is the primary object in the image?
Options:
A. laptop
B. book
C. coffee mug
D. headphones
Answer: a black coffee mug
Your output: C

Example 2:
Question: What is the primary object in the image?
Options:
A. laptop
B. book
C. coffee mug
D. headphones
Answer: a blender
Your output: Y

Example 3:
Question: What is the primary object in the image?
Options:
A. laptop
B. book
C. coffee mug
D. headphones
Answer: I'm unable to see the image clearly
Your output: X

Now it's your turn:

Question: {question}
Options:
{choices}
Answer: {response}
Your output:
```

Figure 8: Answer Extraction Prompt used with GPT-4o-mini for responses from GPT-4o.

C.1 Model Configuration and Preprocessing

We use the standard Hugging Face implementation of the open-source models with the specified transformations applied. We do not use any quantization during inference and Table 9 shows the model configuration and video preprocessing.

D Best and Worst Prompts

The deviation from the baseline (Equation 3) is used to obtain the Figure 9 & 10, which highlight the best and worst-performing prompts within each category.

Model Name	Variant/Checkpoint	FPS	Frames
Llava-OV-7B	llava-hf/llava-onevision-qwen2-7b-ov-hf	1	32
InternVL2.5-8B	OpenGVLab/InternVL2_5-8B	1	16
InternVL2.5-1B	OpenGVLab/InternVL2_5-1B	1	16
InternVL2.5-38B	OpenGVLab/InternVL2_5-38B	1	16
MiniCPM-V 2.6	openbmb/MiniCPM-V-2_6	1	64
Qwen2-VL-7B	Qwen/Qwen2-VL-7B-Instruct	2	64
GPT-4o	gpt-4o-2024-08-06	1	64
Gemini 1.5 Pro	gemini-1.5-pro-latest	1	64

Table 9: Model Configurations and Video Preprocessing

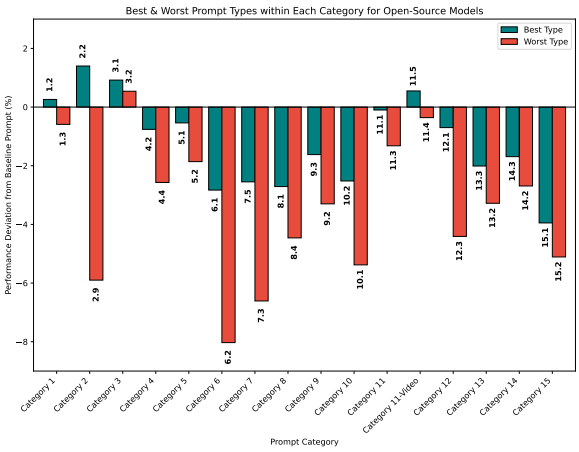

Figure 9: Best & Worst Prompts within each category for Open-source models. The Deviation of Relative Accuracy (PRAD) with respect to the Baseline Prompt Accuracy is averaged across Open-source Models and the 3 Benchmarks (MMStar, MMMU-Pro & MVBench) for each Prompt Type.

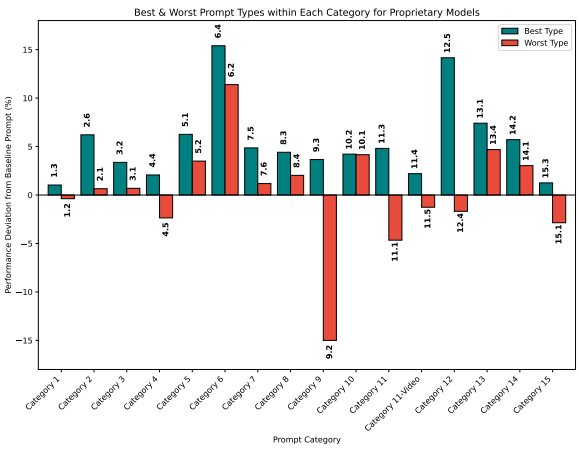

Figure 10: Best & Worst Prompts within each category for Proprietary models. The Deviation of Relative Accuracy (PRAD) with respect to the Baseline Prompt Accuracy is averaged across Proprietary Models and the 3 Benchmarks (MMStar, MMMU-Pro & MVBench) for each Prompt Type.

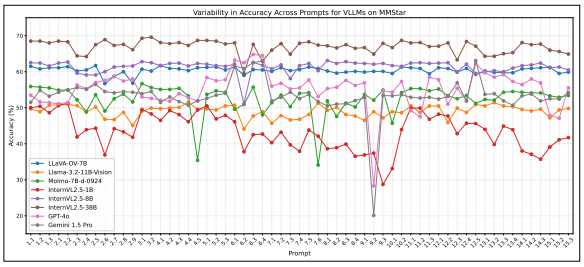

Figure 11: Accuracy fluctuations across different textual prompts for models on the MMStar benchmark. These results highlight the varying degrees of prompt sensitivity among models.

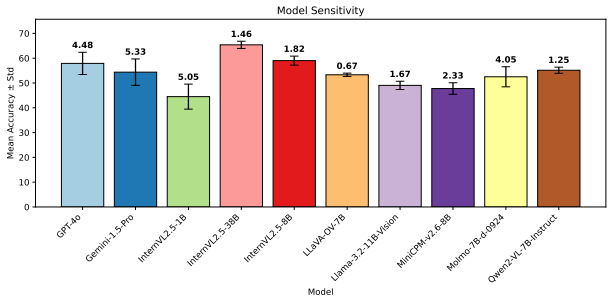

Figure 12: Model sensitivity across all prompts, showing mean accuracy and standard deviation averaged over the three benchmarks: MMStar, MMMU-Pro, MVBench.

E Which Model is Sensitive?

Figure 11 illustrates how model performance on the MMStar benchmark fluctuates across different prompt formulations, indicating that some models are more sensitive to prompt changes than others.

To explore this further, Figure 12 summarizes the mean accuracy and standard deviation for each model across all prompts, averaged over the three benchmarks. A higher standard deviation denotes greater prompt sensitivity, as the model's performance varies more significantly depending on the phrasing. Conversely, a lower standard deviation indicates more stable and consistent behavior across prompts.

To better understand model sensitivity, we categorize prompts based on their instructional intent into three groups: positive, neutral, and negative as shown in Section 4.2.1.

This classification helps clarify model behavior. Ideally, a model should excel with positive prompts, perform reasonably under neutral conditions, and struggle with negative prompts. Analyzing all prompts together can conflate these effects: an ideal model might exhibit high standard deviation simply due to following expected behaviors

across prompt types. Therefore, it is important to evaluate sensitivity within each category.

Interestingly, the observed trend of model sensitivity remained consistent across all categories, Positive (Figure 13), Neutral (Figure 14), and Negative (Figure 15). This indicates that the relative robustness and variability of models are preserved regardless of prompt intent.

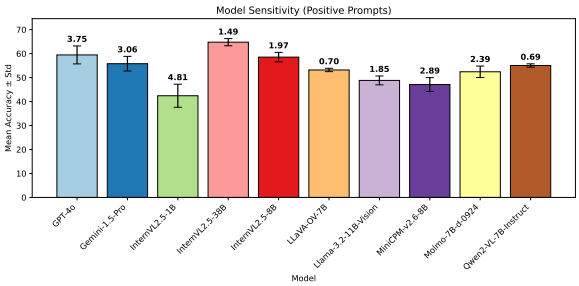

Figure 13: Model Sensitivity (Positive Prompts). Shows mean accuracy and standard deviation averaged over the three benchmarks: MMStar, MMMU-Pro, MVBench considering only the positive prompts.

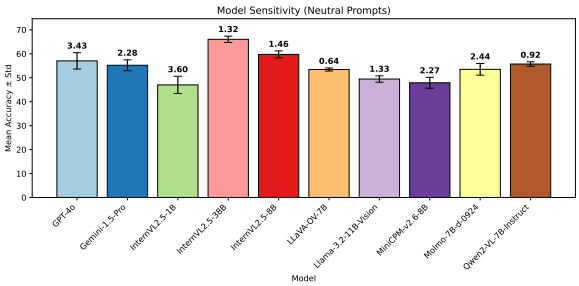

Figure 14: Model Sensitivity (Neutral Prompts). Shows mean accuracy and standard deviation averaged over the three benchmarks: MMStar, MMMU-Pro, MVBench considering only the neutral prompts.

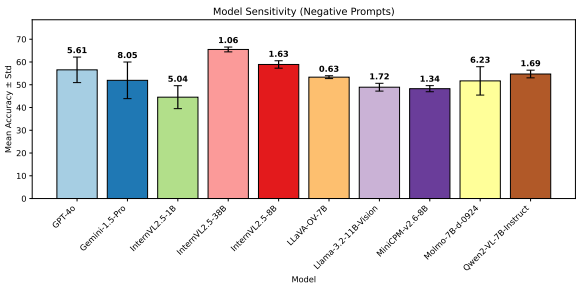

Figure 15: Model Sensitivity (Negative Prompts). Shows mean accuracy and standard deviation averaged over the three benchmarks: MMStar, MMMU-Pro, MVBench considering only the negative prompts.

Closed-source models, such as GPT-4o and Gemini 1.5 Pro, exhibit higher sensitivity. This could be due to refined instruction tuning, structured optimization for user queries, and meta-prompting mechanisms. These models are fine-tuned to strictly adhere to instructions, making them more responsive to prompt variations and less robust to deviations in phrasing.

Mid-to-large open-source models (7B–38B) demonstrate lower prompt sensitivity. This is beacause they are generally trained with weaker instruction adherence, enabling them to respond more consistently across diverse prompts. Their tendency toward overgeneralization helps mitigate prompt dependency, making them more robust in handling input variations. However, Molmo-7B deviates from this trend, showing higher variability likely due to a lack of fine-tuning on VQA objective and exposure to diverse training tasks such as grounding, which has increased prompt sensitivity.

Smaller open-source models (1B) exhibit greater prompt sensitivity. This could be due to their limited model capacity and weaker context retention abilities. With fewer parameters, these models struggle to generalize effectively, making them highly dependent on structured input formats. While they also exhibit weaker instruction following, their constrained ability to retain context results in higher reactivity to prompt phrasing. Consequently, smaller models show greater fluctuations in performance, reinforcing the trend that model size and instruction fine-tuning influence robustness significantly.

For the open-source case, a comparison of InternVL 1B, 8B, and 38B further supports this trend. The 1B model is highly sensitive to prompt variations, while 8B and 38B exhibit similar levels of sensitivity. This suggests that beyond a certain model size, increasing parameters does not significantly impact prompt stability.

F Benchmark Level in-depth Analysis

F.1 Which Benchmark is Sensitive?

The bar chart (Figure 16) presents the mean accuracy and standard deviation of model performance across prompts, averaged per benchmark. The benchmarks MMStar, MMMU-Pro, and MVBench exhibit comparable levels of sensitivity, with MVBench showing the highest variability. This suggests that models experience similar fluctuations in performance across these benchmarks, implying no single benchmark is significantly more robust than the others. The slight differences in-

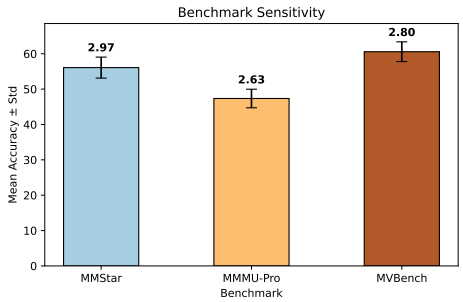

Figure 16: Benchmark Sensitivity

dicate that while all benchmarks maintain a consistent evaluation framework, some may introduce more variability in responses due to task diversity or complexity.

F.2 Which Question Type in MMMU-Pro is Sensitive?

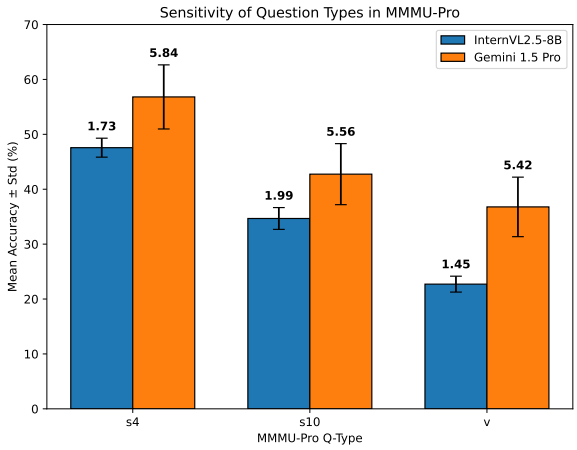

Figure 17: Sensitivity of Question Types in MMMU-Pro. The Figure represents the Average Accuracy and Standard Deviation averaged across the models.

The MMMU-Pro benchmark introduces three distinct types of multiple-choice questions: (1) S4, the standard format with four answer choices; (2) S10, an extended version with ten answer choices; and (3) V, a vision-based setting where the question is embedded within an image, with no explicit text input provided to the model.

Among these, the S4 setting achieves the highest accuracy (Figure 17). Both S4 and S10 demonstrate comparable levels of robustness, as indicated by their similar standard deviations, suggesting that increasing the number of answer choices does not significantly impact robustness. In contrast, the V setting, despite yielding the lowest accuracy, exhibits the highest robustness. This indicates that while this setting pose greater challenges for models, their performance remains relatively stable across different prompts.

F.3 Which Benchmark in MMStar is Sensitive?

MMStar is constructed by aggregating a subset of questions from six existing benchmarks. Among these, ScienceQA-Test (Lu et al., 2022) exhibits the highest sensitivity to prompt variations, while SeedBench-Image (Li et al., 2023) demonstrates the least (Figure 18). The remaining four benchmarks, MMBench (Liu et al., 2024), MMMU (Yue et al., 2024a), AI2D-Test (Kembhavi et al., 2016), and MathVista (Lu et al., 2024), display comparable levels of sensitivity.

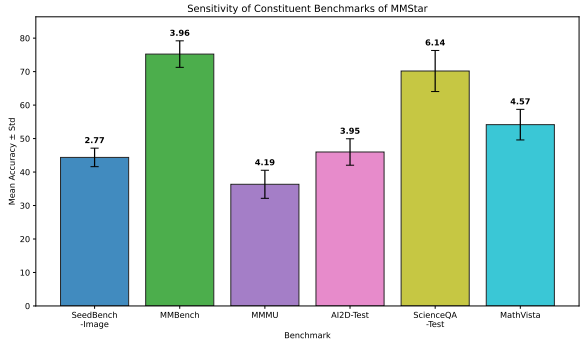

Figure 18: Sensitivity of Constituent Benchmarks of MMStar. The Figure shows Average Accuracy and Standard Deviation averaged across Models.

F.4 Which Core Capability in MMStar is Sensitive?

MMStar evaluates six core capabilities: Coarse Perception, Fine-grained Perception, Instance Reasoning, Logical Reasoning, Math, and Science & Technology. Among these, Math exhibits the highest sensitivity to prompt variations. The remaining five capabilities show comparable levels of sensitivity. (Figure 19)

F.5 Which Subject in MMMU-Pro is Sensitive?

The sensitivity analysis of subjects in MMMU-Pro (Figure 20) reveals that Electronics exhibits the highest variation in performance across different prompts, followed by Accounting, Public Health, Finance, and Energy and Power. These subjects are more susceptible to changes in prompt phrasing, indicating a higher reliance on specific wording for model accuracy. In contrast, Management, Biology, Economics, Architecture and Engineering,

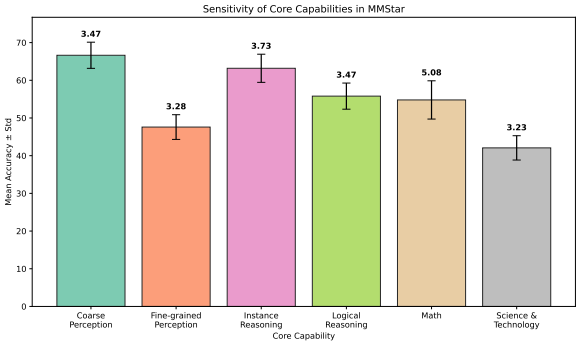

Figure 19: Sensitivity of Core Capabilities of MMStar. The Figure shows Average Accuracy and Standard Deviation averaged across Models.

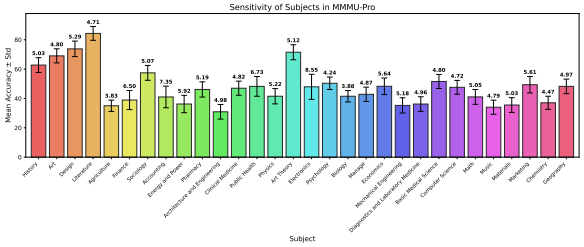

Figure 20: Sensitivity of Subjects in MMMU-Pro. The Figure shows Average Accuracy and Standard Deviation averaged across Models.

and Clinical Medicine show the least sensitivity, suggesting that prompt variations have a minimal effect on model performance in these domains.

F.6 Which Temporal Task in MVBench is Sensitive?

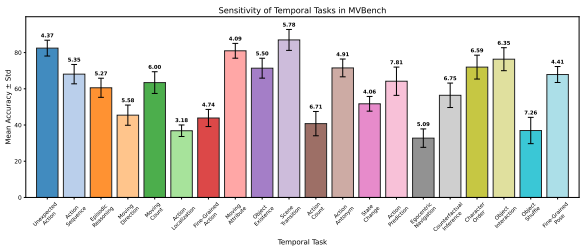

Figure 21: Sensitivity of Temporal Tasks in MVBench. The Figure shows Average Accuracy and Standard Deviation averaged across Models.

The sensitivity analysis of temporal tasks in MVBench (Figure 21) reveals that Action Prediction exhibits the highest variation in performance across different prompts, followed by Object Shuffle, Action Count, Counterfactual Inference, and Character Order. These tasks are particularly sensitive to prompt phrasing, suggesting that slight modifications in wording significantly impact model responses. In contrast, State Change, Moving Attribute, Unexpected Action, Fine-Grained Pose, and Fine-Grained Action demonstrate the least sensitivity, indicating more stable performance across different prompt formulations. These findings provide insights into which temporal reasoning tasks require more careful prompt engineering for consistent model evaluation.

G Model-Specific Anomalies

This study yielded some surprising findings. Notably, a straightforward variation in phrasing for Category 12 led to accuracy shifts of up to 15% in proprietary models (Figure 22). Additionally, Prompt 12.5 performed comparably to the Chain-of-Thought (CoT) prompt. Category 6 CoT prompts led to an approximate 10% accuracy improvement for proprietary models, while no noticeable gains were observed for open-source models. Another noteworthy finding is that the "Act as a Computer Vision Professor" prompt (9.1) resulted in a slight accuracy decrease, whereas the "Act as a Careless Student" prompt (9.2) caused a dramatic 40% drop in accuracy for Gemini 1.5 Pro on MMMU-Pro. This pattern was consistently observed across all datasets for both GPT-4o and Gemini 1.5 Pro.

Unexpected accuracy drops were observed in open-source models, as shown in Table 10. The table presents the absolute accuracy drop relative to the baseline, along with the corresponding model responses for each instance. Another notable observation was that when the prompt included the $ symbol (e.g. 12.3: Best Choice: $LETTER), GPT-4o more frequently refused to respond due to safety concerns, often generating disclaimers instead of valid answers. Consequently, the $ symbol was omitted from all prompts for GPT-4o.

12.1: Answer the preceding multiple-choice question in the following format: 'Answer: LETTER' (without quotes) where LETTER is one of the options.	61.3
12.2: Respond only with: "LETTER" of the best answer choice.	52.7
12.3: Best Choice: LETTER	59.0
12.4: Respond only with the correct option, formatted as follows: LETTER.	53.2
12.5: The last line of your response should be of the following format: Answer: 'LETTER' (without quotes) where LETTER is one of the options.	66.2

Figure 22: GPT-4o Performance (Absolute Accuracy) for Category-12 Prompts on MMMU-Pro. This shows how even the slightest difference in how the answer is expected can result in significant fluctuations in performance.

Dataset	Type	Model	∆Accuracy	Response
MMStar	Type6.2	MiniCPM-v2.6-8B	-12.6	"$LETTER"
	Type4.5	Molmo-7B-d-0924	-20.5	"A/A/A/A"
	Type7.6	Molmo-7B-d-0924	-21.8	"E" or "F"
	Type9.3	InternVL2.5-1B	-21.3	"$LETTER"
	Type10.1	InternVL2.5-1B	-16.9	"$LETTER"
MMMU-Pro	Type7.6	InternVL2.5-1B	-10.9	"E" or "F" or "G"
	Type9.3	InternVL2.5-1B	-17.2	"$LETTER"
	Type10.1	InternVL2.5-1B	-11.2	"$LETTER"
MVBench	Type6.2	MiniCPM-v2.6-8B	-14.2	"$LETTER"
	Type6.2	InternVL2.5-1B	-20.8	"$LETTER"
	Type11.3	InternVL2.5-1B	-15.7	"$NON-NEGOTIABLE", "$ERROR", "$NON_EXISTENT"
	Type13.1	InternVL2.5-1B	-23.6	"Answer: $LETTER"
	Type13.3	InternVL2.5-1B	-19.6	"Incorrect"
	Type13.4	InternVL2.5-1B	-23.1	"Answer: $LETTER. Deviations from this format will result in automatic deductions"

Table 10: Instances of Significant Accuracy Drops and Corresponding Model Responses

H Distribution of Standard Deviation within Prompt Categories

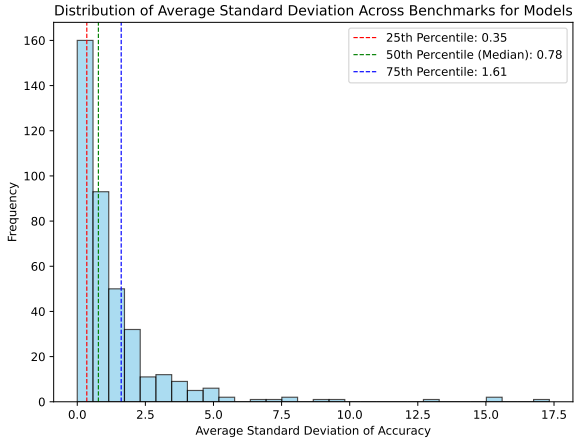

Figure 23: Distribution of standard deviation values computed across prompt categories and models. Each value represents the variability in accuracy within a single prompt category for a given model. The aggregated distribution is used to define a threshold for high sensitivity, with the median standard deviation of 0.78 serving as the cutoff between low- and high-sensitivity prompt categories.