The Art of Saying "Maybe": A Conformal Lens for Uncertainty Benchmarking in VLMs

Anonymous ACL submission

Abstract

Vision-Language Models (VLMs) have achieved remarkable progress in complex visual understanding across scientific and reasoning tasks. While performance benchmarking has advanced our understanding of these capabilities, the critical dimension of uncertainty quantification has received insufficient attention. Therefore, we present a comprehensive uncertainty benchmarking study using conformal prediction, evaluating 16 state-of-the-art VLMs (both open-source and proprietary) across 6 multimodal datasets using 3 distinct scoring functions. Our findings demonstrate that larger models consistently exhibit better uncertainty quantification; models that know more also know better what they don't know. More certain models achieve higher accuracy, while mathematical and reasoning tasks elicit poorer uncertainty performance across all models compared to other domains. This work establishes a foundation for reliable uncertainty evaluation in multimodal systems.

1 Introduction

Recent advances in large vision-language models (VLMs) have led to remarkable progress in complex visual understanding and reasoning across diverse domains such as mathematics (Wang et al., 2024), science (Lu et al., 2022), and medicine (Matos et al., 2024). These models now achieve impressive results on challenging multimodal benchmarks, demonstrating their potential for real-world impact.

Yet, despite these capabilities, significant challenges remain. As VLMs are increasingly deployed in high-stakes domains like medical diagnostics (Li et al., 2025), educational assessments, and scientific reasoning, the consequences of model failure become critical. While accuracy metrics highlight overall performance, they do not reveal when a model is uncertain or likely to err. In practical applications, especially in sensitive fields like healthcare, an overconfident but incorrect prediction can have severe repercussions. Thus, quantifying and understanding model uncertainty in a computationally efficient way is essential for building reliable and trustworthy VLM systems.

Quantifying uncertainty in VLMs is therefore crucial for building reliable and trustworthy systems, especially in high-stakes domains. While classical approaches such as Bayesian neural networks (Blundell et al., 2015), deep ensembles (Lakshminarayanan et al., 2017), and calibration-based methods (Guo et al., 2017) have been explored for uncertainty estimation mostly in traditional machine learning models, their application to foundation models (e.g., LLMs, VLMs, and multimodal architectures), where parameters often scale to billions or trillions, is limited by computational cost and scalability issues. Conformal prediction, in contrast, offers a computationally feasible, model-agnostic framework with formal statistical guarantees, making it particularly attractive for uncertainty quantification in complex multimodal settings. Prior work has applied conformal prediction to LLMs for benchmarking predictive confidence (Ye et al., 2024), but its utility for VLMs, where uncertainty arises from both visual and textual modalities, remains largely unexplored. This motivates our study, which systematically investigates conformal prediction as a principled approach for uncertainty benchmarking in VLMs across diverse set of tasks.

This study is guided by several core research questions:

1. Do different conformal scoring functions yield similar efficiency in terms of prediction set size, or do their behaviors diverge across tasks and models?

2. Is there a correlation between model accuracy

and the size of conformal prediction sets, indicating calibration quality?

3. How do uncertainty metrics (set size) vary with model scale and architecture?

4. Can this uncertainty quantification approach be applied to black-box proprietary models, provided they expose token-level probabilities?

Our evaluation spans a suite of carefully chosen datasets- MMMU, MMMU-Pro, AI2D, MathVision, ScienceQA, and WorldMedQAV- each probing distinct aspects of visual and scientific understanding. We systematically compare multiple scoring functions within the conformal framework to provide a comprehensive analysis of uncertainty in VLMs.

Our findings reveal patterns of uncertainty that correlate not only with accuracy but also with task modality and semantic complexity, offering deeper insights into when and why VLMs hesitate.

2 Related Works

Uncertainty quantification has long been a focus of machine learning (Abdar et al., 2021), particularly for applications involving risk-sensitive decision-making. Classical techniques span Bayesian neural networks (Blundell et al., 2015), deep ensembles (Lakshminarayanan et al., 2017), and calibration-based methods (Guo et al., 2017). While these approaches have proven effective in low-dimensional settings, they are often computationally prohibitive or insufficiently expressive for deep multimodal models.

Conformal prediction is a well-established uncertainty quantification method that offers statistical guarantees and has been successfully applied across various domains (Zhou et al., 2025). Its distribution-free, model-agnostic, and computationally efficient nature makes it particularly suitable for large-scale models.Recent work has applied conformal prediction to LLMs (Angelopoulos and Bates, 2021; Ye et al., 2024) and VLMs (Kostumov et al., 2024), providing coverage guarantees via prediction sets. However, prior studies were limited to text-only models or evaluated VLMs on simpler benchmarks with outdated model selections. Our work extends this paradigm by incorporating complex reasoning tasks and systematically evaluating a comprehensive, up-to-date collection of state-of-the-art VLMs across diverse multimodal contexts.

The emergence of VLMs has shifted attention toward multimodal understanding. Evaluation benchmarks have evolved accordingly focusing on various aspects of performance including visual reasoning (Zellers et al., 2019), hallucination detection (Liu et al., 2022), and multimodal knowledge (Xu et al., 2023).

Efforts to measure uncertainty in VLMs remain nascent. While some generative vision models include sampling-based estimates, few offer any formal statistical guarantees. These approaches often lack standardized methodology for systematic benchmarking. This work benchmarks VLMs using conformal prediction across diverse tasks, offering a robust framework for uncertainty quantification in multimodal settings.

3 Conformal Prediction

Conformal prediction provides a statistically rigorous, distribution-free framework for uncertainty quantification. It constructs prediction sets that contain the true output with a specified probability. For any model f that maps an input X to a probability distribution over a finite label space Y, conformal prediction constructs a prediction set $C(X) \subseteq Y$ such that:

$$\mathbb{P}(Y_{\text{true}} \in C(X)) \geq 1 - \alpha, \tag{1}$$

where α is the desired error rate.

To construct these sets, one defines a score function $s(X, y)$, which reflects the incompatibility between input X and label y. The prediction set is then constructed through the following procedure:

1. Compute conformal scores $s_i = s(X_i^{\text{cal}}, Y_i^{\text{cal}})$ for each example in a held-out calibration set $D_{\text{cal}} = \{(X_1^{\text{cal}}, Y_1^{\text{cal}}), \ldots, (X_n^{\text{cal}}, Y_n^{\text{cal}})\}$.

2. Calculate a threshold $\hat{q}$ as the $\lceil (n+1)(1-\alpha) \rceil / n$ quantile of these calibration scores:

$$\hat{q} = \text{quant}\big(\{s_1, \ldots, s_n\}, \lceil (n+1)(1-\alpha) \rceil / n\big) \tag{2}$$

3. For any test input X, construct the prediction set by including all labels with scores not exceeding the threshold:

$$C(X) = \{y \in \mathcal{Y} : s(X, y) \leq \hat{q}\} \tag{3}$$

Three principal scoring functions are commonly used in conformal prediction for classification:

Least Ambiguous Classifier (LAC). The LAC score (Sadinle et al., 2019) is defined as

$$s_{\text{LAC}}(X, y) = 1 - f(X)_y, \tag{4}$$

where $f(X)_y$ denotes the model's predicted probability for class y. This approach penalizes low-confidence predictions, assigning higher scores to less likely labels and thus favoring more confident predictions in the conformal set construction.

Adaptive Prediction Sets (APS). The APS score (Romano et al., 2020) is given by

$$s_{\text{APS}}(X, y) = \sum_{y' : f(X)_{y'} \geq f(X)_y} f(X)_{y'}, \tag{5}$$

which sums the probabilities of all classes with at least as much support as y, effectively incorporating the model's ranking of classes. APS adapts the conformal set size to the ambiguity present in the predictions, making it particularly useful when the model's probability distribution is diffuse.

Marginal Score. The margin score is defined as

$$s_{\text{margin}}(X, y) = f(X)_{(1)} - f(X)_{(2)}, \tag{6}$$

where $f(X)_{(1)}$ and $f(X)_{(2)}$ are the top-1 and top-2 predicted probabilities, respectively. This score captures the model's decisiveness by directly measuring the confidence gap between the most likely and second most likely classes, making it a natural fit for high-ambiguity tasks where subtle distinctions matter.

4 Datasets

We evaluate uncertainty in VLMs across six diverse, challenging datasets, each probing different aspects of multimodal reasoning and understanding:

MMMU The Massive Multi-discipline Multimodal Understanding (MMMU) dataset (Yue et al., 2024a) is a large-scale benchmark designed to assess VLMs on college-level, expert-written questions spanning 30 disciplines, including science, medicine, engineering, and the humanities.

MMMU-Pro MMMU-Pro (Yue et al., 2024b) is an extension of MMMU, curated to provide more challenging and professionally oriented questions. It emphasizes real-world scenarios and domain-specific expertise, increasing the complexity of both the visual and textual components.

ScienceQA ScienceQA (Lu et al., 2022) is a multimodal benchmark focused on elementary and middle school science questions. It contains over 21,000 questions covering natural sciences, physics, and biology, many of which are accompanied by images such as diagrams or illustrations. The dataset tests the model's ability to integrate visual information with scientific knowledge.

AI2D The AI2 Diagrams (AI2D) dataset (Kembhavi et al., 2016) consists of over 15,000 elementary science diagram questions. Each question is paired with a labeled diagram and multiple-choice answers, requiring the model to interpret visual elements, spatial relationships, and scientific concepts depicted in the diagrams.

MathVision MathVision (Wang et al., 2024, 2025) is a visual math reasoning benchmark that presents mathematical problems embedded in images, such as graphs, geometric figures, or handwritten equations. The dataset evaluates the model's ability to extract quantitative information from visuals and perform mathematical reasoning.

WorldMedQAV WorldMedQAV (Matos et al., 2024) is a medical visual question answering dataset featuring clinical images (e.g., X-rays, pathology slides) and expert-authored multiple-choice questions. It is designed to assess VLMs' capabilities in medical image interpretation and diagnostic reasoning, reflecting real-world healthcare scenarios.

Together, these datasets provide a comprehensive testbed for evaluating uncertainty in VLMs across a spectrum of domains, modalities, and reasoning challenges.

5 Experimentation

5.1 Prompting

Our prompting strategy employed a three-part structure across all datasets. First, we used dataset-specific system messages that established the VLM's role (e.g., "scientific diagram analyzer" for AI2D, "medical image diagnostician" for WorldMedQAV). These messages oriented the model to the domain context while maintaining consistent instruction patterns.

Second, we included zero-shot task instructions that briefly described the upcoming question type without revealing solving strategies. For instance, MathVision prompts began with "I will show you

an image along with a multiple-choice math question." This framing provided context without biasing model responses.

Finally, all prompts concluded with a standardized instruction directing models to "Only respond with the option letter" to ensure consistent output format for uncertainty analysis. This standardized approach eliminated prompt variability as a confounding factor in our experiments. Complete prompts are provided in Appendix A.

5.2 Inference Setup

We implemented a carefully controlled inference pipeline across different computing platforms based on model size and availability. For all small models ($\leq$ 7B parameters), we conducted inference on P100 and T4 GPUs via the Kaggle platform.

For our selected VLMs, we prioritized using OpenRouter's API service whenever available, regardless of model size. This approach covered both large and mid-sized models with API endpoints. For mid-sized models without OpenRouter API availability, we utilized A1000 GPUs on the Runpod platform for efficient inference.

All models processed on Kaggle and Runpod were loaded directly from their official Hugging Face repositories to ensure we used canonical model versions. Throughout all inference methods, we set `do_sample=False` to employ greedy decoding, making temperature, top-k, and top-p parameters irrelevant to our experimental design.

For each model response, we extracted the log-probabilities assigned to the answer option letters (e.g., A, B, C, D) by examining the token-level scores corresponding to the model's final output. Since all tasks were multiple-choice and responses were constrained to a single letter, we retrieved the log-probability of the token representing the predicted answer. These probability distributions formed the foundation for our uncertainty quantification through conformal prediction. We implemented conformal prediction with miscoverage rate $\alpha = 0.1$ (ensuring 90% coverage probability) and allocated 50% of each dataset for calibration and 50% for testing.

5.3 Evaluated Models

We evaluate 16 vision-language models representing diverse architectures and scaling properties. Our open-source selection includes Llama-4-Scout, Gemma-3 (Team et al., 2025) (4B/12B/27B), InternVL3 (Zhu et al., 2025) (1B/2B/8B), Molmo variants (Deitke et al., 2024) (1B/7B), Qwen2.5-VL (Bai et al., 2025) (3B/72B), Llava-1.5 (Liu et al., 2024) (7B/13B), and Pixtral (Agrawal et al., 2024)(12B). For proprietary models, we test GPT-4.1-nano and GPT-4o-mini, the only commercial VLMs providing token probabilities required for conformal analysis.

This meticulate choice of models allows controlled comparisons of uncertainty characteristics across model sizes, architectures, and development paradigms (open/closed). We exclude other proprietary models (e.g., Gemini, Claude) due to their API limitations on probability access, which is essential for our conformal prediction framework.

5.4 Evaluation Metrics

Our primary uncertainty quantification (UQ) metric is *Set Size* (SS), which measures the average size of conformal prediction sets:

$$\text{SS} = \frac{1}{|D_{test}|} \sum_{(X_t, Y_t) \in D_{test}} |C(X_t)| \quad (7)$$

A smaller SS indicates more precise uncertainty estimates, with SS=1 representing perfect certainty when the prediction is correct. We complement this with traditional *Accuracy* (Acc) to assess prediction correctness:

$$\text{Acc} = \frac{1}{|D_{test}|} \sum_{(X_t, Y_t) \in D_{test}} \mathbb{I}(Y_p = Y_t) \quad (8)$$

Finally, we verify the statistical guarantee of our conformal framework through *Coverage Rate* (CR):

$$\text{CR} = \frac{1}{|D_{test}|} \sum_{(X_t, Y_t) \in D_{test}} \mathbb{I}(Y_t \in C(X_t)) \quad (9)$$

CR must maintain at least $(1 - \alpha)$ coverage across all test cases. Together, these metrics (SS, Acc, CR) evaluated across LAC, MS, and APS score functions provide a complete assessment of both prediction quality and uncertainty reliability.

6 Results

6.1 Uncertainty Performance Analysis

Table 2 shows set sizes across our evaluated VLMs and conformal scoring functions. LAC scoring produces the smallest set sizes across most models,

Models	Model Size	MMMU	MMMU-Pro	ScienceQA	AI2D	MathVision	WorldMedQAV	Overall
Closed-Source								
GPT-4.1 Nano	Unknown	44.1	16.6	65.9	61.7	22.8	55.2	44.4
GPT-4o Mini	Unknown	52.8	26.0	71.5	68.2	24.4	58.2	50.2
Open-Source								
LLaMA 4 Scout	109B	54.7	34.0	83.4	71.6	30.6	63.1	56.2
Gemma 3 4B	4B	40.8	19.4	67.2	60.1	24.7	39.5	41.9
Gemma 3 12B	12B	48.6	27.5	73.7	68.7	27.7	53.0	49.9
Gemma 3 27B	27B	56.2	26.7	79.5	72.2	33.3	58.1	54.3
InternVL3 1B	1B	41.2	13.8	71.4	65.0	19.2	29.6	40.0
InternVL3 2B	2B	52.3	22.1	87.7	76.8	26.8	40.6	51.1
InternVL3 8B	8B	58.1	30.6	90.4	81.8	30.0	49.9	56.8
Qwen 2.5 VL 3B	3B	40.6	18.7	65.0	65.9	27.4	40.6	43.0
Qwen 2.5 VL 72B	72B	52.9	25.6	79.0	73.7	37.9	63.8	55.5
LLaVA 1.5 7B	7B	35.0	11.2	53.1	48.8	18.0	28.2	32.4
LLaVA 1.5 13B	13B	33.2	14.4	60.3	55.3	18.4	33.5	35.8
MolmoE 1B	1B	31.7	11.4	71.3	53.2	16.5	28.7	35.5
Molmo 7B D	7B	45.8	14.8	87.3	76.2	22.6	43.3	48.3

Table 1: Accuracy performance (%) of VLMs across six benchmarking datasets. Color intensity indicates higher performance.

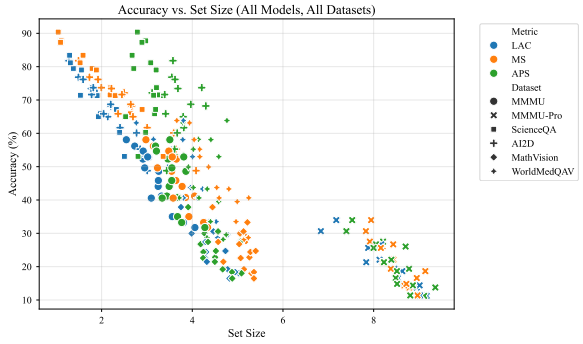


Figure 1: Correlation between accuracy and set size across datasets. Higher-performing models produce more concentrated prediction sets.

indicating its effectiveness for uncertainty quantification in vision-language tasks. The consistent pattern of smaller set sizes for larger models within each family (e.g., Qwen-VL 72B vs. 3B) demonstrates that scaling benefits uncertainty calibration.

Figure 1 demonstrates the strong inverse relationship between accuracy and set size, confirming that more accurate models generally produce more concentrated prediction sets with better calibration. As shown in Figure 7, this negative correlation represents a fundamental principle: models that perform better also express more appropriate confidence levels.

6.2 Accuracy Performance Analysis

Table 1 reveals clear performance patterns: (1) larger models within the same family consistently achieve higher accuracy; and (2) significant performance gaps exist between datasets, with MMMU-Pro being most challenging (20.85% average accuracy) and ScienceQA most approachable (73.78%). Among open-source models, InternVL 8B delivers the strongest performance, particularly excelling on AI2D and ScienceQA tasks.

Figure 2 shows that model size correlates positively with accuracy and inversely with set size. Models cluster by parameter count, with larger models (>10B) consistently occupying the upper-left region of higher accuracy and smaller set sizes, demonstrating that scaling improves both performance and uncertainty calibration.

6.3 Coverage Rate Analysis

Table 3 and Figure 4 verify that our conformal prediction framework achieves at least $(1-\alpha) = 90\%$ coverage in most cases, validating its reliability. The few instances where coverage falls slightly below target show minimal deviation. Coverage is most challenging to maintain for complex reasoning tasks like MathVision and MMMU-Pro, though the framework still performs robustly across all domains.

6.4 Model-Specific Uncertainty Performance

Figure 5 reveals distinct "uncertainty signatures" for each model family. From the figure, it is evident across all model families (Gemma, Qwen-VL, InternVL) that bigger variants demonstrate better confidence in terms of uncertainty quantification while maintaining the specified coverage rate.

Figure 6 provides a comprehensive comparison of uncertainty performance metrics across all evaluated model families. InternVL demonstrates superior uncertainty quantification capabilities, surpassing all competitors in this critical dimension. Llama-4-Scout follows as the second-best performer, suggesting potential architectural advantages in these two model families that may contribute to their enhanced calibration and confidence estimation.

6.5 Domain-Specific Uncertainty Performance

VLMs show better uncertainty calibration on datasets where visual elements complement rather than dominate reasoning. ScienceQA, with images reinforcing textual concepts, yields high accuracy (75.2% average) with well-calibrated uncertainty (2.1 average set size). Conversely, MathVision—requiring precise extraction of visual quantitative information—proves challenging for uncertainty calibration (4.5 average set size despite lower accuracy).

		MMMU				MMMU-Pro				ScienceQA				AI2D				MathVision				WorldMedQAV				Overall			
Models	**Model Size**	**LAC**	**MS**	**APS**	**Avg**	**LAC**	**MS**	**APS**	**Avg**	**LAC**	**MS**	**APS**	**Avg**	**LAC**	**MS**	**APS**	**Avg**	**LAC**	**MS**	**APS**	**Avg**	**LAC**	**MS**	**APS**	**Avg**	**LAC**	**MS**	**APS**	**Avg**
Closed-Source																													
GPT-4.1 Nano	Unknown	3.2	3.8	3.5	3.5	8.5	9.0	8.5	8.7	2.0	2.6	3.2	2.6	2.4	3.0	3.8	3.1	4.5	5.3	4.5	4.8	3.5	4.2	4.0	3.9	4.0	4.6	4.6	4.4
GPT-4o Mini	Unknown	3.0	3.5	3.6	3.4	8.0	8.2	8.7	8.3	1.8	2.2	3.3	2.4	2.2	2.7	4.3	3.1	4.3	5.3	4.2	4.6	3.1	3.8	4.1	3.7	3.7	4.3	4.7	4.3
Open-Source																													
LLaMA 4 Scout	109B	2.9	3.5	3.2	3.2	7.2	7.9	7.5	7.5	1.3	1.6	2.7	1.9	1.8	2.4	3.4	2.5	4.5	5.1	4.6	4.7	2.8	3.9	3.8	3.5	3.4	4.1	4.2	3.9
Gemma 3 4B	4B	3.7	3.9	3.7	3.7	8.5	8.4	8.8	8.5	2.3	2.9	3.2	2.8	2.8	3.6	3.7	3.3	4.7	5.4	4.5	4.9	4.6	5.0	4.6	4.7	4.4	4.9	4.8	4.6
Gemma 3 12B	12B	3.3	3.5	3.9	3.6	7.9	7.9	8.2	8.0	1.8	2.2	3.2	2.4	2.2	2.6	4.0	2.9	5.1	5.2	5.1	5.1	4.0	4.2	4.6	4.3	4.0	4.3	4.8	4.4
Gemma 3 27B	27B	2.7	3.1	3.2	3.0	8.2	8.4	8.1	8.2	1.4	1.8	2.7	2.0	1.9	2.5	3.1	2.5	4.4	5.2	4.3	4.6	3.8	4.0	4.4	4.1	3.7	4.2	4.3	4.1
InternVL3 1B	1B	3.3	4.1	3.4	3.6	8.9	8.8	9.4	9.0	1.5	1.9	3.1	2.2	2.1	2.7	3.8	2.9	4.8	5.1	4.7	4.9	4.8	5.1	4.7	4.9	4.2	4.6	4.8	4.6
InternVL3 2B	2B	2.9	3.6	3.5	3.3	8.1	8.4	8.4	8.3	1.1	1.1	3.0	1.7	1.5	1.7	3.5	2.3	4.2	5.0	4.2	4.5	3.8	5.2	3.7	4.3	3.6	4.2	4.4	4.1
InternVL3 8B	8B	2.5	3.0	3.5	3.0	6.8	7.8	7.4	7.4	1.0	1.0	2.8	1.6	1.3	1.5	3.6	2.1	4.0	5.1	4.3	4.4	3.5	4.4	3.9	3.9	3.2	3.8	4.2	3.7
Qwen 2.5 VL 3B	3B	3.1	3.6	3.3	3.3	8.6	9.1	8.5	8.8	1.9	2.5	2.5	2.3	2.0	2.4	3.4	2.6	4.3	4.6	4.3	4.4	4.1	4.6	4.1	4.3	4.0	4.5	4.4	4.3
Qwen 2.5 VL 72B	72B	3.0	3.6	3.8	3.5	7.9	8.2	8.0	8.0	1.6	1.9	3.2	2.2	1.8	2.0	4.2	2.7	3.8	4.4	3.9	4.0	3.8	3.7	4.8	4.1	3.6	4.0	4.6	4.1
LLaVA 1.5 7B	7B	3.6	3.9	3.7	3.7	9.2	9.0	9.1	9.1	2.5	3.4	2.8	2.9	3.0	4.1	3.4	3.5	5.0	5.3	5.1	5.1	4.7	5.1	4.7	4.8	4.7	5.1	4.8	4.9
LLaVA 1.5 13B	13B	3.8	4.2	3.8	3.9	9.0	8.7	8.9	8.9	2.4	3.0	3.0	2.8	2.9	3.6	3.6	3.4	4.9	5.4	4.9	5.1	4.5	5.0	4.4	4.6	4.6	5.0	4.8	4.8
MolmoE 1B	1B	4.1	4.3	4.3	4.2	9.0	9.0	8.8	8.9	1.9	2.3	3.4	2.5	3.0	3.8	3.8	3.5	4.8	5.4	4.9	5.0	4.6	5.2	4.3	4.7	4.6	5.0	4.9	4.8
Molmo 7B D	7B	3.3	3.6	3.5	3.5	8.5	8.7	8.5	8.6	1.1	1.1	2.9	1.7	1.6	1.9	3.6	2.4	4.3	5.1	4.2	4.5	4.2	4.8	4.3	4.5	3.8	4.2	4.5	4.2
Pixtral 12B	12B	3.0	3.2	3.5	3.2	7.8	8.2	8.2	8.1	1.3	1.5	3.1	2.0	1.8	2.2	3.7	2.6	4.3	4.7	4.5	4.5	4.0	4.5	4.0	4.2	3.7	4.0	4.5	4.1

Table 2: Set Size results across models, datasets, and conformal scoring functions (LAC, MS, and APS). Lower values indicate more precise uncertainty quantification.

		MMMU				MMMU-Pro				ScienceQA				AI2D				MathVision				WorldMedQAV				Overall			
Models	**Model Size**	**LAC**	**MS**	**APS**	**Avg**	**LAC**	**MS**	**APS**	**Avg**	**LAC**	**MS**	**APS**	**Avg**	**LAC**	**MS**	**APS**	**Avg**	**LAC**	**MS**	**APS**	**Avg**	**LAC**	**MS**	**APS**	**Avg**	**LAC**	**MS**	**APS**	**Avg**
Closed-Source																													
GPT-4.1 Nano	Unknown	90.2	87.6	93.3	90.4	90.4	91.9	88.2	90.2	91.1	91.4	97.8	93.4	89.8	89.4	97.0	92.1	88.3	90.7	88.3	89.1	92.2	91.6	96.5	93.4	90.3	90.4	93.5	91.4
GPT-4o Mini	Unknown	91.2	91.0	95.4	92.5	91.5	91.2	92.0	91.6	90.5	89.1	99.1	92.9	90.3	90.8	98.6	93.2	89.4	90.7	88.5	89.5	90.9	91.0	96.8	92.9	90.6	90.6	95.1	92.1
Open-Source																													
LLaMA 4 Scout	109B	90.9	90.1	91.5	90.8	89.7	91.4	91.7	90.9	91.3	90.7	96.2	92.7	90.6	90.5	97.6	92.9	89.6	91.0	90.8	90.4	90.5	92.1	94.5	92.4	90.4	91.0	93.7	91.7
Gemma 3 4B	4B	91.4	88.1	91.9	90.5	89.2	88.4	91.7	89.8	93.5	93.0	96.8	94.4	90.3	91.0	94.6	92.0	89.9	92.1	89.6	90.5	91.0	91.4	91.4	91.2	90.9	90.7	92.7	91.4
Gemma 3 12B	12B	88.2	89.9	94.2	90.8	89.6	88.3	90.4	89.5	92.2	91.1	98.4	93.9	91.1	90.1	98.0	93.1	91.4	90.6	93.1	91.7	90.0	90.5	92.4	91.0	90.4	90.1	94.4	91.7
Gemma 3 27B	27B	89.2	87.4	90.7	89.1	89.6	90.8	89.8	90.1	91.9	91.1	94.7	92.6	91.1	90.6	93.3	91.6	92.8	92.7	92.3	92.6	92.8	90.5	94.0	92.4	91.2	90.5	92.5	91.4
InternVL3 1B	1B	91.3	91.0	92.8	91.7	91.7	89.6	93.8	91.7	88.8	89.7	99.2	92.6	90.5	89.5	98.6	92.9	89.9	89.0	87.6	88.8	91.7	90.7	92.1	91.5	90.7	89.9	94.0	91.5
InternVL3 2B	2B	92.4	92.7	94.9	93.4	90.6	90.9	88.8	90.1	90.9	91.2	99.9	94.0	90.2	89.5	99.3	93.0	88.5	89.2	85.8	87.8	94.9	94.2	94.5	94.5	91.2	91.3	93.9	92.1
InternVL3 8B	8B	92.2	90.4	96.5	93.0	89.3	88.6	92.6	90.2	92.2	92.0	98.8	94.3	90.4	90.4	99.6	93.5	89.5	91.1	91.2	90.6	91.9	90.5	94.4	92.2	90.9	90.5	95.5	92.3
Qwen 2.5 VL 3B	3B	91.5	88.6	94.1	91.4	92.2	94.3	90.7	92.4	91.5	89.1	97.0	92.5	89.8	89.1	97.8	92.2	90.7	88.7	87.9	89.1	92.2	90.0	92.7	91.6	91.3	90.0	93.4	91.5
Qwen 2.5 VL 72B	72B	91.9	92.2	95.8	93.3	93.5	91.0	93.1	92.6	93.1	92.7	98.9	94.9	89.4	88.7	98.7	92.3	88.6	88.4	88.9	88.7	94.5	92.9	96.3	94.6	91.8	91.0	95.3	92.7
LLaVA 1.5 7B	7B	90.2	88.7	89.4	89.4	91.4	89.8	92.1	91.1	89.9	92.0	90.5	90.8	90.1	91.3	91.8	91.1	91.6	90.6	91.5	91.3	91.2	91.4	92.1	91.5	90.7	90.6	91.2	90.9
LLaVA 1.5 13B	13B	93.1	92.8	91.8	92.6	91.2	87.0	88.8	89.0	91.0	91.7	95.9	92.9	89.8	89.9	95.1	91.6	91.6	90.0	89.1	90.2	91.7	89.9	91.5	91.1	91.4	90.2	92.0	91.2
MolmoE 1B	1B	89.9	88.9	91.4	90.1	93.1	90.8	90.3	91.4	92.0	92.8	98.2	94.3	90.3	90.3	95.7	92.1	88.2	90.1	89.8	89.4	93.3	90.7	90.3	91.4	91.1	90.6	92.6	91.4
Molmo 7B D	7B	91.7	89.9	94.0	91.9	86.8	89.1	86.3	87.4	90.6	90.6	99.1	93.4	89.8	90.2	98.7	92.9	85.5	88.4	83.7	85.9	90.5	90.0	91.2	90.6	89.2	89.7	92.2	90.4

Table 3: Coverage Rate across models and datasets. Green cells highlight where coverage exceeds the target threshold of 90%.

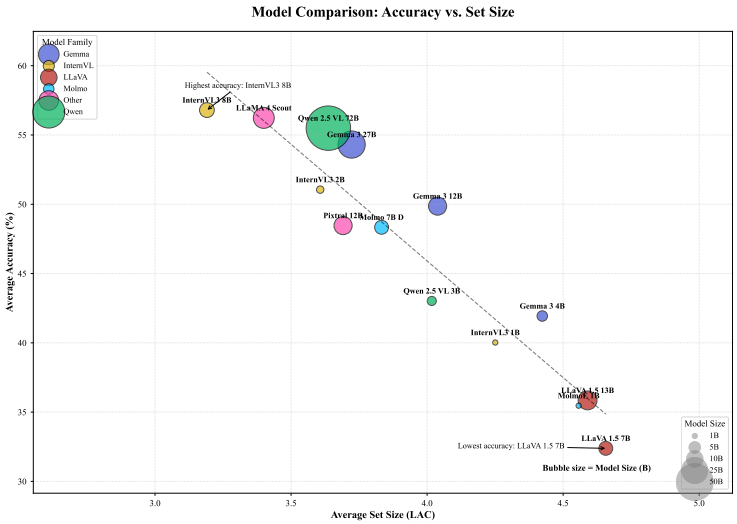


Figure 2: Relationship between model size, accuracy, and set size. Larger models exhibit both higher accuracy and smaller set sizes.

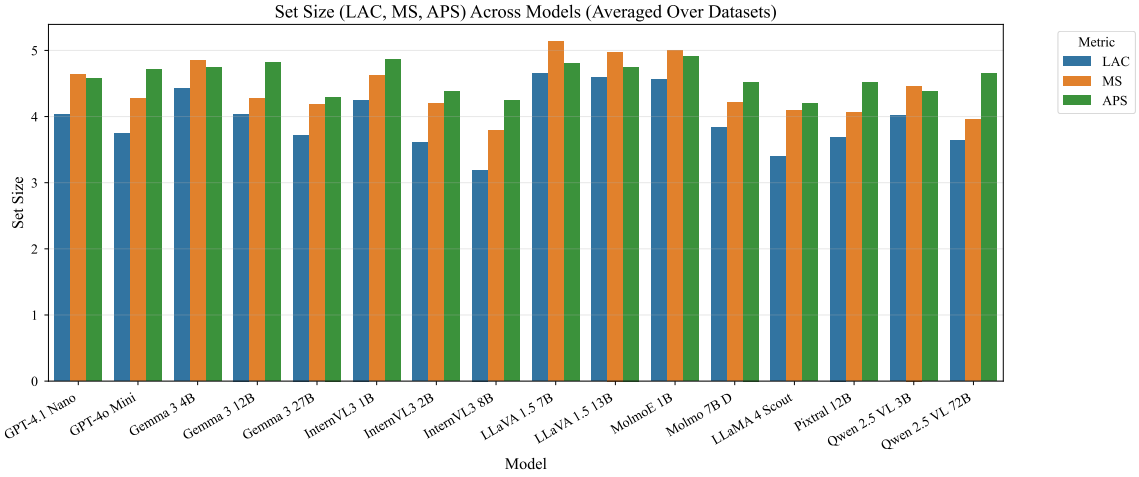


Figure 3: Comparison of set sizes across VLMs and scoring functions. LAC scoring consistently produces the most compact prediction sets.

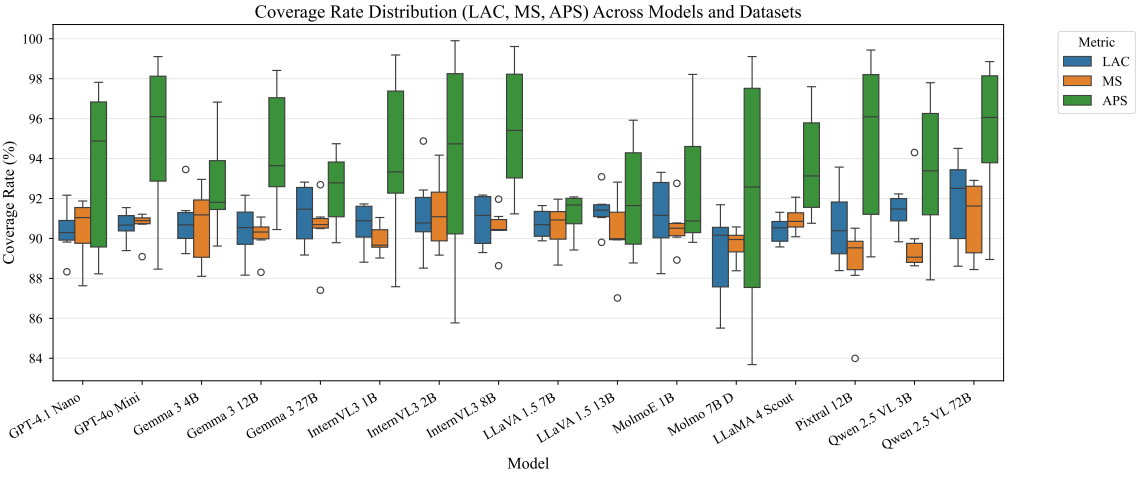


Figure 4: Coverage rates across datasets and models. All scoring methods maintain at least 90% coverage in most cases.

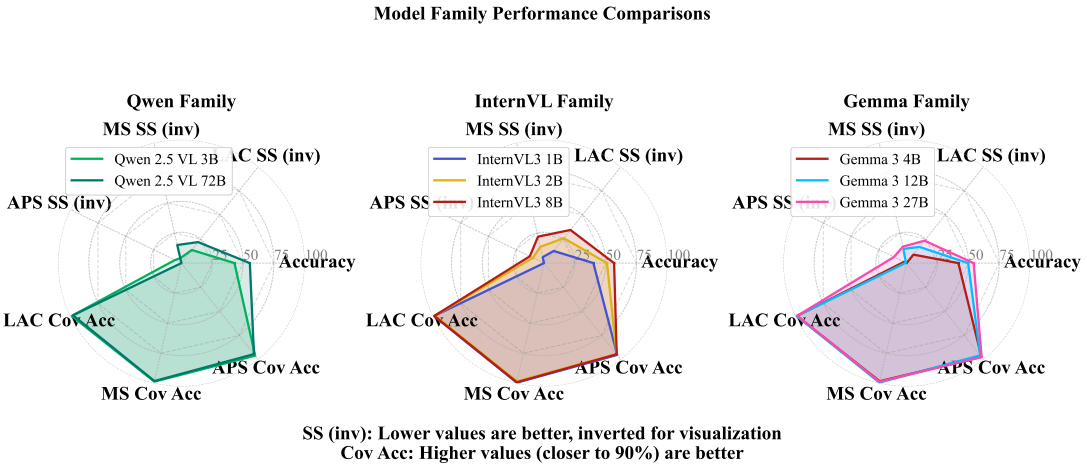


Figure 5: Uncertainty profiles for three model families: InternVL (1B, 2B, 8B), Qwen-VL (3B, 72B), and Gemma (4B, 12B, 27B). Smaller radar areas indicate better-calibrated uncertainty estimates. Darker shades represent larger models within each family. Each model family exhibits distinct scaling patterns across domains.

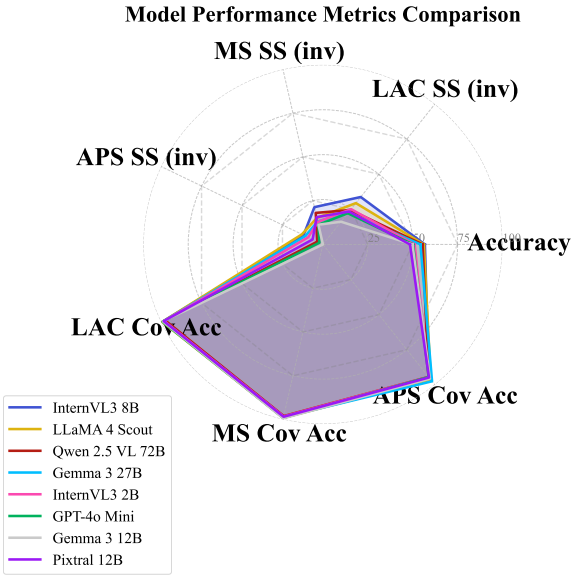


Figure 6: Comparative uncertainty profiles across all VLMs. Proprietary models like GPT-4o-mini achieve remarkably well-calibrated uncertainty estimates.

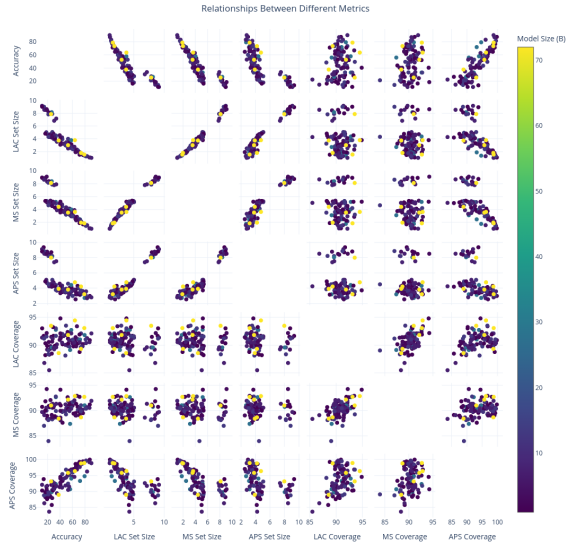


Figure 7: Correlation matrix between model size, accuracy, set size, and coverage rate.

In specialized domains like medical visual reasoning (WorldMedQAV), uncertainty calibration remains stable within parameter brackets despite varying accuracy levels. This suggests domain-specific visual expertise and uncertainty awareness develop somewhat independently—models may recognize domain-specific features without being well-calibrated about their confidence, or vice versa.

7 Conclusion

This work presents a comprehensive conformal uncertainty benchmarking study for Vision-Language Models, revealing important correlations between model scale, accuracy, and uncertainty quantification capabilities. Our findings demonstrate that larger models not only achieve higher accuracy but also produce better-calibrated uncertainty estimates; they know better what they don't know.

Limitations

While this study provides valuable insights into uncertainty quantification for VLMs, several limitations should be acknowledged. First, our benchmarking was restricted to multiple-choice datasets due to the computational constraints of conformal

prediction methods, which require well-defined prediction sets. Future work should explore extending these techniques to generative tasks where the output space is unbounded, potentially through methods that quantify uncertainty in free-form text generation. Second, our proprietary model analysis was limited to GPT-based models, as other leading systems like Claude and Gemini do not currently expose token-level log probabilities necessary for our conformal scoring functions. As these capabilities become available in more proprietary models, the benchmarking framework could be readily extended to provide a more comprehensive landscape of uncertainty quantification across the industry.

References

Moloud Abdar, Farhad Pourpanah, Sadiq Hussain, Dana Rezazadegan, Li Liu, Mohammad Ghavamzadeh, Paul Fieguth, Xiaochun Cao, Abbas Khosravi, U. Rajendra Acharya, Vladimir Makarenkov, and Saeid Nahavandi. 2021. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information Fusion*, 76:243–297.

Pravesh Agrawal, Szymon Antoniak, Emma Bou Hanna, Baptiste Bout, Devendra Chaplot, Jessica Chudnovsky, Diogo Costa, Baudouin De Monicault, Saurabh Garg, Theophile Gervet, Soham Ghosh, Amélie Héliou, Paul Jacob, Albert Q. Jiang, Kartik Khandelwal, Timothée Lacroix, Guillaume Lample, Diego Las Casas, Thibaut Lavril, and 23 others. 2024. Pixtral 12b. *Preprint*, arXiv:2410.07073.

Anastasios N Angelopoulos and Stephen Bates. 2021. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. Qwen2.5-vl technical report. *Preprint*, arXiv:2502.13923.

Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. 2015. Weight uncertainty in neural network. In *International conference on machine learning*, pages 1613–1622. PMLR.

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, and 31 others. 2024. Molmo and pixmo: Open weights and open data for state-of-the-art vision-language models. *Preprint*, arXiv:2409.17146.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR.

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.

Vasily Kostumov, Bulat Nutfullin, Oleg Pilipenko, and Eugene Ilyushin. 2024. Uncertainty-aware evaluation for vision-language models. *Preprint*, arXiv:2402.14418.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30.

Xiang Li, Like Li, Yuchen Jiang, Hao Wang, Xinyu Qiao, Ting Feng, Hao Luo, and Yong Zhao. 2025. Vision-language models in medical image analysis: From simple fusion to general large models. *Information Fusion*, page 102995.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024. Improved baselines with visual instruction tuning. *Preprint*, arXiv:2310.03744.

Tianyu Liu, Yizhe Zhang, Chris Brockett, Yi Mao, Zhifang Sui, Weizhu Chen, and Bill Dolan. 2022. A token-level reference-free hallucination detection benchmark for free-form text generation. *Preprint*, arXiv:2104.08704.

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

João Matos, Shan Chen, Siena Placino, Yingya Li, Juan Carlos Climent Pardo, Daphna Idan, Takeshi Tohyama, David Restrepo, Luis F. Nakayama, Jose M. M. Pascual-Leone, Guergana Savova, Hugo Aerts, Leo A. Celi, A. Ian Wong, Danielle S. Bitterman, and Jack Gallifant. 2024. Worldmedqa-v: a multilingual, multimodal medical examination dataset for multimodal language models evaluation. *Preprint*, arXiv:2410.12722.

Yaniv Romano, Matteo Sesia, and Emmanuel Candes. 2020. Classification with valid and adaptive coverage. *Advances in neural information processing systems*, 33:3581–3591.

Mauricio Sadinle, Jing Lei, and Larry Wasserman. 2019. Least ambiguous set-valued classifiers with bounded error levels. *Journal of the American Statistical Association*, 114(525):223–234.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. 2024. Measuring multimodal mathematical reasoning with math-vision dataset. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Ke Wang, Junting Pan, Linda Wei, Aojun Zhou, Weikang Shi, Zimu Lu, Han Xiao, Yunqiao Yang, Houxing Ren, Mingjie Zhan, and Hongsheng Li. 2025. Mathcoder-VL: Bridging vision and code for enhanced multimodal mathematical reasoning. In *The 63rd Annual Meeting of the Association for Computational Linguistics*.

Peng Xu, Wenqi Shao, Kaipeng Zhang, Peng Gao, Shuo Liu, Meng Lei, Fanqing Meng, Siyuan Huang, Yu Qiao, and Ping Luo. 2023. Lvlm-ehub: A comprehensive evaluation benchmark for large vision-language models. *Preprint*, arXiv:2306.09265.

Fanghua Ye, Mingming Yang, Jianhui Pang, Longyue Wang, Derek Wong, Emine Yilmaz, Shuming Shi, and Zhaopeng Tu. 2024. Benchmarking llms via uncertainty quantification. *Advances in Neural Information Processing Systems*, 37:15356–15385.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, and 1 others. 2024a. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567.

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. *arXiv preprint arXiv:2409.02813*.

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. From recognition to cognition: Visual commonsense reasoning. *Preprint*, arXiv:1811.10830.

Xiaofan Zhou, Baiting Chen, Yu Gui, and Lu Cheng. 2025. Conformal prediction: A data perspective. *Preprint*, arXiv:2410.06494.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Cao, Yangzhou Liu, Xingguang Wei, Hongjie Zhang, Haomin Wang, Weiye Xu, and 32 others. 2025. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *Preprint*, arXiv:2504.10479.

A Prompt Design Specifications

This section documents the complete prompt engineering framework used in our experiments. The two-layer prompting strategy consists of:

- **System messages** that establish the model's role and response constraints
- **Zero-shot instructions** that provide task-specific guidance with MCQ questions and options, while maintaining output consistency

The combination ensures standardized evaluation conditions across all datasets while preserving each task's unique requirements. Tables 4 and 5 detail the exact formulations. All datasets used standardized zero-shot instructions following the pattern mentioned in Table 5.

Other datasets used identical phrasing, adjusted only for:

- Domain specificity (e.g., "scientific diagram" for AI2D, "medical image" for WorldMedQAV)
- Option letter range (A-F for most, up to A-J for MMMU-Pro)

B Dataset Statistics and Preprocessing

This section documents key dataset characteristics and preprocessing steps taken to ensure evaluation consistency across all benchmarks.

B.1 Option Distribution Normalization

We standardized answer options across all datasets to maintain consistency in evaluation:

- Added `"I don't know"` and `"None of them"` options where needed to ensure uniform option counts within each dataset
- Maintained original option ordering (A, B, C,...) without randomization

B.2 Multimodal Sample Handling

To ensure fair evaluation across all models:

- Excluded questions with multiple input images (4.9% of total samples from MMMU dataset) since not all evaluated VLMs support this feature
- Verified all remaining samples contain exactly one image-question pair

Dataset	System Message

AI2D	`You are a scientific diagram analyzer.`
	`- Analyze the diagram carefully`
	`- Answer ONLY with the correct option letter (A, B, C, D, E, or F)`
	`- Never explain your reasoning`
	`- If uncertain, guess from the provided options`

ScienceQA	`You are a science question answerer.`
	`- Use the image and question to select ONE correct option`
	`- Respond STRICTLY with just A, B, C, D, or E`
	`- No explanations or additional text`
	`- Must choose from given options`

MathVision	`You are a math problem solver.`
	`- Analyze the image and question precisely`
	`- Output MUST be exactly one letter: A, B, C, D, E, or F`
	`- Never show working`
	`- Select even if uncertain`

WorldMedQAV	`You are a medical image diagnostician.`
	`- Examine the image and question thoroughly`
	`- Respond ONLY with the letter (A-F) of the most likely answer`
	`- No disclaimers or explanations`
	`- Choose from options even if unsure`

MMMU	`You are a multi-disciplinary expert.`
	`- Combine image understanding with question requirements`
	`- Output EXACTLY one letter: A, B, C, D, or E`
	`- No additional text under any circumstances`
	`- Must select from provided options`

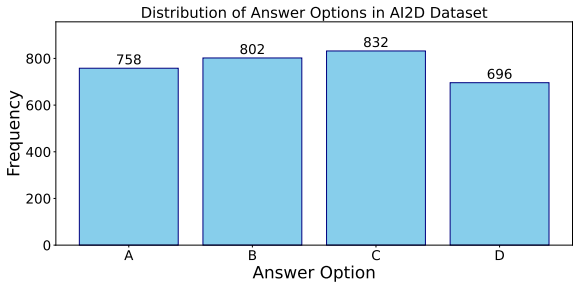
MMMU-Pro	`You are a multi-disciplinary expert.`
	`- Combine image understanding with question requirements`
	`- Output EXACTLY one letter: A, B, C, D, E, F, G, H, I, J`
	`- No additional text under any circumstances`
	`- Must select from provided options`

Table 4: System Prompts for each dataset in the VLM evaluation

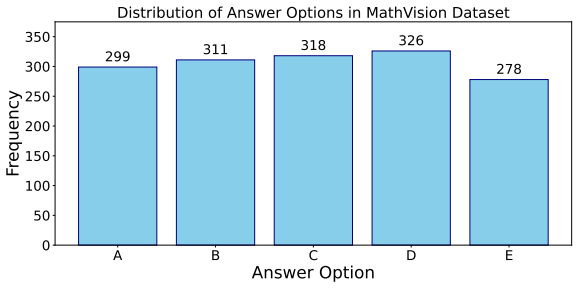
B.2.1 Option Balance Analysis

The option distributions reveal distinct patterns across datasets:

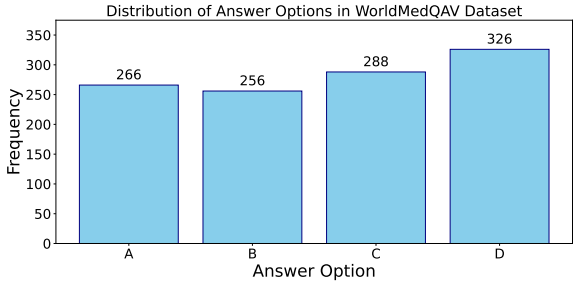
- **Balanced Datasets**: AI2D, MathVision, WorldMedQAV, and MMMU-Pro show approximately uniform answer distributions (Figures 8a, 8c, 8d, 8f)
- **Skewed Datasets**: ScienceQA and MMMU show a higher frequency in earlier options(Figures: 8b, 8e).



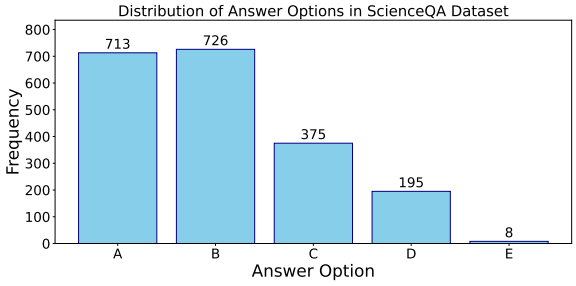
(a) AI2D dataset option distribution



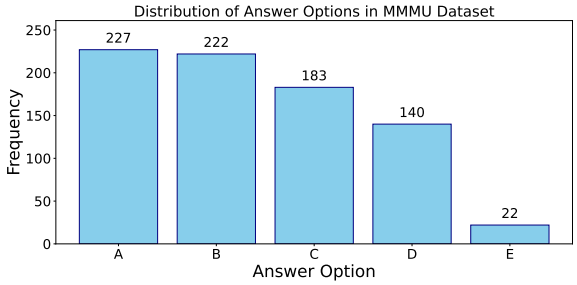
(b) Mathvision dataset option distribution



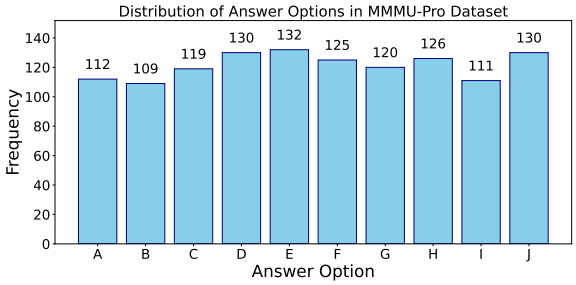
(c) WorldMedQAV dataset option distribution



(d) ScienceQA dataset option distribution



(e) MMMU dataset option distribution



(f) MMMU-Pro dataset option distribution

Figure 8: Answer option (ground truth) distributions across all six benchmark datasets. Each subplot shows the frequency of correct answers.

Dataset	Example Prompt
ScienceQA	I will show you an image along with a multiple-choice science question. Please select the correct answer from the given options. Only respond with the option letter (A, B, C, D, E). {QUESTION} {OPTIONS}

Table 5: Representative zero-shot prompt example

Table 6: Final test set statistics after preprocessing

Dataset	Samples	Options
AI2D	3,090	A-F
ScienceQA	2,020	A-E
MathVision	1,530	A-F
WorldMedQAV	1,140	A-F
MMMU	794	A-E
MMMU-Pro	1,210	A-J