Unveiling Uncertainty: A Deep Dive into Calibration and Performance of Multimodal Large Language Models *

Zijun Chen[1,2], Wenbo Hu[1], Guande He[3], Zhijie Deng[4], Zheng Zhang[2], Richang Hong[1]

[1]Hefei University of Technology, Hefei, China
[2]Data Space Research Institute, Hefei, China
[3]UT Austin, Austin, USA
[4]Shanghai Jiao Tong University, Shanghai, China

Abstract

Multimodal large language models (MLLMs) combine visual and textual data for tasks such as image captioning and visual question answering. Proper uncertainty calibration is crucial, yet challenging, for reliable use in areas like healthcare and autonomous driving. This paper investigates representative MLLMs, focusing on their calibration across various scenarios, including before and after visual fine-tuning, as well as before and after multimodal training of the base LLMs. We observed miscalibration in their performance, and at the same time, no significant differences in calibration across these scenarios. We also highlight how uncertainty differs between text and images and how their integration affects overall uncertainty. To better understand MLLMs' miscalibration and their ability to self-assess uncertainty, we construct the IDK (I don't know) dataset, which is key to evaluating how they handle unknowns. Our findings reveal that MLLMs tend to give answers rather than admit uncertainty, but this self-assessment improves with proper prompt adjustments. Finally, to calibrate MLLMs and enhance model reliability, we propose techniques such as temperature scaling and iterative prompt optimization. Our results provide insights into improving MLLMs for effective and responsible deployment in multimodal applications. Code and IDK dataset: https://github.com/hfutml/Calibration-MLLM.

1 INTRODUCTION

Multimodal large language models (MLLMs) represent a significant advancement in artificial intelligence by merging the capabilities of processing both visual and textual inputs (Khan and Fu, 2021; Lei et al., 2021; Radford et al., 2021; OpenAI, 2023; Liu et al., 2024). These models, trained on large datasets containing paired images and texts, excel in tasks such as image captioning, visual question answering, and cross-modal retrieval by correlating visual elements with corresponding textual descriptions.

However, like LLMs, MLLMs also suffer from issues such as hallucinations and unreliability (Bai et al., 2024). To identify and address these problems, quantifying and then calibrating the uncertainty of these models is an important approach. This process aligns the model's uncertainty with its actual predictive accuracy, similar to how humans can roughly evaluate their confidence in a particular matter. Accurate calibration is essential for MLLMs in safety-critical fields like autonomous driving (Yang et al., 2023), medical care (Buddenkotte et al., 2023), drug discovery (Li et al., 2024) and weather prediction (Price et al., 2024) to avoid overconfidence that can compromise reliability. Overconfidence in model predictions can lead to false assurances, increasing the risk of catastrophic errors, such as misdiagnoses or inaccurate weather forecasts, ultimately endangering lives and safety.

Recent studies have raised concerns about model fine-tuning and alignment of large language models inducing overconfidence (Kadavath et al., 2022; Tian et al., 2023), leading to suboptimal calibration (OpenAI, 2023; He et al., 2023; Zhao et al., 2023). These works have shown notable success in identifying and addressing overconfidence in unimodal models using calibration techniques, thereby improving the reliability of model predictions in single-modality tasks. Despite these advancements, the effect of miscalibration on MLLMs, remains underexplored. Proper calibration of MLLMs unlocks their full potential, enabling reliable use across various applications (Murphy and Winkler, 1977), especially in more rich image-text settings. Compared with unimodal models, MLLM needs multimodal information input, making the analysis of model confidence more complex and challenging.

*The first two authors contributed equally to this work. Corresponding author: Wenbo Hu (wenbohu@hfut.edu.cn.)

Proceedings of the 31st International Conference on Computational Linguistics, pages 3095–3109
January 19–24, 2025. ©2025 Association for Computational Linguistics

Whether there is a difference in the confidence for different modal information and how the model confidence changes during the integration of different modal information are all worthy of exploration.

Our research investigates the uncertainty calibration of state-of-the-art MLLMs, such as LLaVA (Liu et al., 2023a) and Qwen-VL (Bai et al., 2023). We first observed the calibration differences in several models across various settings, specifically focusing on before and after fine-tuning, and performance of linguistic tasks (compared to their corresponding base LLM). While the calibration differences of this two scenarios, were not substantial, we found that, overall, MLLMs consistently exhibited miscalibration. We then delved deeper into uncertainty quantification for MLLMs, exploring the differences in uncertainty exhibited by MLLMs when processing information from images versus text, and examining how the continuous integration of these two modalities affects uncertainty. Following this, we constructed the IDK dataset to further investigate the overconfidence issue in MLLMs and assess whether it can be mitigated with simple prompts. Finally, we introduced several calibration techniques to achieve more accurate calibration of MLLMs.

The contribution of this paper can be summarized as follows:

- **Calibration stability, yet persistent miscalibration:** We show that MLLMs maintain relatively consistent calibration before and after fine-tuning, alleviating concerns of degraded calibration. Our findings also suggest that MLLMs have minimal impact on the calibration of linguistic tasks after training from the base LLM, enabling visual data integration without significantly affecting original linguistic capabilities. However, the overall calibration of MLLMs is still suboptimal.

- **Differentiated uncertainty integration:** We observed that compared to images, MLLMs have lower uncertainty in the information of text, and the information of the two modalities can be integrated together to reduce the uncertainty of the model.

- **IDK Dataset and OOD assessment:** We constructed the IDK dataset by having the model repeatedly answer multiple times and created the OOD (out of distribution) dataset using recent news and GPT-3.5. MLLMs often answer questions even when unsure, but prompt-based encouragement can mitigate this, as seen in OOD data.

- **Advanced calibration techniques:** We propose and validate advanced calibration strategies, including temperature scaling and iterative prompt optimization, which significantly improve the reliability and effectiveness of MLLMs in diverse applications.

2 PRELIMINARIES

In this section, we briefly introduce the basics and the training process of MLLMs, followed by an explanation of how to quantify uncertainty under the multiple-choice setting.

2.1 Multimodal Large Language Models

With the rise of LLMs, many works have begun to focus on MLLMs. Most MLLMs are retrained by adding an image encoding part to LLMs (Liu et al., 2023a). This gives LMMs the ability to process vision. Liu et al. (2023a) proposed a new training method for multimodal models: training through instruction-following data. Sun et al. (2023) argued that MLLMs are constructed across multiple modalities, and the misalignment between the two modalities may lead to "hallucination". Previous work has explored the issue of calibration in LMs. Kadavath et al. (2022) demonstrated that advanced large-scale pretrained language models exhibit good calibration, while aligned language models (using fine-tuning or reinforcement learning to align human usage habits) are poorly calibrated due to being overly confident in logit-based multiple-choice questions. In the field of uncertainty, some researchers have also made significant contributions. They have decomposed the uncertainty of models into finer components. For instance, Kendall and Gal (2017) and Malinin and Gales (2018) decomposed uncertainty into model uncertainty and data uncertainty, while He et al. (2023) proposed the existence of format uncertainty and answer uncertainty in the logit-based multiple-choice setting. Their work provides new insights and approaches for more rational and accurate assessment of model calibration.

2.2 The Training Process of MLLMs

In this section, we will use LLaVA as an example to illustrate its training process.

Stage 1: Pre-training for Feature Alignment. During training, LLaVA maintain the visual encoder and LLM weights in a frozen state, focusing on maximizing the likelihood with only trainable parameters represented by the projection matrix. This approach allows for aligning the image features with the pre-trained LLM word embedding. Essentially, this stage can be viewed as training a compatible visual tokenizer specifically designed for the frozen LLM.

Stage 2: Fine-tuning End-to-End. Always keep the visual encoder weights frozen, and continue to update both the pre-trained weights of the projection layer and LLM in LLaVA. To enhance LLaVA's ability to follow instructions more effectively, the authors fine-tuned the model using language-image instruction-following dataset.

2.3 Uncertainty Quantification and Calibration of MLLMs

Uncertainty quantification and calibration in machine learning models are crucial for evaluating the alignment between a model's confidence in its predictions and the actual accuracy of those predictions (Guo et al., 2017; Kendall and Gal, 2017; Cui et al., 2020; Liu et al., 2023b). Misalignment, where a model is overly confident in incorrect predictions or uncertain about accurate ones, undermines its reliability and trustworthiness, especially in real-world applications that depend on the model's certainty for decision-making. Recent studies have seen the development and use of various effective uncertainty quantification methods, such as logit-based likelihood (He et al., 2023), semantic entropy (Kuhn et al., 2023), and self-expression (Xiong et al., 2023). These approaches offer valuable insights for assessing uncertainty in both LLMs and MLLMs.

To assess how well a model's confidence or uncertainty aligns with its accuracy, Expected Calibration Error (ECE) serves as a key metric. The goal of ECE is to quantify the difference between a model's predicted confidence and its actual performance. This is done by dividing predictions into bins, calculating the average accuracy and confidence for each bin, and summing the weighted differences across all bins:

$$\mathrm{ECE} = \sum_{m=1}^{M} \frac{N_m}{N} \left|\mathrm{acc}(B_m) - \mathrm{conf}(B_m)\right|, \quad (1)$$

where M is the number of bins, N is the total number of samples, N_m is the number of samples in the bin B_m, and $\mathrm{acc}(B_m)$ and $\mathrm{conf}(B_m)$ represent the actual accuracy and predicted confidence in the bin B_m. A smaller ECE indicates better model calibration. We use 10 equal-sized bins for ECE calculation.

We also use two ECE variants:

- **Maximum Calibration Error (MCE):** the maximum difference between accuracy and confidence across all bins. This provides a measure of the worst-case scenario that model calibration errors may reach.

$$\mathrm{MCE} = \max_{m=1}^{M} \left|\mathrm{acc}(B_m) - \mathrm{conf}(B_m)\right|$$

- **Normalized Expected Calibration Error (ENCE):** normalizes the error by dividing by the predicted confidence of each bin, giving more balanced calibration across varying confidence levels.

$$\mathrm{ENCE} = \sum_{m=1}^{M} \frac{N_m}{N} \frac{\left|\mathrm{acc}(B_m) - \mathrm{conf}(B_m)\right|}{\mathrm{conf}(B_m)}$$

3 ANALYSIS OF CALIBRATION ACROSS DIFFERENT SCENARIOS

In this section, we adopt logits-based likelihood as the uncertainty quantification method and primarily examine whether there are significant differences in the calibration of MLLMs at different scenarios. Additionally, we aim to observe if any miscalibration occurs.

3.1 Calibration Differences Between Before and After Fine-tuning MLLMs

In LLMs, some studies have shown that LLMs have good calibration during the Pre-Trained phase, but their calibration decreases after fine-tuning (He et al., 2023). To demonstrate how the fine-tuning process in MLLMs might influence calibration performance, we use some VQA (Visual Question Answering) datasets to test calibration of them at different stages.

Experiment Setup:

Model: The pre-trained MLLMs we selected include the LLaVA-v1.5-7B (Liu et al., 2023a) and LLaVA-v1.5-13B trained after Stage 1 (without fine-tuning), and Qwen-VL. The corresponding aligned MLLMs for these models are the LLaVA

trained after Stage 2 and Qwen-VL-Chat (Bai et al., 2023)

Dataset: We selected same VQA task datasets covering a wide range, including MMBench (Liu et al., 2023c), SEED-Bench (Li et al., 2023), BilbaoQA2 (Hugging Face, 2023), RealworldQA (Hugging Face, 2024b), MathVers (Zhang, 2024), Creature (biological habits generated by GPT-4), and ScreenShot (manually annotated screenshots from films). These datasets involve evaluations of various abilities such as perception and reasoning. We randomly sample some subsets from these datasets for testing.

Stability of Calibration After Fine-tuning: As can be seen in Table 1, we list the accuracy, confidence and ECE (MCE and ENCE) of the MLLMs before and after fine-tuning, namely the stage 1 and stage 2 (The remaining data results are shown in the Appendix A.) We observed that after fine-tuning, the accuracy and the average confidence of the model improved on various datasets, this is similar to LLMs (Kadavath et al., 2022; He et al., 2023). However, the changes in ECE (MCE and ENCE) were inconsistent. In some datasets, calibration decreased after fine-tuning, while in others, it improved, with the latter being more common. This suggests that calibration changes in MLLMs may not be directly tied to fine-tuning.

Model	Acc	Conf	ECE	MCE	ENCE
LLaVA-7B-Stage1	0.542	0.448	0.118	0.34	0.234
LLaVA-7B-Stage2	0.741	0.757	**0.075**	**0.246**	**0.11**
LLaVA-13B-Stage1	0.636	0.423	0.235	**0.383**	0.306
LLaVA-13B-Stage2	0.741	0.836	**0.101**	0.387	**0.181**
Qwen-VL	0.742	0.604	0.137	0.28	0.196
Qwen-VL-Chat	0.762	0.755	**0.079**	**0.193**	**0.151**

Table 1: ECE (MCE and ENCE), accuracy, and confidence of several MLLMs tested on MMBench

3.2 Calibration Changes in MLLMs for Linguistic Tasks

Many MLLMs like LLaVA, update the base LLM weights during training. Ideally, these models should retain their original linguistic task calibration after gaining the ability to process images. In this section, we compare several model pairs on linguistic tasks to assess calibration changes.

Experiment Setup:

Model: On the model side, we selected several pairs of models for comparison, namely Qwen-VL and Qwen-7B, LLaVA-v1.5-7B and Vicuna-v1.5-7B, LLaVA-v1.5-13B and Vicuna-v1.5-13B, LLaVA-llama-2-13b-chat-lightning-preview (Liu et al., 2023a) and LLaMA-2-13B-Chat (Touvron et al., 2023)

Dataset: The linguistic task datasets we selected include: ARC (Clark et al., 2018), CommonsenseQA (Talmor et al., 2019), MMLU (Hendrycks et al., 2021), OpenBookQA (Mihaylov et al., 2018), and RACE (Lai et al., 2017). We randomly sample some subsets from these datasets for testing. All these datasets are configured in a multiple-choice format.

Model	Acc	Conf	ECE	MCE	ENCE
Vicuna-7B	0.375	0.589	**0.213**	**0.349**	0.377
LLaVA-7B	0.421	0.640	0.224	0.705	**0.355**
Vicuna-13B	0.414	0.667	**0.253**	**0.455**	**0.404**
LLaVA-13B	0.431	0.739	0.308	0.490	0.443
LLaMA2-13B-Chat	0.400	0.667	0.267	0.571	0.435
LLaVA-LLaMA2	0.407	0.636	**0.229**	**0.349**	**0.399**
Qwen-7B	0.424	0.550	0.134	0.701	0.251
Qwen-VL-Chat	0.500	0.599	**0.099**	**0.224**	**0.199**

Table 2: ECE (MCE and ENCE), accuracy, and confidence of several models tested on MMLU, the LLaVA models used here are fully trained models after Stage 2

Minimal Impact of Linguistic Tasks: We report the accuracy and ECE (MCE and ENCE) of the MLLMs in terms of the linguistic QA tasks in Table 2. Remaining results are shown in the Appendix B. The results are compared between the uni-modal LLM and multimodal models. It can be seen that, for these datasets, calibration in some cases slightly increases, while in others it slightly decreases. This indicates that the calibration of MLLMs for linguistic tasks did not significantly deteriorate with the update of model parameter weights.

Analysis of Two Experiments: Although we did not observe significant calibration differences in MLLMs across different settings in the previous two experiments, we still identified instances of miscalibration in many tasks. For example, in Table 2, LLaVA-13B has an accuracy of 0.431, but its confidence reaches 0.739, indicating clear overconfidence. This is evident in several datasets where the model's accuracy and confidence are not aligned, and the ECE (MCE and ENCE) values are often relatively high.

4 UNCERTAINTY ANALYSIS AND MULTIMODAL INTEGRATION

In the context of MLLMs, which incorporates the visual domain, the sources of uncertainty become even more diverse, encompassing elements from

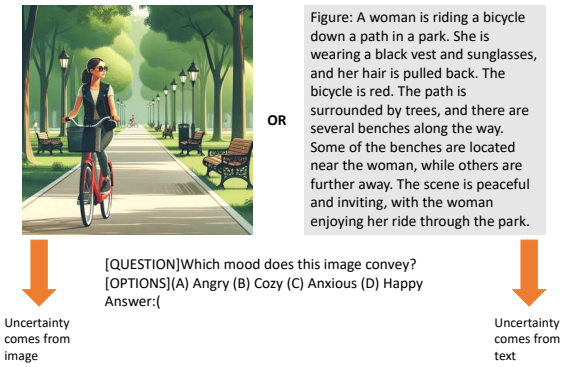

Figure 1: Replaces images with text descriptions, the descriptions are generated by GPT-4V and can accurately describe the images

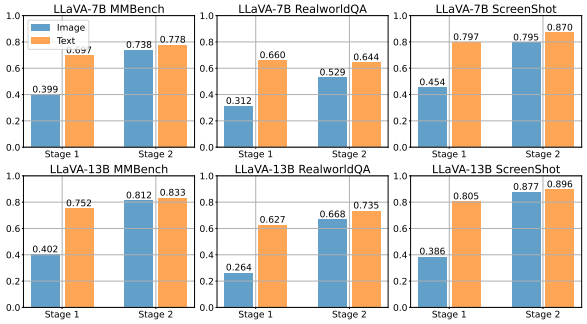

Figure 2: Use logits-based likelihood to quantify model uncertainty, where higher confidence means lower uncertainty. Stage 1 refers to the Pre-Trained MLLMs, while Stage 2 follows visual fine-tuning

both textual and visual modalities. This line of inquiry not only advances our comprehension of MLLMs' operational dynamics but also highlights the complexity of managing uncertainty in a multimodal setting.

4.1 Uncertainty Originating from Different Modalities

In the VQA task, the model's uncertainty about basic information reflects its uncertainty about the image content.

Replacing images with text descriptions. Images can be described in text, which introduces another form of uncertainty. For text uncertainty experiments, we generate textual descriptions to replace the images. As shown in Fig.1, we use both image and text as the problem's basic information. MLLMs must process these two modalities to answer the question.

Beyond comparing uncertainty across different modalities, we also explore how the model's uncertainty changes across training stages under this setup.

Results and Findings. We tested the performance of MLLMs including LLaVA-7B, and LLaVA-13B on their different stages. We use logits-based likelihood to quantify the uncertainty of these models. From Fig.2, we can see that in the Pre-Trained stage, MLLMs have a much higher confidence in the text than in the image, while after visual fine-tuning, the model has a slightly higher confidence in the text than in the image. This indicates two issues:

(i) Compared to text, MLLMs exhibit high uncertainty in information derived from images.

(ii) After visual fine-tuning, MLLMs reduce the uncertainty of the image, which means they will trust the information of the image more.

4.2 Integrating Multimodal Information to Reduce Uncertainty

Humans can boost their confidence in answers by gathering additional information when solving problems. Similarly, we aim to determine if MLLMs can integrate information from different modalities to reduce uncertainty, which is crucial for evaluating a model's performance and reliability. In this section, we extend our uncertainty quantification beyond logits-based likelihood to also include *semantic entropy* (Kuhn et al., 2023) for open-ended responses. Semantic entropy quantifies uncertainty by clustering a model's responses based on semantic similarity. If the model's answers vary significantly in meaning, it indicates high uncertainty. This provides an intuitive measure: the more varied the answers, the less certain the model is. This method bridges language understanding with information theory.

We have created a VQA dataset where images and text both describe a type of organism. The model's task is to infer the organism from either modality and answer questions about its habits. The dataset is divided into two question types: multiple-choice and open-ended, with model uncertainty quantified using logits-based likelihood and semantic entropy, respectively.

We will add varying levels of Gaussian noise to the images to obscure some of the information they convey. Meanwhile, the textual description will be split into several independent sentences, each describing different characteristics of the organism. We will then progressively add these textual descriptions to the images with different levels of noise. More details in Appendix C. During this pro-

cess, we will observe how the model's uncertainty changes with the increase in textual information at specific noise levels and compare the uncertainty change curves across different noise levels.

Logits-Based vs. Text-Based Uncertainty Quantification. Logits-based and text-based approaches are two methods for quantifying uncertainty in language models. Logits-based methods are more precise and fine-grained, while text-based methods, which handle discrete outputs, often require multiple inferences. However, text-based methods offer more practical insights for black-box models.

Results and Findings. As shown in Fig.3, We observed that for images with a fixed level of noise, we continuously increase text information, whether it is multiple-choice questions or open answers, resulting in an increase in MLLMs confidence and a decrease in semantic entropy. This indicates that MLLMs can continuously integrate the information of text modalities while obtaining fixed image information, thereby reducing their uncertainty. In addition, images with higher levels of noise often have higher uncertainty in MLLMs, as can be seen from the points in the image that do not include textual information. With the integration of text information, the model can reduce the uncertainty of images with higher noise more quickly. However, the uncertainty reflected by the model may still be higher than those of clear images. This is because clearer images can also make the model more confident compared to noisy images. This experiment demonstrates that the model can complement the uncertainty of two modalities, and they will simultaneously affect the uncertainty of MLLMs in the final answer.

5 CAN MLLMS KNOW WHAT THEY DON'T KNOW?

LLMs are criticized for their tendency to generate hallucinations. To investigate whether LLMs can recognize when they lack sufficient knowledge on a particular question and express this in natural language, Cheng et al. (2024) constructed a "I don't know" dataset and tested several models to observe the phenomenon. We referred to their method to construct a VQA dataset for MLLMs

The IDK dataset is divided into two parts: the model-specific dataset segments the existing dataset into what the model "knows" and "doesn't know" by having the model repeatedly answer questions. The OOD dataset that is outside the model's training scope, containing questions for which it couldn't possibly know the answers.

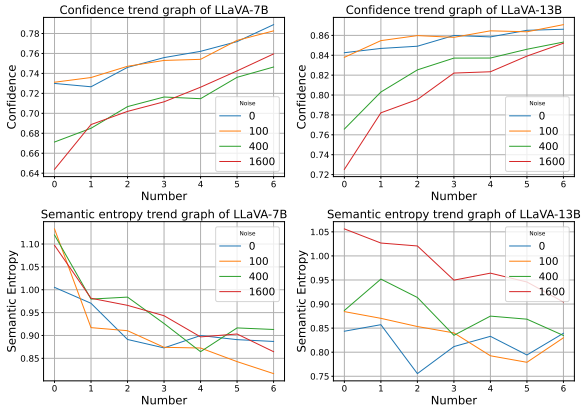

Figure 3: The change in uncertainty of images with different levels of noise as the text description increases. Noise=0 means no noise is added. $NoisyImage$=$Image$+$N(0, Noise)$

Construction of dataset:

Model-specific dataset: For a given question, a model answers 10 times. Based on an accuracy threshold (we use 1, meaning all correct), we classify whether the model "knows" or "doesn't know." This method segments the existing dataset into IDK and IK categories. For example, in Table 3, using MMBench to query LLaVA-7B, we segmented the data into IDK and IK, with $2,292$ and $2,085$ entries, respectively.

OOD dataset: We constructed the July24-NewsVQA dataset (total of 20968). This dataset consists of news and pictures from July 2024 (after the models were released) that we scraped, and it was used to create multiple-choice question via GPT-3.5. We assume the model cannot know the OOD answers, so there is no IK portion for this dataset. Some examples of IDK datasets and more construction details are shown in Appendix D. We will opensource this set of the data and the construction details.

We aim to explore whether MLLMs exhibit hallucinations and how well they recognize and express them. To test this, we used the IDK dataset and categorized the results into four domains:
(1) MLLMs know they don't know (IK-IDK),
(2) MLLMs don't know they know (IDK-IDK),
(3) MLLMs know they know (IK-IK),
(4) MLLMs don't know they know (IDK-IK).

Results and Findings. From Table 3, we observed that without prompting, MLLMs always

Datasets	LLaVA-7B					LLaVA-13B				
	IK-IDK	IDK-IDK	IK-IK	IDK-IK	TRUTHFUL	IK-IDK	IDK-IDK	IK-IK	IDK-IK	TRUTHFUL
MMBench	0	2292	2085	0	47.64%	0	1846	2529	0	57.81%
MMBench (Prompting)	627	1665	1935	150	58.53%	151	1695	2507	22	60.75%
SEED-Bench	0	9923	4310	0	30.28%	0	8671	5561	0	39.07%
SEED-Bench (Prompting)	2126	7797	3861	449	42.06%	273	8398	5510	51	40.63%
MobileVQA	0	827	44	0	5.05%	0	829	42	0	4.82%
MobileVQA (Prompting)	526	301	34	10	64.29%	622	207	17	25	73.36%
PathVQA	0	2610	198	0	7.05%	0	2474	334	0	11.89%
PathVQA (Prompting)	1532	1078	103	95	58.23%	1267	1207	235	99	53.49%
July24-NewsVQA	0	20968	/	/	0%	0	20968	/	/	0%
July24-NewsVQA (Prompting)	11990	8978	/	/	42.82%	5098	15870	/	/	24.31%

Table 3: In this experiment, MMBench, SEED-Bench, July24-NewsVQA are multiple-choice questions, MobileVQA (Hugging Face, 2024a), and PathVQA (He et al., 2020) are open-ended answer questions. IK-IDK, IDK-IDK, IK-IK, IDK-IK correspond to the number of four domains. *TRUTHFUL* = (*IK-IDK* + *IK-IK*) / (*IK-IDK* + *IDK-IDK* + *IK-IK* + *IDK-IK*)

Model	IK-IDK	IDK-IDK	TRUTHFUL
GPT-4o	61	939	6.10%
GPT-4o (Prompting)	697	303	69.70%
Claude-3-haiku	133	867	13.30%
Claude-3-haiku (Prompting)	677	323	67.70%

Table 4: Test two closed source models using 1000 sampled data from the July24-NewsVQA dataset.

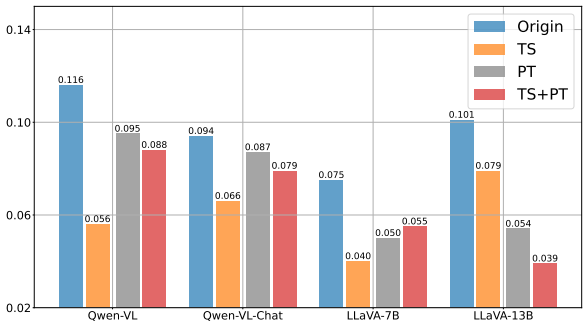

Figure 4: Changes in ECE after calibration for different models tested on MMBench.

provide answers regardless of whether they know the answer or not, indicating an serious overconfidence phenomenon in MLLMs. When using prompts such as "If you don't know the answer, please say..." to encourage MLLMs to handle questions more cautiously, although the models still tend to answer rather than refuse, their accuracy in self-assessment (TRUTHFUL) improves. In addition, compared to multiple-choice questions, open-ended questions have a higher rejection rate, which is intuitive because open-ended answers are more difficult and the accuracy is even lower. This suggests that a well-designed prompt can help alleviate MLLMs' hallucinations to some extent. For OOD, similar results are observed, even for questions they are unlikely to know, MLLMs still tend to provide answers. However, after being encouraged by prompts, this effect can also generalize to OOD.

To analyze the truthfulness of closed-source multimodal models, we sampled 1,000 instances from the July24-NewsVQA dataset and tested them on two closed-source models: GPT-4o-2024-05-13 and Claude-3-haiku-20240307. Since these models were released before the July24-NewsVQA dataset, they should theoretically be unaware of it. As shown in Table 4, when using prompting, larger models has better results(69.7%, 67.7%) compared with the smaller models, indicating that the prompting might be more effective to larger models. When prompting is not used, larger models still has better results(6.1%, 13.3%) compared with the smaller models(0%), indicating that the larger models has a better robust to the IDK items.

6 CALIBRATION TECHNIQUES FOR MLLMS

To alleviate miscalibration and enhance the reliability of MLLMs, we propose two techniques specifically for these models.

6.1 Temperature Scaling with MLLMs' Predictive Distribution

Temperature Scaling (TS) is a widely used and effective technique for improving the confidence calibration of neural network classification models (Guo et al., 2017). It works by adjusting the output probabilities of the softmax function, dividing them by a scalar known as the temperature parameter (T). This parameter plays a crucial role in controlling the smoothness of the resulting probability distribution: higher values of T produce a smoother, more evenly spread distribution, while lower values result in a sharper, more confident

distribution. By adjusting T, TS can help align the model's confidence with its actual accuracy, making it a valuable method for addressing overconfidence and miscalibration in neural networks.

TS optimizes the temperature T through the following objective:

$$\min_{T} -\sum_{i=1}^{M}\sum_{j=1}^{|y|} \mathbf{1}_{y_i=j} \log\left[\text{softmax}\left(\boldsymbol{l}_i/T\right)\right]_j . \quad (2)$$

Here, M is the number of samples, $|y|$ is the number of classes, y_i is the true label of the i-th sample, $\mathbf{1}_{y_i=j}$ is an indicator function, and $\boldsymbol{l}_i$ are the logits for the i-th sample. The temperature parameter T scales the logits $\boldsymbol{l}_i/T$, controlling the probability distribution's smoothness. When $T = 1$, the output is unchanged; when $T > 1$, the distribution is smoother (less confident); and when $T < 1$, it is sharper (more confident).

Despite its simplicity, TS is highly effective in calibrating MLLMs, especially in models with poor initial calibration, as shown in Fig.4.

Algorithm 1 APE for calibration

Input: Seed prompts S, evaluation function f, top prompts k, similar prompts m, iterations n
Output: Best prompt

Initialize $G \leftarrow S$
for $i = 1$ **to** n **do**
  Generate similar prompts for G using GPT: $G_{\text{new}} \leftarrow \bigcup_{p \in G} \text{generate}(p, m)$
  Evaluate G_{new}: $E \leftarrow \{(p, f(p)) \mid p \in G_{\text{new}}\}$
  Sort E by accuracy bands and ECE
  Update G with top k prompts from E
  Update best prompt if needed
  Record current best prompt and top k prompts
end for
return best prompt

6.2 Prompt Tuning

Some researchers have shown that model calibration is influenced by the prompts used (Jiang et al., 2023). In the previous section, we also observed that prompt can encourage models to better express their uncertainty. Simply adjusting the prompt can help calibrate the model without additional processing. For instance, adding phrases like "This answer might be..." for overly confident models or "This answer must be..." for models lacking confidence can achieve self-calibration.

To achieve self-calibration, we have applied some existing prompt optimization frameworks. The process of prompt tuning has been widely studied, and there are many applicable frameworks that can be adopted, such as APE (Zhou et al., 2022), APO (Pryzant et al., 2023) and LongPO (Hsieh et al., 2023). We specifically tuned the prompt suffix, optimizing phrases like "Answer:" to more calibration-friendly versions such as "This answer might be:". The process is outlined in Algorithm 1, which iteratively refines suffixes to improve model calibration (Detail in Appendix E). Prompt tuning can be combined with techniques like TS. Our tests show that in some cases, this combination leads to better calibration results.

7 CONCLUSIONS

In this work, we investigated the uncertainty and calibration of MLLMs, focusing on several key areas: the calibration differences between pre-trained and fine-tuned models, the comparison between MLLMs and base LLMs in linguistic tasks, the models' uncertainty when handling text versus images, the integration of uncertainty from both modalities, performance in IDK settings, and calibration techniques.

Our findings reveal that fine-tuned MLLMs do not show significant deterioration in calibration compared to their pre-trained counterparts. Similarly, the multimodal training process, which adjusts the base language model to handle images, does not substantially affect the calibration of linguistic tasks. However, MLLMs still exhibit miscalibration. Specifically, we found that MLLMs display greater uncertainty with image information than with text, and the integration of both modalities affects the overall uncertainty. When tested on the IDK dataset, MLLMs showed a significant overconfidence issue, but strategic prompts improved their ability to self-assess uncertainty. Lastly, we explored calibration techniques like temperature scaling and prompt tuning, demonstrating that these methods effectively enhance MLLM calibration.

The future work of this study includes exploring the integration of the robustness and adversial security of MLLMs (Zhao et al., 2024; Wang et al., 2023a,b). Additionally, it will explore the uncertainty and calibration of MLLMs in modalities be-

yond images.

We hope our work provides valuable insights for developing more reliable and robust MLLMs.

Limitations and Ethical Considerations

In this study, we introduced several widely used uncertainty quantification methods. However, given the vast number of available techniques and our time constraints, we were unable to explore the impact of other potentially more innovative methods. We also employed standard metrics like ECE for model calibration, but recognize that additional metrics could provide a deeper and more nuanced understanding of the phenomena we observed.

Ethical AI development relies heavily on precise uncertainty quantification and reliability, ensuring that model predictions align with real-world confidence. In high-stakes fields such as healthcare, maintaining human oversight of AI is essential for managing uncertainties, ensuring accountability, and minimizing risks. This oversight is critical not only for mitigating potential errors but also for fostering trust in AI systems. Ethical AI must also prioritize transparency, fairness, and the protection of user rights, ensuring that models behave responsibly under uncertainty and that their limitations are well understood. Balancing AI capabilities with these ethical considerations is key to building systems that are both safe and trusted by society.

Acknowledgements

This work is jointly supported by National Natural Science Foundation of China (No. 62306098), the Open Projects Program of State Key Laboratory of Multimodal Artificial Intelligence Systems, the Fundamental Research Funds for the Central Universities (No. JZ2024HGTB0256), the SMP-IDATA Open Youth Fund (No. SMP2023-iData-009) and the Open Project of Anhui Provincial Key Laboratory of Multimodal Cognitive Computation, Anhui University (No. MMC202412).
Part of this work was done when the first author was interning at Data Space Research Institute, Hefei, China.

References

Jinze Bai, Song Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Jung Fu Lin, Chen Zhou, and Jingren Zhou. 2023. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond.

Zechen Bai, Pichao Wang, Tianjun Xiao, Tong He, Zongbo Han, Zheng Zhang, and Mike Zheng Shou. 2024. Hallucination of multimodal large language models: A survey. *arXiv preprint arXiv:2404.18930*.

Thomas Buddenkotte, Lorena Escudero Sanchez, Mireia Crispin-Ortuzar, Ramona Woitek, Cathal McCague, James D Brenton, Ozan Öktem, Evis Sala, and Leonardo Rundo. 2023. Calibrating ensembles for scalable uncertainty quantification in deep learning-based medical image segmentation. *Computers in Biology and Medicine*, 163:107096.

Qinyuan Cheng, Tianxiang Sun, Xiangyang Liu, Wenwei Zhang, Zhangyue Yin, Shimin Li, Linyang Li, Kai Chen, and Xipeng Qiu. 2024. Can ai assistants know what they don't know? *arXiv preprint arXiv:2401.13275*.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv:1803.05457v1*.

Peng Cui, Wenbo Hu, and Jun Zhu. 2020. Calibrated reliable regression using maximum mean discrepancy. *Advances in Neural Information Processing Systems*, 33:17164–17175.

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.

Guande He, Peng Cui, Jianfei Chen, Wenbo Hu, and Jun Zhu. 2023. Investigating uncertainty calibration of aligned language models under the multiple-choice setting. *arXiv preprint arXiv:2310.11732*.

Xuehai He, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. 2020. Pathvqa: 30000+ questions for medical visual question answering. *arXiv preprint arXiv:2003.10286*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Cho-Jui Hsieh, Si Si, Felix X. Yu, and Inderjit S. Dhillon. 2023. Automatic engineering of long prompts. abs/2311.10117.

Hugging Face. 2023. BilbaoQA2. [BilbaoQA2].

Hugging Face. 2024a. MobileVQA. [MobileVQA].

Hugging Face. 2024b. realworldqa. [realworldqa].

Mingjian Jiang, Yangjun Ruan, Sicong Huang, Saifei Liao, Silviu Pitis, Roger Baker Grosse, and Jimmy Ba. 2023. Calibrating language models via augmented prompt ensembles.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.

Alex Kendall and Yarin Gal. 2017. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems*, 30.

Zaid Khan and Yun Fu. 2021. Exploiting bert for multimodal target sentiment classification through input space translation. In *Proceedings of the 29th ACM international conference on multimedia*, pages 3034–3042.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. *arXiv preprint arXiv:2302.09664*.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. RACE: Large-scale ReAding comprehension dataset from examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 785–794, Copenhagen, Denmark. Association for Computational Linguistics.

Chenyi Lei, Shixian Luo, Yong Liu, Wanggui He, Jiamang Wang, Guoxin Wang, Haihong Tang, Chunyan Miao, and Houqiang Li. 2021. Understanding chinese video and language via contrastive multimodal pre-training. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 2567–2576.

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

Peiyao Li, Lan Hua, Zhechao Ma, Wenbo Hu, Ye Liu, and Jun Zhu. 2024. Conformalized graph learning for molecular admet property prediction and reliable uncertainty quantification. *Journal of Chemical Information and Modeling*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023a. Visual instruction tuning. abs/2304.08485.

Jing Liu, Sihan Chen, Xingjian He, Longteng Guo, Xinxin Zhu, Weining Wang, and Jinhui Tang. 2024. Valor: Vision-audio-language omni-perception pretraining model and dataset. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Ying Liu, Peng Cui, Wenbo Hu, and Richang Hong. 2023b. Deep ensembles meets quantile regression: Uncertainty-aware imputation for time series. *arXiv preprint arXiv:2312.01294*.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. 2023c. Mmbench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*.

Andrey Malinin and Mark Gales. 2018. Predictive uncertainty estimation via prior networks. *Advances in neural information processing systems*, 31.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *EMNLP*.

Allan H Murphy and Robert L Winkler. 1977. Reliability of subjective probability forecasts of precipitation and temperature. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 26(1):41–47.

OpenAI. 2023. Gpt-4 technical report. abs/2303.08774.

Ilan Price, Alvaro Sanchez-Gonzalez, Ferran Alet, Tom R Andersson, Andrew El-Kadi, Dominic Masters, Timo Ewalds, Jacklynn Stott, Shakir Mohamed, Peter Battaglia, et al. 2024. Probabilistic weather forecasting with machine learning. *Nature*, pages 1–7.

Reid Pryzant, Dan Iter, Jerry Li, Yin Tat Lee, Chenguang Zhu, and Michael Zeng. 2023. Automatic prompt optimization with "gradient descent" and beam search. abs/2305.03495:7957–7968.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, et al. 2023. Aligning large multimodal models with factually augmented rlhf. *arXiv preprint arXiv:2309.14525*.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D Manning. 2023. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. *arXiv preprint arXiv:2305.14975*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Youze Wang, Wenbo Hu, Yinpeng Dong, Hanwang Zhang, Hang Su, and Richang Hong. 2023a. Exploring transferability of multimodal adversarial samples for vision-language pre-training models with contrastive learning. *arXiv preprint arXiv:2308.12636*.

Youze Wang, Wenbo Hu, and Richang Hong. 2023b. Iterative adversarial attack on image-guided story ending generation. *IEEE Transactions on Multimedia*.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. 2023. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. *arXiv preprint arXiv:2306.13063*.

Mingliang Yang, Kun Jiang, Junze Wen, Liang Peng, Yanding Yang, Hong Wang, Mengmeng Yang, Xinyu Jiao, and Diange Yang. 2023. Real-time evaluation of perception uncertainty and validity verification of autonomous driving. *Sensors*, 23(5):2867.

Zhang. 2024. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems?

Theodore Zhao, Mu Wei, J. Samuel Preston, and Hoifung Poon. 2023. Automatic calibration and error correction for large language models via pareto optimal self-supervision. abs/2306.16564.

Yunqing Zhao, Tianyu Pang, Chao Du, Xiao Yang, Chongxuan Li, Ngai-Man Man Cheung, and Min Lin. 2024. On evaluating adversarial robustness of large vision-language models. *Advances in Neural Information Processing Systems*, 36.

Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. 2022. Large language models are human-level prompt engineers.

Appendix A

Table 5, 6, 7, 8, 9, 10 show the ECE (MCE and ENCE), accuracy, and confidence of several MLLMs tested on datasets, respectively. These tables are for observing the calibration differences between MLLMs before and after fine-tuning. It can be observed that there is no consistent change in the calibration before and after fine-tuning. But we can still observe the phenomenon of miscalibration.

Model	Acc	Conf	ECE	MCE	ENCE
LLaVA-7B-Stage1	0.363	0.36	**0.075**	0.438	**0.115**
LLaVA-7B-Stage2	0.636	0.645	0.105	**0.281**	0.147
LLaVA-13B-Stage1	0.432	0.351	**0.101**	0.294	**0.219**
LLaVA-13B-Stage2	0.636	0.776	0.139	**0.215**	0.22
Qwen-VL	0.504	0.482	0.072	**0.101**	**0.133**
Qwen-VL-Chat	0.613	0.632	**0.071**	0.46	0.245

Table 5: ECE (MCE and ENCE), accuracy, and confidence of several MLLMs tested on SEED-Bench

Model	Acc	Conf	ECE	MCE	ENCE
LLaVA-7B-Stage1	0.412	0.357	**0.058**	0.453	0.162
LLaVA-7B-Stage2	0.606	0.686	0.081	**0.129**	**0.104**
LLaVA-13B-Stage1	0.577	0.369	0.207	0.445	0.314
LLaVA-13B-Stage2	0.741	0.826	**0.096**	**0.364**	**0.151**
Qwen-VL	0.564	0.544	**0.113**	0.192	0.136
Qwen-VL-Chat	0.683	0.601	0.145	**0.174**	**0.082**

Table 6: ECE (MCE and ENCE), accuracy, and confidence of several MLLMs tested on BilbaoQA2

Model	Acc	Conf	ECE	MCE	ENCE
LLaVA-7B-Stage1	0.39	0.309	0.086	0.817	0.169
LLaVA-7B-Stage2	0.48	0.532	**0.058**	**0.288**	**0.128**
LLaVA-13B-Stage1	0.35	0.264	**0.091**	0.244	0.253
LLaVA-13B-Stage2	0.54	0.671	0.131	**0.236**	**0.231**
Qwen-VL	0.45	0.413	**0.063**	0.394	0.135
Qwen-VL-Chat	0.54	0.579	0.068	**0.148**	**0.113**

Table 7: ECE (MCE and ENCE), accuracy, and confidence of several MLLMs tested on RealworldQA

Model	Acc	Conf	ECE	MCE	ENCE
LLaVA-7B-Stage1	0.238	0.235	**0.059**	**0.061**	**0.122**
LLaVA-7B-Stage2	0.215	0.414	0.202	0.721	0.414
LLaVA-13B-Stage1	0.288	0.182	**0.106**	**0.352**	**0.252**
LLaVA-13B-Stage2	0.258	0.474	0.197	0.921	0.38
Qwen-VL	0.232	0.285	**0.055**	**0.443**	**0.115**
Qwen-VL-Chat	0.314	0.453	0.146	0.959	0.253

Table 8: ECE (MCE and ENCE), accuracy, and confidence of several MLLMs tested on MathVerse

Model	Acc	Conf	ECE	MCE	ENCE
LLaVA-7B-Stage1	0.635	0.497	0.162	0.703	0.23
LLaVA-7B-Stage2	0.77	0.747	**0.092**	**0.369**	**0.144**
LLaVA-13B-Stage1	0.53	0.379	0.157	0.316	0.267
LLaVA-13B-Stage2	0.86	0.853	**0.04**	**0.252**	**0.057**
Qwen-VL	0.77	0.62	0.173	**0.374**	0.217
Qwen-VL-Chat	0.74	0.893	**0.158**	0.699	**0.199**

Table 9: ECE (MCE and ENCE), accuracy, and confidence of several MLLMs tested on Creature

Model	Acc	Conf	ECE	MCE	ENCE
LLaVA-7B-Stage1	0.6	0.454	0.145	**0.354**	0.195
LLaVA-7B-Stage2	0.79	0.795	**0.061**	0.618	**0.089**
LLaVA-13B-Stage1	0.64	0.386	0.275	0.456	0.412
LLaVA-13B-Stage2	0.8	0.877	**0.077**	**0.372**	**0.112**
Qwen-VL	0.611	0.75	0.179	0.446	0.245
Qwen-VL-Chat	0.788	0.81	**0.062**	**0.374**	**0.096**

Table 10: ECE (MCE and ENCE), accuracy, and confidence of several MLLMs tested on ScreenShot

Model	Acc	Conf	ECE	MCE	ENCE
Vicuna-7B	0.700	0.687	0.033	**0.214**	0.057
LLaVA-7B	0.785	0.794	**0.022**	0.292	**0.040**
Vicuna-13B	0.831	0.826	**0.042**	**0.207**	**0.056**
LLaVA-13B	0.846	0.891	0.050	0.217	0.069
LLaMA2-13B-Chat	0.784	0.855	0.075	**0.181**	0.097
LLaVA-LLaMA2	0.755	0.805	**0.051**	0.732	**0.078**
Qwen-7B	0.844	0.763	**0.088**	0.278	**0.111**
Qwen-VL-Chat	0.654	0.558	0.102	**0.219**	0.161

Table 11: ECE (MCE and ENCE), accuracy, and confidence of several models tested on ARC

Model	Acc	Conf	ECE	MCE	ENCE
Vicuna-7B	0.490	0.559	0.069	**0.143**	0.131
LLaVA-7B	0.626	0.658	**0.058**	0.293	**0.091**
Vicuna-13B	0.620	0.684	**0.064**	0.293	**0.097**
LLaVA-13B	0.668	0.840	0.177	**0.282**	0.227
LLaMA2-13B-Chat	0.592	0.700	0.109	0.226	0.161
LLaVA-LLaMA2	0.596	0.675	**0.079**	**0.108**	**0.133**
Qwen-7B	0.664	0.605	0.058	**0.108**	0.094
Qwen-VL-Chat	0.626	0.641	**0.043**	0.133	**0.065**

Table 12: ECE (MCE and ENCE), accuracy, and confidence of several models tested on OpenbookQA

Model	Acc	Conf	ECE	MCE	ENCE
Vicuna-7B	0.560	0.579	**0.041**	**0.148**	**0.072**
LLaVA-7B	0.708	0.852	0.152	0.707	0.188
Vicuna-13B	0.634	0.736	0.108	0.180	0.167
LLaVA-13B	0.624	0.682	**0.063**	**0.148**	**0.084**
LLaMA2-13B-Chat	0.600	0.700	0.102	0.294	0.142
LLaVA-LLaMA2	0.616	0.694	**0.079**	**0.148**	**0.125**
Qwen-7B	0.778	0.672	0.105	**0.188**	**0.137**
Qwen-VL-Chat	0.628	0.545	**0.097**	0.204	0.165

Table 13: ECE (MCE and ENCE), accuracy, and confidence of several models tested on CommonsenseQA

Model	Acc	Conf	ECE	MCE	ENCE
Vicuna-7B	0.632	0.705	0.075	0.208	0.103
LLaVA-7B	0.740	0.755	**0.056**	**0.118**	**0.078**
Vicuna-13B	0.704	0.755	**0.066**	**0.286**	**0.096**
LLaVA-13B	0.708	0.836	0.136	0.288	0.174
LLaMA2-13B-Chat	0.696	0.786	0.090	0.191	0.137
LLaVA-LLaMA2	0.682	0.734	**0.066**	**0.110**	**0.101**
Qwen-7B	0.806	0.689	0.116	0.204	0.151
Qwen-VL-Chat	0.637	0.623	**0.047**	**0.108**	**0.086**

Table 14: ECE (MCE and ENCE), accuracy, and confidence of several models tested on RACE

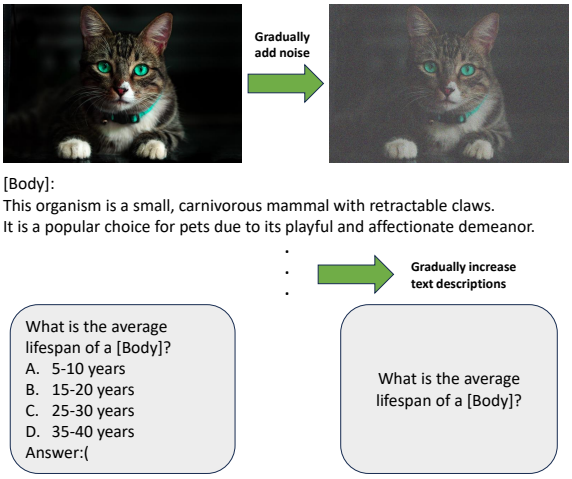

Figure 5: Gradually add text descriptions on images with different levels of noise, and observe the changes in uncertainty of information integration models for the two modalities

Appendix B

Table 11, 12, 13, 14 show the accuracy and ECE (MCE and ENCE) of several models tested on CommonsenceQA and RACE, respectively. After training the visual modality, the calibration of the model for linguistic tasks remained relatively stable, But we can still observe the phenomenon of miscalibration.

Appendix C

We further explain how to continuously integrate information from different modalities to observe changes in uncertainty. Fig.5 shows the overall method

Specifically:

1. Add Gaussian noise with different standard deviations σ_i to the original image I, obtaining the set of noisy images $\{I_{\sigma_1}, I_{\sigma_2}, \ldots, I_{\sigma_n}\}$.

2. Split the textual description into several independent sentences $S = \{S_1, S_2, \ldots, S_m\}$.

3. For each noise level σ_i, progressively add the textual descriptions $S_{\leq k}$, where $S_{\leq k} = S_1 + S_2 + \ldots + S_k$, and calculate the model's uncertainty $U(I_{\sigma_i}, S_{\leq k})$.

4. Plot and analyze the change in model uncertainty U with increasing textual descriptions $S_{\leq k}$ for different noise levels σ_i.

Appendix D

Here are some examples of IDK datasets we constructed.

model-specific dataset: Fig.6 shows LLavVA-7B repeatedly answering the same question 10 times in the MMBench dataset, but failing to get all answers correct. LLaVA-7B uses a sampling strategy with a temperature of 1 and top-p set to 0.95. Fig.9 shows the process of constructing the IDK dataset. When the accuracy of multiple answers is less than the threshold, it is considered that the model does not know. Specifically:

(i) For a single question-answer data point, have the model answer 10 times and record the accuracy.

(ii) For data items with accuracy below a specified threshold, label them as 'I do not Know'.

This dataset is selected from previously public datasets by choosing items that the model does not know, and is used to evaluate the performance of MLLMs on questions they do not know. Our contribution lies in applying Cheng et al. (2024)'s construction method for LLMs to MLLMs and constructing a VQA dataset.

OOD dataset: Fig.7 and Fig.8 show the news from July 24 that we crawled and constructed the question using GPT-3.5 (prompt is showed in Fig.10). We scraped the news from `https://www.chinanews.com.cn`. Specifically:

(i) Scrape the article contents and accompanying images from news websites for July 2024.

(ii) Provide the article content to GPT-3.5 and use the aforementioned prompt to generate several multiple-choice questions.

Our contribution lies in proposing a method to construct an OOD (Out-of-Distribution) dataset for MLLMs by transforming the latest news article contents and accompanying images through language models. We provide a dataset used to evaluate the performance of MLLMs on unknown questions. It contains 6,774 images and 20,968 multiple-choice questions (each image corresponds to multiple question-and-answer items). The dataset con-

Two magnets are placed as shown.
Hint: Magnets that attract pull together.
Magnets that repel push apart. Will these magnets attract or repel each other?
A:Repel.
B:Attract.
Please select one of the options above.

Figure 6: This question was answered 10 times by LLaVA-7B but was not answered correctly each time. We believe this model does not know the answer.

Who did President Biden announce his support for in the 2024 presidential election?
A:Donald Trump
B:Kamala Harris
C:Joe Biden
D:Hillary Clinton
Please select one of the options above.

Figure 7: The models we tested were trained before July 2024, so they absolutely cannot know who Biden announced his support for in the 2024 election.

Who recently held a farewell reception in the embassy?
A:Ambassador Xing Haiming
B:Ambassador Tan Yujun
C:Ambassador Xing Haiming and his wife
D:Ambassador Tan Yujun and his wife
Please select one of the options above.

Figure 8: The models we tested were trained before July 2024, so they absolutely cannot know who recently held a farewell reception in the embassy.

tains 7 fields, namely question, option A, option B, option C, option D, answer, and image path.

Appendix E

The process, outlined in Algorithm 1, begins with a set of initial suffixes. In each iteration, GPT-3.5 generates similar suffixes, which are evaluated for accuracy and ECE. Suffixes are then grouped by accuracy and sorted by ECE in ascending order to prioritize well-calibrated options. The top k suffixes are selected for the next iteration. After the specified iterations, the algorithm returns the best-performing suffix.

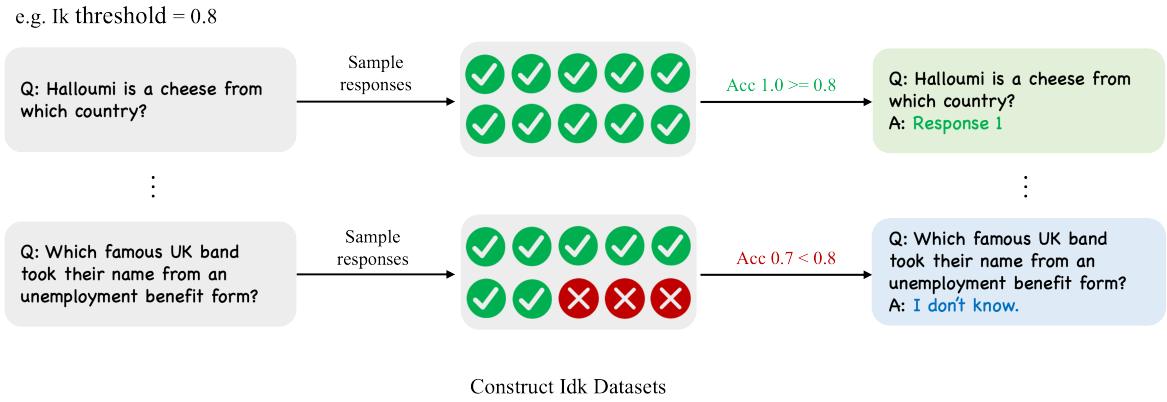

Figure 9: This figure comes from (Cheng et al., 2024). We have the model answer multiple times and determine whether it knows the answer to the question based on a pre-set threshold.

This is a news from 2024: {content}
You need to construct multiple-choice questions based on this news article. You must simultaneously meet the following five requirements:

1. This multiple-choice question is used to test the timeliness of the model. Therefore, the questions you construct must be timely. For example, if the model was trained in 2023, then the questions you ask should theoretically be ones that were not known before or in 2023.
2. The return format has several fields: question, A, B, C, D, Answer.
3. You need to generate 1 to 5 multiple-choice questions based on this news. Organize each one into a dictionary and finally return them strictly in the form of a list.
4. Note that the questions should not mention descriptions like "According to the news content" because the news content will not be provided to the test model. Just ask the question directly. However, your answer must be found in the news and not fabricated out of thin air.

Figure 10: This figure shows the prompt for constructing multiple-choice questions from crawled news by GPT.