UNCERTAINTY QUANTIFICATION FOR MULTIMODAL LARGE LANGUAGE MODELS WITH COHERENCE-ADJUSTED SEMANTIC VOLUME

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal Large Language Models (MLLMs) hold promise in tackling tasks comprising multiple input modalities, but may produce seemingly plausible but erroneous output, making them hard to trust and deploy. Accurate uncertainty metrics during inference could enable efficient escalation of queries from MLLMs to human experts or larger models for improved performance. However, existing uncertainty metrics are designed and tested only for specific modalities, and require external verifiers, additional training, or high computational resources, while struggling to handle scenarios such as out-of-distribution (OOD) or adversarial settings. To overcome these limitations, we propose UMPIRE, a training-free framework to estimate MLLM uncertainty for tasks involving various input modalities at inference time without external tools, based on the diversity of the MLLM's responses computed by its enclosed semantic volume that is adjusted with internal indicators of each response's coherence. UMPIRE does not require external modality-specific interventions and instead rely on the MLLM's own internal modality features, allowing it to generalize across modalities. We provide theoretical analysis to offer intuition on how UMPIRE could satisfy key desiderata, and empirically show that it outperforms baselines in predicting incorrect responses and providing calibrated uncertainty estimates across different input modality tasks involving text, image, text and video, including for OOD, adversarial and domain-specific data settings. We also show that UMPIRE performs well for uncertainty quantification on generation tasks beyond text, such as image and audio generation.

1 INTRODUCTION

Building on Large Language Models' (LLMs) capabilities in handling a wide variety of text-based tasks (OpenAI et al., 2024), Multimodal Large Language Models (MLLMs) are LLM-based models that can process the input of different modalities such as text along with images, audio or video via modality-specific encoders, aligned with an LLM text-decoder to produce text output, allowing them to perform important multimodal tasks such as question answering involving different modalities (Liu et al., 2023c; Hartsock & Rasool, 2024). However, while MLLMs have shown impressive capabilities, deploying them reliably in important practical settings (e.g., medical imaging analysis (Liu et al., 2023a; Tian et al., 2024; Lee et al., 2025)) may be challenging as they may produce seemingly plausible but erroneous output such as object hallucination (Bai et al., 2024), potentially more so than text-only LLMs given additional complexities from processing multimodal input. While there are works that attempt to directly mitigate such errors or hallucinations during model training by adjusting the training data (Liu et al., 2023b; Yu et al., 2024; Wang et al., 2024; Yue et al., 2024), model architecture (Liu et al., 2024; Tong et al., 2024; Zhai et al., 2023), or training process (Jiang et al., 2024; Yue et al., 2024), these errors cannot be completely eliminated, given real-world data that is noisy and ambiguous.

A complementary approach to tackling this challenge is uncertainty quantification for MLLMs, where we estimate how uncertain an MLLM is on a given query and consequently how likely it is that the MLLM would get that query correct. This would allow users to apply triaging and escalate queries that an MLLM is uncertain about to other more expensive but accurate models or human experts. However, existing uncertainty quantification works largely focus on text-only LLM settings (Kuhn

et al., 2023; Malinin & Gales, 2021) or aim to fix modality-specific hallucination (e.g., image-text input), and rely on either external verifiers (Liu et al., 2023b; Sun et al., 2023) or methods that involve relatively expensive computation (Zhang et al., 2024a; Khan & Fu, 2024b) to do so, which may not be practical in many settings with resource limitations. Given the inherently multimodal nature of our environment, we would likely need to deal with an increasing number of data modalities, and it would not be scalable nor effective to develop separate modality-specific uncertainty metrics requiring specially-engineered features for each modality. On the model architecture front, most MLLM works have adopted general frameworks comprising a core LLM and corresponding encoders/decoders for each modality (Li et al., 2022; Wu et al., 2024). This raises the question: *Can we similarly achieve an effective training-free uncertainty quantification method that can be applied generally across input modalities without designing modality-specific mechanisms, or using external tools?*

In our work, we present UMPIRE, a training-free framework to estimate the uncertainty of MLLM output for task queries involving multimodal input. Rather than using external tools or modality-specific methods, UMPIRE uses a simple but effective *modality-agnostic* method that rely on the MLLM's own multimodal feature spaces to compute a metric that indicates how likely it may get a query wrong. Our key contributions are as follows: we (1) proposed a set of clear desiderata that MLLM uncertainty metrics should satisfy (Sec. 2); (2) developed our novel UMPIRE framework , inspired by the quality-diversity kernel decomposition of determinantal point processes (Kulesza et al., 2012), that estimates uncertainty based on the semantic volume enclosed by sampled responses adjusted by each response's coherence scores (Sec. 4); (3) provide theoretical analysis to build intuition and insights on how UMPIRE relates to the desiderata (Sec. 5; (4) empirically show that UMPIRE consistently outperforms baselines in various multimodal input question-answering tasks such as text-image, text-only, text-audio and text-video settings and image/audio generation tasks despite not needing explicit modality-specific mechanisms (Sec. 6.1, 6.2, 6.3) ; and (5) show that UMPIRE can perform well in settings involving deferring uncertain task instances to larger models or uncertainty estimation for blackbox API-access MLLM models (Sec. 6.4).

2 Problem formulation and desiderata

Problem formulation. Consider a whitebox MLLM $\mathcal{M}$ that takes in multi-modality input, including text q and other modality I (image, audio, video, etc), and autoregressively produce text output $y = [w_i]_{i=1}^N$ that are sequences of tokens w from the MLLM decoder's vocab space. Specifically, MLLMs can be represented as conditional probability distributions $\mathcal{M}(I, q) \coloneqq p_{\mathcal{M}}(y|I, q) = p_{\mathcal{M}}(w_1|I, q)p_{\mathcal{M}}(w_2|I, q, w_1) \dots p_{\mathcal{M}}(w_n|I, q, w_{1:n-1})$.

We have tasks $\mathcal{T}$ with task instances $t \in \mathcal{T}$, where $t \coloneqq (I_t, q_t)$ represents the multimodal input query, and we explicitly denote task instances with known text ground truth output y_t^*, as $t^* \coloneqq (t; y_t^*)$. The MLLMs' response $\hat{y}_t$ to a task instance t can be sampled (e.g., with low temperature) from $\mathcal{M}(t) = \mathcal{M}(I_t, q_t)$, and its performance evaluated by how well the response matches the ground truth, i.e., a utility function $a(\mathcal{M}, t^*) \coloneqq v(\hat{y}, y^*) \in [0, 1]$. Unless otherwise stated, we consider a utility that is a binary indicator $v(\hat{y}_t, y_t^*) \coloneqq \mathbb{I}\{\hat{y}_t = y_t^*\}$. We define the overall MLLM performance on task $\mathcal{T}$ as the expected performance over its labeled task instances, i.e., $a(\mathcal{M}, \mathcal{T}) \coloneqq \mathbb{E}_{t^* \in \mathcal{T}} a(\mathcal{M}, t^*)$.

Our goal is to develop a framework that computes a task instance-specific uncertainty metric $u(\mathcal{M}; t)$ for any $t \in \mathcal{T}$ at inference time that is highly indicative of the expected accuracy $a(\mathcal{M}, t^*)$. Note that we are developing a metric for overall uncertainty (rather than aleatoric or epistemic uncertainty), which is task-specific (i.e., uncertainty associated with a given input query) rather than response-specific (i.e., confidence score for a given sampled instance of MLLM output). Our metric can be used to assess whether a task instance is likely to be answered wrongly by an MLLM, and hence should be discarded, and escalated to a more capable MLLM model or human expert instead.

Desiderata. Given the above setting, we propose a non-exhaustive list of key desiderata that an uncertainty metric should satisfy. First, the metric should meet three *effectiveness* desiderata:

R1 Classification. The metric should be able to distinguish between task instances that the MLLM will get correct ($t_c \in \mathcal{C} \coloneqq \{t \in \mathcal{T} \mid a(\mathcal{M}, t^*) = 1\}$) or wrong ($t_w \in \mathcal{W} \coloneqq \{t \in \mathcal{T} | a(\mathcal{M}, t^*) = 0\}$ with high probability. Specifically, for randomly sampled pairs of task instances t_c and t_w,

$$\mathbb{P}[u(\mathcal{M}, t_w) > u(\mathcal{M}, t_c)] \approx 1 \tag{1}$$

where the goal is for Eq. (1) to be as close to 1 as possible (which in practice may be limited by the task and model), implying that the metric can classify well whether the model will get task instances wrong, using just $\mathcal{M}$ and instance input t. This means that there exist a threshold γ such that $u(\mathcal{M}, t) > \gamma$ indicates that it is likely that $t \in \mathcal{W}$, and smaller values indicate that $t \in \mathcal{C}$. Note that Eq. (1) can be evaluated by computing the Area under the Receiver Operating Characteristic Curve (AUROC) of the metric, which we do in Sec. 6.1.

R2a Proportionality. The metric should be proportional to the probability that the MLLM will get the task instance wrong, i.e.,

$$u(\mathcal{M}, t) \propto \mathbb{P}[a(\mathcal{M}, t^*) = 0]. \tag{2}$$

Compared to **R1**, satisfying **R2a** allows the metric to provide a meaningful continuous score of how likely the MLLM will get a task instance wrong, rather than be used only for classification.

R2b Calibration. Given a small sample of *unlabeled* task instances from $\mathcal{T}$, the metric u should be easily adjustable to $\tilde{u} \in [0, 1]$ (e.g., using min-max scaling) such that it provides well calibrated estimates of how confident the MLLM is in answering $t \in \mathcal{T}$ correctly, (Guo et al., 2017), i.e.,

$$\mathbb{P}(a(\mathcal{M}, t^*) = 1 \mid \tilde{u}(\mathcal{M}, t) = p) \approx p, \quad \forall p \in [0, 1]. \tag{3}$$

R2bis a stricter version of **R2a** where the metric is properly scaled to estimate the probability that the MLLM would get a given task instance correct. The probability estimates enable downstream applications such as risk management, and more interpretable choices of the classification threshold γ for **R1**. Note that for ease of comparability with past works (Tian et al., 2023), **R2b** and $\tilde{u}$ are formulated based on response accuracy $a(\mathcal{M}, t^*) = 1$, while **R2a** and u are based on error $a(\mathcal{M}, t^*) = 0$ which is more natural for an uncertainty measure. In B.3, we provide further discussion on the differences among these desiderata and why they are needed.

We also consider design desiderata related to common practical requirements of metric deployment:

R3 Multimodal generalizability. The metric should be computed with the same modality-agnostic framework, without additional modality-specific engineering or tools, and still satisfy all other desiderata across a wide range of input modalities (e.g., text, image, audio, video). A stricter version of this desideratum, **R3'**, is for the metric to all apply across different output modalities (e.g., image generation), even though many LLM works focus on multimodality input with text output. While not the focus of this work, we provide some results and analysis on **R3'** in Sec. 6.3.

R4 Multimodal coherence. The metric should consider the coherence of each sampled response with respect to the various input modalities in the task instance query (e.g. images and text), instead of only one modality. Having a modality-agnostic framework to satisfy **R3** *does not mean ignoring modality information*. Rather, there should ideally be a general method that can *make use of modality information already trained into the MLLM*.

R5 Computational Efficiency. The metric could be efficiently computed, with (a) fast computational runtime, and (b) no strict requirements of external pre-trained models or separately trained reward models as they incur additional costs and may not be feasible for some inference pipelines. When the MLLM under study is a blackbox, it may be necessary to relax condition (b) to use a proxy whitebox model, but the proxy model should be small and cheap to run.

3 Related Works

Modality-specific methods. Although MLLMs' hallucination and miscalibration problems are well known (Chen et al., 2025; Rohrbach et al., 2018; Bai et al., 2024), research on task instance-specific uncertainty quantification for MLLMs is relatively underdeveloped. Most works are focused on (image-text input, text output) modalities, with image-specific approaches and external tools, violating **R3**. This include works that rely on the use of external reference/entailment models (Zhang et al., 2024a; Sun et al., 2023; Liu et al., 2023b), supervised training of classifiers (Li et al., 2024), or large numbers of modality-specific query perturbations (Khan & Fu, 2024a; Zhang et al., 2024a) to test model consistency (also violating **R5**). Furthermore, even by relaxing the design desiderata by allowing access to external models or more computation time and focusing on just modalities that these methods are designed for, they do not satisfy the effectiveness desiderata (**R1**-**R2b**) compared to our method UMPIRE, as we see later in Sec. 6.

LLM uncertainty methods. If we focus on only text-output settings, we found that existing LLM uncertainty metrics designed only for text input-settings could be adapted for multimodality ouput and potentially achieve better effectiveness (e.g., **R1** on classification) compared to modality-specific methods (see Sec. 6). This includes methods based on lexical (token probability) distributions (Malinin & Gales, 2021), semantic clusters/graphs derived from external text entailment models (Kuhn et al., 2023; Nikitin et al., 2024; Lin et al., 2024), semantic embeddings of sampled responses (Chen et al., 2024; Qiu & Miikkulainen, 2024), and prompting (Xiong et al., 2024a). However, these approaches still typically violate several desiderata (e.g., **R4** by not considering response coherence with multimodal input) and *underperform UMPIRE even for the text-only LLM scenario*.

Entropy-based approaches. Many MLLM and LLM uncertainty works also rely on computing discrete entropy measures (Malinin & Gales, 2021; Nikitin et al., 2024; Zhang et al., 2024a). However, it is unclear how to compare entropy values across different support sets (e.g., distributions defined on 2 versus 5 classes), especially when the support set is determined by external models, making them potentially hard to use in practice. Eigen (Chen et al., 2024) considers differential entropy in the sentence embedding space, and by making a relatively strong Gaussian assumption ends up with a final metric form that bears some similarity to UMPIRE (in fact, Eigen can be seen as a special case). However, key differences (see App. B.2) result in Eigen consistently underperforming UMPIRE as can be seen in both multimodal input and text-only tasks (Table 1), and not meeting key desiderata such as modality coherence (**R4**).

In contrast to existing works, we can combine several insights to design a framework that meets all desiderata. First, many MLLMs adopt an encoder/decoder + LLM core architecture (Zhan et al., 2024; Wu et al., 2024), where rich modality-specific features have already been trained into the model. Hence, it may be possible to develop a *modality-agnostic framework using an MLLM's inherent embeddings without the use of external tools* (**R3**, **R5**). Second, trained MLLMs would have the capability to *take into account all input modalities and response coherence via their model-generated probabilities* (**R4**), which though typically uncalibrated would contain useful relative information among responses. Finally, our goal is to develop an uncertainty metric for a task instance t, rather than a response-specific confidence score. Although there are settings where response-wise indicators would be useful, our framework has been developed for the task-instance settings as uncertainty is conventionally understood as a property of the response distribution rather than a sample from it. Thus, we consider the *response distribution for t which we can probe with efficient sampling* via accelerated batch inference (Kwon et al., 2023) without external tools.

4 METHOD

Combining the insights above, we present the framework and theoretical analysis of our proposed metric UMPIRE, a simple but effective framework with a task instance-specific uncertainty metric that jointly considers (a) the volume enclosed by sampled responses in the MLLM's semantic embedding space capturing global diversity, adjusted by (b) local measures of the MLLM's per-response coherence based on all input modalities. Drawing inspiration from the quality-diversity kernel decomposition of determinantal point processes (DPP) (Kulesza et al., 2012), our UMPIRE framework has the following steps (summarized in Fig. 1) that aim to satisfy **R3**-**R5**:

U1 Sampling. Given a task instance $t \in \mathcal{T}$, the MLLM generate k responses $\mathcal{Y}_t = \{\hat{y}_i\}_{i=1}^{k}$.

U2 Semantic embedding. For each response $\hat{y}_i$, we extract the last embedding layer vector of the last response token (more analysis on layer selection in App. D.1) $\phi_i \in \mathbb{R}^d$, and normalize it if not already so (Reimers & Gurevych, 2019) With k samples, we form the $k \times d$ embedding matrix Φ_t, where using the MLLM's rich multimodal semantic embeddings satisfy **R3**.

U3 Incoherence score. Concurrently, we extract the model-generated probability scores p_i for each response $\hat{y}_i$ which has been generated conditional on all input modalities, capturing multimodal coherence signals (**R4**). We compute the incoherence score $c_i \in \mathbb{R}^+$, $c_i \coloneqq \exp \alpha(1 - p_i)$, where α is a scaling hyperparameter that is fixed across $\mathcal{T}$ and can be set heuristically (App. D.2). This score captures how incoherent each response is: e.g., a response deemed fully coherent by the MLLM will have $p_i = 1$ and the smallest value $c_i = 1$, while low probability responses will have large c_i. With k samples, we will have a $k \times k$ incoherence score diagonal matrix C_t.

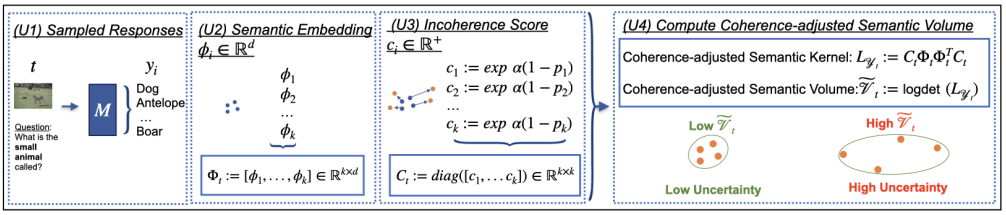

Figure 1: Schematic describing the UMPIRE framework

U4 Coherence-adjusted semantic volume. Given only the above and without external tools (**R5**), we compute its coherence-adjusted semantic kernel $L_{\mathcal{Y}_t} := C_t \Phi_t \Phi_t^T C_t$, similar to quality-adjusted kernels used in DPPs. We can then compute the final UMPIRE uncertainty metric $\widetilde{V}_t := \log\det(L_{\mathcal{Y}_t})$, where in practice a small jitter term is added to $L_{\mathcal{Y}_t}$ for numerical stability and to avoid degeneracy. $\widetilde{V}_t$ captures coherence-adjusted semantic volume, since from geometry, $\det(L_{\mathcal{Y}_t}) = \text{Vol}^2(C_t \Phi_t)$, the squared volume spanned by the coherence-adjusted response semantic embedding vectors. Note that high incoherence responses have magnified contributions, resulting in the metric having highest values when responses are both diverse and incoherent.

Beyond the design desiderata, UMPIRE is also constructed to satisfy the effectiveness desiderata **R1**-**R2b** well. We first provide some theoretical analysis on UMPIRE given simplifying assumptions that provide some intuition and interpretation of the metric's components (Sec. 5 and App. A for details on assumptions and proofs), This complements our empirical results (Sec. 6) on how UMPIRE achieves good performance over a wide range of settings and outperforms baselines.

5 Theoretical Analysis and Intuition

For our theoretical analysis, we consider the metric normalized by sample number k, which we first show can be decomposed into two interpretable terms: an unadjusted semantic-volume diversity term and a Monte-Carlo estimate of the average incoherence based on the model's internal assessment,

$$\bar{V}_t := \tfrac{1}{k}\tilde{V}_t = \tfrac{1}{k}\log\det(\Phi\Phi^T) + \tfrac{2\alpha}{k}\sum_{i=1}^{k}(1 - p_i). \tag{4}$$

Proposition 1 (Coherence-adjusted semantic volume decomposition and Monte-Carlo coherence term)**.** *Eq.* (4) *holds exactly, and under the simplifying assumption of i.i.d. response samples, the second term can be interpreted as a Monte-Carlo estimator of average incoherence* $2\alpha\mathbb{E}[(1 - p_i)]$ *whose standard error scales as* $O(1/\sqrt{k})$.

For the first semantic volume term, we further build intuition by considering the simplifying assumption (A5 in App. A) that the MLLM response embeddings follow a finite mixture distribution (i.e., some number of semantic clusters), relating the term to model uncertainty arising from the global diversity of semantic response clusters. Together with Proposition 1, it suggests how the proposed metric with both terms are correlated with the MLLM's correctness probability for a given task instance and may satisfy **R2a**, which we show empirically in Sec. 6.2.

Proposition 2 (Semantic volume)**.** *Let the MLLM's response embedding distribution be a finite mixture model with total covariance* $\Sigma_{\text{mix}} = \Sigma_{\text{within}} + \Sigma_{\text{between}}$. *If the between-cluster covariance increases such that* $\Sigma'_{\text{between}} \succeq \Sigma_{\text{between}}$ *in the positive semidefinite (PSD) order, let the new total covariance be* $\Sigma'_{\text{mix}} = \Sigma_{\text{within}} + \Sigma'_{\text{between}}$. *Then, the determinant is non-decreasing:* $\det(\Sigma'_{\text{mix}}) \geq \det(\Sigma_{\text{mix}})$.

Consequently, for large k where sample covariance $S_k = \frac{1}{k}\Phi^T\Phi \approx \Sigma_{\text{mix}}$, the unadjusted empirical semantic volume term in Eq. (4) is expected to increase with the between-cluster spread. As discussed in App. A, together with assumption A5 regarding how responses with higher probability of being correct will tend to concentrate on a subset of cluster, this will relate the semantic volume to the MLLM's correctness probability.

We now show the statistical stability of our proposed metric over sample size k, which complements our empirical results that small values of k tend to provide reliable estimates. This also provides some

Dataset	AUROC ↑					CPC ↑					ECE ↓				
	NC	LN-Ent.	Sem.Ent.	Eigen	Ours	NC	LN-Ent.	Sem.Ent.	Eigen	Ours	NC	LN-Ent.	Sem.Ent.	Eigen	Ours
Image-text															
VQAv2	0.769	0.781	0.848	0.868	**0.882**	0.784	0.553	0.916	0.938	**0.946**	0.326	0.046	0.046	0.047	**0.038**
OKVQA	0.528	0.705	0.716	0.738	**0.755**	0.778	0.851	0.277	0.893	**0.966**	0.504	0.041	0.144	0.162	**0.036**
AdVQA	0.657	0.647	0.763	0.774	**0.787**	0.562	0.916	0.759	0.888	**0.979**	0.344	0.068	0.161	0.217	**0.042**
MathVista	0.763	0.667	0.805	0.814	**0.822**	0.721	0.909	0.856	0.797	**0.945**	0.078	0.116	0.220	0.312	**0.071**
VQA-RAD	0.706	0.614	0.767	**0.803**	0.802	0.733	0.892	0.690	0.656	**0.908**	0.138	0.111	0.359	0.366	**0.067**
Avg (image)	0.685	0.683	0.780	0.799	**0.810**	0.716	0.824	0.700	0.834	**0.949**	0.278	0.076	0.186	0.221	**0.051**
Text-only															
CoQA	-	0.640	0.739	0.738	**0.791**	-	0.430	0.790	0.495	**0.880**	-	0.148	0.312	0.242	**0.081**
TriviaQA	-	0.552	0.653	0.650	**0.720**	-	0.413	0.298	0.531	**0.923**	-	0.166	0.332	0.282	**0.054**
NQ	-	0.755	0.850	0.824	**0.853**	-	0.566	0.693	0.727	**0.892**	-	0.130	0.397	0.287	**0.022**
Audio-text															
SLUE-P2-SQA5	-	0.755	0.794	0.783	**0.819**	-	0.685	0.824	0.831	**0.940**	-	0.098	0.235	0.175	**0.058**
Spoken SQuAD	-	0.710	0.765	0.775	**0.797**	-	0.855	0.897	0.877	**0.985**	-	0.071	0.279	0.229	**0.030**
Video-text															
Video-MME-short	-	0.717	0.706	0.821	**0.823**	-	0.618	0.683	0.666	**0.697**	-	**0.089**	0.226	0.334	0.156
Avg (all)	-	0.686	0.764	0.781	**0.805**	-	0.699	0.698	0.754	**0.915**	-	0.099	0.246	0.241	**0.060**

Table 1: Effectiveness of various uncertainty metrics across multimodal datasets (details in App. C.1). The evaluation covers (i) **uncertain response classification** (**R1**) using AUROC (↑ better), (ii) **uncertainty proportionality** (**R2a**) using CPC (↑ better), and (iii) **calibration** (**R2b**) using ECE (↓ better). Overall, UMPIRE achieves the best or second-best performance across all baselines and modalities, with only marginal differences when not ranked first. For a fair comparison with NC, we report the average on all image-text datasets (Avg image) and on all datasets (Avg all).

intuition on how if metric values for correct and wrong task instances have a large enough expectation gap (possibly contributed partially by factors related to the first 2 propositions), our metric could then **R1**. In Sec. 6.1, we see how UMPIRE shows strong **R1** empirical performance.

Proposition 3 (Concentration and Ranking Consistency)**.** *Assuming bounded embedding norms, there exist constants c, C such that for any $\eta > 0$, the metric concentrates around its mean with sub-exponential tails in k: $\Pr\left(|\bar{V}_t - \mathbb{E}\bar{V}_t| > \eta\right) \leq C\exp(-ck\eta^2)$. Moreover, if two instances t_a and t_b satisfy a true mean gap $\mathbb{E}\bar{V}_{t_a} - \mathbb{E}\bar{V}_{t_b} \geq \Delta > 0$, then the empirical ordering $\bar{V}_{t_a} > \bar{V}_{t_b}$ holds with probability at least $1 - 2C\exp(-ck(\Delta/2)^2)$.*

Finally, we show that **R2b** can be satisfied by scaling our metric via isotonic regression if the metric is monotone. While this assumption does not always hold exactly, our empirical results suggest that this is a reasonable approximation (Fig. 4 in App. C.4), and that min-max scaling *without labeled data* can also typically achieve good **R2b** results (Sec. 6.2).

Proposition 4 (Calibration under Monotonicity)**.** *If the function $u \mapsto \Pr(\text{correct} \mid \bar{V}_t = u)$ is monotone, then isotonic regression fitted on a development set of size n yields a consistent calibrated estimator of the true correctness probability. Standard non-parametric worst-case convergence rates apply (e.g., $O(n^{-2/3})$ for the mean squared error).*

6 Experimental results

We empirically evaluate whether UMPIRE satisfies the desiderata (Sec. 2), compare its performance against baselines, and demonstrate its utility in practical scenarios. As literature on image-text input modality tasks is relatively more developed with benchmarks and baselines, we primarily conduct in-depth studies in this setting, though we also analyzed other modalities such as audio-text and video-text input modality tasks, as well as the text-only single-modality setting to assess the modality generalizability **R3** desiderata. We also evaluated metrics on the stricter **R3'** desideratum by considering image and audio generation tasks (details on datasets are in App. C.1.1).

We use `Llava-v1.5-13b` (Liu et al., 2023c), `Phi-4` (Abdin et al., 2024), `LLaVA-NeXT-Video-7b-hf` (Zhang et al., 2024b) for image-text, audio-text, and video-text experiments respectively. We compare UMPIRE against baselines representative of different approaches, including a modality(image)-specific metric Neighborhood Consistency (NC) (Khan & Fu, 2024a), and 3 text-only LLM uncertainty metrics that we adapt to the multimodal input setting: LN-Entropy (LN-Ent) (Malinin & Gales, 2021), Semantic Entropy (Sem.Ent) (Kuhn et al., 2023), and Eigenscore (Eigen) (Chen et al., 2024). NC and Sem.Ent use external tools and hence violate **R5**,

but we still run them to analyze any performance issues beyond this violation. Details on experiment settings, additional baseline methods and ablation studies to highlight UMPIRE's robustness across parameters can be found in the Appendix (App. C, App. D).

6.1 **R1**: CLASSIFICATION OF UNCERTAIN RESPONSES

We first evaluate metrics on **R1**, i.e., whether the metrics can classify tasks that the MLLM will get correct (t_c) or wrong (t_w), measured via AUROC. Table 1 shows that UMPIRE **consistently achieves the best or competitive performance** with an average AUROC around 0.810 on image-text datasets, excelling particularly in challenging datasets like OKVQA and AdVQA, where multimodal-specific methods like NC struggle due to adversarial and out-of-distribution scenarios. This robustness highlights UMPIRE's ability to handle diverse and incoherent predictions in these datasets. Beyond vision, UMPIRE also **demonstrates robust classification performance across text-only, audio-text, and video-text tasks**, underscoring its modality generalizability **R3**. Moreover, in practice, users will need to set thresholds based on their use cases to target some minimum requirements, such as False Positive Rates (FPR). In App. C.3 Table 6, we also show how UMPIRE framework's better AUROC performance for **R1** translates to consistently higher True Positive Rates (TPR) given various FPR requirements. The consistent improvement across all scenarios and modalities suggests that UMPIRE robustly satisfies **R1**, and can be deployed more reliably in real-world question-answering applications where high-stakes decisions depend on model uncertainty (see Sec. 6.4).

6.2 **R2A**, **R2B**: PROPORTIONALITY AND CALIBRATION

Similar to past calibration works (Guo et al., 2017), we sort instances in a given task $t \in \mathcal{T}$ by the computed uncertainty metric $u(\mathcal{M}, t)$, and put them in equally-sized bins b_j. Each bin is associated with its highest metric value u_j, and the estimated probability that instances within will be answered correctly that is computed by its expected response accuracy $\bar{a}_j = \sum_{t_j \in b_j} a(\mathcal{M}, t_j)/|b_j|$.

Method	R1	R2ab	R3	R3'	R4	R5
NC	✗	✗	✗	✗	✗	✗
LN-Ent.	✗	✓	✓	✗	✗	✓
Sem.Ent.	✓	✗	✗	✗	✓	✗
Eigen	✓	✗	✓	✗	✗	✓
Ours	✓	✓	✓	✓	✓	✓

Table 2: Summary of whether UMPIRE and baselines satisfy the proposed desiderata.

Calibration Pearson Correlation (CPC)(R2a). We define CPC score as the negative Pearson correlation between u_j and $\bar{a}_j$ across bins (Eq. (2)). Higher CPC indicates that the metric is more linearly correlated to the estimated probability that the MLLM will answer the instance wrongly. Table 1 shows that UMPIRE **consistently performs better than baselines across all settings**, achieving an average CPC of ~ 0.95 for image-text tasks and 0.915 across all modality tasks (**R3**), more than 20% higher than the next best metric. Note that UMPIRE also produces **more stable and reliable results with consistently high CPC**, unlike other baselines with performance that fluctuates greatly depending on the specific task. This is critical for uncertainty metrics, which themselves are meant to assess model reliability.

Expected Calibration Error (ECE) (R2b). The strong linear relationship indicated by UMPIRE's CPC score suggests that a simple scaling process would be sufficient to make the UMPIRE metric well-calibrated and satisfy **R2b**. We evaluate the ECE (Guo et al., 2017) of metrics by using an **unlabeled** development set of instances (5% of dataset) to compute $\tilde{u}$ via min-max scaling before computing the ECE. UMPIRE achieves a very low ECE on almost all datasets with an average of 0.060 (see Table 1), and is **significantly lower than baselines** with up to $2-6\times$ lower ECE than the next best baselines for the more challenging tasks.

6.3 **R3**,**R4**,**R5**: MULTIMODAL GENERALIZABILITY, COHERENCE AND COMPUTE EFFICIENCY

Multimodal generalizability (R3). As seen in table 1, UMPIRE, without modality-specific modifications or external tools, **consistently performs well in effectiveness** (**R1**-**R2b**) across multiple input modality tasks (image-text, text-only, audio-text, video-text) and hence satisfy **R3**, empirically supporting our approach in using MLLMs' inherent multimodality capabilities to achieve **R3**. In contrast, metrics that rely on modality-specific input, such as NC, cannot be directly applied to other input modalities to satisfy **R3**. Text-only modality baselines (LN-Ent, Sem.Ent, Eigen) could

Method	Image		Audio
	AnyGPT	NExTGPT	NExTGPT
PUNC	0.460	0.227	-
LN-Ent.	0.235	0.443	0.641
Eigen	0.341	0.581	0.215
Ours	**0.839**	**0.755**	**0.706**

Table 3: Pearson Correlation between uncertainty metrics and CLIP/CLAP score (image/audio) in image and audio generation tasks. UMPIRE achieves the highest correlation with the continuous evaluation criterion CLIP/CLAP score.

Method	Overhead (s)
NC	9.0
Sem.Ent.	9.1
Eigen	1.8e-3
LN-Ent.	3.4e-4
Ours	8.3e-4

Table 4: Comparison overhead runtime ($k = 50$) metrics, averaged on 3000 VQAv2 samples and run on an L40.

be adapted to multimodal input tasks as these tasks still involve text output, but they have worse effectiveness compared to UMPIRE, even for the text-only setting as they did not fully capture both global semantic diversity and local coherence scores unlike UMPIRE.

Non-text output generation tasks (R3'). To demonstrate that UMPIRE can also satisfy the stricter desiderata **R3**' that requires applicability to various modality output (beyond text) generation tasks, we ran experiments on image (MS-COCO caption Chen et al. (2015)) and audio (name and cite) generation using any-to-any MLLMs NExT-GPT (Wu et al., 2024) and AnyGPT (Zhan et al., 2024)). As the utility for such tasks are continuous scores, we evaluate the uncertainty metrics on 2a, by computing the Pearson correlation between the metrics and quality scores (image: CLIP score (Hessel et al., 2021); audio: CLAP score (Elizalde et al., 2023)). Sem.Ent is designed to require text-specific external tools, and hence cannot be applied. For these settings, we also compare with PUNC (Franchi et al., 2025), an image-specific uncertainty metric. Table 3 shows that **UMPIRE consistently has strong correlation with image and audio quality, outperforms all baselines** across different modalities, and hence satisfy **R3**'. This allows UMPIRE to also be applied to assess whether a given task instance might be challenging for MLLMs to produce high quality image/audio generations for, which to our knowledge has not been well-studied in past works.

Multimodal coherence (R4). To further demonstrate that UMPIRE considers multimodality information despite not using modality-specific external tools or methods, we analyze post-generation whether metric performance degrades when image-input information is (1) corrupted with noise, (2) replaced with a black image, or (3) removed. A metric that satisfies **R4** should show performance degradation for (1), which worsens for (2) and (3), and is comparable between (2) and (3) since all useful signals would be removed in both cases. Fig. 2 **shows that UMPIRE exhibits this behavior well and satisfies R4**, along with Sem.Ent to a lesser extent. In contrast, among other input modality-agnostic baselines, LN-Ent exhibits large degradation but with inconsistent trends (e.g., no image has less degradation than noisy image), and Eigen remains unchanged implying that it does not consider multimodal coherence at all.

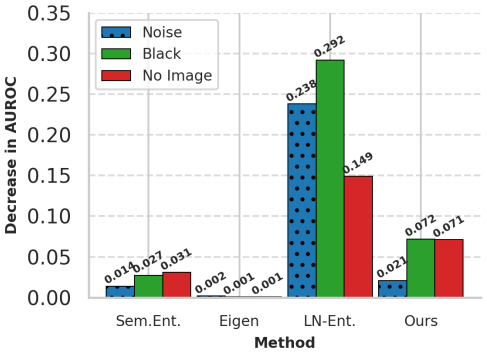

Figure 2: Decrease in AUROC when image-input information is (1) corrupted with noise, (2) replaced with a black image, or (3) removed.

Computational efficiency (R5). Table 4 shows the computational overhead time introduced for each method. Note that both NC and Sem.Ent violates **R5** as they require external tools and expensive computation (e.g., pairwise response evaluation) and the former requires training as well. UMPIRE and the other more efficient metrics satisfying **R5** take up to 4 orders of magnitude less computational overheads compared to them. We also show that UMPIRE achieves significant performance margins over single-sample methods (App. C.5) and consistently outperforms baselines regardless of sampling budget k (App. D.3) with as few as $k = 5$ generations (which can be sped up by accelerated LLM batch inference (Kwon et al., 2023)).

Method	Image					Avg (image)	Text-only			Audio		Video	Avg (all)
	VQAv2	OKVQA	AdVQA	MathVista	VQA RAD		CoQA	TriviaQA	NQ	SLUE-P2 SQA5	Spoken SQuAD	Video-MME short	
NC	0.934	0.664	0.722	0.344	0.502	0.633	-	-	-	-	-	-	-
LN-Ent.	0.949	0.787	0.751	0.301	0.514	0.660	0.543	0.534	0.328	0.708	0.677	0.182	0.570
Sem.Ent.	0.948	0.773	0.782	0.370	0.549	0.684	0.599	0.629	0.356	0.701	0.678	0.174	0.596
Eigen	0.963	0.800	0.800	0.382	0.589	0.707	0.594	0.597	0.350	0.704	0.704	**0.209**	0.608
Ours	**0.966**	**0.807**	**0.809**	**0.388**	**0.600**	**0.714**	**0.656**	**0.691**	**0.372**	**0.737**	**0.719**	**0.209**	**0.632**

Table 5: Comparison of AURAC across datasets for different uncertainty metrics, including NC, LN-Ent, Sem.Ent, Eigen, and UMPIRE (Ours).

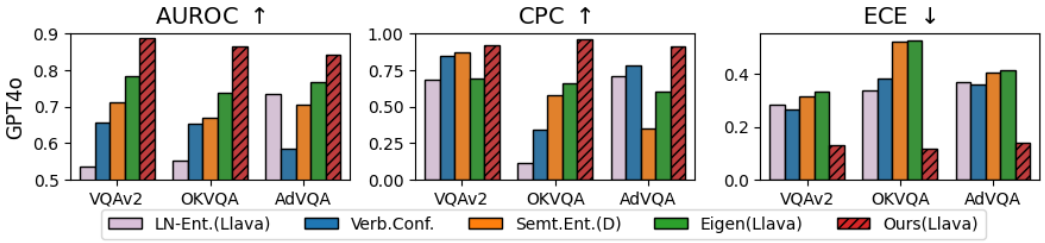

Figure 3: Performance of uncertainty metrics in blackbox settings across image-text QA datasets. Sem.Ent (D) indicates its discrete version, (Llava) indicates Llava as the white-box proxy model.

6.4 Practical applications

Selective answering. We consider a practical scenario where a user has a small local MLLM for a task, but also a limited budget to route/escalate some task instances to a more capable but expensive MLLM or human expert to answer. A good uncertainty metric like UMPIRE could help select the instances that the small MLLM could answer with the lowest uncertainty while escalating the rest, and hence improve overall accuracy. We evaluate this via the Area Under the Rejection-Accuracy Curve (AURAC) (Hüllermeier & Waegeman, 2021), which summarizes the improvement in accuracy across varying rejection thresholds. UMPIRE consistently achieves the highest AURAC across all datasets (Table 5), indicating more reliable uncertainty estimation for selective answering. To show that UMPIRE's AURAC performance gains over baselines are statistically significant, we performed statistical tests across various MLLM model-dataset combinations to show the difference is statisticaly significant (App. C.6).

Blackbox Models. In practice, users might need to estimate uncertainty for blackbox MLLMs that can only be accessed via APIs and do not provide internal model semantic embeddings or response probabilities. To apply UMPIRE, we can employ a much smaller whitebox proxy MLLM that processes the blackbox MLLM's responses to generate embeddings and probabilities for computing our uncertainty metric. For this setting, we also compared with the Verbalized Confidence (Verb.Conf.) baseline (Xiong et al., 2024a) that can be run with SOTA blackbox API models. In Fig. 3, we see that UMPIRE consistently and significantly outperform baselines when assessing GPT4o's (Hurst et al., 2024) uncertainty on VQAv2, OKVQA, and AdVQA using `Llava-v1.5-13b` as the whitebox proxy model. Experiments with other blackbox (e.g., Claude, GPT4o Mini) and proxy models show similar results where UMPIRE has large performance gains over baselines (App. D.7). This shows how UMPIRE remains practical for blackbox settings given its ease of use and speed (**R5**).

7 Conclusion

We propose UMPIRE, a novel inference-time framework that provides efficient and effective MLLM uncertainty estimates that generalizes across multiple input and output modalities, without the need for modality-specific tools or interventions. We presented desiderata that MLLM uncertainty metrics should satisfy, provided some theoretical analysis with simplifying assumptions to build intuition on how UMPIRE could be interpreted, and empirically showed how UMPIRE consistently outperforms baselines across an extensive range of multimodal question-answering and generation tasks. Future work could analyze and extend the framework for multimodal reasoning tasks.

REFERENCES

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, et al. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*, 2024.

Anthropic. The claude 3 model family: Opus, sonnet, haiku. 2024. URL https://api.semanticscholar.org/CorpusID:268232499.

Zechen Bai, Pichao Wang, Tianjun Xiao, Tong He, Zongbo Han, Zheng Zhang, and Mike Zheng Shou. Hallucination of Multimodal Large Language Models: A Survey, April 2024.

R.E. Barlow. *Statistical Inference Under Order Restrictions: The Theory and Application of Isotonic Regression*. Out-of-print Books on demand. J. Wiley, 1972. ISBN 9780471049708. URL https://books.google.com/books?id=DEamySUDBWcC.

Sabyasachi Chatterjee, Adityanand Guntuboyina, and Bodhisattva Sen. On risk bounds in isotonic and other shape restricted regression problems. 2015.

Chao Chen, Kai Liu, Ze Chen, Yi Gu, Yue Wu, Mingyuan Tao, Zhihang Fu, and Jieping Ye. INSIDE: LLMs' internal states retain the power of hallucination detection. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Zj12nzlQbz.

X. Chen, H. Fang, TY Lin, R. Vedantam, S. Gupta, P. Dollár, and C. L. Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv:1504.00325*, 2015.

Zijun Chen, Wenbo Hu, Guande He, Zhijie Deng, ZHeng ZHang, and Richang Hong. Unveiling uncertainty: A deep dive into calibration and performance of multimodal large language models. In Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert (eds.), *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 3095–3109, Abu Dhabi, UAE, January 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.coling-main.208/.

Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. Clap learning audio concepts from natural language supervision. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630, June 2024. ISSN 1476-4687. doi: 10.1038/s41586-024-07421-0.

Marina Fomicheva, Shuo Sun, Lisa Yankovskaya, Frédéric Blain, Francisco Guzmán, Mark Fishel, Nikolaos Aletras, Vishrav Chaudhary, and Lucia Specia. Unsupervised quality estimation for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:539–555, 2020.

Gianni Franchi, Nacim Belkhir, Dat Nguyen Trong, Guoxuan Xia, and Andrea Pilzer. Towards understanding and quantifying uncertainty for text-to-image generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 8062–8072, 2025.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 24108–24118, 2025.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

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

Adityanand Guntuboyina and Bodhisattva Sen. Nonparametric shape-restricted regression. *Statistical Science*, 33(4):568–594, 2018.

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

Iryna Hartsock and Ghulam Rasool. Vision-language models for medical report generation and visual question answering: A review. *Frontiers in artificial intelligence*, 7:1430984, 2024.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: A reference-free evaluation metric for image captioning. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 7514–7528, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.595. URL https://aclanthology.org/2021.emnlp-main.595/.

Eyke Hüllermeier and Willem Waegeman. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine learning*, 110(3):457–506, 2021.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Chaoya Jiang, Haiyang Xu, Mengfan Dong, Jiaxing Chen, Wei Ye, Ming Yan, Qinghao Ye, Ji Zhang, Fei Huang, and Shikun Zhang. Hallucination Augmented Contrastive Learning for Multimodal Large Language Model, February 2024.

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *arXiv preprint arXiv:1705.03551*, 2017.

Zaid Khan and Yun Fu. Consistency and uncertainty: Identifying unreliable responses from black-box vision-language models for selective visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10854–10863, 2024a.

Zaid Khan and Yun Fu. Consistency and Uncertainty: Identifying Unreliable Responses From Black-Box Vision-Language Models for Selective Visual Question Answering, April 2024b.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic Uncertainty: Linguistic Invariances for Uncertainty Estimation in Natural Language Generation, April 2023.

Alex Kulesza, Ben Taskar, et al. Determinantal point processes for machine learning. *Foundations and Trends® in Machine Learning*, 5(2–3):123–286, 2012.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466, 2019.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*, 2023.

Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. A dataset of clinically generated visual questions and answers about radiology images. *Scientific data*, 5(1):1–10, 2018.

S. Lee, J. Youn, H. Kim, et al. Cxr-llava: a multimodal large language model for interpreting chest x-ray images. *European Radiology*, 2025. doi: 10.1007/s00330-024-11339-6. URL https://doi.org/10.1007/s00330-024-11339-6.

Chia-Hsuan Li, Szu-Lin Wu, Chi-Liang Liu, and Hung-yi Lee. Spoken squad: A study of mitigating the impact of speech recognition errors on listening comprehension. *arXiv preprint arXiv:1804.00320*, 2018.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation, 2022. URL https://arxiv.org/abs/2201.12086.

Linjie Li, Jie Lei, Zhe Gan, and Jingjing Liu. Adversarial vqa: A new benchmark for evaluating the robustness of vqa models. In *International Conference on Computer Vision (ICCV)*, 2021.

Qing Li, Chenyang Lyu, Jiahui Geng, Derui Zhu, Maxim Panov, and Fakhri Karray. Reference-free Hallucination Detection for Large Vision-Language Models, August 2024.

Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. Generating with confidence: Uncertainty quantification for black-box large language models. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=DWkJCSxKU5.

F. Liu, T. Zhu, X. Wu, et al. A medical multimodal large language model for future pandemics. *NPJ Digital Medicine*, 6:226, 2023a. doi: 10.1038/s41746-023-00952-2. URL https://doi.org/10.1038/s41746-023-00952-2.

Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. Aligning large multi-modal model with robust instruction tuning. *arXiv preprint arXiv:2306.14565*, 2023b.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023c.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26296–26306, 2024.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*, 2023.

Andrey Malinin and Mark Gales. Uncertainty estimation in autoregressive structured prediction. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=jN5y-zb5Q7m.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Alexander Nikitin, Jannik Kossen, Yarin Gal, and Pekka Marttinen. Kernel Language Entropy: Fine-grained Uncertainty Quantification for LLMs from Semantic Similarities, May 2024.

OpenAI. Gpt-4o mini: Advancing cost-efficient intelligence, July 2024. URL https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun

Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. Gpt-4 technical report, 2024. URL https://arxiv.org/abs/2303.08774.

Xin Qiu and Risto Miikkulainen. Semantic density: Uncertainty quantification for large language models through confidence measurement in semantic space. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=LOH6qzI7T6.

Siva Reddy, Danqi Chen, and Christopher D Manning. Coqa: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266, 2019.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019. URL http://arxiv.org/abs/1908.10084.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. *arXiv preprint arXiv:1809.02156*, 2018.

Suwon Shon, Siddhant Arora, Chyi-Jiunn Lin, Ankita Pasad, Felix Wu, Roshan Sharma, Wei-Lun Wu, Hung-yi Lee, Karen Livescu, and Shinji Watanabe. Slue phase-2: A benchmark suite of diverse spoken language understanding tasks. *arXiv preprint arXiv:2212.10525*, 2022.

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

D. Tian, S. Jiang, L. Zhang, X. Lu, and Y. Xu. The role of large language models in medical image processing: a narrative review. *Quantitative Imaging in Medicine and Surgery*, 14(1):1108–1121, 2024. doi: 10.21037/qims-23-892. URL https://doi.org/10.21037/qims-23-892.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D Manning. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 5433–5442, 2023.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9568–9578, 2024.

Joel A Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of computational mathematics*, 12(4):389–434, 2012.

Lei Wang, Jiabang He, Shenshen Li, Ning Liu, and Ee-Peng Lim. Mitigating fine-grained hallucination by fine-tuning large vision-language models with caption rewrites. In *International Conference on Multimedia Modeling*, pp. 32–45. Springer, 2024.

Hermann Weyl. The asymptotic distribution law for the eigenvalues of linear partial differential equations (with applications to the theory of black body radiation). *Math. Ann*, 71(1):441–479, 1912.

Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. NExT-GPT: Any-to-any multimodal LLM. In *Proceedings of the International Conference on Machine Learning*, pp. 53366–53397, 2024.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms, 2024a. URL https://arxiv.org/abs/2306.13063.

Miao Xiong, Andrea Santilli, Michael Kirchhof, Adam Golinski, and Sinead Williamson. Efficient and effective uncertainty quantification for llms. In *Neurips Safe Generative AI Workshop 2024*, 2024b.

Qifan Yu, Juncheng Li, Longhui Wei, Liang Pang, Wentao Ye, Bosheng Qin, Siliang Tang, Qi Tian, and Yueting Zhuang. Hallucidoctor: Mitigating hallucinatory toxicity in visual instruction data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12944–12953, 2024.

Zihao Yue, Liang Zhang, and Qin Jin. Less is more: Mitigating multimodal hallucination from an eos decision perspective. *arXiv preprint arXiv:2402.14545*, 2024.

Bohan Zhai, Shijia Yang, Chenfeng Xu, Sheng Shen, Kurt Keutzer, and Manling Li. Halle-switch: Controlling object hallucination in large vision language models. *arXiv e-prints*, pp. arXiv–2310, 2023.

Jun Zhan, Junqi Dai, Jiasheng Ye, Yunhua Zhou, Dong Zhang, Zhigeng Liu, Xin Zhang, Ruibin Yuan, Ge Zhang, Linyang Li, Hang Yan, Jie Fu, Tao Gui, Tianxiang Sun, Yu-Gang Jiang, and Xipeng Qiu. AnyGPT: Unified multimodal LLM with discrete sequence modeling. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9637–9662, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.521. URL https://aclanthology.org/2024.acl-long.521/.

Ruiyang Zhang, Hu Zhang, and Zhedong Zheng. Vl-uncertainty: Detecting hallucination in large vision-language model via uncertainty estimation, 2024a.

Yuanhan Zhang, Bo Li, haotian Liu, Yong jae Lee, Liangke Gui, Di Fu, Jiashi Feng, Ziwei Liu, and Chunyuan Li. Llava-next: A strong zero-shot video understanding model, April 2024b. URL https://llava-vl.github.io/blog/2024-04-30-llava-next-video/.

A THEORETICAL ANALYSIS AND INTUITION

In this section we providing additional details on the assumptions, propositions and proofs for Sec. 5. As mentioned in Sec. 5, the analysis below hold under simplifying assumptions and are not intended as formal guarantees for MLLM distributions in practice, but rather is meant to help provide intuition for UMPIRE and how its components relate to uncertainty and the desired desiderata as described in Sec. 2.

A.1 A.1 NOTATION AND ASSUMPTIONS

We first recap the problem setup and state the assumptions used in proofs for our propositions.

Setup. For a given task instance $t \in \mathcal{T}$, we sample k responses from the model. Let $\phi_i \in \mathbb{R}^d$ denote the embedding (row vector) for sample i, and $\Phi \in \mathbb{R}^{k\times d}$ with rows $\phi_1^\top, \ldots, \phi_k^\top$ be the semantic embedding matrix. Let $p_i \in [0,1]$ be the model-generated probability for each response, and the coherence matrix and scores are $C = \mathrm{diag}(c_1, \ldots, c_k)$, $c_i := \exp(\alpha(1-p_i))$ respectively, for scalar $\alpha \geq 0$. Our unnormalized UMPIRE score is

$$\tilde{V}_t := \log\det(C\Phi\Phi^\top C),$$

and we use its normalized form over the number of samples k:

$$\bar{V}_t := \frac{1}{k}\tilde{V}_t = \frac{1}{k}\log\det(\Phi\Phi^\top) + 2\alpha \cdot \frac{1}{k}\sum_{i=1}^{k}(1-p_i). \tag{5}$$

Assumptions. We make the following assumptions:

A1 (Bounded embeddings). There exists $B > 0$ such that $\|\phi_i\|_2 \leq B$ for all sampled responses i. This is typical for most MLLMs.

A2 (Approximate i.i.d. sampling). For a fixed instance t, the sampled responses (ϕ_i, p_i) are conditionally i.i.d. draws from the model's conditional response distribution, based on the multimodal task instance input queries. We write expectations over this sampling as $\mathbb{E}$.

A3 (Bounded incoherence variance). The incoherence variables $1-p_i \in [0,1]$, and $\mathrm{Var}(1-p_i) \leq \sigma_p^2 < \infty$.

A4 (Non-Degenerate Covariance). We assume the population covariance matrix of the embeddings, $\Sigma^\star = \mathbb{E}[\phi\phi^T]$, is strictly positive definite, i.e., its minimum eigenvalue $\lambda_{\min}(\Sigma^\star)$ is strictly greater than zero.

This assumption can be met without loss of generality. If the true covariance is singular, we can analyze a regularized matrix $\Sigma^\star + \epsilon I$ for an arbitrarily small $\epsilon > 0$, ensuring that all quantities in our proofs are well-defined. As mentioned in the main paper, we also implement this empirically for numerical stability in the computation of our metric. For notational simplicity, we will proceed using the standard notation ($\Sigma^\star$, $\bar{V}_t$, etc.) to refer to these potentially regularized quantities.

A5 (Clustered posterior structure for volume analysis). (Only used in proposition 2) The model's conditional response distribution for task instance t is well approximated by a finite mixture of m semantic clusters with weights w_j, cluster means μ_j and within-cluster covariances Σ^j. Let $\Sigma_{\text{within}} = \sum_j w_j\Sigma^j$ and $\Sigma_{\text{between}} = \sum_j w_j(\mu_j - \bar{\mu})(\mu_j - \bar{\mu})^\top$; then $\Sigma_{\text{mix}} = \Sigma_{\text{within}} + \Sigma_{\text{between}}$. Correctness concentrates in a subset of clusters (see discussion in App. A.6).

A6 (Informative internal model probabilities). The MLLM's internal generated probabilities for each response, while not necessarily calibrated, is positively correlated with the probability of the response being correct.

A.2 USEFUL LEMMAS

We state two lemmas used in the proofs: (i) a matrix Bernstein tail inequality (Tropp, 2012) in a form we use, and (ii) a common log-determinant perturbation bound via the integral representation.

Lemma 1 (Matrix Bernstein ((Tropp, 2012) Theorem 1.6)). *Let $X_1, \ldots, X_k$ be independent, mean-zero, symmetric random matrices in $\mathbb{R}^{d\times d}$ with $\|X_i\| \le R$ almost surely. Define the matrix variance parameter*

$$\sigma^2 := \left\| \sum_{i=1}^{k} \mathbb{E}[X_i^2] \right\|.$$

Then for all $t \ge 0$,

$$\Pr\left(\left\| \sum_{i=1}^{k} X_i \right\| \ge t \right) \le d \cdot \exp\left(-\frac{t^2/2}{\sigma^2 + Rt/3} \right).$$

Lemma 2 (Log-determinant perturbation). *Let $A \in \mathbb{R}^{d\times d}$ be symmetric positive definite and $\Delta \in \mathbb{R}^{d\times d}$ be symmetric. Suppose*

$$\|\Delta\|_2 < \lambda_{\min}(A),$$

so that $A + t\Delta \succ 0$ for every $t \in [0,1]$. Then

$$\log\det(A+\Delta) - \log\det(A) \;=\; \int_0^1 \operatorname{tr}\big((A+t\Delta)^{-1}\Delta\big)\, dt. \tag{6}$$

In particular, the following uniform bound holds:

$$\big|\log\det(A+\Delta) - \log\det(A)\big| \;\le\; \frac{d}{\lambda_{\min}(A) - \|\Delta\|_2}\,\|\Delta\|_2 \;\le\; \frac{2d}{\lambda_{\min}(A)}\,\|\Delta\|_2, \tag{7}$$

where the last inequality is valid whenever $\|\Delta\|_2 \le \lambda_{\min}(A)/2$.

Proof. Define the scalar function $f(t) := \log\det(A + t\Delta)$ for $t \in [0,1]$. Each $A + t\Delta$ is positive definite by the assumption $\|\Delta\|_2 < \lambda_{\min}(A)$, so f is differentiable on $[0,1]$. By Jacobi's formula (derivative of a determinant) or the standard matrix-differentiation identity,

$$f'(t) \;=\; \frac{d}{dt}\log\det(A+t\Delta) \;=\; \operatorname{tr}\big((A+t\Delta)^{-1}\Delta\big).$$

Integrating f' from 0 to 1 yields the integral identity equation 6.

To obtain the bound equation 7, we can take absolute values in equation 6 and use the elementary inequalities $|\operatorname{tr}(M)| \le d\|M\|_2$ and $\|XY\|_2 \le \|X\|_2\|Y\|_2$:

$$\begin{aligned}\big|\log\det(A+\Delta) - \log\det(A)\big| &\le \int_0^1 \big|\operatorname{tr}((A+t\Delta)^{-1}\Delta)\big|\, dt \\ &\le \int_0^1 d\,\|(A+t\Delta)^{-1}\Delta\|_2\, dt \le d\,\|\Delta\|_2 \int_0^1 \|(A+t\Delta)^{-1}\|_2\, dt.\end{aligned}$$

For each $t \in [0,1]$ we have $\lambda_{\min}(A+t\Delta) \ge \lambda_{\min}(A) - t\|\Delta\|_2$ (by Weyl's inequality or elementary eigenvalue perturbation), so

$$\|(A+t\Delta)^{-1}\|_2 \;=\; \frac{1}{\lambda_{\min}(A+t\Delta)} \;\le\; \frac{1}{\lambda_{\min}(A) - t\|\Delta\|_2} \;\le\; \frac{1}{\lambda_{\min}(A) - \|\Delta\|_2}.$$

Finally, substituting this bound into the integral gives

$$\big|\log\det(A+\Delta) - \log\det(A)\big| \le \frac{d\,\|\Delta\|_2}{\lambda_{\min}(A) - \|\Delta\|_2},$$

which is the first inequality in equation 7. If $\|\Delta\|_2 \le \lambda_{\min}(A)/2$ then $\lambda_{\min}(A) - \|\Delta\|_2 \ge \lambda_{\min}(A)/2$, and the second displayed inequality follows. □

A.2 Propositions and proofs

We now state and prove the formal propositions that substantiate the main-text claims. Each proposition includes a brief discussion of where assumptions are used and what the practical implications are.

A.2.1 PROPOSITION 1 (COHERENCE-ADJUSTED SEMANTIC VOLUME DECOMPOSITION AND MONTE-CARLO COHERENCE TERM)

The coherence-adjusted semantic volume can be decomposed two the two terms below:

$$\bar{V}_t = \frac{1}{k}\log\det(C\Phi\Phi^\top C) = \frac{1}{k}\log\det(\Phi\Phi^\top) + 2\alpha\cdot\frac{1}{k}\sum_{i=1}^{k}(1-p_i).$$

Furthermore, under A2-A3, the coherence term $2\alpha\cdot\frac{1}{k}\sum_{i=1}^{k}(1-p_i)$ concentrates about its population mean at an $O(1/\sqrt{k})$ rate: precisely, by Hoeffding there exist constants c_1, C_1 depending on α and the incoherence range such that for any $\eta > 0$,

$$\Pr\left(\left|2\alpha\cdot\frac{1}{k}\sum_{i=1}^{k}(1-p_i) - 2\alpha\,\mathbb{E}[1-p]\right| > \eta\right) \le C_1\exp(-c_1 k\eta^2).$$

Proof. As C and $\Phi\Phi^\top$ are $k\times k$ square matrices, we have

$$\det(C\Phi\Phi^\top C) = \det(C)\det(\Phi\Phi^\top)\det(C) = \det(\Phi\Phi^\top)\det(C)^2,$$

hence

$$\log\det(C\Phi\Phi^\top C) = \log\det(\Phi\Phi^\top) + 2\log\det(C).$$

Substituting C which is a diagonal matrix with $c_i = \exp(\alpha(1-p_i))$, we have

$$2\log\det(C) = 2\sum_{i=1}^{k}\log c_i = 2\alpha\sum_{i=1}^{k}(1-p_i).$$

We can obtain the final claimed decomposition by dividing throughout by k. The concentration result follows from standard Hoeffding bounds for bounded scalar i.i.d. variables $1-p_i$ (Assumptions A2-A3). □

A.2.2 PROPOSITION 2 (SEMANTIC VOLUME)

As stated in assumption A5, Let the MLLM's response embedding distribution be a finite mixture model with total covariance $\Sigma_{\text{mix}} = \Sigma_{\text{within}} + \Sigma_{\text{between}}$. If the between-cluster covariance increases such that $\Sigma'_{\text{between}} \succeq \Sigma_{\text{between}}$ in the positive semidefinite (PSD) order, let the new total covariance be $\Sigma'_{\text{mix}} = \Sigma_{\text{within}} + \Sigma'_{\text{between}}$. Then, the determinant is non-decreasing: $\det(\Sigma'_{\text{mix}}) \ge \det(\Sigma_{\text{mix}})$.

Consequently, for large k where sample covariance $S_k = \frac{1}{k}\Phi^T\Phi \approx \Sigma_{\text{mix}}$, $\det(S_k)$, or the unadjusted semantic volume term in our proposed metric, increases with Σ_{between}. When correctness probability concentrates in a subset of clusters, increasing Σ_{between} (more spread / more separated modes) implies a decrease in $\Pr(\text{correct}|t)$.

Proof. Given $\Sigma_{\text{mix}} = \Sigma_{\text{within}} + \Sigma_{\text{between}}$, and suppose that between-cluster spread increases such that the new covariance $\Sigma'_{\text{between}} \succeq \Sigma_{\text{between}}$, which by definition means the difference matrix D is positive semidefinite. Then $\Sigma'_{\text{mix}} = \Sigma_{\text{mix}} + D$, and by Weyl's inequality (Weyl, 1912), $\lambda_i(\Sigma_{\text{mix}} + D) \ge \lambda_i(\Sigma_{\text{mix}})\forall i$ where i is the i-th largest eigenvalue. Since the determinant of a matrix is the product of its eigenvalues, we have that $\det(\Sigma'_{\text{mix}}) \ge \det(\Sigma_{\text{mix}})$.

For finite k, let $S_k := \frac{1}{k}\Phi^\top\Phi$ be the empirical per-sample second moment. Under A2 and A1, S_k concentrates about Σ_{mix} (matrix Bernstein; see Proposition A.3 below). Hence with high probability for large k, $\det(S_k)$ is close to $\det(\Sigma_{\text{mix}})$ and inherits its monotonicity under positive semidefinite increases of Σ_{between}.

Finally, consider the assumption that correctness is concentrated in a subset of the posterior's semantic clusters (Assumption A5). Let the set of cluster indices J be partitioned into J_{correct} and $J_{\text{incorrect}}$, and let $q_j = p(\text{correct}|\text{cluster } j)$ be the correctness probability for cluster j. The total correctness probability is then the weighted average $p(\text{correct}|t) = \sum_{j\in J} w_j q_j$, where w_j is the posterior probability mass on cluster j.

An MLLM's increase in uncertainty involves a transfer of posterior probability mass from the high-correctness clusters to the low-correctness clusters, thereby decreasing the total weight $\sum_{j\in J_{\text{correct}}} w_j$.

This causes (a) weight to be moved from high-q_j terms to low-q_j terms, causing the value of the weighted average $p(\text{correct}|t)$ to decrease, and (b) the between-cluster covariance matrix, $\Sigma_{\text{between}} = \sum_j w_j(\mu_j - \bar{\mu})(\mu_j - \bar{\mu})^T$, to increase in the PSD order. A posterior concentrated on a few modes has a smaller second moment than one spread more uniformly across many modes.

Since an increase in Σ_{between} in the PSD order leads to a non-decreasing determinant of the total covariance approximated by our proposed semantic volume term, an increase in model uncertainty corresponds to a non-decreasing semantic volume and a non-increasing probability of correctness, illustrating the relationship between the volume term and the expected error. □

A.2.3 Proposition 3 (Concentration and ranking consistency)

Under assumptions A1-A4, there exist constants c, C such that for any $\eta > 0$, the score concentrates around its mean with sub-exponential tails in k:

$$\Pr\left(|\bar{V}_t - \mathbb{E}\bar{V}_t| > \eta\right) \leq C\exp(-ck\eta^2).$$

Moreover, if two instances t_a and t_b satisfy a true mean gap $\mathbb{E}\bar{V}_{t_a} - \mathbb{E}\bar{V}_{t_b} \geq \Delta > 0$, then the empirical ordering $\bar{V}_{t_a} > \bar{V}_{t_b}$ holds with probability at least $1 - 2C\exp(-ck(\Delta/2)^2)$.

Proof. We will compute bounds for each of the two decomposed terms (semantic volume and coherence scores) in our proposed metric separately, before combining them.

(i) Coherence term concentration. As already shown in App. A.2.1, there exist constants $c_1, C_1 > 0$ (depending on α and the incoherence range) such that for any $\eta > 0$,

$$\Pr\left(\left|2\alpha \cdot \frac{1}{k}\sum_{i=1}^{k}(1 - p_i) - 2\alpha\mathbb{E}[1 - p]\right| > \eta\right) \leq C_1\exp(-c_1 k\eta^2).$$

(ii) Volume term concentration. Consider $S_k = \frac{1}{k}\sum_{i=1}^{k}\phi_i\phi_i^\top$ and $\Sigma^\star = \mathbb{E}[\phi\phi^\top]$. Note that our semantic volume term

$$\frac{1}{k}\log\det(\Phi\Phi^\top) = \frac{1}{k}\log\det(kS_k) = \frac{\log k}{k} + \frac{1}{k}\log\det(S_k).$$

has an additional $\frac{\log k}{k}$ term that vanishes as $k \to \infty$, so it suffices to consider $\frac{1}{k}\log\det(S_k)$ in our analysis.

We define centered symmetric random matrices:

$$Y_i := \phi_i\phi_i^\top - \Sigma^\star, \quad i = 1, \ldots, k,$$

so $S_k - \Sigma^\star = \frac{1}{k}\sum_{i=1}^{k} Y_i$ and $\mathbb{E}[Y_i] = 0$. From A1, we have that $\|\phi_i\phi_i^\top\| \leq \|\phi_i\|_2^2 \leq B^2$, and hence $\|Y_i\| \leq B^2 + \|\Sigma^\star\| \leq 2B^2$ (we absorbed $\|\Sigma^\star\| \leq B^2$ for simplicity). We can then apply Lemma 1, the matrix Bernstein inequality: there are constants $c_2, C_2 > 0$ such that for all $\delta > 0$,

$$p(\|S_k - \Sigma^\star\| \geq \delta) \leq C_2 d\exp(-c_2 k\delta^2). \tag{8}$$

We now translate the matrix norm bound to log-determinant bounds for our metric. Assumption A4 states that $\Sigma^\star$ is positive definite with $\lambda_{\min}(\Sigma^\star) > 0$. For clarity and to cover the practical case where $\Sigma^\star$ may be nearly singular, we denote regularized matrices

$$S_k(\varepsilon) := S_k + \varepsilon I, \quad \Sigma^\star(\varepsilon) := \Sigma^\star + \varepsilon I,$$

where a small $\varepsilon > 0$ is chosen. Then $\lambda_{\min}(\Sigma^\star(\varepsilon)) = \lambda_{\min}(\Sigma^\star) + \varepsilon =: \lambda_0 > 0$. ε could be treated as for example the empirical jitter used in numerical computation.

Using the integral representation for log-determinant difference (Lemma 2), for positive definite A and symmetric Δ with $\|\Delta\|_{\text{spec}} < \lambda_{\min}(A)$ we have

$$\log\det(A + \Delta) - \log\det(A) = \int_0^1 \operatorname{tr}((A + t\Delta)^{-1}\Delta)\, dt.$$

With this identity, we can get the bound

$$|\log\det(A+\Delta) - \log\det(A)| \le \frac{d\,\|\Delta\|_{\text{spec}}}{\lambda_{\min}(A) - \|\Delta\|_{\text{spec}}}. \tag{9}$$

By applying Eq. (9) with $A = \Sigma^\star(\varepsilon)$ and $\Delta = S_k(\varepsilon) - \Sigma^\star(\varepsilon) = S_k - \Sigma^\star$, on the high-probability event $\|S_k - \Sigma^\star\|_{\text{spec}} \le \delta$ and when $\delta < \lambda_0/2$ we obtain

$$|\log\det(S_k(\varepsilon)) - \log\det(\Sigma^\star(\varepsilon))| \le \frac{2d}{\lambda_0}\,\|S_k - \Sigma^\star\|_{\text{spec}}.$$

This converts Eq. (8) into a log-determinant difference: there exist constants (re-labelled) $c_3, C_3 > 0$ such that for any $\eta > 0$

$$\Pr\left(\left|\frac{1}{k}\log\det(S_k(\varepsilon)) - \frac{1}{k}\log\det(\Sigma^\star(\varepsilon))\right| > \eta\right) \le C_3\, d\, \exp\left(-c_3 k \left(\frac{\eta\lambda_0}{2d}\right)^2\right).$$

Absorbing factors of d, λ_0 into constants gives a sub-exponential-in-k tail of the same qualitative form as before.

Hence, given additive $\frac{\log k}{k}$ term that vanishes and the regularized bound above, our semantic volume term $\frac{1}{k}\log\det(\Phi\Phi^\top)$ concentrates around its expectation at the above rate.

(iii) Combining with union bound Finally, we combine with union bound the concentration events for both the coherence and semantic volume term. There exist constants $c, C > 0$ depending on B, d, λ_0, α such that for any $\eta > 0$:

$$\Pr\left(\left|\bar{V}_t - \mathbb{E}\bar{V}_t\right| > \eta\right) \le C\exp(-ck\eta^2).$$

To consider desiderata **R1** we will additionally need to analyze the bounds for ranking. Given two task instances t_a and t_b that satisfy $\mathbb{E}\bar{V}_{t_a} - \mathbb{E}\bar{V}_{t_b} \ge \Delta > 0$, then the event that $\bar{V}_{t_a} \le \bar{V}_{t_b}$ implies at least either $|\bar{V}_{t_a} - \mathbb{E}\bar{V}_{t_a}|$ or $|\bar{V}_{t_b} - \mathbb{E}\bar{V}_{t_b}|$ exceeds $\Delta/2$. Applying the concentration bound to both instances and the union bound yields the desired bound on the probability that the metrics will be misordered:

$$\Pr(\bar{V}_{t_a} \le \bar{V}_{t_b}) \le 2C\exp\left(-ck(\Delta/2)^2\right).$$

□

As mentioned in assumption A4, where we consider regularized matrices, and on the high-probability event $\|S_k - \Sigma^\star\| \le \delta$ with $\delta \le \lambda_0/2$ (using A4 with $\lambda_0 = \lambda_{\min}(\Sigma^\star_\varepsilon)$), by Lemma 2 with $A = \Sigma^\star_\varepsilon$ and $\Delta = S^\varepsilon_k - \Sigma^\star_\varepsilon$ we get

$$|\log\det(S^\varepsilon_k) - \log\det(\Sigma^\star_\varepsilon)| \le \frac{d}{\lambda_0 - \|\Delta\|}\|\Delta\| \le \frac{2d}{\lambda_0}\|\Delta\|.$$

Since $\|\Delta\| = \|S_k - \Sigma^\star\| \le \delta$, we obtain

$$\Pr\left(|\log\det(S^\varepsilon_k) - \log\det(\Sigma^\star_\varepsilon)| \ge \eta\right) \le C_2 d\exp\left(-c_2 k(\eta\lambda_0/(2d))^2\right),$$

rearranging $\eta \propto d\delta/\lambda_0$. Because $\frac{1}{k}\log\det(\Phi\Phi^\top + k\varepsilon I) = \frac{1}{k}\log k + \frac{1}{k}\log\det(S^\varepsilon_k)$ and $\frac{1}{k}\log k \to 0$ we get the same exponential-in-k concentration rate (constants adjusted) for the normalized volume term $\frac{1}{k}\log\det(\Phi\Phi^\top + k\varepsilon I)$.

(iii) Combine term tails. Union-bounding the coherence term tail and volume term tail yields the stated exponential concentration for $\bar{V}^\varepsilon_t$ with constants c, C depending on B, d, λ_0, α.

(iv) Ranking bound. Suppose two instances t_a, t_b satisfy $\mathbb{E}\bar{V}^\varepsilon_{t_a} - \mathbb{E}\bar{V}^\varepsilon_{t_b} \ge \Delta$. The event that $\bar{V}^\varepsilon_{t_a} \le \bar{V}^\varepsilon_{t_b}$ implies either $|\bar{V}^\varepsilon_{t_a} - \mathbb{E}\bar{V}^\varepsilon_{t_a}| \ge \Delta/2$ or $|\bar{V}^\varepsilon_{t_b} - \mathbb{E}\bar{V}^\varepsilon_{t_b}| \ge \Delta/2$. Applying the concentration bound and union bound yields the exponential bound on misordering probability claimed. □

PROPOSITION A.4 (CALIBRATION UNDER MONOTONICITY)

If the function $u \mapsto \Pr(\text{correct} \mid \bar{V}_t = u)$ is monotone, then isotonic regression fitted on a development set of size n yields a consistent calibrated estimator of the true correctness probability.

Standard non-parametric worst-case convergence rates apply (e.g., $O(n^{-2/3})$ for the mean squared error).

This is a standard result in isotonic regression and shape-constrained estimation (Barlow, 1972; Chatterjee et al., 2015; Guntuboyina & Sen, 2018). We refer readers to the cited literature for full technical statements and proofs.

B Discussion

B.1 Assessing multimodal query input coherence (**R4**)

For a metric to satisfy **R4**, it should consider the coherence of each sampled response with respect to the multimodal task instance query, rather than just a single modality (e.g., text). We design an experimental setting on image-text modalities to assess this by computing uncertainty metrics based on (1) both image and text portions of the query ($t = (I_t, q_t)$), (2) image I_t with additive noise $\mathcal{N}(0, 0.5^2)$, (3) I_t is entirely black, and (4) no image, i.e only the text portion of the query (only q_t) with the MLLM text response $\hat{y}_t$. A metric that satisfies **R4** should perform significantly better under (1), while a metric that does not will produce similar performance regardless of (1)-(4).

Specifically, for (2)-(4), after the MLLM has generated responses $\hat{y}_t$ based on (I_t, q_t), we recompute the various metrics LN-Ent, Sem.Ent, Eigen, and UMPIRE based on the query-answer pair lower quality to no image, e.g., based on recomputing the response logits and embedding vectors of text-only query-answer pairs $[q_t, \hat{y}_t]$, on a subset of the VQAv2 validation set. In Fig. 2, we observe that LN-Ent and UMPIRE, and to a smaller extent Sem.Ent, are sensitive to the lack of multi-modality information, with their performance increasing once the image queries are provided during the computation of the metrics. On the other hand, Eigen is insensitive to whether the image query is provided or not. This may be because Eigen measures only the diversity of responses through the covariance matrix of text response sentence embeddings across multiple generations, which is not affected by the image query bias. On the contrary, logit signals are more sensitive to the coherence of the multimodal input query and the generated response, hence metrics that use some form of that such as LN-Ent and UMPIRE can better satisfy **R4**.

B.2 Comparisons with Eigenscore

As mentioned in App. C.1, the Eigen (Chen et al., 2024) metric involves computing the log determinant of the covariance matrix of sampled sentence embeddings. At first glance, this metric may seem similar to that of UMPIRE. However, there are key differences that lead to Eigen consistently underperforming our proposed UMPIRE metric, as can be seen in both the MLLM (Sec. 6) and LLM case (App. C.7), and Eigen could be interpreted as a special case of UMPIRE.

A major distinction, among others, is that Chen et al. (2024) analyzed only the LLM setting, and proposed Eigen by considering the differential entropy of sentence embeddings, assuming that the embeddings form a multivariate Gaussian distribution – this motivated the log determinant term of the metric, which bears similarity to UMPIRE. However, our UMPIRE framework considers the more general MLLM setting, and adopts a different approach inspired by the quality-diversity kernel decomposition of determinantal point processes (DPP), which naturally factors the incoherence scores when computing the UMPIRE metric to adjust the semantic volume enclosed by the responses' semantic embeddings. This inclusion of the incoherence scores help (1) satisfy **R4**, as we can see in App. B.1 that Eigen does not, and (2) significantly improve metric performance (App. D.2). Eigen could possibly be interpreted as the special case of UMPIRE where all responses have incoherence scores of 1 (i.e., the model-generated probabilities of all responses $p_i = 1 \; \forall i$), or when the hyperparameter α is set to 0. Note that the incoherence scores also boost performance in the LLM setting (App. C.7), indicating that while incoherence scores help in addressing App. B.1, its weighting of different responses in the computation of UMPIRE also helps in single modality settings.

B.3 Additional discussion on the effectiveness desiderata

In this section, we provide further discussion on the various effectiveness desiderata, such as the differences and relevance of **R1**, **R2a** and **R2b**. For ease of discussion, we focus on comparing **R1** and **R2b**, which is a stricter form of **R2a**.

The classification desiderata **R1** and the calibration desiderata **R2b** are primarily motivated by different considerations. In the former, we are concerned about classifying whether a task instance t will be answered correctly or not by the MLLM. As represented in Eq. (1), for this desiderata the metric should be able to successfully rank the task instances that the MLLM will get wrong higher than those that it will get correct, which can be evaluated by the AUROC of the metric. Such evaluations are used in many MLLM and LLM uncertainty quantification works (Farquhar et al., 2024; Malinin & Gales, 2021; Chen et al., 2024; Xiong et al., 2024a) to assess the performance of their metrics. While useful, note that the desiderata does not consider a quantitative, continuous measure of the uncertainty associated with each task response, since classification of correct/wrong responses is a binary task.

However, in the latter, we are concerned about providing an accurate, calibrated estimate of whether the MLLM will get a task instance correct, conditional on the uncertainty metric (as in Eq. (3)), which can be evaluated via the expected calibration error (ECE). Note that in this scenario, we are not concerned about classifying whether a task instance will be answered correctly (**R1**), but instead are focused on being accurate about the *probability* that a task instance will be answered correctly given an associated metric value.

To illustrate the difference, consider an extreme example where an MLLM will definitely get 50% of the task instances correct, and the rest wrong. The vacuous metric that assigns the same uncertainty score to all task instances might satisfy **R2b** since it will output the average accuracy, 0.5, as the score for all task instances. This metric would violate **R1** and fail to classify the correct from wrong task instances. Instead, a better metric might strive to assign 1 to all task instances that can be answered correctly and 0 to the rest, satisfying both **R1** and **R2b**.

In practice, we would likely not have perfect information prior to evaluation on whether a task instance will be correct or wrong. That is why for **R1** the goal is only for the metric to get as close to 1 as possible, as the best possible AUROC would depend on the model and task. However, given two metrics that can achieve the same AUROC, a poor metric might only obtain the right relative ranking of task instances, while a good metric would not only achieve the same AUROC but also provide calibrated probabilities on how likely a task instance would be answered correctly or not. Hence, both the **R1** and **R2b** should be considered when evaluating uncertainty metrics, as we described in Sec. 2. In the absence of a small development set of unlabeled task instances before deployment, metrics satisfying **R2a** would at least provide interpretable relative information regarding how likely a task instance would be answered correctly compared to another.

C Experimental settings and other results

C.1 Benchmarks

C.1.1 Datasets

For our experiments, we utilize a diverse set of general multi-modality question-answering baseline datasets to ensure a comprehensive evaluation across different scenarios. Specifically, for image-text understanding, we use VQAv2 (Goyal et al., 2017), OKVQA (Marino et al., 2019), and AdVQA (Li et al., 2021), which include challenging cases such as out-of-distribution and adversarial settings. Besides, we also try to use the domain-specific visual QA datasets, including VQA-RAD, a dataset of question-answer pairs on radiology images, and MathVista, a consolidated Mathematical reasoning baseline within visual contexts. We evaluate our method using the first 15,000 samples from the validation split of VQAv2, along with the full validation sets of OKVQA (5,000 samples) and AdVQA (10,000 samples), the test split of VQA-RAD (Lau et al., 2018) (450 samples), and MathVista (Lu et al., 2023) (testmini split - 1,000 samples).

These datasets provide a robust test bed for assessing the effectiveness of our approach across different types of visual QA tasks. Besides image-text understanding, we also show the effectiveness of various

uncertainty metrics on audio-text and video-text understanding via audio QA datasets, including the test split of SLUE-P2-SQA5 (Shon et al., 2022), Spoken SQuAD (Li et al., 2018), and video QA datasets with Video-MME-short (Fu et al., 2025) (we convert multi-choice answers into free-form answers by taking the correct choice). For text-only datasets, we have CoQA (Reddy et al., 2019), TriviaQA, (Joshi et al., 2017), and NQ (Kwiatkowski et al., 2019).

C.1.2 BASELINES

The details of each baseline are as follows.

- **Neighborhood Consistency (NC) (Khan & Fu, 2024a).** This method tries to examine the reliability of the model via the consistency of the model's responses over the visual rephrased questions generated by a small proxy Visual Question Generation (VQG) model. We implement this method by training `BLIP` (Li et al., 2022) as the VQG model with its default setting. To ensure a fair comparison, we use `Llava-v1.5-13b` as the VQA model, aligning with the model used in our experiments.
- **Length-normalized Entropy (LN-Entropy) Malinin & Gales (2021).** This approach normalizes the joint log-probability of each sequence by dividing it by the sequence length and is proposed by Malinin & Gales (2021) for uncertainty quantification in LLM. Following Kuhn et al. (2023), we also apply multinomial sampling instead of using an ensemble of models.
- **Semantic Entropy Kuhn et al. (2023).** This method introduces a concept of semantic entropy, which measures the uncertainty over different meanings. Following their algorithms, we try to cluster the generated sequences by `Deberta` as the text entailment model and then compute the entropy based on these clusters.
- **EigenScore Chen et al. (2024).** We follow their default settings and compute the log determinant of the covariance matrix by Eigenvalues via Singular Value Decomposition (SVD), with the exception of the jitter term value – we found that using a jitter term of 10^{-8} rather than their default setting of 10^{-3} improves their performance, hence we applied that and reported the improved performance.
- **Verbalized Confidence Xiong et al. (2024a).** This method is applied specifically to blackbox models where we instruct it to provide a measure of its own confidence. For a single instance, we sample generations k times and return the most frequent answer along with the average reported confidence by the model.
- **Image generation UQ methods** . We implement PUNC Franchi et al. (2025) as the image generation uncertainty metric. This approach tries to generate the caption from the generated image and compute the text similarity between the new generated caption and the input caption through text similarity metrics such as ROUGE or BertScore. We use their default settings with the `Llava-v1.5-13b` as the caption generation Vision-Language model and ROUGE as the text similarity metric.

C.1.3 EXPERIMENTAL SETTINGS

- **Models and parameters.** We primarily use `Llava-v1.5-13b` as our image-text MLLM, with further analysis on other models provided in App. D.6, `Phi-4` as audio-text MLLM, and `LLaVA-NeXT-Video-7b-hf` as video-text ones. Following past work Kuhn et al. (2023), for each task instance t, the MLLM generates the most likely answer using a low-temperature setting ($T = 0.01$) and we use this answer $\hat{y}_t$ to evaluate the correctness of the model when answering this pair. For the computation of the various uncertainty metrics that require multiple samples, we apply Monte Carlo sampling to generate n samples from the MLLM using $T = 1$ and top_p $= 0.9$. In the main paper, we use the number of generated samples $n = 50$, and ablation results on the impact of this hyperparameter are presented and discussed in App. D.3.
- **Evaluation.** We use ROUGE-L and exact match as the evaluation functions $a(\mathcal{M}, t^*)$, given the model answer $\hat{y}_t$ and ground truth answer y_t^*, to assess the model performance. In the main paper, we report results using exact match, while additional results with ROUGE-L with varying parameters can be found in App. D.4.

- **Blackbox APIs.** For OpenAI's GPT models, we used $n = 50$ generations per prompt. For Anthropic's Claude 3.5 Haiku model, we used the same model parameters as specified above but a smaller number of generations $n = 20$ due to limitations on API credits.
- **Image/Audio Generation settings.** In this setting, we use NExT-GPT and AnyGPT models as the image/audio generation models, and their default configurations. To make it consistent with MLLM understanding tasks, we also use low-temperature generation for computing CLIP score, and multi-sampling with $n = 10$, temperature $T = 1$, and top_p $= 0.9$, We conduct experiments on 500 task instances of the MS-COCO caption validation set (Chen et al., 2015) for the image generation task, and the full Audiocap test set for the audio generation task. We compare UMPIRE to image generation uncertainty quantification method PUNC (Franchi et al., 2025). To evaluate the performance of multimodal generation uncertainty methods, instead of introducing the in-distribution and out-of-distribution datasets and trying to let uncertainty metrics classify these sets as in Franchi et al. (2025), we show that these uncertainty metrics satisfy R2a by computing the Pearson Correlation between the uncertainty metric and the quality scores, including continuous CLIP score (Hessel et al., 2021), or CLAP score (Elizalde et al., 2023) for image or audio generation tasks, respectively. These quality scores compute the similarity between the generated image/audio and its corresponding input caption from a real image/audio.

C.1.4 PROMPTS

Following Liu et al. (2023c), we use the following prompt for all baseline tasks:

```
 <modality>.  Answer this question in a word or a phrase.
{question}
```

The prompt used to elicit verbalized confidence from the blackbox API models are slightly different, such that they output their confidence in the answer along with the response. In accordance with Xiong et al. (2024a), we use the following prompt to extract verbalized confidence:

```
 <modality>.  Read the question, provide your answer, and your
confidence in this answer.  Note:  The confidence indicates how
likely you think your answer is true.  Use the following format
to answer:  "'Answer and Confidence (0-100):  [ONLY a word or a
phrase; not a complete sentence], [Your confidence level, please
only include the numerical number in the range of 0-100]%"' Only
give me the reply according to this format, don't give me any
other words.  Now, please answer this question and provide your
confidence level.  Question:  {question}
```

C.2 ADDITIONAL STATE-OF-THE-ART LLM UQ BASELINES

We have conducted experiments to evaluate UMPIRE against some other LLM UQ baselines, including Kernel Language Entropy (Nikitin et al., 2024), Degree (Lin et al., 2024), and Semantic Density (Qiu & Miikkulainen, 2024) on OKVQA and VQAv2, and AdVQA. Table 9 shows the performance across these datasets, demonstrating the robustness of our method, which achieves the highest AUROC (0.808), lowest ECE (0.039), and highest AURAC (0.861), with near-top CPC (0.964). UMPIRE outperforms these baselines in general over the effectiveness desiderata. While these three baselines similarly rely on multiple response samples to compute their uncertainty metrics, they are formulated based on premises and design considerations that are different from UMPIRE, leading to distinct algorithms and hence different performances. Degree and KLE extended Semantic Entropy (Kuhn et al., 2023) by computing semantic graphs derived from NLI models and deriving uncertainty metrics from them (e.g. via spectral graph analysis, graph kernels), while Semantic Density focuses on response-wise metrics, building on the intuition that a response that is semantically closer to more highly probable samples should be more trustworthy by estimating a semantic kernel density based on output from NLI models and response token probabilities. UMPIRE adopts a different approach where, inspired by DPPs, directly computes the coherence-adjusted (quality term in DPP literature) semantic volume enclosed by responses in the MLLM's embedding space, without help from external tools like NLI.

C.3 Evaluating TPR under FPR Constraints

Metric	Method	Image					Audio		Video	Avg
		VQAv2	OKVQA	AdVQA	MathVista	VQA-RAD	SLUE-P2 SQA5	Spoken SQuAD	Video-MME short	
TPR@10% FPR ↑	NC	0.362	0.095	0.189	0.408	0.189	-	-	-	0.249
	LN-Ent	0.282	0.244	0.168	0.347	0.127	0.299	0.248	0.287	0.250
	Sem.Ent	0.574	0.327	0.419	0.437	0.511	**0.557**	0.464	0.311	0.450
	Eigen	0.602	0.340	0.466	0.483	**0.601**	0.443	0.435	0.534	0.488
	Ours	**0.629**	**0.369**	**0.477**	**0.497**	0.587	0.522	**0.490**	**0.539**	**0.514**
TPR@1% FPR ↑	NC	0.049	0.008	0.019	0.030	0.023	-	-	-	0.026
	LN-Ent	0.057	0.030	0.066	0.075	0.065	0.025	0.023	0.022	0.045
	Sem.Ent	0.177	0.057	0.125	**0.136**	0.286	0.095	**0.169**	0.054	0.137
	Eigen	0.215	0.074	0.171	0.086	0.304	0.134	0.118	0.274	0.172
	Ours	**0.230**	**0.091**	**0.185**	0.131	**0.326**	**0.154**	0.162	**0.287**	**0.196**

Table 6: True Positive Rates (TPR) at different False Positive Rate (FPR) thresholds for various uncertainty quantification methods across multimodal tasks. Results are reported at FPR levels of 10% and 1%.

In addition to the AUROC metric reported in Sec. 6.1, we also provide results on the True Positive Rate (TPR) achievable for a given False Positive Rate (FPR), which users might have different minimum requirements for based on their application. As shown in Table 6, we provide True Positive Rate (TPR) at 10% and 1% FPR levels, and the results generally align with AUROC trends reported in the main text, where UMPIRE consistently performs well compared to baselines across datasets.

C.4 Plots for Calibration **R2a**

To better visualize the performance of the various metrics for proportionality **R2a**, we plot the error rate ($\mathbb{P}[a(\mathcal{M}, t^*) = 0]$) v.s. uncertainty score u on the AdVQA validation set in Fig. 4. UMPIRE manages to achieve the strongest linear correlation with error rate compared to all other metrics. This satisfies the desiderata of **R2a**.

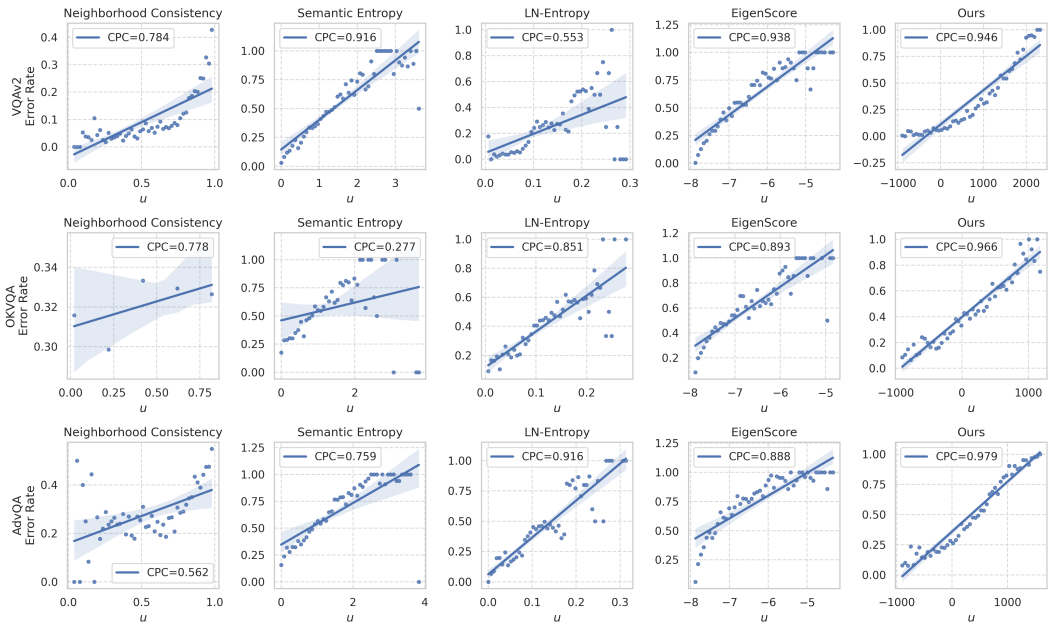

Figure 4: Pearson correlation plots of the uncertainty scores u on the VQAv2, OKVQA, and AdVQA, which demonstrate UMPIRE's strong correlation compared to other metrics.

C.5 SINGLE SAMPLING METHOD

We have also run experiments on basic uncertainty metrics that use only a single MLLM response, rather than a sampled set of MLLM responses. We ran the single-sample methods listed in Xiong et al. (2024b): Sequence Probability, Mean Token Entropy (Fomicheva et al., 2020), and Perplexity.

Table 10 shows these methods' results for the various MLLM datasets, along with UMPIRE, based on five response generations. Note that while the single-sampled methods may be cheaper to compute, they also produce significantly worse performance results compared to UMPIRE with $k = 5$. The appropriate metric to use would depend on the application requirements. For settings that require better uncertainty metric performance, UMPIRE would likely be a good choice especially since accelerated batched response generation (Kwon et al., 2023) is fast and typically not a computational resource bottleneck, while single-sample methods may be more suitable for very time-sensitive applications.

C.6 STATISTICAL T-TEST ON AURAC

To determine whether the difference in AURAC between UMPIRE and other baselines is statistically significant, we performed t-tests and Wilcoxon signed-rank tests (non-parametric) to ensure the results agree.

- H_0: $X_{UMPIRE} - X_{other} = 0$
- H_1: $X_{UMPIRE} - X_{other} > 0$

where X_{UMPIRE} is the set of AURAC scores from UMPIRE and X_{other} is the set of AURAC scores from baseline methods, including length-normalized entropy. and EigenScore.

Test	LN-Ent.	EigenScore	SE
t-test pvalue	0.000413	0.00791	0.0000342
Wilcoxon test pvalue	0.00390	0.00391	0.00390

Table 7: Statistical tests calculated based on 9 datapoints for the 9 model-dataset combinations (i.e. Models: Llava-13B, Llava-7B and Mllama-11B with datasets: VQAv2, AdVQA and OKVQA). Each data and model combination is treated as a single data point. The p-value of both tests are $\ll 0.01$, thus there is sufficient evidence at the 1% level of significance to conclude that $X_{UMPIRE} - X_{other} > 0$.

The results in Table 7 show that the AURAC between UMPIRE and other baselines are statistically significant. However, the hypothesis tests were each performed with only 9 data points, which may not be sufficient. Thus, we treat each model and dataset as a population, and perform sampling with replacement instead (sample size 1000 of 100 observations each). The differences are approximately normally distributed according to Central Limit Theorem, which would allow some robustness to non-normality. The results are shown in Table 8, where both the t-test and the Wilcoxon sign-rank test agree, indicating that the difference in AURAC between UMPIRE and other baselines is statistically significant.

Test	**VQAv2**			**AdVQA**			**OKVQA**		
	LN-Ent.	EigenScore	SE	LN-Ent.	EigenScore	SE	LN-Ent.	EigenScore	SE
Llava-13B									
t-test pvalue	1.157E-283	5.99E-47	7.58E-118	0	4.85E-59	3.97E-127	1.1E-197	3.28E-66	2.92E-180
Wilcoxon test pvalue	6.506E-162	5.31E-67	1.55E-108	4.1E-164	1.89E-53	6.73E-106	2.6E-141	5.53E-63	1.34E-137
Llava-7B									
t-test pvalue	9.883E-296	3.64E-49	1.41E-98	0	3.34E-79	6.46E-153	4.6E-207	2.32E-69	1.19E-187
Wilcoxon test pvalue	9.072E-163	3.09E-64	7.49E-87	4.4E-165	5E-72	4.09E-119	9.4E-143	1.33E-62	2.87E-139
Mllama-11B									
t-test pvalue	0	2.5E-223	3.05E-184	1.044E-314	1.01E-35	5.79E-164	0	3.7E-113	9.07E-24
Wilcoxon test pvalue	9.854E-165	2.9E-146	2.15E-133	1.3E-162	7.42E-35	3.32E-124	3.5E-165	1.38E-95	7.44E-22

Table 8: Statistical tests calculated based on bootstrap sampling for each model-dataset combination.

Metric	Method	VQAv2	OKVQA	AdVQA	Avg
AUROC ↑	KLE	0.868	0.702	0.777	0.782
	Degree	0.871	0.711	0.782	0.788
	Sem.Dens.	0.860	0.702	0.767	0.776
	Ours	**0.882**	**0.755**	**0.787**	**0.808**
CPC ↑	KLE	0.861	0.704	0.844	0.803
	Degree	0.936	0.934	0.972	0.947
	Sem.Dens.	**0.980**	0.947	**0.986**	**0.971**
	Ours	0.946	**0.966**	0.979	0.964
ECE ↓	KLE	0.133	0.260	0.206	0.200
	Degree	0.044	0.098	**0.013**	0.052
	Sem.Dens.	**0.026**	0.119	0.043	0.063
	Ours	0.038	**0.036**	0.042	**0.039**
AURAC ↑	KLE	0.963	0.778	0.806	0.849
	Degree	0.963	0.768	0.806	0.846
	Sem.Dens.	0.961	0.759	0.799	0.840
	Ours	**0.966**	**0.807**	**0.809**	**0.861**

Table 9: Comparision of UMPIRE with SOTA LLM uncertainty quantification methods, including Semantic Density (Sem.Dens.), Degree, Kernel Language Entropy (KLE), and UMPIRE (Ours) across VQA datasets (VQAv2, OKVQA, AdvQA). We report AUROC (↑ better), CPC (↑ better), ECE (↓ lower is better), and AURAC (↑ better). UMPIRE achieves consistently strong results across all metrics, often outperforming or matching the best baselines.

C.7 Single modality experiment

We also tested our UMPIRE metric on purely textual datasets. To generate the embeddings and answers, we used the `Llama-3.1-8B-Instruct` model (Grattafiori et al., 2024) instead of MLLMs. The datasets tested include Conversational Question Answering (CoQA), TriviaQA, Natural Questions (NQ) and Stanford Question Answering Dataset (SQuAD). We performed tuning of the weighting parameter $\tilde{\alpha}$ for each dataset.

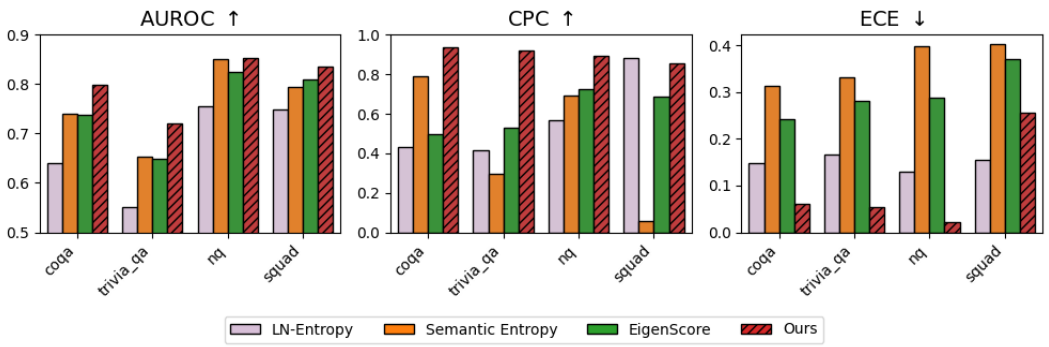

Figure 5: Performance comparison of different uncertainty quantification methods across LLM tasks. The metrics include AUROC (higher is better), CPC (higher is better), and ECE (lower is better).

As shown in Fig. 5, UMPIRE managed to outperform other metrics in most cases, except for the CPC score on the SQuAD dataset, where LN-Entropy performs slightly better. Thus, as noted in Sec. 2, our method is not reliant on modality-specific characteristics when computing the metric, as it is capable of working well in textual tasks of a single modality. UMPIRE is a general framework that can also perform well in the special case of text-only LLM settings.

Dataset	Method	AUROC ↑	ECE ↓	CPC ↑
VQAv2	Seq Prob	0.632	0.121	0.374
	Mean Token Entropy	0.628	0.129	0.046
	Perplexity	0.629	0.131	0.125
	Ours (k=5)	**0.873**	**0.067**	**0.923**
ADVQA	Seq Prob	0.595	0.303	0.372
	Mean Token Entropy	0.590	0.302	0.170
	Perplexity	0.592	0.336	0.151
	Ours (k=5)	**0.774**	**0.055**	**0.959**
OKVQA	Seq Prob	0.581	0.304	0.463
	Mean Token Entropy	0.580	0.303	0.039
	Perplexity	0.581	0.335	0.225
	Ours (k=5)	**0.740**	**0.097**	**0.944**
MathVista	SingleProb	0.628	0.539	0.322
	Mean Token Entropy	0.606	0.601	0.334
	Perplexity	0.616	0.643	0.224
	Ours (k=5)	**0.791**	**0.087**	**0.706**
VQA-RAD	Seq Prob	0.540	0.525	0.140
	Mean Token Entropy	0.535	0.534	0.118
	Perplexity	0.537	0.550	0.168
	Ours (k=5)	**0.806**	**0.090**	**0.828**

Table 10: Comparison of the performance of single sampling methods and UMPIRE across various VQA datasets.

D Ablation Studies

D.1 Embedding Layer Selection

We analyzed the impact of the layer index when extracting the embedding vectors by computing the AUROC performance on different embedding matrices extracted from different layer indices. As shown in Fig. 6 (b), the change in the layer indices makes the AUROC vary slightly. The last layer still yields the best performance, so we adopt it for all of our experiments.

D.2 Weighting Parameter $\tilde{\alpha}$

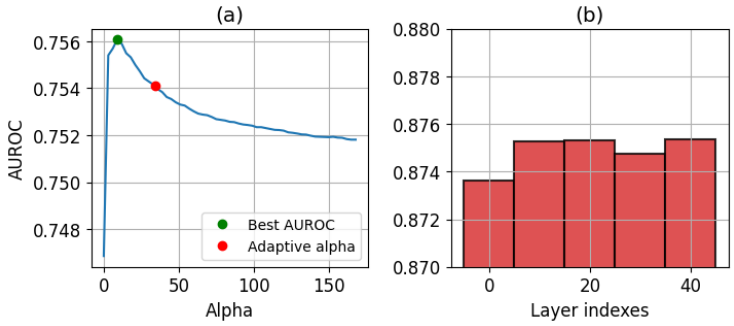

Figure 6: (a) Tuning of the weighting parameter $\tilde{\alpha}$ with respect to AUROC on the small development set (10%) of AdVQA. The 'adaptive alpha' value set without the need for hyperparameter tuning produces good performance. $\tilde{\alpha} = 0$ is suboptimal, reflecting the importance of the incoherence scores. (b) Ablation study on choosing the layer index to extract embedding vectors. Results show that different layer indices only have slight variations in the AUROC performance.

Subset Size	AUROC	ECE	CPC	AURAC
1%	0.882	0.04	0.948	0.966
5%	0.882	0.038	0.947	0.966
10%	0.882	0.038	0.946	0.966

Table 11: UMPIRE performance is robust to the size of unlabeled set of task instance that we use to compute adaptive α.

As mentioned in **U3** in Sec. 4, the incoherence score in UMPIRE has a scaling hyperparameter α that is related to the hyperparameter $\tilde{\alpha} = 2k\alpha$ that controls the balance between two terms in Eq. (4): the unadjusted semantic volume metric and the expectation value of the model-generated probabilities of getting the task instances wrong. In most of our experiments, we did not tune the hyperparameter based on a labeled development set but instead set $\tilde{\alpha}$ such that both terms have the same expected contribution (e.g. based on an unlabeled sample of task instances). We also show that UMPIRE performance is also robust to the size of this unlabeled set of task instance that we use to set by conducting an experiment of using different subset sizes (1%, 5%, and 10%) of the unlabeled evaluation set to set α on the VQAv2 validation set. As can be seen in the table Table 11, UMPIRE maintains consistent and high performance across all metrics, with minimal variation in AUROC, ECE, CPC, and AURAC, underscoring its robustness to the subset of unlabeled data used.

However, in practice, users could potentially search for a better hyperparameter value for their task, such as via grid search or AutoML methods like Bayesian Optimization. In Fig. 6 (a), we provide an illustration of a plot of AUROC v.s. $\tilde{\alpha}$ values from tuning $\tilde{\alpha}$ for the AdVQA dataset, based on a development set consisting of randomly sampled 10% of the full dataset. Note that while using grid search would yield a higher AUROC (green dot), the 'adaptive alpha' approach of setting $\tilde{\alpha}$ to balance both terms in Eq. (4) will not be very far off from the optimum. In addition, an alpha value of 0 has a significantly lower performance, indicating that the incoherence score contributes to the good performance of UMPIRE.

D.3 Number of generations analysis

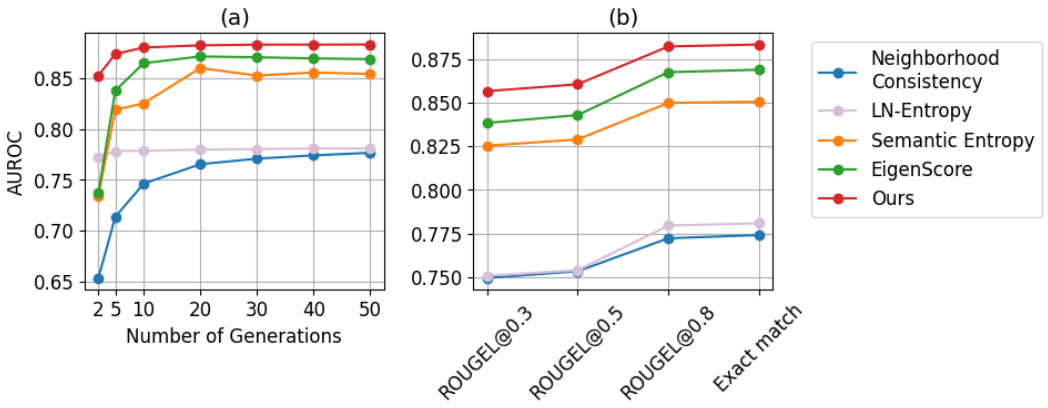

Figure 7: Ablation study on the (a) number of generations for our method and (b) evaluation methods. (a) shows the AUROC performance as the number of generations increases, demonstrating the impact of additional generations. UMPIRE is able to achieve high performance with few generations. (b) consistently outperforms baseline approaches regardless of the chosen evaluation functions.

To analyze the impact of the number of generations on the various metrics' performance, we conduct an ablation study by varying the number of generated responses (from 2 to 50) per task instance for a VQAv2 validation subset. As shown in Fig. 7 (a), while increasing the number of generations generally improves AUROC across all methods, UMPIRE achieves higher performance with significantly fewer generations compared to baselines. This indicates that our method is more efficient, requiring fewer samples to reach strong performance, whereas other methods continue to rely on additional

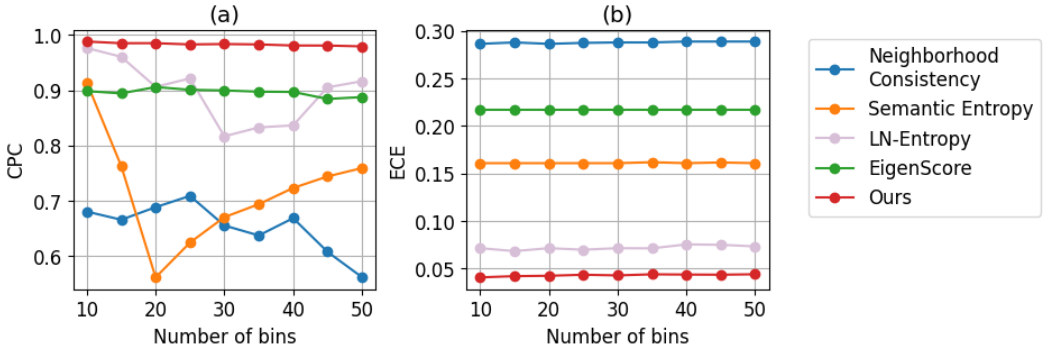

Figure 8: Results for the effect of number of bins on (a) CPC and (b) ECE. Both measures show that UMPIRE consistently outperforms baselines.

generations for improvement. The results highlight the robustness of our approach in capturing correctness signals effectively, even with a limited number of generations.

D.4 Ablation on evaluation parameters

Evaluation function $a(\mathcal{M}, t^*)$ Following the setting in Kuhn et al. (2023), we further evaluate the performance of our method and baselines under various levels of the ROUGE-L. Fig. 7(b) presents the AUROC scores across different evaluation functions $a(\mathcal{M}, t^*)$ on a subset of the VQAv2 validation set, demonstrating that our method consistently outperforms baseline approaches regardless of the chosen evaluation functions. These results highlight the versatility and robustness of our approach across different correctness evaluation criteria.

Effect of number of bins in ECE and CPC. We also analyzed the effect of the number of bins when computing ECE and CPC by randomly trying on a subset of AdVQA dataset. Fig. 8 illustrates that UMPIRE still achieves the best and consistent performance across all bin values.

D.5 Sampling temperature

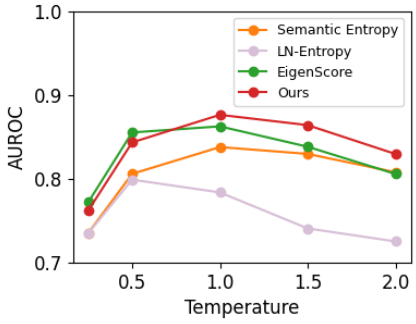

Figure 9: Impact of temperature during the generation process on evaluation performance.

Besides the number of generations in App. D.3, we analyzed the impact of temperature during the generation process on the evaluation performance. We conducted an ablation study by varying the generation temperature (from 0.25 to 2) on a subset of the VQAv2 validation set. As shown in Fig. 9, the temperature of 1 helps UMPIRE achieve the best performance and outperforms the best performance of other baselines (Eigen and LN-Ent).

D.6 Model sizes and families analysis

We analyze the impact of model size and architecture family on evaluation performance by comparing different models across various sizes and families on a subset of the VQAv2 validation set for the image-text understanding task and the SLUE-P2-SQA5 test set for the audio-text ones. As shown in Fig. 10, we observe a slight increase in AUROC as the model size increases within the same family. This suggests that larger models tend to generate more informative and reliable outputs. Additionally, our method show a strong performance AUROC across all tested models, demonstrating its robustness regardless of model size or architecture. These findings highlight that while larger models can enhance performance, our approach remains effective across different model scales and families.

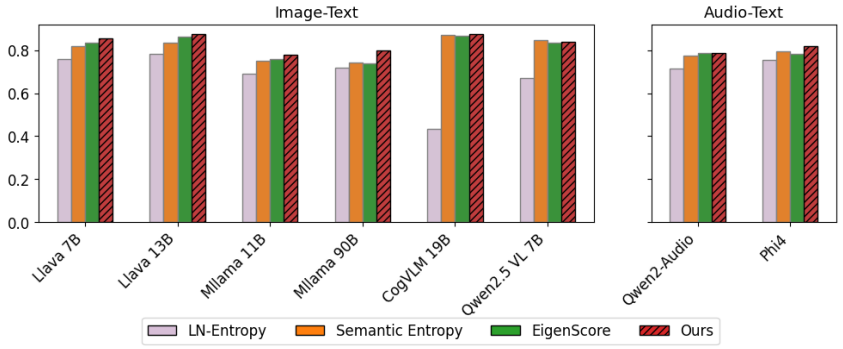

Figure 10: Ablation study across different models on image-text (VQAv2) and audio-text (SLUE-P2-SQA5) understanding datasets, evaluating AUROC performance for LN-Ent, Sem.Ent, Eigen, and UMPIRE. The results indicate that UMPIRE consistently achieves strong AUROC across various models, including `Llava-v1.5-7b`, `Llava-v1.5-13b`, `Llama-3.2-11B-Vision`, `Llama-3.2-90B-Vision`, `cogvlm2-llama3-chat-19B`, `Qwen2.5-VL-7B-Instruct` for image-text, and `Qwen2-Audio-7B`, `Phi4` for audio-text. This highlights the robustness and effectiveness of our approach across different model architectures.

D.7 Results of other blackbox and whitebox proxy models in blackbox settings.

D.7.1 Blackbox models

As in Fig. 11, we find that UMPIRE also outperforms other baselines using different black-box models, including Claude 3.5 Haiku (Anthropic, 2024), GPT4o-mini (OpenAI, 2024).

D.7.2 White-box proxy models

The experiments in Sec. 6.4 use a simple approach of applying a vanilla `Llava-v1.5-13b` proxy model for all blackbox models. As seen in our empirical results Fig. 3, UMPIRE consistently outperforms baselines without any fine-tuning of the proxy model. In general, we observe that performant models tend to produce similar semantic volume, while variations in incoherence scores introduce noise, they do not have a significant adverse impact on overall performance. The table Table 12 shows new ablation results where using different whitebox proxy models for GPT-4o still yield good performance on 3000 samples of the VQAv2 validation set.

Method	**AUROC** ↑	**CPC** ↑	**ECE** ↓
UMPIRE (`Llava-v1.5-13b`)	0.890	0.904	0.094
UMPIRE (`Llava-v1.5-7b`)	0.893	0.900	0.087
UMPIRE (`Llama-3.2-11B-Vision`)	0.839	0.943	0.091

Table 12: Results of UMPIRE in blackbox settings with different proxy models.

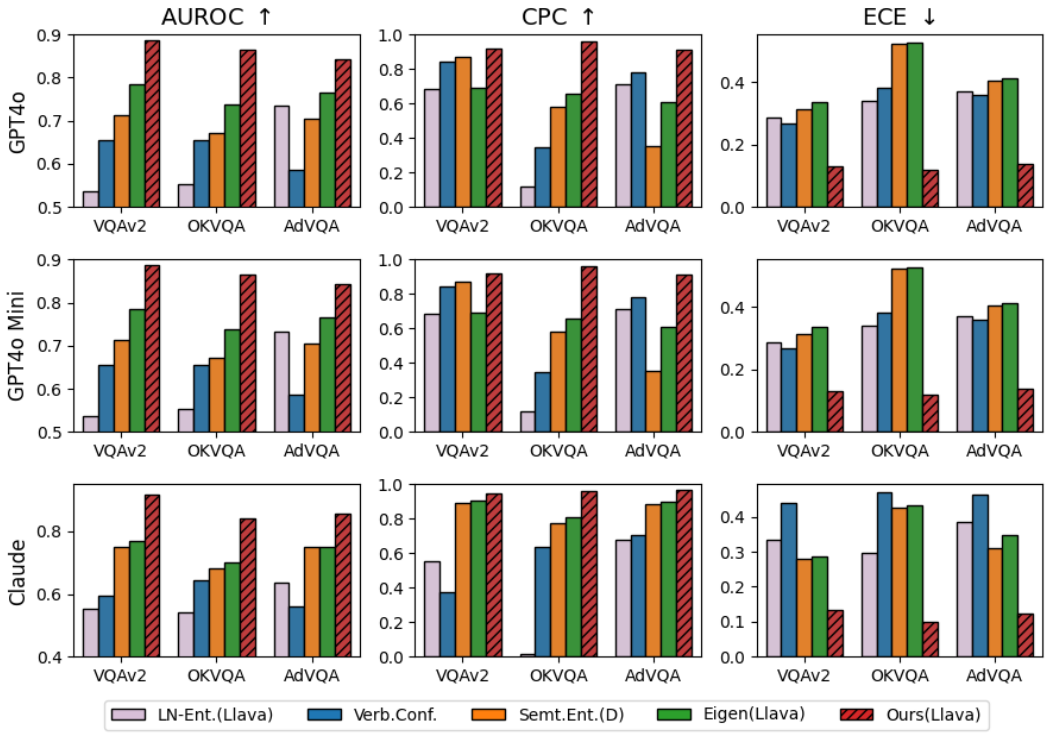

Figure 11: UMPIRE metric consistently outperforms other baselines across various black-box models, including GPT4o, GPT4o-mini, and Claude 3.5 Haiku.

D.8 Length-normalized effect

In prior work on uncertainty estimation and related scoring functions, length normalization has often been applied to adjust for biases introduced by varying response lengths (Kuhn et al., 2023; Malinin & Gales, 2021). Motivated by this, we explored whether length normalization could also benefit the quality term of UMPIRE, i.e, length-normalized incoherence score. Empirically, as shown in Table 13, we observed that applying length normalization does not consistently improve performance across most MLLM baselines. In fact, the normalized variant frequently underperforms in terms of AUROC and CPC, and yields better ECE in several datasets. However, in the pure LLM setting as seen in App. C.7, length normalization appears to offer some advantages (see Table 14), suggesting that its effectiveness may be setting-dependent.

Metric	Method	VQAv2	AdVQA	OKVQA	MathVista	VQA-Rad
AUROC ↑	Without Length Normalized	**0.882**	**0.787**	0.755	0.822	**0.802**
	Length Normalized	0.875	0.779	**0.756**	**0.825**	0.792
CPC ↑	Without Length Normalized	0.946	**0.979**	**0.966**	**0.945**	0.908
	Length Normalized	**0.986**	0.978	0.946	0.936	**0.935**
ECE ↓	Without Length Normalized	**0.038**	0.042	0.036	0.071	**0.067**
	Length Normalized	0.062	**0.019**	**0.034**	**0.056**	0.068

Table 13: Comparison of UMPIRE with and without length normalization across various VQA datasets.

Metric	Method	CoQA	TriviaQA	NQ	SQuAD
AUROC ↑	Without Length Normalized	0.749	0.641	0.844	0.813
	Length Normalized	**0.799**	**0.720**	**0.853**	**0.836**
CPC ↑	Without Length Normalized	0.876	0.850	0.780	**0.888**
	Length Normalized	**0.937**	**0.923**	**0.892**	0.855
ECE ↓	Without Length Normalized	0.068	0.098	0.076	**0.117**
	Length Normalized	**0.061**	**0.054**	**0.022**	0.257

Table 14: Comparison of UMPIRE with and without length normalization across various text datasets.