FESTA: Functionally Equivalent Sampling for Trust Assessment of Multimodal LLMs

Debarpan Bhattacharya* Indian Institute of Science Bangalore, India debarpanb@iisc.ac.in

Apoorva Kulkarni* University of Maryland College Park, USA apoorvak@umd.edu

Sriram Ganapathy Indian Institute of Science Bangalore, India sriramg@iisc.ac.in

Abstract

The accurate trust assessment of multimodal large language models (MLLMs) generated predictions, which can enable selective prediction and improve user confidence, is challenging due to the diverse multi-modal input paradigms. We propose **F**unctionally **E**quivalent **S**ampling for **T**rust **A**ssessment (FESTA), a multimodal input sampling technique for MLLMs, that generates an uncertainty measure based on the equivalent and complementary input samplings. The proposed task-preserving sampling approach for uncertainty quantification expands the input space to probe the consistency (through equivalent samples) and sensitivity (through complementary samples) of the model. FESTA uses only input-output access of the model (black-box), and does not require ground truth (unsupervised). The experiments are conducted with various off-the-shelf multi-modal LLMs, on both visual and audio reasoning tasks. The proposed FESTA uncertainty estimate achieves significant improvement (33.3% relative improvement for vision-LLMs and 29.6% relative improvement for audio-LLMs) in selective prediction performance, based on area-under-receiver-operating-characteristic curve (AUROC) metric in detecting mispredictions. The code implementation is open-sourced [1].

1 Introduction

Large language models (LLMs) have achieved remarkable performance across a wide array of natural language processing tasks (Brown et al., 2020; Touvron et al., 2023; OpenAI, 2023). However, their performance in integration with other modalities- called, multimodal LLMs (MLLMs), is often inferior compared to text-only LLMs (Li et al., 2024; Fu et al., 2024b).

1.1 Selective prediction

The works on selective prediction (SP) propose to prevent a model from answering in highly uncertain settings, thereby abstaining from potentially wrong predictions (abstention) (Wen et al., 2025; Madhusudhan et al., 2025). Selective prediction is highly desirable for building safe AI deployments (Amodei et al., 2016; Hendrycks et al., 2021b) in:
(a) Tasks where a model offers low accuracy.
(b) Safety-critical scenarios like finance, medicine and autonomous driving, where incorrect predictions are very expensive.
An efficient SP algorithm offers *low selective risk* (high accuracy in the subset of questions it chooses to answer) in deployment.

1.2 Selective prediction for multimodal LLMs

One of the most common approaches to SP for LLMs is based on quantifying the variance in the output in the form of an entropy measure (Kuhn et al., 2023; Farquhar et al., 2024; Ling et al., 2024). The model predictions with high uncertainty (and entropy) are more prone to errors, and hence are abstained from prediction. However, in some cases, inaccurate predictions by MLLMs may arise from their insensitivity to the input (referred to as mode collapse). Such mis-predictions have low predictive entropy, making the entropy an unreliable measure of prediction accuracy. Other works also find LLMs to generate low-uncertainty hallucinations (Simhi et al., 2025). In many of these works, the log-probability has been the most common means to obtain output confidence in traditional neural networks. However, for MLLMs,
(a) Log-probability might not be accessible for closed-source models.
(b) The calibration of log-probability may be upset during the instruction tuning phase (Tian et al., 2023).

[1] https://github.com/iiscleap/mllm-uncertainty-estimation
* Equal contribution.

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 12277–12295
November 4-9, 2025 ©2025 Association for Computational Linguistics

Figure 1: An example of multi-modal reasoning input (top-panel). An equivalent sample (middle panel) with same gray-scale image and rephrased prompt question is expected to keep the MLLM prediction unchanged, whereas a complementary input sample (bottom panel) is expected to alter the prediction. The proposed FESTA uses equivalent and complementary samples to generate the uncertainty measure.

(c) Calibration of log-probability deteriorates as model performance degrades (Guo et al., 2017; Wang et al., 2023; Kürbis et al., 2024).

Hence, the need for uncertainty and confidence estimation for MLLMs, in black-box settings and low-performing tasks, is paramount. Recently published works (Fu et al., 2024b; Bhattacharya et al., 2025) report that MLLMs are astonishingly poor in simple multimodal reasoning tasks. Although this makes multimodal reasoning a suitable application for evaluating abstention algorithms, computation of predictive uncertainty of MLLMs remains challenging (Hendrycks et al., 2022; Li et al., 2023).

1.3 Contribution

Based on the above discussion, and with the advent of multimodal LLMs such as large vision language models (LVLM) (Liu et al., 2023; Agrawal et al., 2024; Bai et al., 2025) and large audio language models (LALM) (Chu et al., 2024; Tang et al., 2023), we make the following observations:

(O1) The log-likelihood based quantification of model confidence is inapt for MLLMs because of its unavailability in black-box settings and its upset calibration during instruction tuning.

(O2) The existing output entropy-based uncertainty measures for black-box models fail in the case of low-entropy hallucinations.

(O3) Abstention performance of such entropy-based prior works also degrades in low accuracy tasks like reasoning.

In this paper, we propose *functional equivalence sampling (FES)* and *functional complementary sampling (FCS)* for abstaining from predictions in multimodal LLMs. FES samples are equivalent to the original input for the given task, and hence, ideally should not change the model prediction from the original input. FCS samples are complementary samples for the same task that are expected to alter the model prediction. The examples are illustrated in Figure 1. The MLLM predictions for FES and FCS samples are used to compute the FESTA uncertainty, which acts as an indicator of mis-prediction. We restrict this work to multiple-choice reasoning involving audio/visual prompts. The key contributions are:

- We propose equivalent (FES) and complementary (FCS) input samplings to identify the consistency and sensitivity of model outputs. They contribute to obtain the FESTA uncertainty score that is used to abstain from MLLM mispredictions in unsupervised and black-box settings.
- We quantify FESTA score as a KL-divergence from an ideally consistent and sensitive model, that improves over the entropy-based measures.
- FESTA successfully mitigates low uncertainty hallucinations, where the other baselines fail.
- Extensive benchmark comparisons with various other prior works on audio/visual reasoning tasks and multiple open-source MLLMs.

Figure 2 illustrates the working of FESTA.

2 Problem statement

While FESTA is applicable to general multi-modal outputs, we make a few assumptions and restrictions. We restrict this paper to the purely textual output case, in a multi-choice question-answering (MCQA) setting. Further, it is also assumed that the MLLMs under study are instruction-tuned well enough to follow instructions to generate single-token MCQA outputs. Now, we define the problem statement, followed by formulating the FESTA

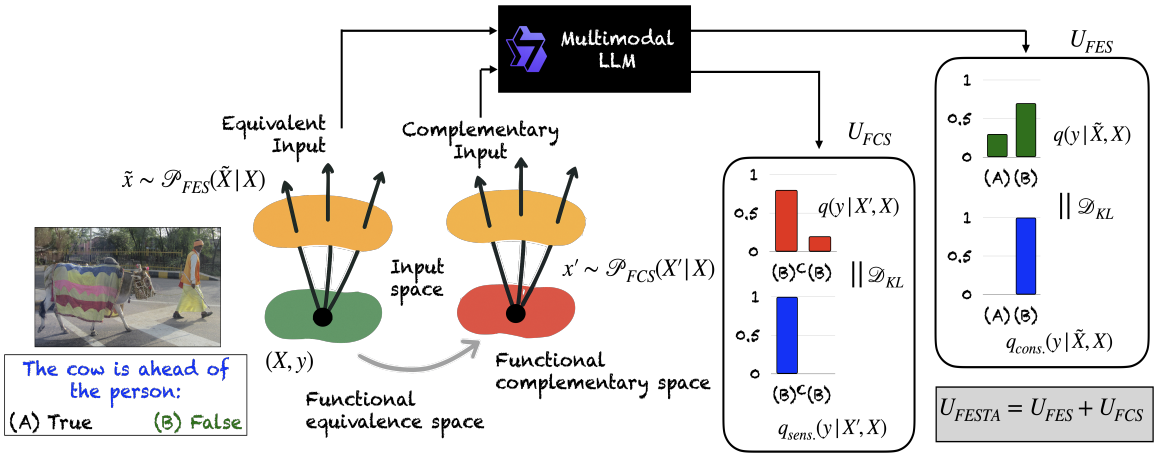


Figure 2: Schematic illustration of the proposed FESTA uncertainty quantification approach. Given a multimodal MCQ input, we generate functional equivalent samples (FES) and functional complementary samples (FCS). We compute divergence of model predictive uncertainty from an ideally consistent model (for FES) and an ideally sensitive model (for FCS), and then combine these measures to generate the FESTA uncertainty score.

algorithm. A summary of the notations used is present in Appendix Table 3 .

Let $\boldsymbol{X} = [\boldsymbol{X_O}, \boldsymbol{X_T}] \in \boldsymbol{\mathcal{X}}$ denote a multi-modal input instance (e.g., an audio/image + textual prompt) to an MLLM with ground-truth response $\mathbf{y}_{\texttt{target}}$, and let $\mathcal{S}$ denote the finite set of possible MCQA outputs. The MLLM defines a predictive distribution over a fixed set of outputs: $q(\mathbf{y}|\boldsymbol{X}) \in \Delta^{|\mathcal{S}|}$. The model's prediction is (greedy sampling):

$$\hat{\mathbf{y}} := \arg\max_{\mathbf{y} \in \mathcal{S}} q(\mathbf{y}|\boldsymbol{X}).$$

Problem statement: Given multi-modal input $\boldsymbol{X}$ and model output $\hat{\mathbf{y}}$ generated from inherent predictive distribution $q(\mathbf{y}|\boldsymbol{X})$, estimate the predictive uncertainty of $\hat{\mathbf{y}}$ without access to $q(\mathbf{y}|\boldsymbol{X})$ in a black-box setting.

We resort to the following directions to develop the uncertainty estimator,

- Sampling FES to estimate the epistemic uncertainty (consistency).
- Sampling FCS to measure counterfactual uncertainty (sensitivity).
- Combining the two measures to compute the FESTA uncertainty score.

3 FESTA Uncertainty Estimator

In this section, we describe the sampling processes, and the computation of FESTA uncertainty score.

3.1 Functional equivalent samples (FES)

Given an input–output pair $(\boldsymbol{X}, \mathbf{y}_{\texttt{target}})$, let $T(\cdot)$ denote the task that the model must solve to generate the correct response, and let $\mathbf{M}_{\texttt{ideal}}$ denote the hypothetical model which has the ideal behavior.

Def: A transformation $\tilde{\mathbf{X}} = \mathcal{E}(\mathbf{X})$ is said as a *functionally equivalent sample* of $\mathbf{X}$ if:

$$T(\tilde{\mathbf{X}}) = T(\mathbf{X}) \text{ and } M_{\texttt{ideal}}(\tilde{\mathbf{X}}) = M_{\texttt{ideal}}(\mathbf{X})$$

We define a distribution $\mathcal{P}_{\text{FES}}(\tilde{\mathbf{X}}|\mathbf{X})$ over all possible equivalent transformations $(\{\mathcal{E}_i(\mathbf{X})\}_i)$ and denote the sampling process as:

$$\tilde{\mathbf{X}} \sim \mathcal{P}_{\text{FES}}(\tilde{\mathbf{X}}|\mathbf{X}) \text{ or } \tilde{\mathbf{X}} \sim_{\mathcal{E}} \mathbf{X}$$

For example, in Figure 1, the task a model must perform to answer $\mathbf{X}$ is spatial reasoning involving the objects - car and the cat. An equivalent sample (FES), $\tilde{\mathbf{X}}$ is a transformation of the input $\mathbf{X}$ (image, question) so that the spatial relationship of the cat and the car are equivalent to the one in the original image, or the rephrased prompt preserves the query part of the original question semantically. Thus, the FES transformed image and question remain consistent with the original input. An example FES sample ($\tilde{\mathbf{X}}$) in Figure 1 involves the original image and the paraphrased question. In an ideal setting, the output of the MLLM should be unaltered. In an analogous fashion, the audio version of FES, for different temporal reasoning tasks, are shown in Figure 7.in Appendix. Also, the FES samples hold formal equivalence among them (Appendix A.1).

3.2 Functional complementary samples (FCS)

Def: A transformation $\mathbf{X}' = \mathcal{C}(\mathbf{X})$ is a *functionally complementary sample* of $\mathbf{X}$ if:

$$T(\mathbf{X}') = T(\mathbf{X}) \text{ and } M_{\texttt{ideal}}(\mathbf{X}') \neq M_{\texttt{ideal}}(\mathbf{X})$$

So, an FCS sample is task-equivalent but functionally divergent. We define a distribution $\mathcal{P}_{\text{FCS}}(\mathbf{X}'|\mathbf{X})$ over all complementary transformations ($\{\mathcal{C}_i(\mathbf{X})\}_i$). We sample $\mathbf{X}'$ as,

$$\mathbf{X}' \sim \mathcal{P}_{\text{FCS}}(\mathbf{X}'|\mathbf{X}) \text{ or } \mathbf{X}' \sim_{\mathcal{C}} \mathbf{X}$$

For example, in Figure 1, the complementary sample $\mathbf{X}'$ is a transformation of the image that doesn't alter the spatial relationship between the cat and the car, and the question is semantically reversed. The combination ensures that a robust model will modify its response with respect to the original prediction. Alternatively, the image can be transformed complementarily as well. In an analogous fashion, the FCS perturbations of the audio examples are shown in Section A.5. Note that, negation words like "not" are avoided while generating complementary questions to avoid model sensitivity to negation words (Truong et al., 2023). Although a complementary sample is not equivalent to the original input, it can be shown that all complementary samples $\mathbf{X}'$ have formal equivalence among them (Appendix A.2). Further details on the generation of FES and FCS samples are discussed in Sections 5.5 and 5.6.

3.3 Ideal behavior under FES and FCS

Before uncertainty quantification, we first introduce the notion of a consistent and sensitive model, $M_{\texttt{cons.}}$ and $M_{\texttt{sens.}}$, respectively. Their definitions do not rely on the ground truth $\mathbf{y}_{\texttt{target}}$, making the approach fully unsupervised. They have a subset of the properties that $M_{\texttt{ideal}}$ has.

- **Consistency under FES**: The model $M_{\texttt{cons.}}$ generates predictions that remain consistent and unaltered with respect to the original prediction, i.e., $(\mathbf{y} = \hat{\mathbf{y}})$, under FES.
- **Sensitivity to FCS**: The model $M_{\texttt{sens.}}$ is sensitive to counterfactual negations and its predictions for FCS are complementary to the original predictions, i.e., $(\mathbf{y} \neq \hat{\mathbf{y}})$.

We later define the consistency entropy ($U_{\texttt{FES}}$) and sensitivity entropy ($U_{\texttt{FCS}}$), based on the deviation of a given model M from $M_{\texttt{cons.}}$ and $M_{\texttt{sens.}}$, respectively. Note that, an ideal model $M_{\texttt{ideal}}$ satisfies properties of both $M_{\texttt{cons.}}$ and $M_{\texttt{sens.}}$ ($M_{\texttt{ideal}} \subset M_{\texttt{cons.}}$, $M_{\texttt{ideal}} \subset M_{\texttt{sens.}}$). $M_{\texttt{ideal}}$ is both consistent and sensitive, while the converse is false.

3.4 Uncertainty Estimation from FES

For an input $\boldsymbol{X} = \mathbf{x}$, denote its FES samples as: $\{\tilde{\mathbf{x}}_1, \tilde{\mathbf{x}}_2, \ldots, \tilde{\mathbf{x}}_{K_1}\}$. Let $q(\mathbf{y}|\tilde{\mathbf{x}}_k)$ be the predictive distribution of model M for $\tilde{\mathbf{x}}_k$ and $\mathbf{y}$ is the output random variable sampled using the stochastic decoding of the MLLMs. The uncertainty measure of the model M is measured as the deviation from the grounded model $M_{\texttt{cons.}}$.

- **Predictive distribution of $M_{\texttt{cons.}}$**: The definition of $M_{\text{cons.}}$ implies that its predictive distribution is a Kronecker delta function with respect to equivalent sampling. The entire weight of the delta is concentrated at the greedy output $\hat{\mathbf{y}}$.

$$q_{\texttt{cons.}}(\mathbf{y} \mid \tilde{\mathbf{x}}_k, \mathbf{x}) = \delta_{\mathbf{y}\hat{\mathbf{y}}} \quad \forall\, \tilde{\mathbf{x}}_k \sim_{\mathcal{E}} \mathcal{P}_{\text{FES}},\ \mathbf{y} \in \mathcal{S}$$

- **Predictive distribution of M**: This is given as:

$$q_{FES}(\mathbf{y} \mid \mathbf{x}) = \mathbb{E}_{\tilde{\mathbf{x}}_k \sim \mathcal{P}_{\text{FES}}}\, q(\mathbf{y} \mid \tilde{\mathbf{x}}_k, \mathbf{x}), \mathbf{y} \in \mathcal{S}$$

Now, we pose the uncertainty of model M as the deviation of its predictive distribution from that of $M_{\texttt{cons.}}$, i.e.,

$$U_{FES}(M|\mathbf{x}) = D_{KL}(q_{\texttt{cons.}}(\mathbf{y} \mid \mathbf{x}) || q_{FES}(\mathbf{y} \mid \mathbf{x}))$$

Proposition 3.1. *The $U_{FES}(M|\mathbf{x})$ simplifies to:*

$$U_{FES}(M \mid \mathbf{x}) := -\log q_{FES}(\mathbf{y} = \hat{\mathbf{y}} \mid \mathbf{x}), \mathbf{y} \in \mathcal{S}$$

Proof. The proof is given in the Appendix A.3. □

It is interesting to note that, resorting to $M_{\text{cons.}}$ makes the quantification self-supervised. While it is solely dependent on model predictions without access to ground truth (unsupervised), it is based on the deviation from a hypothetically ideal model.

3.5 Uncertainty Estimation from FCS

The uncertainty from complementary samples: $\{\mathbf{x}'_1, \mathbf{x}'_2, \ldots, \mathbf{x}'_{K_2}\}$, drawn from $\mathbf{x}'_k \sim \mathcal{P}_{\text{FCS}}$, is,

- **Predictive distribution of $M_{\texttt{sens.}}$**: Based on the definition of FCS, they alter the predictions - $M_{\texttt{sens.}}(\mathbf{x}') \neq M_{\texttt{sens.}}(\mathbf{x})$. Clearly, under the original support $\mathcal{S}$, the predictive distribution $q(\mathbf{y} \mid \mathbf{x}'_k, \mathbf{x})$ can be any distribution. But, we restrict the support to have only two members as $\mathcal{S}' = \{\hat{\mathbf{y}}, \hat{\mathbf{y}}^c\}$ where $\hat{\mathbf{y}}^c = \{\mathbf{y} : \mathbf{y} \in \mathcal{S}, \mathbf{y} \neq \hat{\mathbf{y}}\}$.

Now, the predictive distribution becomes a Kronecker delta with entire mass on $\hat{\mathbf{y}}^c$.

$$q_{\text{sens.}}(\mathbf{y} \mid \mathbf{x}'_k, \mathbf{x}) = \delta_{\mathbf{y}\hat{\mathbf{y}}^c} \quad \forall\, \mathbf{x}'_k \sim_{\mathcal{C}} \boldsymbol{X},\ \mathbf{y} \in \mathcal{S}'$$

- **Predictive distribution of** M: In this case,

$$q_{FCS}(\mathbf{y} \mid \mathbf{x}) = \mathbb{E}_{\mathbf{x}'_k \sim \mathcal{P}_{\text{FCS}}}\, q(\mathbf{y} \mid \mathbf{x}'_k, \mathbf{x}), \mathbf{y} \in \mathcal{S}'$$

Now, the uncertainty of model M from FCS is the deviation from $M_{\text{sens.}}$ in predictive distribution:

$$U_{FCS}(M|\mathbf{x}) = D_{KL}(q_{\text{sens.}}(\mathbf{y} \mid \mathbf{x}) || q_{FCS}(\mathbf{y} \mid \mathbf{x}))$$

Proposition 3.2. *The* $U_{FCS}(M|\mathbf{x})$ *simplifies to:*

$$U_{FCS}(M \mid \mathbf{x}) := -\log\left(\sum_{\mathbf{y}} q_{FCS}(\mathbf{y} \neq \hat{\mathbf{y}} \mid \mathbf{x})\right)$$

Proof. The proof is similar to Appendix A.3. □

FES samples do not provide the right framework for quantifying the low-uncertainty hallucinations like insensitivity to the prompt. However, U_{FCS} helps to abstain from such low-uncertainty mis-predictions as such models show no sensitivity to the complementary samples. We formally show this for a single attention block in Appendix A.4.

3.6 FESTA uncertainty estimate

The FESTA uncertainty estimate combines two axes of uncertainty quantification - U_{FES}, which measures the model consistency for equivalent samples, and U_{FCS}, which measures the model sensitivity for complementary samples. The combination helps to abstain both high uncertainty and low uncertainty mis-predictions, as evident in Section 6. FESTA is detailed in Figure 2 and Algorithm 1.

Algorithm 1 FESTA Uncertainty Estimator

Require: Input $\boldsymbol{X}$, original prediction $\hat{\mathbf{y}} = \arg\max_{\mathbf{y}} q(\mathbf{y}|\boldsymbol{X})$, no. of samples $K = K_1 + K_2$.

FES Sampling:

Generate K_1 FES samples: $\{\tilde{\mathbf{x}}_k\}_{k=1}^{K_1} \sim P_{\text{FES}}(\tilde{\mathbf{x}}|\boldsymbol{X})$ by sampling text and non-text modalities $(\boldsymbol{X}_T, \boldsymbol{X}_O)$ for (K_{11}, K_{12}) times and using all combinations ($K_1 = K_{11} \times K_{12}$).

Compute $q_{\text{FES}}(\mathbf{y}|\boldsymbol{X})$ using $\{\tilde{\mathbf{x}}_k\}_{k=1}^{K_1}$

Compute FES uncertainty for prediction $\hat{\mathbf{y}}$:

$$U_{\text{FES}} = -\log q_{\text{FES}}(\mathbf{y} = \hat{\mathbf{y}}|\boldsymbol{X})$$

FCS Sampling:

Generate K_2 FCS samples: $\{\mathbf{x}'_k\}_{k=1}^{K_2} \sim P_{\text{FCS}}(\mathbf{x}'|\boldsymbol{X})$ by complementary sampling either of the text or non-text modalities (e.g. $\boldsymbol{X}_T$) for K_{21} times and equivalent sampling the other (e.g. $\boldsymbol{X}_O$) for K_{22} times, and use all combinations ($K_2 = K_{21} \times K_{22}$).

Compute $q_{\text{FCS}}(\mathbf{y}|\boldsymbol{X})$ using $\{\mathbf{x}'_k\}_{k=1}^{K_2}$.

Compute FCS uncertainty for prediction $\hat{\mathbf{y}}$:

$$U_{\text{FCS}} = -\log\left(\sum_{\mathbf{y} \neq \hat{\mathbf{y}}} q_{\text{FCS}}(\mathbf{y}|\boldsymbol{X})\right)$$

FESTA: $U_{\text{FESTA}} = U_{\text{FES}} + U_{\text{FCS}}$

4 Related Prior Work

Uncertainty estimation for LLMs: Uncertainty estimation for Large Language Models (LLMs) is understudied, yet an important area as the uncertainty measure seldom correlates with their task accuracy (Ye et al., 2024).

There can be black-box or white-box approaches to LLM uncertainty estimation (Shorinwa et al., 2025; Wen et al., 2025). White-box methods leverage internal model access to quantify uncertainty (Chen et al., 2024). However, such approaches are restrictive for closed-source models (Lin et al., 2024). Black-box methods are suitable for closed-source models as they assume only query access (Gao et al., 2024; Kadavath et al., 2022), such as input ensembling, augmenting, or rephrasing to estimate predictive entropy (Hou et al., 2024; Yang et al., 2024; Jiang et al., 2023). Other approaches for uncertainty estimation for LLMs use conformal prediction (Yadkori et al., 2024; Wang et al., 2025; Silva-Rodríguez et al., 2025). Critical gaps remain in uncertainty estimation for LLMs regarding robustness and efficient sampling strategies (Shorinwa et al., 2025; Abbasi-Yadkori et al., 2024; Yona et al., 2024).

Uncertainty estimation for MLLMs: Although the multimodal LLMs suffer more from erroneous predictions and hallucinations, effective abstention algorithms for them are scarce. The amplified hallucinations for MLLMs can be attributed to the ineffective grounding (Favero et al., 2024) and fusion (Kang et al., 2025). A handful of other studies, that reduce MLLM hallucination, model the problem as a causal graph (Li et al., 2025), mutual-information decoding with direct preference optimization (Favero et al., 2024), and visual

attention redistribution (Kang et al., 2025). However, they need training or access to internal model parameters/log-probabilities. Other works employ uncertainty-aware agentic framework (Zhi et al., 2025), semantic perturbation (Zhao et al., 2025; Khan and Fu, 2024) or self-assessment to compute uncertainty (Chen et al., 2025). In contrast, our work is a fully unsupervised, black-box model-based and post-hoc uncertainty estimation method. The predictive uncertainty estimation can be based on open-ended or multiple-choice (MCQA) predictions. Because of its limited output, simplicity, and faster inference (Zhang et al., 2024), the MCQA framework is widely used in accuracy and uncertainty evaluation benchmarks for LLMs (Hendrycks et al., 2021a; Wang et al., 2024; Yang et al., 2025b) and MLLMs (Liu et al., 2024; Fu et al., 2024a; Sakshi et al., 2025). However, the free-text applications require uncertainty estimation for open-ended predictions, like semantic entropy (Kuhn et al., 2023; Farquhar et al., 2024) and iterative prompting (Abbasi-Yadkori et al., 2024).

5 Experimental Setup

5.1 Tasks and Datasets

We use 3 datasets on positional reasoning - spatial reasoning for the vision-LLMs and temporal reasoning for audio-LLMs, in this study.

BLINK: The BLINK dataset (Fu et al., 2024b) points to the limitations of modern vision-LLMs with binary questions, where relative positions between different objects in an image are queried. The spatial reasoning samples from the validation split (143 samples) are chosen for evaluation.

VSR: The Visual spatial reasoning (VSR) dataset (Liu et al., 2023) also has MCQ questions with two choices. It is entirely focused on diverse spatial reasoning samples. The validation partition with 100 randomly samples are used for evaluation.

TREA: The Temporal Reasoning Evaluation of Audio (TREA) dataset (Bhattacharya et al., 2025) is a comprehensive audio-temporal reasoning dataset on which audio-LLMs perform poorly (Kuan and Lee, 2025). It has MCQ with four answer choices. It further divides the temporal reasoning task into 3 categories - ordering, duration, and event counting. We use a subset of 300 samples (100 per task).

5.2 Multimodal LLMs

Visual spatial reasoning: We use large vision language models (LVLM)- `Gemma-3` (Team et al., 2025), `LLaVA-1.6` (Liu et al., 2023), `Qwen-2.5VL` (Yang et al., 2025a), `Phi4` (Abdin et al., 2024) and `Pixtral` (Agrawal et al., 2024) for the evaluation. All of these models show significantly lower performance (Table 1) compared to humans performance of $> 95\%$ (Fu et al., 2024b). Further details and the model cards are in Appendix A.10.

Audio temporal reasoning: We evaluate two open-source audio-LLMs `Qwen2-audio` (Chu et al., 2024) and `SALMONN` (Tang et al., 2023). We observe their temporal reasoning performance to be poor for the tasks (Table 2). We have also experimented with generating audio captions using the audio-LLMs and then passing the text captions to a text-only LLM `Qwen-2` (Bai et al., 2025) as suggested in (Bhattacharya et al., 2025), for improved accuracy (Table 2). More details are in Appendix A.10.

5.3 Comparison with Baseline Systems

- **Output entropy (OE)**: The predictive distribution of the models is estimated using stochastic decoding, and the entropy is measured as (Kuhn et al., 2023): $\mathcal{H}\left(q(\mathbf{y}|\mathbf{x})\right) = -\sum_{\mathbf{y}} q(\mathbf{y}|\mathbf{x}) \log q(\mathbf{y}|\mathbf{x})$.
- **Verbalized confidence (VC)**: LLMs are good estimators of their own confidence (Tian et al., 2023), when verbalized through prompting.
- **Input augmentations (IA)**: Based on the approach in (Bahat and Shakhnarovich, 2020) to obtain predictions using input augmentations and computing entropy from the ensemble. Apart from image augmentations (IA-I), we performed text augmentations (IA-T), using paraphrasing. Finally, we report performance of combined augmentations (IA-IT).
- **Rephrase uncertainty (RU)**: This system (Yang et al., 2024) uses text rephrasing and measures the answer consistency.
- **Black-box uncertainty (BU)**: The work is based on using a combination of top-K prompting and random output sampling that yields the most stable performance (Xiong et al., 2024). We use top-4 prompting with outputs sampling 5 times.
- **Semantic uncertainty**: Another popular uncertainty estimation method is the concept of semantic entropy (Kuhn et al., 2023; Farquhar et al., 2024). It is designed primarily to cater to the open-text answers. However, in the MCQA settings, semantic entropy boils down to Output

entropy (OE) (Appendix A.7).

5.4 Evaluation metric

An uncertainty measure should correlate with the probability of a prediction being incorrect. The performance of uncertainty methods is evaluated using Area-Under-Receiver-Operating-Curve (AUROC):

$$\texttt{AUROC} = \texttt{AUC}\left(\frac{1}{U}, \mathbb{1}_{\{\hat{\mathbf{y}}=\mathbf{y}_{\texttt{target}}\}}\right)$$

where U is the uncertainty, $\mathbf{y}_{\texttt{target}}$ denotes the ground-truth, $\hat{\mathbf{y}}$ denotes the model output.

5.5 Equivalent Samples (FES)

Adding to the description provided in Section 3.1, FES samples are the transformations of the original input that keep the task objective of the original question unchanged ideally. Transformations used on images to create equivalent spatial reasoning samples are the RGB to grayscale transformation, adding Gaussian noise, and mild levels of blurring. Equivalent transformations on the text (the question) related to multiple para-phrasings. Different combinations of such image and text equivalent samples form the multimodal FES samples.

For the audio tasks, FES samples are transformations on the original input that maintain the task and functional equivalence. The transformations are done on both modalities: audio input and textual question. While some transformations, such as adding noise, varying the loudness, paraphrasing the textual question, etc, are generic, FES transformations can be task-specific too. For example, for the order task, where the goal is to answer which audio event occurred after the event A, changing the duration of event A by trimming is an FES transformation. However, this is an invalid FES transformation with respect to the event duration task. The complete description of equivalent sampling (FES) transforms used along with the sampling process, and a few examples are detailed in Appendix A.5 and Figure 7.

5.6 Complementary samples (FCS)

Based on the discussion in Section 3.2, FCS samples are the transformations of the original input that keep the task objective of the original question same, but alter the direction of the spatial/temporal reasoning. In an ideal setting, FCS samples force the MLLM to alter its response with respect to the original input.

For example, for audio tasks, a complementary transform can be the addition of a new audio event at the start or end of the input audio clip for the audio event counting task. For images, it can be flipping an image horizontally for a "at the left of" or "at the right of" type of questions. For the text, it can be altering the spatial relationship from "in front of" to "at the behind of". The FCS samples used in Table 1 use complemented text, as generalized complementary transform on images is difficult. However, as proof of the concept, we generate complementary image samples for the "left of" and "right of" questions, and note consistent improvements (Appendix A.6). The FCS transforms used are in Appendix A.5 and Figure 7.

6 Results

6.1 FESTA uncertainty evaluation

The performance of FESTA and other baselines on vision-LLMs and audio-LLMs are reported in Tables 1 and 2. We make the following observations:

- The abstention AUROC of FESTA uncertainty score outperforms all the black-box baseline approaches significantly, achieving 24.6% and 41.9% relative improvements over the second-best approaches on BLINK and VSR datasets, respectively, averaged over different models. Also, for audio-LLMs, it achieves 25.4%, 30.5%, and 32.8% relative improvements for order, duration and count tasks, respectively.
- Although the audio-LLM performances were poor for temporal reasoning (accuracy: order - 52%, duration - 43% and count - 30%), the AUROC achieved by FESTA of 0.89, 0.77, and 0.77 respectively, shows its effectiveness for the low-performing reasoning tasks.
- For vision-LLMs, FESTA shows effectiveness across low and high accuracy models. The largest improvements were observed for `Phi-4` (the model with the least size of 5.6B).
- For both audio-LLMs and vision-LLMs, the baseline systems perform inconsistently. None of them consistently provide second-best performance across models and tasks.

6.2 Ablations

To further probe the improvements provided by FESTA, we conduct the following analyses:

- FESTA uncertainty is a quantification of both equivalent and complementary input sam-

Dataset	Model	Pred.	Baseline Results (AUROC)							Ours (AUROC)
		Acc.	OE	VC	IA-I	IA-T	IA-IT	RU	BU	FESTA
BLINK	Gemma-3	0.80	0.53	0.61	0.55	0.68	0.63	0.71	0.69	**0.81** (14.1%)
	LLaVA-1.6	0.71	0.67	0.56	0.47	0.62	0.63	0.62	0.51	**0.77** (14.9%)
	Qwen-2.5-VL	0.88	0.86	0.65	0.77	0.78	0.80	0.77	0.60	**0.93** (8.1%)
	Phi-4	0.71	0.63	0.49	0.56	0.60	0.64	0.58	0.38	**0.87** (35.9%)
	Pixtral	0.78	0.72	0.59	0.75	0.70	0.73	0.75	0.58	**0.90** (20.0%)
	Avg.	0.78	0.68	0.58	0.62	0.68	0.69	0.69	0.55	**0.86** (24.6%)
VSR	Gemma-3	0.74	0.53	0.56	0.60	0.58	0.63	0.59	0.66	**0.88** (33.3%)
	LLaVA-1.6	0.60	0.57	0.52	0.57	0.58	0.66	0.56	0.53	**0.74** (12.1%)
	Qwen-2.5-VL	0.95	0.65	0.47	0.61	0.65	0.61	0.63	0.59	**0.92** (41.5%)
	Phi-4	0.68	0.58	0.49	0.48	0.50	0.60	0.50	0.56	**0.94** (56.7%)
	Pixtral	0.76	0.57	0.59	0.55	0.62	0.60	0.68	0.61	**0.91** (33.8%)
	Avg.	0.75	0.58	0.53	0.56	0.59	0.62	0.59	0.59	**0.88** (41.9%)

Table 1: Results for vision-LLMs. The final column in green (%) reports the relative improvement of FESTA over the best baseline result (underlined).

Dataset	Model	Pred.	Baseline Results (AUROC)							Ours (AUROC)
		Acc.	OE	VC	IA-A	IA-T	IA-AT	RU	BU	FESTA
TREA-O	Qwen2-Audio	0.51	0.66	0.67	0.65	0.68	0.63	0.70	0.53	**0.91** (30.0%)
	SALMONN	0.31	0.54	0.54	0.55	0.55	0.66	0.64	0.39	**0.86** (30.3%)
	SALMONN desc. + LLM	0.75	0.68	0.74	0.83	0.64	0.85	0.68	0.65	**0.90** (5.8%)
	Avg.	0.52	0.63	0.65	0.68	0.62	0.71	0.67	0.52	**0.89** (25.4%)
TREA-D	Qwen2-Audio	0.45	0.61	0.75	0.55	0.62	0.59	0.66	0.56	**0.75** (0.0%)
	SALMONN	0.35	0.54	0.48	0.50	0.56	0.50	0.54	0.50	**0.76** (35.7%)
	SALMONN desc. + LLM	0.49	0.57	0.49	0.67	0.55	0.69	0.56	0.54	**0.80** (15.9%)
	Avg.	0.43	0.57	0.57	0.57	0.58	0.59	0.59	0.53	**0.77** (30.5%)
TREA-C	Qwen2-Audio	0.21	0.49	0.68	0.48	0.47	0.45	0.47	0.50	**0.83** (22.1%)
	SALMONN	0.20	0.34	0.40	0.46	0.29	0.32	0.27	0.43	**0.66** (43.5%)
	SALMONN desc. + LLM	0.50	0.61	0.66	0.54	0.54	0.45	0.55	0.62	**0.81** (22.7%)
	Avg.	0.30	0.48	0.58	0.49	0.43	0.41	0.43	0.52	**0.77** (32.8%)

Table 2: Results for audio-LLMs using Qwen-2-Audio (Chu et al., 2024), SALMONN (Tang et al., 2023) and audio captions generated by SALMONN followed by text-only Qwen-2 model (LLM). The final column in green (%) reports the relative improvement of FESTA over the best baseline result (underlined).

plings. We separately analyze the AUROCs from equivalent (FES) and complementary (FCS) samples, as shown in Figures 3 and 4. For the order and duration tasks in audio reasoning, AUROC for Qwen2-Audio is more influenced by the FCS, but SALMONN and SALMONN desc.+LLM models are more influenced by the FES. It shows the complementary nature of FES and FCS inputs under different scenarios. For the poorly performing event count task, FCS uncertainty contributes the most, showing its robustness in such low-accuracy scenarios.

- Appendix Figure 5 shows the scatter plot of the confidence scores ($\frac{1}{\text{uncertainty}}$). Figure 6 shows the scatter plot of the detection scores for the output sampling baseline approach (OE). It is evident that poor baseline AUROCs for different models are primarily caused by low uncertainty mis-predictions, which are majorly detected using FESTA.

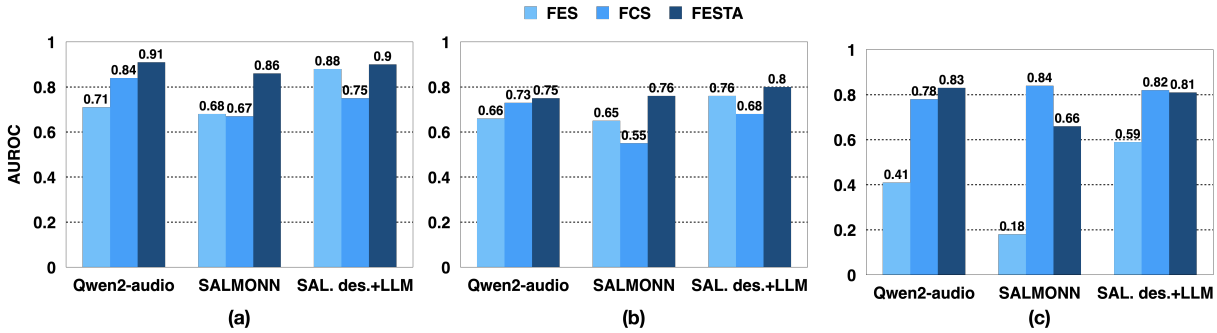


Figure 3: AUROC for uncertainty based on FES, FCS and FESTA on (a) TREA-O, (b) TREA-D, and (c) TREA-C.

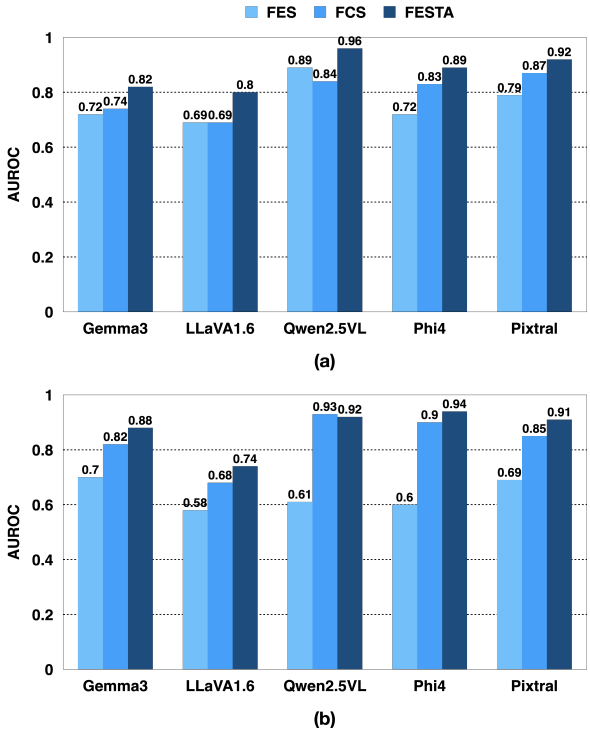


Figure 4: AUROC for uncertainty based on FES, FCS and FESTA on (a) BLINK, (b) VSR data.

- We also compare the effectiveness of our proposed uncertainty quantification method, which uses the KL-divergence distance from the predictive distribution of an ideally consistent and sensitive model. To compare, we compute the AUROCs by replacing the KL-divergence with the standard entropy measure and the results are reported in Appendix Tables 10 and 11. They highlight the superiority of using the KL distance in a pseudo-supervised fashion, over simply the entropy.

6.3 Number of equivalent samples

For the vision tasks, the value of K was varied from 8 to 112, including FES and FCS samples, while for audio tasks, K was varied from 10 to 120 (Appendix Tables 5, 6, 7, 8 and 9). The results from this analysis highlight that, even with $K = 16$ (vision tasks) and $K = 20$ (audio tasks), FESTA can provide a robust estimate of model uncertainty.

7 Conclusion

The proposed multimodal uncertainty estimation algorithm, FESTA, is presented as a principled and formally grounded framework for selective prediction by MLLMs. Firstly, it introduces a novel input sampling paradigm based on functional equivalence (FES) and complementarity (FCS), a previously unexplored space, and demonstrates its effectiveness for abstention in the unsupervised and black-box settings. Secondly, it proposes the KL-distance from a hypothetically ideal model as an uncertainty score over the widely used entropy.
FESTA enables accurate abstention from incorrect predictions in both audio-LLMs and vision-LLMs, elicited by improved AUROC values. Given the limited grounding of current MLLMs (especially lightweight MLLMs), they tend to produce biased, low-uncertainty hallucinations. A key contribution of FESTA is its ability to detect and abstain from such hallucinations. This acts as the core reason behind the substantial improvements in selective prediction performance across models.
Building on these findings, we plan to extend FESTA to support open-ended generation and multimodal outputs in MLLMs, beyond MCQA tasks.

8 Acknowledgements

We thank TCS for supporting this work through the TCS Research Scholar Program and Google for providing compute grants through the GCP Credit Award.

9 Limitations

Despite its effectiveness, the current formulation of FESTA has a few limitations and opens several avenues for future work:

- **Computational Overhead:** Unlike standard LLMs, FESTA relies on input samples from both text and non-text modalities, increasing computational demand. This also means that about K additional inferences are made per sample, thereby, increasing the computational demand for computing the uncertainty metric. While the current implementation prioritizes predictive performance—suitable for high-stakes scenarios such as safety-critical applications—reducing latency using FES/FCS samples passed through quantized versions of the models may be undertaken as future research to further reduce the computational demand.
- **Generation of FES and FCS:** The audio and visual samples for FES and FCS are generated using standard audio/image augmentation tools as well as text rephrasing using LLMs, with limited human intervention. While this process is currently semi-automated with limited human supervision and performed seamlessly for the tasks considered in this work, fully automating the process of FES and FCS generations for more complex audio/image tasks might require additional work.
- **Improving LLM Behavior with FES and FCS:** In the current work, the FES and FCS were used to analyze the uncertainty, without attempting to improve the base model performance. However, the dual-space sampling strategy (functional equivalence and complementarity) could also be incorporated as in-context prompts to nudge the LLMs to ensure equivalence/complementarity in the outputs, thereby improving base LLM accuracy. For example, one could input K FES and prompt the LLM to ensure that the answer to the original prompt should also match the response to all the K FES. With this constraint, the LLM would be forced to generate a consistent response that matches the response to the FES, thereby improving the prediction accuracy.
- **Extension to open-text Generation:** In the current work, the scope was limited to multi-choice question answering tasks only. Many real-world applications of multimodal LLMs require open-ended, free-form text outputs. A significant future direction is to extend FESTA beyond classification tasks, enabling uncertainty-aware abstention in generative settings. Further, audio and image generation tasks open up new avenues for uncertainty estimation, which is not addressed in this work.

References

Yasin Abbasi-Yadkori, Ilja Kuzborskij, András György, and Csaba Szepesvari. 2024. To believe or not to believe your LLM: Iterativeprompting for estimating epistemic uncertainty. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, and 1 others. 2024. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*.

Pravesh Agrawal, Szymon Antoniak, Emma Bou Hanna, Baptiste Bout, Devendra Chaplot, Jessica Chudnovsky, Diogo Costa, Baudouin De Monicault, Saurabh Garg, Theophile Gervet, and 1 others. 2024. Pixtral 12b. *arXiv preprint arXiv:2410.07073*.

Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. 2016. Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*.

Yuval Bahat and Gregory Shakhnarovich. 2020. Classification confidence estimation with test-time data-augmentation. *arXiv preprint arXiv:2006.16705*.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Debarpan Bhattacharya, Apoorva Kulkarni, and Sriram Ganapathy. 2025. Benchmarking and Confidence Evaluation of LALMs For Temporal Reasoning. In *Interspeech 2025*, pages 2068–2072.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901.

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

Zijun Chen, Wenbo Hu, Guande He, Zhijie Deng, Zheng Zhang, and Richang Hong. 2025. Unveiling uncertainty: A deep dive into calibration and performance

of multimodal large language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 3095–3109.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, and 1 others. 2024. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. 2024. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630.

Alessandro Favero, Luca Zancato, Matthew Trager, Siddharth Choudhary, Pramuditha Perera, Alessandro Achille, Ashwin Swaminathan, and Stefano Soatto. 2024. Multi-modal hallucination control by visual information grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14303–14312.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, and Rongrong Ji. 2024a. Mme: A comprehensive evaluation benchmark for multimodal large language models. *Preprint*, arXiv:2306.13394.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith, Wei-Chiu Ma, and Ranjay Krishna. 2024b. Blink: Multimodal large language models can see but not perceive. In *European Conference on Computer Vision*, pages 148–166. Springer.

Xiang Gao, Jiaxin Zhang, Lalla Mouatadid, and Kamalika Das. 2024. SPUQ: Perturbation-based uncertainty quantification for large language models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2336–2346, St. Julian's, Malta. Association for Computational Linguistics.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*. PMLR.

Dan Hendrycks, Steven Basart, Mantas Mazeika, Mohammadreza Mostajabi, Jacob Steinhardt, and Dawn Song. 2022. Scaling out-of-distribution detection for real-world settings. *arXiv preprint arXiv:2208.09143*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021a. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

Dan Hendrycks, Nicholas Carlini, John Schulman, and Jacob Steinhardt. 2021b. Unsolved problems in ml safety. *arXiv preprint arXiv:2109.13916*.

Bairu Hou, Yujian Liu, Kaizhi Qian, Jacob Andreas, Shiyu Chang, and Yang Zhang. 2024. Decomposing uncertainty for large language models through input clarification ensembling. In *International Conference on Machine Learning*, pages 19023–19042. PMLR.

Mingjian Jiang, Yangjun Ruan, Sicong Huang, Saifei Liao, Silviu Pitis, Roger Baker Grosse, and Jimmy Ba. 2023. Calibrating language models via augmented prompt ensembles. In *International Conference on Machine Learning*.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, and 1 others. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.

Seil Kang, Jinyeong Kim, Junhyeok Kim, and Seong Jae Hwang. 2025. See what you are told: Visual attention sink in large multimodal models. In *The Thirteenth International Conference on Learning Representations*.

Zaid Khan and Yun Fu. 2024. 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*, pages 10854–10863.

Chun-Yi Kuan and Hung-yi Lee. 2025. Can large audio-language models truly hear? tackling hallucinations with multi-task assessment and stepwise audio reasoning. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *The Eleventh International Conference on Learning Representations*.

Lukas Kürbis, Luke Ciolek, Daniel Beck, and Iryna Gurevych. 2024. Uncertainty quantification for large language models: A survey. *arXiv preprint arXiv:2402.16120*.

Shawn Li, Jiashu Qu, Yuxiao Zhou, Yuehan Qin, Tiankai Yang, and Yue Zhao. 2025. Treble counterfactual vlms: A causal approach to hallucination. *Preprint*, arXiv:2503.06169.

Shengzhi Li, Rongyu Lin, and Shichao Pei. 2024. Multimodal preference alignment remedies degradation of visual instruction tuning on language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14188–14200.

Yifei Li, Zhou Zhao, Xiaohan Yu, and Deng Cai. 2023. Multimodal uncertainty estimation for deep learning models. *arXiv preprint arXiv:2304.02637*.

Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. 2024. Generating with confidence: Uncertainty quantification for black-box large language models. *Transactions on Machine Learning Research*.

Chen Ling, Xujiang Zhao, Xuchao Zhang, Wei Cheng, Yanchi Liu, Yiyou Sun, Mika Oishi, Takao Osaki, Katsushi Matsuda, Jie Ji, and 1 others. 2024. Uncertainty quantification for in-context learning of large language models. *arXiv preprint arXiv:2402.10189*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, and 1 others. 2024. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pages 216–233. Springer.

Nishanth Madhusudhan, Sathwik Tejaswi Madhusudhan, Vikas Yadav, and Masoud Hashemi. 2025. Do llms know when to not answer? investigating abstention abilities of large language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 9329–9345.

OpenAI. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

S Sakshi, Utkarsh Tyagi, Sonal Kumar, Ashish Seth, Ramaneswaran Selvakumar, Oriol Nieto, Ramani Duraiswami, Sreyan Ghosh, and Dinesh Manocha. 2025. MMAU: A massive multi-task audio understanding and reasoning benchmark. In *The Thirteenth International Conference on Learning Representations*.

Ola Shorinwa, Zhiting Mei, Justin Lidard, Allen Z Ren, and Anirudha Majumdar. 2025. A survey on uncertainty quantification of large language models: Taxonomy, open research challenges, and future directions. *ACM Computing Surveys*.

Julio Silva-Rodríguez, Ismail Ben Ayed, and Jose Dolz. 2025. Conformal prediction for zero-shot models. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pages 19931–19941.

Adi Simhi, Itay Itzhak, Fazl Barez, Gabriel Stanovsky, and Yonatan Belinkov. 2025. Trust me, i'm wrong: High-certainty hallucinations in llms. *arXiv preprint arXiv:2502.12964*.

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. 2023. Salmonn: Towards generic hearing abilities for large language models. *arXiv preprint arXiv:2310.13289*.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, and 1 others. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

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

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timoth ’ee Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Thinh Hung Truong, Timothy Baldwin, Karin Verspoor, and Trevor Cohn. 2023. Language models are not naysayers: an analysis of language models on negation benchmarks. In *Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (* SEM 2023)*, pages 101–114.

Xunzhi Wang, Zhuowei Zhang, Gaonan Chen, Qiongyu Li, Bitong Luo, Zhixin Han, Haotian Wang, Hang Gao, Mengting Hu, and 1 others. 2024. Ubench: Benchmarking uncertainty in large language models with multiple choice questions. *arXiv preprint arXiv:2406.12784*.

Yuxin Wang, Jason Wei, and et al. 2023. Llm-as-a-judge: Are large language models reliable evaluators? In *NeurIPS*.

Zhiyuan Wang, Qingni Wang, Yue Zhang, Tianlong Chen, Xiaofeng Zhu, Xiaoshuang Shi, and Kaidi Xu. 2025. SConU: Selective conformal uncertainty in large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 19052–19075, Vienna, Austria. Association for Computational Linguistics.

Bingbing Wen, Jihan Yao, Shangbin Feng, Chenjun Xu, Yulia Tsvetkov, Bill Howe, and Lucy Lu Wang. 2025. Know your limits: A survey of abstention in large language models. *Transactions of the Association for Computational Linguistics*, 13:529–556.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, YIFEI LI, Jie Fu, Junxian He, and Bryan Hooi. 2024. Can LLMs express their uncertainty? an empirical evaluation of confidence elicitation in LLMs. In *The Twelfth International Conference on Learning Representations*.

Yasin Abbasi Yadkori, Ilja Kuzborskij, David Stutz, András György, Adam Fisch, Arnaud Doucet, Iuliya Beloshapka, Wei-Hung Weng, Yao-Yuan Yang, Csaba Szepesvári, and 1 others. 2024. Mitigating llm hallucinations via conformal abstention. *arXiv preprint arXiv:2405.01563*.

Adam X. Yang, Chen Chen, and Konstantinos Pitas. 2024. Just rephrase it! uncertainty estimation in closed-source language models via multiple rephrased queries. In *Neurips Safe Generative AI Workshop 2024*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025a. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Yongjin Yang, Haneul Yoo, and Hwaran Lee. 2025b. MAQA: Evaluating uncertainty quantification in LLMs regarding data uncertainty. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 5846–5863, Albuquerque, New Mexico. Association for Computational Linguistics.

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

Gal Yona, Roee Aharoni, and Mor Geva. 2024. Can large language models faithfully express their intrinsic uncertainty in words? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7752–7764.

Ziyin Zhang, Zhaokun Jiang, Lizhen Xu, Hongkun Hao, and Rui Wang. 2024. Multiple-choice questions are efficient and robust llm evaluators. *arXiv preprint arXiv:2405.11966*.

Yunpu Zhao, Rui Zhang, Junbin Xiao, Ruibo Hou, Jiaming Guo, Zihao Zhang, Yifan Hao, and Yunji Chen. 2025. Object-level verbalized confidence calibration in vision-language models via semantic perturbation. *Preprint*, arXiv:2504.14848.

Zhuo Zhi, Chen Feng, Adam Daneshmend, Mine Orlu, Andreas Demosthenous, Lu Yin, Da Li, Ziquan Liu, and Miguel R. D. Rodrigues. 2025. Seeing and reasoning with confidence: Supercharging multimodal llms with an uncertainty-aware agentic framework. *Preprint*, arXiv:2503.08308.

A Appendix

A.1 Equivalence of input and FES samples

Proof of Equivalence relation for FES - We show that $\sim_{\mathcal{E}}$ satisfies the three properties of an equivalence relation.

Proposition A.1. *The relation $\sim_{\mathcal{E}}$ defined over inputs $\mathbf{X} \sim_{\mathcal{E}} \tilde{\mathbf{X}}$ such that $T(\mathbf{X}) = T(\tilde{\mathbf{X}})$ and $M_{\texttt{ideal}}(\mathbf{X}) = M_{\texttt{ideal}}(\tilde{\mathbf{X}})$ is an equivalence relation on the input space.*

Proof. **(Reflexivity):** For any input $\mathbf{X}$, we have

$$T(\mathbf{X}) = T(\mathbf{X}), \quad M_{\text{ideal}}(\mathbf{X}) = M_{\text{ideal}}(\mathbf{X}),$$

so $\mathbf{X} \sim_{\mathcal{E}} \mathbf{X}$.

(Symmetry): Suppose $\mathbf{X} \sim_{\mathcal{E}} \tilde{\mathbf{X}}$. Then,

$$T(\mathbf{X}) = T(\tilde{\mathbf{X}}), \quad M_{\text{ideal}}(\mathbf{X}) = M_{\text{ideal}}(\tilde{\mathbf{X}}).$$

By symmetry of equality, this implies:

$$T(\tilde{\mathbf{X}}) = T(\mathbf{X}), \quad M_{\text{ideal}}(\tilde{\mathbf{X}}) = M_{\text{ideal}}(\mathbf{X}),$$

so $\tilde{\mathbf{X}} \sim_{\mathcal{E}} \mathbf{X}$.

(Transitivity): Suppose $\mathbf{X} \sim_{\mathcal{E}} \tilde{\mathbf{X}}$ and $\tilde{\mathbf{X}} \sim_{\mathcal{E}} \hat{\mathbf{X}}$. Then,

$$T(\mathbf{X}) = T(\tilde{\mathbf{X}}), \quad M_{\text{ideal}}(\mathbf{X}) = M_{\text{ideal}}(\tilde{\mathbf{X}}),$$

and

$$T(\tilde{\mathbf{X}}) = T(\hat{\mathbf{X}}), \quad M_{\text{ideal}}(\tilde{\mathbf{X}}) = M_{\text{ideal}}(\hat{\mathbf{X}}).$$

By transitivity of equality, we get:

$$T(\mathbf{X}) = T(\hat{\mathbf{X}}), \quad M_{\text{ideal}}(\mathbf{X}) = M_{\text{ideal}}(\hat{\mathbf{X}}),$$

so $\mathbf{X} \sim_{\mathcal{E}} \hat{\mathbf{X}}$.

Hence, $\sim_{\mathcal{E}}$ is reflexive, symmetric, and transitive, and thus an equivalence relation. □

A.2 Equivalence between different FCS samples

Proposition A.2. *Let $\mathcal{C}_{\mathbf{X}} := \{\mathbf{X}' : T(\mathbf{X}') = T(\mathbf{X}),\ M_{ideal}(\mathbf{X}') \neq M_{ideal}(\mathbf{X})\}$ denote the set of functionally complementary samples of input $\mathbf{X}$.*

Define a relation $\sim_{\mathcal{C}}$ over this set such that:

$$\mathbf{X}_1 \sim_{\mathcal{C}} \mathbf{X}_2 \iff T(\mathbf{X}_1) = T(\mathbf{X}_2)$$

and

$$M_{ideal}(\mathbf{X}_1) = M_{ideal}(\mathbf{X}_2).$$

Then $\sim_{\mathcal{C}}$ is an equivalence relation over the set of functionally complementary samples $\mathcal{C}_{\mathbf{X}}$.

Proof. We verify the three properties of equivalence:

(Reflexivity): For any $\mathbf{X}_1 \in \mathcal{C}_{\mathbf{X}}$, clearly $T(\mathbf{X}_1) = T(\mathbf{X}_1)$ and $M_{\text{ideal}}(\mathbf{X}_1) = M_{\text{ideal}}(\mathbf{X}_1)$, so $\mathbf{X}_1 \sim_{\mathcal{C}} \mathbf{X}_1$.

(Symmetry): If $\mathbf{X}_1 \sim_{\mathcal{C}} \mathbf{X}_2$, then:

$$T(\mathbf{X}_1) = T(\mathbf{X}_2), \quad M_{\text{ideal}}(\mathbf{X}_1) = M_{\text{ideal}}(\mathbf{X}_2).$$

By symmetry of equality, the reverse also holds:

$$T(\mathbf{X}_2) = T(\mathbf{X}_1), \quad M_{\text{ideal}}(\mathbf{X}_2) = M_{\text{ideal}}(\mathbf{X}_1),$$

so $\mathbf{X}_2 \sim_{\mathcal{C}} \mathbf{X}_1$.

(Transitivity): If $\mathbf{X}_1 \sim_{\mathcal{C}} \mathbf{X}_2$ and $\mathbf{X}_2 \sim_{\mathcal{C}} \mathbf{X}_3$, then:

$$T(\mathbf{X}_1) = T(\mathbf{X}_2) = T(\mathbf{X}_3)$$

and

$$M_{\text{ideal}}(\mathbf{X}_1) = M_{\text{ideal}}(\mathbf{X}_2) = M_{\text{ideal}}(\mathbf{X}_3)$$

so $\mathbf{X}_1 \sim_{\mathcal{C}} \mathbf{X}_3$.

Thus, $\sim_{\mathcal{C}}$ is reflexive, symmetric, and transitive over $\mathcal{C}_{\mathbf{X}}$, and therefore an equivalence relation. □

A.3 FES uncertainty closed-form expression

Proposition A.3. *Then the KL divergence from the certain model to q_{FES} simplifies to:*

$$U_{FES}(M \mid \mathbf{x}) := -\log q_{FES}(y = \hat{y} \mid \mathbf{x}).$$

Proof. Recall that KL divergence between distributions $q_{\text{certain}}(y|x)$ and $q_{\text{FES}}(y|x)$ is defined as:

$$D_{\text{KL}}(q_{\text{certain}}(y|\mathbf{x}) \| q_{\text{FES}}(y|\mathbf{x})) = \sum_y q_{\text{certain}}(y|\mathbf{x}) \log \frac{q_{\text{certain}}(y|\mathbf{x})}{q_{\text{FES}}(y|\mathbf{x})}$$

Substituting $q_{\text{certain}}(y|\mathbf{x}) = \delta_{y,\hat{y}}$, we have:

$$\begin{aligned} D_{\text{KL}}(\delta_{y,\hat{y}} \,\|\, q_{\text{FES}}(y \mid \mathbf{x})) &= \sum_y \delta_{y,\hat{y}} \\ &\quad \cdot \log \frac{\delta_{y,\hat{y}}}{q_{\text{FES}}(y \mid \mathbf{x})} \\ &= \log \frac{1}{q_{\text{FES}}(y = \hat{y} \mid \mathbf{x})} \\ &= -\log\big(\\ &\quad q_{\text{FES}}(y = \hat{y} \mid \mathbf{x})\big) \end{aligned}$$

□

A.4 FCS for low-uncertainty hallucinations

We focus on he functional complementary sampling. We show that a hallucinating model has tendency to not react to the complementary transformations of the input. The proof sketch is provided below, for a single attention head.

Theorem A.4 (Negation Invariance in Hallucinating Attention Blocks)**.** *Consider a single-head attention block over input $\mathbf{X} = [\mathbf{X}_{\mathcal{D}}, \mathbf{X}_{\mathcal{N}}]$ with attention weights*

$$\alpha_{ij} := \text{softmax}_j\left(\frac{Q_i^\top K_j}{\sqrt{d_k}}\right), \qquad \text{Attn}_i := \sum_j \alpha_{ij} V_j.$$

Let negation act as $\mathbf{X}_{\mathcal{D}}^{\text{neg}} := f_{\text{neg}}(\mathbf{X}_{\mathcal{D}}), \mathbf{X}_{\mathcal{N}}^{\text{neg}} := \mathbf{X}_{\mathcal{N}}$.

Under two hallucination models:

(1) Missing attention: *$\sum_{j\in\mathcal{D}} \alpha_{ij} \ll \sum_{j\in\mathcal{N}} \alpha_{ij}$,*

(2) Over-reliance on prior: *$K_j \approx K^*, V_j \approx V^*$, independent of $\mathbf{X}$,*

the attention output satisfies:

$$\text{Attn}_i^{\text{neg}} \approx \text{Attn}_i^{\text{orig}},$$

and thus the prediction remains unchanged under negation.

A.5 Details of FES and FCS

The generic functionally equivalent transforms applied to image data include:

- **Contrast:** Adjusts image contrast to simulate lighting variation.
- **Blur:** Applies slight blurring to the image
- **Noise:** Adding a small amount of noise
- **Masking:** Hides a few random pixels
- **Rotate:** Rotates the image slightly to simulate viewpoint changes.
- **Shift:** Translates the image slightly in space.
- **Greyscale:** Removes color information while preserving structure.

The generic functionally equivalent transforms applied to audio data include:

- **Silence:** Adding small duration of silence in between sound events
- **Volume:** Minor adjustment to the volume of different sound events

Table 3: Summary of Notations Used in FESTA

Notation	Description
$X = [X_O, X_T]$	Multimodal input (Non-text and text modalities)
Y	Ground truth
$T(X)$	Task to be performed to answer X
$q(y\|X)$	Predictive distribution of model outputs given X
$\hat{y}$	Model's predicted output, $\hat{y} = \arg\max_{y \in Y} q(y\|X)$
$\tilde{X} \sim P_{\text{FES}}(\tilde{X}\|X)$	Functionally equivalent samples (FES) from X
$X' \sim P_{\text{FCS}}(X'\|X)$	Functionally complementary samples (FCS) from X
$\mathcal{M}_{\text{ideal}}$	Ideal model achieving task objective perfectly
$\mathcal{M}_{\text{consistent}}$	A perfectly consistent model under FES
$\mathcal{M}_{\text{sensitive}}$	A perfectly sensitive model under FCS
$q_{\text{FES}}(y\|X)$	Predictive distribution over FES samples
$q_{\text{FCS}}(y\|X)$	Predictive distribution over FCS samples
U_{FES}	Uncertainty estimate from FES
U_{FCS}	Uncertainty estimate from FCS
U_{FESTA}	Combined FESTA uncertainty estimate
S	Finite set of output sequences
S'	Modified predictive support ($\{\hat{y}, \hat{y}^c\}$)

The generic functionally equivalent transforms applied to text data include:

- **Rephrase:** Paraphrasing the question such that the meaning remains unchanged

Functionally Complementary Transformation for image-text datasets is done by negating the textual question such that the answer changes.

Functionally Complementary Transformation for audio-text datasets is done by negating the audio such that the answer changes. This is task-specific.

- **Count:** Adding new sound events to the original audio
- **Duration:** Replace the longest or shortest sound event in the audio with a sound event not originally present.
- **Order:** Swap the positions of the sound events in the audio

Examples of Functionally Equivalent Transform and Functionally Complementary Transform for both audio and image are given in 7.

A.6 Image-based FCS samples

We considered the complementary operation on images instead of on text, for the 'left-of' and 'right-of' questions (20 image-text pairs). The Table 4 contains the results on 3 models and on the BLINK dataset. It is interesting to note that image-based FCS may potentially offer further performance gains. Similar experiments on other models and on the VSR dataset will be taken up in future.

Model	ROC-AUC (text FCS)	ROC-AUC (image FCS)
Qwen-2.5VL	0.59	0.91
LLaVA-1.6	0.66	0.57
Phi-4	0.92	0.98
Mean	**0.72**	**0.82**

Table 4: Comparison of the performance of image complementary transform-based FCS samples vs text-based FCS samples, from the BLINK dataset.

A.7 Semantic entropy in MCQA setting

Semantic Entropy (SE) (Kuhn et al., 2023; Farquhar et al., 2024) for an input x is defined over semantic clusters $c \in \mathcal{C}$ of free-text outputs s as

$$\begin{aligned} SE(x) &= -\sum_{c \in \mathcal{C}} p(c \mid x) \, \log p(c \mid x) \\ &= -\sum_{c \in \mathcal{C}} \left(\sum_{s \in c} p(s \mid x) \right) \\ &\quad \cdot \log \left(\sum_{s \in c} p(s \mid x) \right). \end{aligned}$$

where (i) s are generated sequences (answers), (ii) c denotes a semantic cluster of sequences with the same meaning, and (iii) $p(s \mid x)$ is the model probability of sequence s given x. (All logarithms are natural.)

MCQ setting: Suppose there are exactly K fixed answer choices $a_1, \ldots, a_K$. The standard (Shannon) entropy of the predictive distribution over these choices is

$$H(x) \;=\; -\sum_{i=1}^{K} p(a_i \mid x)\, \log p(a_i \mid x).$$

In the MCQ scenario, each answer choice is semantically distinct and explicitly enumerated. Consequently:

- The number of semantic clusters equals the number of choices: $|\mathcal{C}| = K$.
- There is a one-to-one correspondence $c_i \leftrightarrow a_i$.
- Hence, for all $i \in \{1, \ldots, K\}$, $p(c_i \mid x) = p(a_i \mid x)$.

Substituting this identification into the definition of semantic entropy,

$$\begin{aligned} SE(x) &= -\sum_{c \in \mathcal{C}} p(c \mid x)\, \log p(c \mid x) \\ &= -\sum_{i=1}^{K} p(a_i \mid x)\, \log p(a_i \mid x) \\ &= H(x). \end{aligned}$$

Conclusion: In the MCQ setting with fixed choices, semantic entropy reduces exactly to standard entropy.

A.8 Hyperparameters

FESTA has the minimal number of hyperparameters- only the number of samples to be used. This makes it easily deployable and devoid of heavy tuning. We have used the same number of equivalent (K_1) and complementary (K_2) samples ($K = K_1 = K_2$).
Vision-LLMs: For vision LLMs, $K_1 = K_2 = 56$ is used. Within modalities, for each multimodal data point, 14 image samplings ($K_{11} = 14$) and 4 text samplings are used ($K_{12} = 4$).
Audio-LLMs: For vision LLMs, $K_1 = K_2 = 60$ is used. Within modalities, for each multimodal data point, 15 audio samplings ($K_{11} = 15$) and 4 text samplings are used ($K_{12} = 4$).

A.9 Notations

The symbols and their meanings are noted in Table 3.

A.10 Model details and License

The models below are used as per the suggested guidelines and only for research purposes.
Gemma-3: The $12B$ model is used from https://huggingface.co/google/gemma-3-12b-it with license [2].
LLaVa-1.6: The $7B$ model is used from https://huggingface.co/llava-hf/llava-v1.6-mistral-7b-hf with license [3].
Phi-4: The $5.6B$ model is used from https://huggingface.co/microsoft/Phi-4-multimodal-instruct with license [4].
Pixtral: The $12B$ model is used from https://huggingface.co/mistralai/Pixtral-12B-2409 with license [5].
Qwen-2.5VL: The $7B$ model is used from https://huggingface.co/Qwen/Qwen2.5-VL-7B-Instruct with license [6].
Qwen2-Audio: The $7B$ model is used from https://huggingface.co/Qwen/Qwen2-Audio-7B-Instruct with license [7].
SALMONN: The $13B$ model is used from https://github.com/bytedance/SALMONN with license [8].
Qwen-2: The $7B$ text-only LLM in Table 2 is used from https://huggingface.co/Qwen/Qwen2-7B-Instruct with license [9].

A.11 Computation budget and hardware

We have used 8 Nvidia RTX A6000 GPU cards for all our experiments.

A.12 Number of Equivalent and complementary Samples

The results for varying the number of samples K are given in Tables 5, 7, 8, 9,and 6.

A.13 Usage of AI assitants

We have used ChatGPT, only restricted to section summarization and for textual rewriting of parts of the paper.

[2] https://ai.google.dev/gemma/terms
[3] http://www.apache.org/licenses/LICENSE-2.0
[4] https://huggingface.co/microsoft/Phi-4-multimodal-instruct/resolve/main/LICENSE
[5] http://www.apache.org/licenses/LICENSE-2.0
[6] http://www.apache.org/licenses/LICENSE-2.0
[7] http://www.apache.org/licenses/LICENSE-2.0
[8] http://www.apache.org/licenses/LICENSE-2.0
[9] http://www.apache.org/licenses/LICENSE-2.0

Sample_Size	2*4	2*8	2*12	2*16	2*20	2*24	2*28	2*32	2*36	2*40	2*44	2*48	2*52	2*56
Gemma-3	0.79	0.82	0.81	0.81	0.81	0.82	0.81	0.81	0.81	0.81	0.81	0.81	0.81	0.81
LLaVa	0.76	0.77	0.74	0.76	0.78	0.76	0.76	0.76	0.77	0.77	0.77	0.77	0.77	0.77
Phi4	0.85	0.86	0.87	0.87	0.87	0.87	0.87	0.87	0.87	0.87	0.87	0.87	0.87	0.87
Pixtral	0.90	0.90	0.89	0.90	0.90	0.91	0.90	0.91	0.90	0.90	0.90	0.90	0.90	0.90
Qwen-2.5-VL	0.92	0.92	0.92	0.92	0.93	0.93	0.92	0.93	0.92	0.93	0.92	0.93	0.93	0.93

Table 5: AUC performance of different models across increasing sample sizes on BLINK dataset.

Sample_Size	2*4	2*8	2*12	2*16	2*20	2*24	2*28	2*32	2*36	2*40	2*44	2*48	2*52	2*56
Gemma-3	0.85	0.87	0.88	0.89	0.88	0.88	0.88	0.88	0.88	0.88	0.88	0.88	0.88	0.88
LLaVa	0.67	0.72	0.69	0.72	0.72	0.70	0.73	0.74	0.73	0.75	0.76	0.74	0.73	0.74
Phi4	0.93	0.94	0.93	0.94	0.94	0.93	0.94	0.94	0.94	0.94	0.94	0.94	0.94	0.94
Pixtral	0.88	0.89	0.91	0.90	0.91	0.90	0.90	0.90	0.90	0.91	0.91	0.91	0.91	0.91
Qwen-2.5-VL	0.91	0.88	0.93	0.92	0.92	0.92	0.91	0.93	0.91	0.91	0.92	0.92	0.92	0.92

Table 6: AUC performance of different models across increasing sample sizes on VSR dataset.

Sample_Size	2*5	2*10	2*15	2*20	2*25	2*30	2*35	2*40	2*45	2*50	2*55	2*60
Qwen2-audio	0.81	0.84	0.83	0.81	0.83	0.83	0.84	0.84	0.83	0.83	0.84	0.83
SALMONN	0.55	0.65	0.59	0.65	0.64	0.64	0.64	0.65	0.65	0.63	0.67	0.66
SALMONN desc+LLM	0.80	0.81	0.80	0.80	0.80	0.82	0.81	0.81	0.81	0.81	0.81	0.81

Table 7: AUC performance for the Audio Event Counting task.

Sample_Size	2*5	2*10	2*15	2*20	2*25	2*30	2*35	2*40	2*45	2*50	2*55	2*60
Qwen2-audio	0.73	0.75	0.77	0.75	0.75	0.75	0.75	0.75	0.75	0.75	0.75	0.75
SALMONN	0.74	0.72	0.76	0.79	0.74	0.76	0.77	0.75	0.74	0.77	0.77	0.76
SALMONN desc+LLM	0.79	0.80	0.80	0.80	0.79	0.80	0.79	0.79	0.79	0.79	0.79	0.80

Table 8: AUC performance for the Audio Event Duration task.

Sample_Size	2*5	2*10	2*15	2*20	2*25	2*30	2*35	2*40	2*45	2*50	2*55	2*60
Qwen2-audio	0.91	0.92	0.91	0.91	0.91	0.91	0.91	0.91	0.91	0.91	0.91	0.91
SALMONN	0.81	0.83	0.83	0.86	0.84	0.86	0.84	0.85	0.86	0.86	0.86	0.86
SALMONN desc+LLM	0.89	0.90	0.90	0.90	0.90	0.90	0.90	0.90	0.90	0.90	0.90	0.90

Table 9: AUC performance for the Audio Event Ordering task.

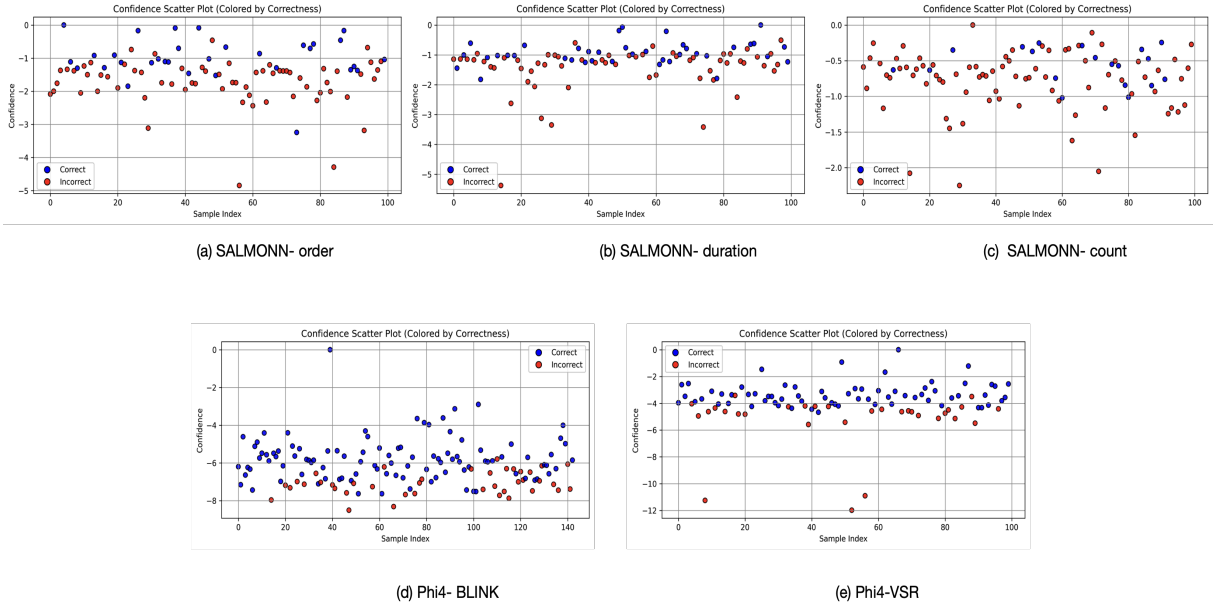


Figure 5: FESTA log(score) plots for best improvement models where score is reciprocal of FESTA uncertainty.

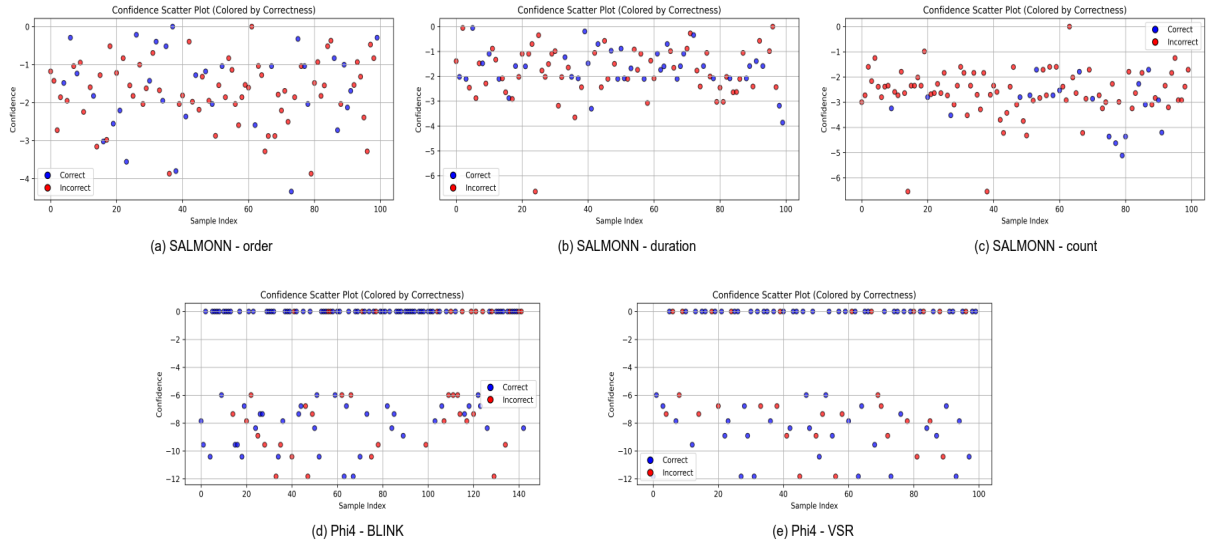


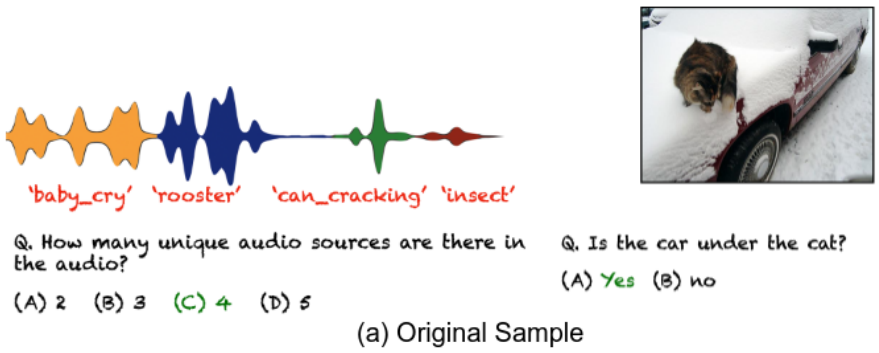
Figure 6: FESTA log(score) plots for output sampling baseline where score is reciprocal of FESTA uncertainty.

Model	**TREA-O**		**TREA-D**		**TREA-C**	
	Entropy	KL-div	Entropy	KL-div	Entropy	KL-div
Qwen2-Audio	0.59	0.91	0.67	0.75	0.38	0.83
SALMONN	0.58	0.86	0.60	0.76	0.27	0.66
SAL. des+LLM	0.76	0.90	0.73	0.80	0.63	0.81
Avg.	0.64	0.89 (39.1%)	0.67	0.77 (14.9%)	0.43	0.77 (79.1%)

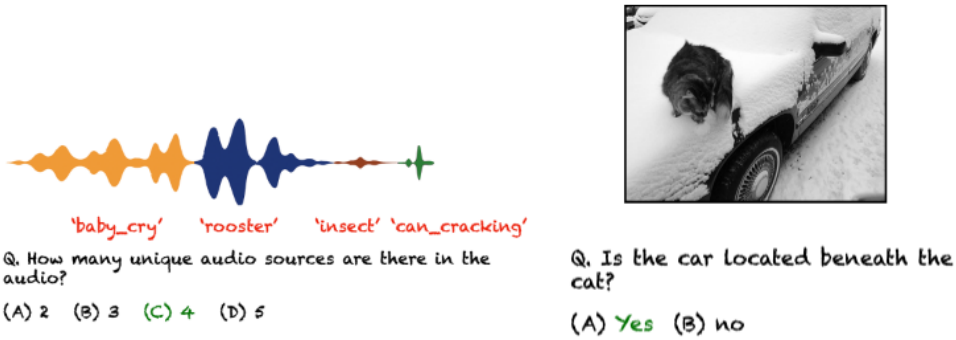
Table 10: Average performance of audio-LLMs using standard entropy measure compared with the proposed KL-div based measure.

Dataset	Gemma3	LLaVA1.6	Qwen2.5VL	Phi4	Pixtral	Avg.
BLINK (Entropy)	0.57	0.66	0.79	0.65	0.73	0.68
BLINK (KL-div)	0.81	0.77	0.93	0.87	0.90	0.86 (26.5%)
VSR (Entropy)	0.61	0.55	0.56	0.41	0.58	0.54
VSR (KL-div)	0.88	0.74	0.92	0.94	0.91	0.88 (63.0%)

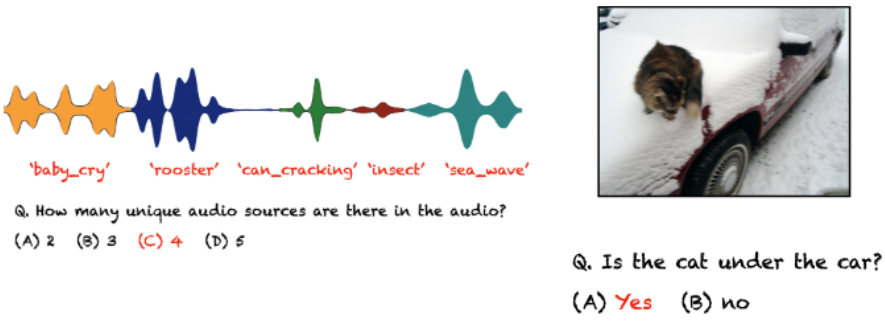
Table 11: Average performance of vision-LLMs using standard entropy measure compared with the proposed KL-div based measure.



(a) Original Sample



(b) Functionally Equivalent Transform



(c) Functionally Complementary Transform

Figure 7: Examples of Functionally Equivalent Transform and Functionally Complementary Transform for both audio-text and image-text questions.