FESTA: Functionally Equivalent Sampling for Trust Assessment of Multimodal LLMs

Anonymous ACL submission

Abstract

The accurate trust assessment of large language models (LLMs), 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), an input sampling technique for multimodal models, which generates an uncertainty measure based on the equivalent and complementary input sampling. The sampling approach expands the input space to measure the consistency (through equivalent samples) and sensitivity (through complementary samples) properties of the model. These two uncertainty measures are combined to form the final FESTA estimate. Our approach only requires black-box access, and is unsupervised. The experiments are conducted with various off-the-shelf multi-modal LLMs, on visual and audio reasoning tasks. The proposed FESTA approach is shown to significantly improve (33.3% relative improvement for vision-LLMs and 29.6% relative improvement for audio-LLMs) the area-under-receiver-operating-curve (AUROC) metric on these reasoning tasks.

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), yet their predictive uncertainty, especially in multimodal reasoning scenarios, remains poorly understood and inadequately quantified (Hendrycks et al., 2022; Li et al., 2023). Selective prediction (SP) is an area of work that attempts to prevent a model from making wrong predictions. In safety-critical scenarios like finance, medicine and autonomous driving, incorrect predictions are very expensive and selective prediction algorithms are highly sought after as part of building safe AI deployment (Amodei et al., 2016; Hendrycks et al., 2021). An efficient SP algorithm helps in providing low *selective risk* (high accuracy) with high *coverage* (the subset of questions for which the model chooses to answer). Further, trust assessment of models is important to increase the user confidence of large models in sensitive applications.

One of the most common approaches to SP is based on uncertainty measures like output entropy. The model predictions with high uncertainty are more prone to errors, and hence abstained from prediction. Several prior works have proposed entropy computation approaches for uncertainty estimation (Kuhn et al., 2023a; Farquhar et al., 2024; Ling et al., 2024).

The SP algorithms are critical when the LLM is operated on out-of-domain or low-resource regions of the input space, as the models have a higher tendency for erroneous generation. Previous studies have demonstrated that calibration typically deteriorates as model performance declines (Guo et al., 2017; Wang et al., 2023b; Kürbis et al., 2024). In some cases, the inaccurate predictions by multimodal LLMs may arise from their insensitivity to the input (referred as mode collapse). Such predictions have low predictive entropy and low accuracy.

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 existing output entropy-based measures are largely developed for text-based LLMs, (O2) Such measures have limited abstention capabilities, especially for challenging tasks like reasoning, (O3) The prior works may fail to abstain from low-entropy mis-predictions.

In this paper, we propose *functional equivalence sampling* and *functional complementary sampling* for multimodal LLMs, that generate samples based on their equivalence/complementarity to the original input and task objective. We assume the exis-

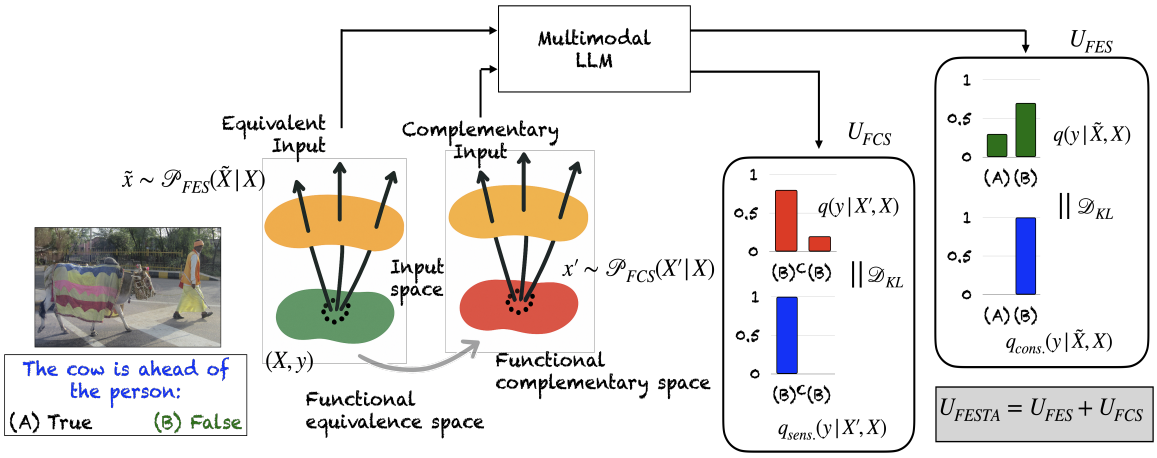


Figure 1: Schematic illustration of the proposed FESTA uncertainty quantification approach. Given a multimodal MCQ input, we generate functional equivalent samples (FES) and functional complementary samples (FCS). We compute divergence of model predictive uncertainty from an ideally consistent model (for FES) and sensitive model (for FCS) and combine these measure to generate the FESTA uncertainty score.

tence of such a functional space and explore their utility in quantifying uncertainty. We restrict this work to multiple-choice reasoning tasks involving audio/visual prompts. The key contributions are:

- We propose *functional equivalence sampling* (FES) and *functional complementary sampling* (FCS) to identify the consistency and sensitivity of model outputs - components which contribute to the uncertainty metric that are measurable for unsupervised and black-box model settings.
- Mathematical quantification of the proposed FESTA score as a KL-divergence from an ideally consistent and sensitive model, that improves over the standard entropy based measures.
- Effective uncertainty estimation for low uncertainty hallucinations, where the other baselines fail, through complementary sampling.
- Extensive performance benchmarking with various other prior works on audio/visual reasoning tasks with multiple open-source LLMs to illustrate the effectiveness of FESTA.

The schematic illustration of the proposed FESTA [1] approach is shown in Figure 1.

2 Problem statement

Let $\boldsymbol{X} = [\boldsymbol{X}_{\mathbf{O}}, \boldsymbol{X}_{\mathbf{T}}] \in \boldsymbol{\mathcal{X}}$ denote a multi-modal input instance (e.g., an audio/image + textual prompt) to a large language model (LLM) with ground-truth response $\mathbf{y}_{\texttt{target}}$, and let $\mathcal{S}$ denote the finite set of possible model outputs. While the FESTA approach is applicable to general multi-modal outputs, we restrict our discussion to the purely textual output case, in a multi-choice question-answering (MCQA) setting. Further, it is also assumed that the LLM is quite competitive in instruction following and generation capabilities, which limits the output sample space in MCQA settings. The LLM defines a predictive distribution over 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 predictive distribution $q(\mathbf{y}|\boldsymbol{X})$, estimate the predictive uncertainty of $\hat{\mathbf{y}} = \arg\max_{\mathbf{y} \in \mathcal{S}} q(\mathbf{y}|\boldsymbol{X})$ in a black-box setting.

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

- Using functionally equivalent sampling to estimate the epistemic uncertainty (consistency).
- Using functionally complementary sampling to measure counterfactual uncertainty (sensitivity).
- Combining the two measures to compute the FESTA measure.

3 FESTA Uncertainty Estimator

3.1 Functional equivalent samples (FES)

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

[1] The code and evaluation pipeline will be released upon paper acceptance.

Def: A transformation $\tilde{\mathbf{X}} = \mathcal{E}(\mathbf{X})$ is said to be 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:

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

Proposition 3.1. *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. The detailed proof is outlined in Appendix A.1. □

3.2 Functional complementary samples (FCS)

We formally define FCS below based on previously defined notations:

Def: A transformation $\mathbf{X}' = \mathcal{C}(\mathbf{X})$ is a *functionally complementary sample* of $\mathbf{X}$ if it is task-equivalent but functionally divergent, i.e.,

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

Similarly, we define a second distribution $\mathcal{P}_{\text{FCS}}(\mathbf{X}'|\mathbf{X})$ over all complementary transformations $(\{\mathcal{C}_i(\mathbf{X})\}_i)$. For example, negations or counter-factual transformations of the original inputs constitute the FCS, which should alter the ideal model prediction. We sample $\mathbf{X}'$ as,

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

Although a complementary sample is not equivalent to original input, it can be shown, similar to Proposition 3.1, that all complementary samples $\mathbf{X}'$ have formal equivalence among multiple samples generated from $\mathcal{P}_{\text{FCS}}$ (proof in Appendix A.2).

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, which do not rely on the underlying labels (unsupervised).

- **Consistency under FES**: The model $M_{\texttt{cons.}}$ generates predictions that remain consistent and unaltered under FES. We define the consistency entropy, $U_{\texttt{FES}}$, based on the deviation of a given model M from $M_{\texttt{cons.}}$.
- **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 define the complementary sensitivity entropy, $U_{\texttt{FCS}}$, based on the deviation of a given model M from $M_{\texttt{sens.}}$.

Note that, $M_{\texttt{ideal}} \subset M_{\texttt{sens.}}$ and $M_{\texttt{ideal}} \subset M_{\texttt{sens.}}$, i.e., an ideal model is both consistent and sensitive, while the converse is not true.

3.4 Uncertainty Estimation from FES

Given a specific input $\boldsymbol{X} = \mathbf{x}$, consider a set of its functional equivalent inputs: $\{\tilde{\mathbf{x}}_1, \tilde{\mathbf{x}}_2, \ldots, \tilde{\mathbf{x}}_{K_1}\}$, drawn from $\tilde{\mathbf{x}}_k \sim \mathcal{P}_{\text{FES}}$. Let $q(\mathbf{y}|\tilde{\mathbf{x}}_k)$ denote the predictive probability distribution of model M. Note that, the response $\mathbf{y}$ is a random sequence which can be sampled using stochastic decoding of LLMs. The uncertainty measure of the model M is measured as the deviation from the grounded model $M_{\texttt{cons.}}$.

- **Predictive distribution of $M_{\texttt{cons.}}$**: In our problem setting, the definition of $M_{\text{cons.}}$ implies that its predictive distribution is a Kronecker delta function with respect to equivalent sampling.

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

- **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 it's 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 3.2. *The KL divergence from the consistent model $M_{\texttt{cons.}}$ to q_{FES} simplifies to $(\mathbf{y} \in \mathcal{S})$:*

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

Proof. The proof follows from the definition of KL divergence, as shown in the Appendix A.3. □

It is interesting to note that, resorting to $M_{\text{cons.}}$ makes the uncertainty quantification solely based on model predictions, and makes it an unsupervised approach.

3.5 Uncertainty Estimation from FCS

We can formally conclude the properties with respect to the predictive distribution for complementary samples: $\{\mathbf{x}'_1, \mathbf{x}'_2, \ldots, \mathbf{x}'_{K_2}\}$, drawn from $\mathbf{x}'_k \sim \mathcal{P}_{\text{FCS}}$ as below:

- **Predictive distribution of $M_{\mathsf{sens.}}$:** Based on the definition of FCS, they alter the predictions - $M_{\mathsf{sens.}}(\mathbf{x}') \neq M_{\mathsf{sens.}}(\mathbf{x})$. 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 $\hat{\mathbf{y}}^c$.

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

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

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

Finally, similar to the case of equivalent samples (FES), we pose the uncertainty of model M computed from complementary samples (FCS) as:

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

Proposition 3.3. *The KL divergence 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 predictive distributions can be used in the definition of KL divergence and follows from the steps similar to Appendix A.3. □

Although the prior works on uncertainty based SP work under the hypothesis that incorrect predictions are associated with high uncertainty, LLMs are sometimes associated with biased mis-predictions. As a naive example, if a model predicts the same choice as the answer to MCQA, stochastic decoding or equivalence sampling does not provide the right framework for quantifying this misbehavior. However, U_{FCS} helps to abstain from such low-uncertainty mis-predictions. when the model predictions show no sensitivity to the complementary sampling. We formally show this for a single attention block in Appendix A.4.

Algorithm 1 FESTA Uncertainty Estimator

Require: Input $\boldsymbol{X}$, predictive distribution $q(\mathbf{y}|\boldsymbol{X})$, original prediction $\hat{\mathbf{y}} = \arg\max_{\mathbf{y}} q(\mathbf{y}|\boldsymbol{X})$, number of samples $K = K_1 + K_2$.

1: **FES Sampling:**
2: Generate K_1 functionally equivalent 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 $(K_1 = K_{11} \times K_{12})$.
3: Compute predictive distribution $q_{\text{FES}}(\mathbf{y}|\boldsymbol{X})$ using $\{\tilde{\mathbf{x}}_k\}_{k=1}^{K_1}$
4: Compute FES uncertainty for prediction $\hat{\mathbf{y}}$:

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

5: **FCS Sampling:**
6: Generate K_2 functionally complementary 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 finally using all combinations $(K_2 = K_{21} \times K_{22})$.
7: Compute predictive distribution $q_{\text{FCS}}(\mathbf{y}|\boldsymbol{X})$ for complementary samples $\{\mathbf{x}'_k\}_{k=1}^{K_2}$.
8: Compute FCS uncertainty:

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

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

3.6 FESTA uncertainty estimate

The FESTA uncertainty estimate finally combines the two axes of uncertainty quantification - U_{FES}, which measures the model response consistency across equivalent sampling, and U_{FCS}, which measures the model sensitivity to complementary sampling, by taking their sum. The combination helps to abstain from both high uncertainty and low uncertainty mis-predictions, as evident in Section 6. The FESTA approach is summarized in Algorithm 1.

4 Related Prior Work

White-box Uncertainty Estimation: Uncertainty estimation methods for Large Language Models (LLMs) can broadly be categorized into white-box and black-box approaches (Gawlikowski et al.,

2023). White-box methods leverage direct access to internal model parameters, gradients, or intermediate activations to quantify uncertainty (Kadavath et al., 2022; Wang et al., 2024). Such approaches, although potentially precise and informative, require internal model details, which are unavailable for closed-source models (Xiao et al., 2023).

Black-box Uncertainty Estimation: Black-box methods do not assume any access to model internals and typically rely on externally observable behaviors such as output distributions under varied prompts or input perturbations (Xiao et al., 2023; Jiang et al., 2024). One popular paradigm involves employing input paraphrases or perturbations and observing variations in the model's output distribution to estimate uncertainty (Kuhn et al., 2023b; Xiong et al., 2024b; Wang et al., 2023a). For instance, Kuhn et al. (Kuhn et al., 2023b) introduced input clarification ensembling, generating multiple clarified versions of the same input to measure predictive entropy and mutual information, effectively decomposing uncertainty into aleatoric and epistemic components. Similarly, Xiong et al. (Xiong et al., 2024b) proposed generating uncertainty estimates through paraphrasing queries.

Another line of work focuses on semantic entropy, quantifying uncertainty based on semantic coherence across generated responses. Tian et al. (Tian et al., 2023b) leveraged semantic entropy as an effective metric to detect hallucinations in model outputs. Wang et al. (Wang et al., 2023a) calibrated model predictions using augmented prompt ensembles, reducing overconfidence and enhancing the reliability of black-box LLMs without needing access to their internal structures. Likewise, Jiang et al. (Jiang et al., 2024) proposed a perturbation-based uncertainty estimation framework that systematically perturbs inputs and evaluates the model response variability.

Most of the prior works are purely based on text-only settings. Recently, uncertainty estimation for multimodal large language models (MLLMs) has gained attention. Huang et al. (Huang et al., 2023; Li et al., 2024) empirically analyzed uncertainty under multimodal scenarios, especially highlighting cases with misleading visual or textual cues, finding that multimodal integration can introduce unique dynamics that require separate evaluation. However, critical gaps remain regarding interpretability, robustness against adversarial perturbations, and efficient sampling strategies to maintain computational feasibility (Gawlikowski et al., 2023; Yin et al., 2023; Mielke et al., 2022).

5 Experimental Setup

5.1 Tasks and Datasets

We use 3 datasets on positional reasoning - 2 for vision-LLMs (BLINK, VSR) and 1 for audio-LLMs (TREA).

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

VSR: The VSR dataset (Liu et al., 2023) is analogous to the BLINK dataset, with binary questions. This is a dataset 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-reasoning dataset, focused on audio-temporal reasoning 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 main sub-categories - ordering, duration, and event counting. We use a subset of 300 samples (100 per task) for the evaluations in this paper.

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-2.5` (Yang et al., 2025), `Phi4` (Abdin et al., 2024) and `Pixtral` (Agrawal et al., 2024) for the evaluation. All of these models show significantly lower performance compared to humans performance of $> 95\%$ (Fu et al., 2024). The model accuracies are reported in Table 1.

Audio temporal reasoning: We have evaluated using `Qwen2-audio` (Chu et al., 2024) and `SALMONN` (Tang et al., 2023), two open-source audio-LLMs. We observe that their temporal reasoning performance is relatively weak for the tasks. We have also experimented with generating audio captions using the audio-LLMs and then passing the text captions to an LLM (Qwen 2.5 (Bai et al., 2025)) along with the MCQ prompt. This approach shows an improvement in accuracy for all three TREA tasks. Model accuracies are reported in Ta-

ble 2.

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 (Kuhn et al., 2023a) as: $\mathcal{H}\left(q(\mathbf{y}|\mathbf{x})\right) = -\sum_{\mathbf{y}} q(\mathbf{y}|\mathbf{x}) \log q(\mathbf{y}|\mathbf{x})$.
- **Verbalized confidence (VC)**: LLMs are good estimators of their own confidence (Tian et al., 2023a), 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 (Xiong et al., 2024b) uses text rephrasing and measures the answer consistency.
- **Black-box uncertainty (BU)**: The work reported in (Xiong et al., 2024a) is reproduced for comparison. The results show that using a combination of top-K prompting and random sampling yields the most stable performance. We have used top-4 prompting with outputs sampled 5 times.

5.4 Evaluation metric

An uncertainty measure should correlate with the probability of incorrect predictions. The performance of uncertainty methods is evaluated using Area-Under-Receiver-Operating-Curve (AUROC), defined as:

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

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

5.5 Equivalent Sample Generation

For the audio tasks, equivalent 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 equivalence transformations such as adding noise, varying loudness of the audio, paraphrasing the textual question, etc, are generic, equivalence 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 in the audio, changing the duration of event A is an equivalence transformation. However, this is not a valid equivalence transformation with respect to the event duration task. Similarly, for vision tasks, task-invariant transformations include adding noise to image, and task-specific transformations include RGB to grayscale transform or changes in object orientation. The complete description of equivalent sampling transforms used along with examples are detailed in Appendix A.5 and Figure 6.

5.6 Complementary samples (FCS)

By definition, the complementary transformation samples inputs which should alter the model predictions. 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 event counting task. The complementary transformations used along with examples are detailed in in Appendix A.5 and Figure 6.

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:

- FESTA based uncertainty measure outperforms all the black-box baselines approaches significantly, achieving 24.6% and 41.9% relative improvements over 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, averaged over different models.
- 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 challenging 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

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 approach 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%)
	Capt. + LLaMa	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%)
	Capt. + LLaMa	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%)
	Capt. + LLaMa	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-based Qwen-2.5 model. The final column in green (%) reports the relative improvement of FESTA approach over the best baseline result (underlined).

them consistently provide second-best performance across models and tasks.

6.2 Ablations

To probe the performance of FESTA uncertainty measure, we conduct the following analyses.

- FESTA uncertainty is a quantification of both equivalent and complementary input samplings. We separately analyze the AUROCs from equivalent (FES) and complementary (FCS) samples, as shown in Figures 2 and 3. For the order and duration tasks in audio reasoning, AUC for Qwen2 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 challenging event count task, FCS uncertainty contributes the most, showing it's

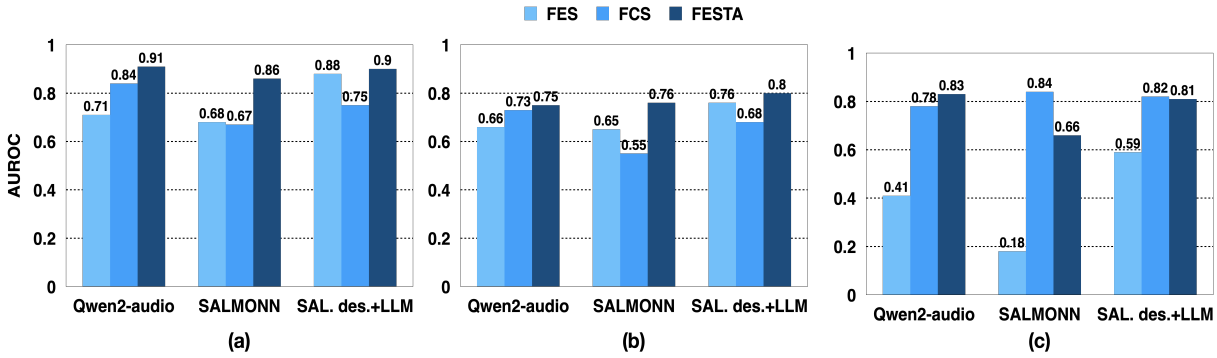


Figure 2: AUROC for uncertainty based on FES, FCS and FESTA on (a) TREA-O, (b) TREA-D, and (c) TREA-C.

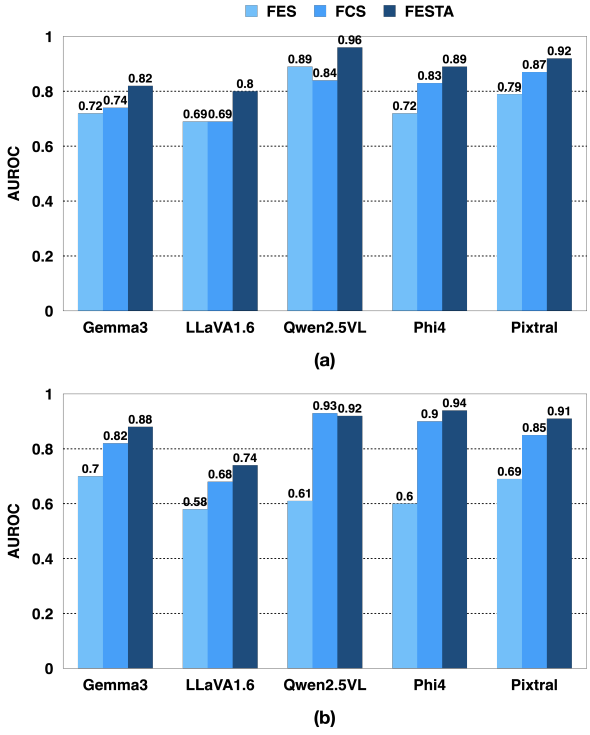


Figure 3: AUROC for uncertainty based on FES, FCS and FESTA on (a) BLINK, (b) VSR data.

robustness in low accuracy scenarios.

- Appendix Figure 4 shows the scatter plot of the confidence scores ($\frac{1}{\text{uncertainty}}$) for a model choice. Figure 5 shows the scatter plot of the detection scores for the output sampling baseline approach for the same model. It is evident that baseline AUCs for these models are affected by low uncertainty mis-predictions which are majorly detected using FESTA.

- We also compare the effectiveness of our proposed uncertainty quantification method, which uses the KL-divergence distance from the predictive distribution of an ideally certain model. To compare, we compute the AUCs by replacing the KL-divergence with the standard entropy measure and the results are reported in Appendix Tables 9 and 10. These results highlight the superiority of the KL measure over 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 4, 5, 6, 7 and 8). The results from this analysis highlight that, even with $K = 16$ samples for vision tasks and $K = 20$ for audio tasks, the proposed 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 trust assessment of LLMs. It introduces a novel input sampling paradigm based on functional equivalence and complementarity, a previously unexplored space, and demonstrates its effectiveness for unsupervised black-box settings. The proposal enables accurate abstention from incorrect predictions in both audio-LLMs and vision-LLMs, elicited by improved AUROC values. Given the limited grounding of current multimodal LLMs (especially lightweight LLMs), the models tend to produce biased, low-uncertainty hallucinations. A key contribution of FESTA is its ability to detect and abstain from such hallucinations. This capability leads to substantial improvements in selective prediction performance. Building on these promising results, we plan to extend FESTA to support natural language generation and multimodal outputs in LLMs, moving beyond the current focus on MCQA tasks.

8 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. While this process is currently automated and performed seamlessly for the tasks considered in this work, more complex audio/image tasks might require careful prompting to generate the appropriate perturbations needed for generating the FES and FCS inputs.

- **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 with the response to the FES, thereby improving the prediction accuracy.

- **Extension to Natural Language 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

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. *Preprint*, arXiv:2505.13115.

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.

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.

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

Jakob Gawlikowski, Cedrique Rovile Njieutcheu Tassi, Mohsin Ali, and 1 others. 2023. A survey on uncertainty quantification of large language models: Taxonomy, open research challenges, and future directions. *arXiv preprint arXiv:2309.09031*.

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

Jiaxing Huang and 1 others. 2023. Exploring response uncertainty in mllms: An empirical evaluation under misleading scenarios. *arXiv preprint arXiv:2312.03223*.

Ruichen Jiang and 1 others. 2024. Spuq: Perturbation-based uncertainty quantification for large language models. *arXiv preprint arXiv:2403.17436*.

Saurav Kadavath, Tom Conerly, Amanda Askell, and 1 others. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.

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. 2023a. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. *arXiv preprint arXiv:2302.09664*.

Martin Kuhn and 1 others. 2023b. Decomposing uncertainty for large language models through input clarification ensembling. *arXiv preprint arXiv:2302.04014*.

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

Jiachen Li and 1 others. 2024. Unveiling uncertainty: A deep dive into calibration and performance of multimodal large language models. *arXiv preprint arXiv:2401.10942*.

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

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.

Sabrina J Mielke and 1 others. 2022. Can large language models faithfully express their intrinsic uncertainty in words? *arXiv preprint arXiv:2210.09117*.

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

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. 2023a. 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.

Yao Tian and 1 others. 2023b. Detecting hallucinations in large language models using semantic entropy. *arXiv preprint arXiv:2308.05698*.

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*.

Yuhan Wang and 1 others. 2023a. Calibrating language models via augmented prompt ensembles. *arXiv preprint arXiv:2305.14988*.

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

Zhenlin Wang and 1 others. 2024. Unveiling uncertainty: A deep dive into calibration and performance of multimodal large language models. *arXiv preprint arXiv:2401.10942*.

Zhisheng Xiao and 1 others. 2023. Generating with confidence: Uncertainty quantification for black-box large language models. *arXiv preprint arXiv:2302.03568*.

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

Shiyue Xiong and 1 others. 2024b. Just rephrase it! uncertainty estimation in closed-source language models via multiple rephrased queries. *arXiv preprint arXiv:2402.03858*.

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

Da Yin and 1 others. 2023. To believe or not to believe your llm. *arXiv preprint arXiv:2305.19847*.

A Appendix

A.1 Equivalence of input and FES samples

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

(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.1. *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.2. *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:

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

□

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.3 (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*

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

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 pixel-level noise
- **Masking:** Hides small number of 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

The generic functionally equivalent transforms applied to text data include:

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

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 ($\{\hat{y}, \hat{y}^c\}$)

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 6.

A.6 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.7 Notations

The symbols and their meanings are noted in Table 3.

A.8 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].

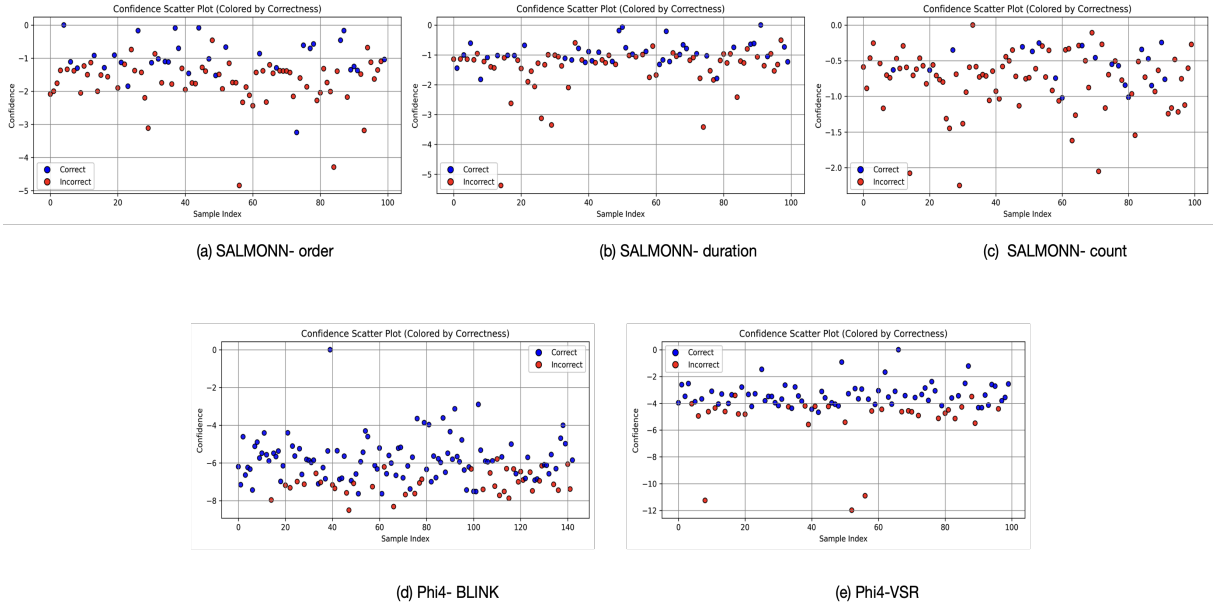
LLaVa-1.6: The $7B$ model is used from https://huggingface.co/llava-hf/llava-v1.6-mistral-7b-hf with license [3].

Phi-4: The $5.6B$ model is used from https://huggingface.co/microsoft/Phi-4-multimodal-instruct with license [4].

[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



(a) SALMONN- order (b) SALMONN- duration (c) SALMONN- count

(d) Phi4- BLINK (e) Phi4-VSR

Figure 4: FESTA log(score) plots for best improvement models where score is reciprocal of FESTA uncertainty.

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 $12B$ model is used from https://github.com/bytedance/SALMONN with license [8].

A.9 Computation budget and hardware

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

A.10 Number of Equivalent and complementary Samples

The results for varying the number of samples K are given in Tables 4, 6, 7, 8,and 5.

A.11 Usage of AI assitants

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

[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

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 4: 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 5: 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 6: 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 7: AUC performance for the 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 8: AUC performance for the Ordering task.

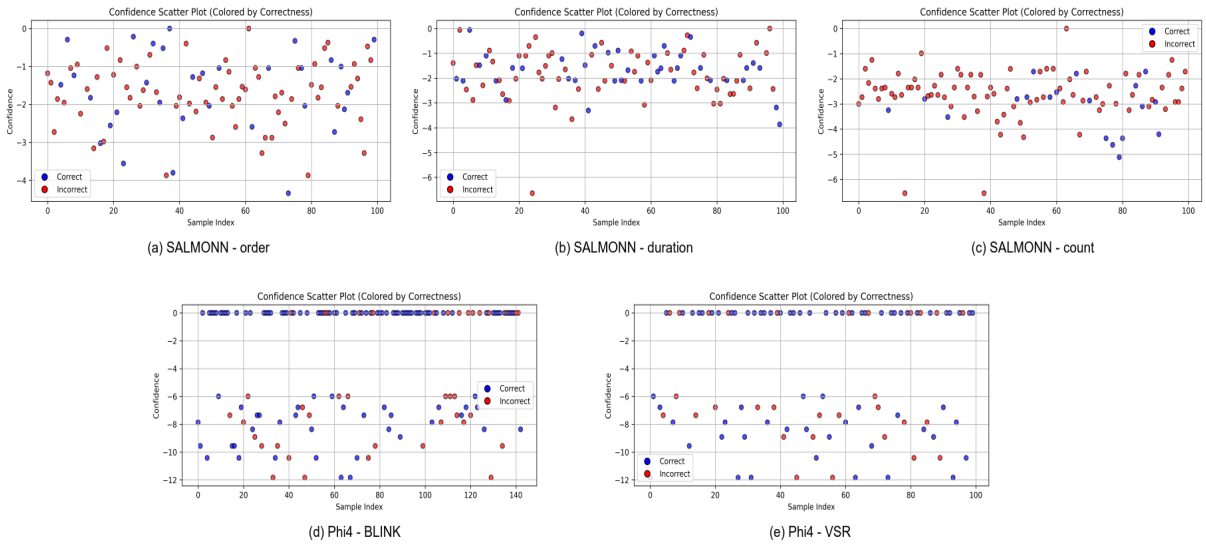


Figure 5: FESTA log(score) plots for output sampling baseline where score is reciprocal of FESTA uncertainty.

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 9: 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 10: Average performance of vision-LLMs using standard entropy measure compared with the proposed KL-div based measure.

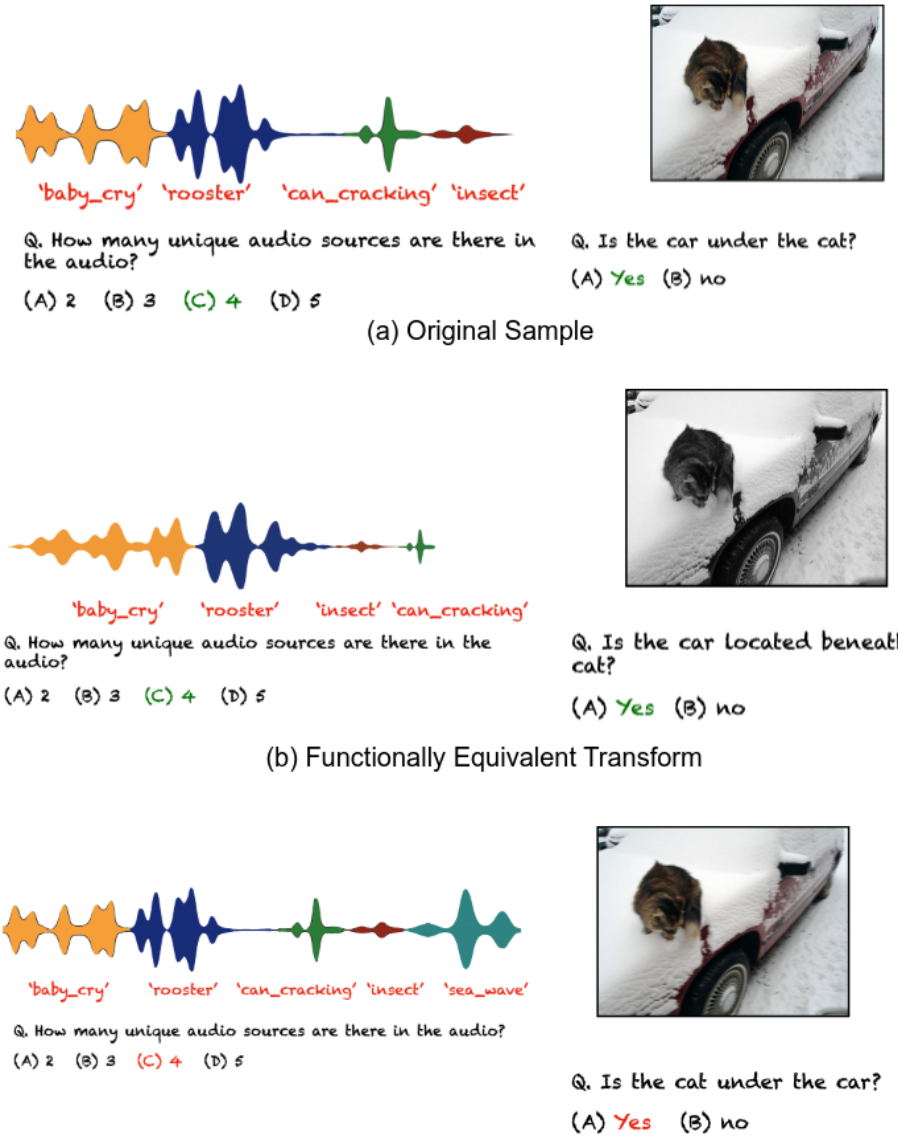


Figure 6: Examples of Functionally Equivalent Transform and Functionally Complementary Transform for both audio-text and image-text questions.