Qualcomm

ICLR2026

# Enhancing Hallucination Detection through Noise Injection

Litian Liu[1], Reza Pourreza[1], Sunny Panchal[1], Apratim Bhattacharyya[1], Yubing Jian[1], Yao Qin[2], Roland Memisevic[1]

1. Qualcomm AI Research 2. UC Santa Babara

# Background

- Dispersion across Samples ⟺ LLM Uncertainty ⟺ Hallucination

prompt

A robe takes 2 bolts of blue fiber and half that much white fiber. How many bolts in total does it take?

sample 1

Half of 2 bolts of white fiber is 2/2 = 1 bolt. So, it takes 2 + 1 = 3 bolts in total. The answer is **3**.

sample 2

2 bolts of blue fiber and half that much white fiber is 2 + half of 2 = 2 + 1 = 3 bolts. The answer is **3**.

sample 3

2 bolts of blue fiber and half that much white fiber is 2 x 2 = 4 bolts of blue fiber. The answer is **4**.

# Background

- Dispersion across Samples ⟺ LLM Uncertainty ⟺ Hallucination

prompt

A robe takes 2 bolts of blue fiber and half that much white fiber. How many bolts in total does it take?

sample 1

Half of 2 bolts of white fiber is 2/2 = 1 bolt. So, it takes 2 + 1 = 3 bolts in total. The answer is **3**.

sample 2

2 bolts of blue fiber and half that much white fiber is 2 + half of 2 = 2 + 1 = 3 bolts. The answer is **3**.

sample 3

2 bolts of blue fiber and half that much white fiber is 2 x 2 = 4 bolts of blue fiber. The answer is **4**.

$$H_{ans} = -\frac{2}{3} \times log\frac{2}{3} - \frac{1}{3} \times log\frac{1}{3}$$

# Background

- Dispersion across Samples ⟺ LLM Uncertainty ⟺ Hallucination

prompt

A robe takes 2 bolts of blue fiber and half that much white fiber. How many bolts in total does it take?

sample 1

Half of 2 bolts of white fiber is 2/2 = 1 bolt. So, it takes 2 + 1 = 3 bolts in total. The answer is **3**.

sample 2

2 bolts of blue fiber and half that much white fiber is 2 + half of 2 = 2 + 1 = 3 bolts. The answer is **3**.

sample 3

2 bolts of blue fiber and half that much white fiber is 2 x 2 = 4 bolts of blue fiber. The answer is **4**.

$$H_{ans} = -\frac{2}{3} \times log\frac{2}{3} - \frac{1}{3} \times log\frac{1}{3}$$

High dispersion → Hallucination.

Thresholding to enable detection:

$$H\,(Y) \begin{cases} > \theta & \text{Hallucination} \\ < \theta & \text{Not Hallucination} \end{cases}$$

# Background

- Existing work: Dispersion Metrics across Multiple Samples.

  **[Token-level]** *(Xiao et al., 2021)* Predictive uncertainty of LLM.

  **[Lexical-level]** *(Manakul et al, 2023)* Dispersion measured from n-gram model.

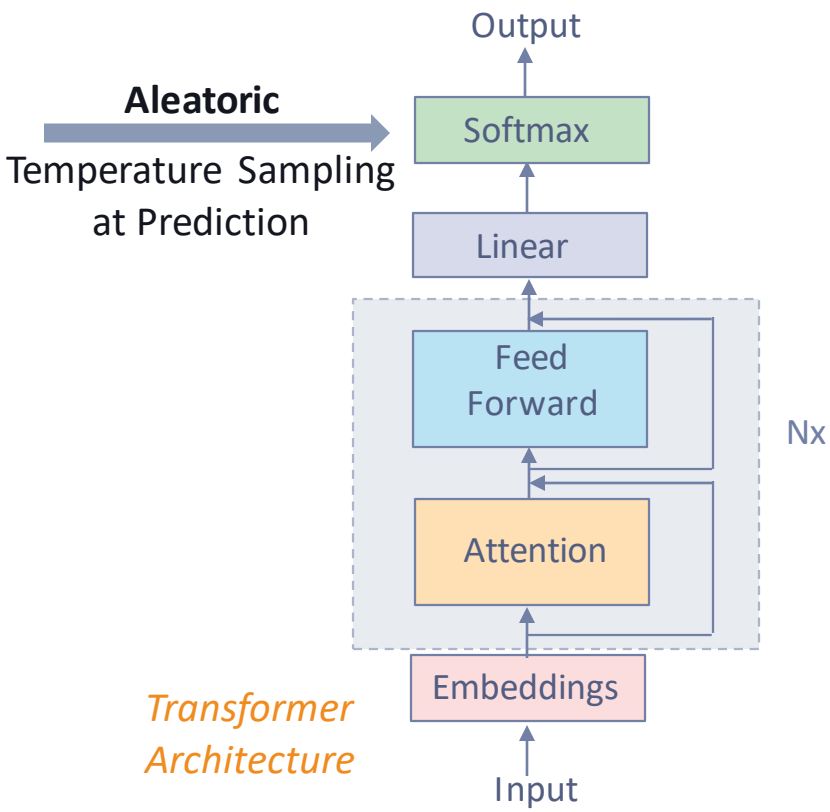
  **[Semantic-level]** *(Kuhn et al, 2023)* Dispersion over semantic groups.

  **[Embedding-level]** *(Chen et al, 2024)* Dispersion of the matrix made from embeddings.

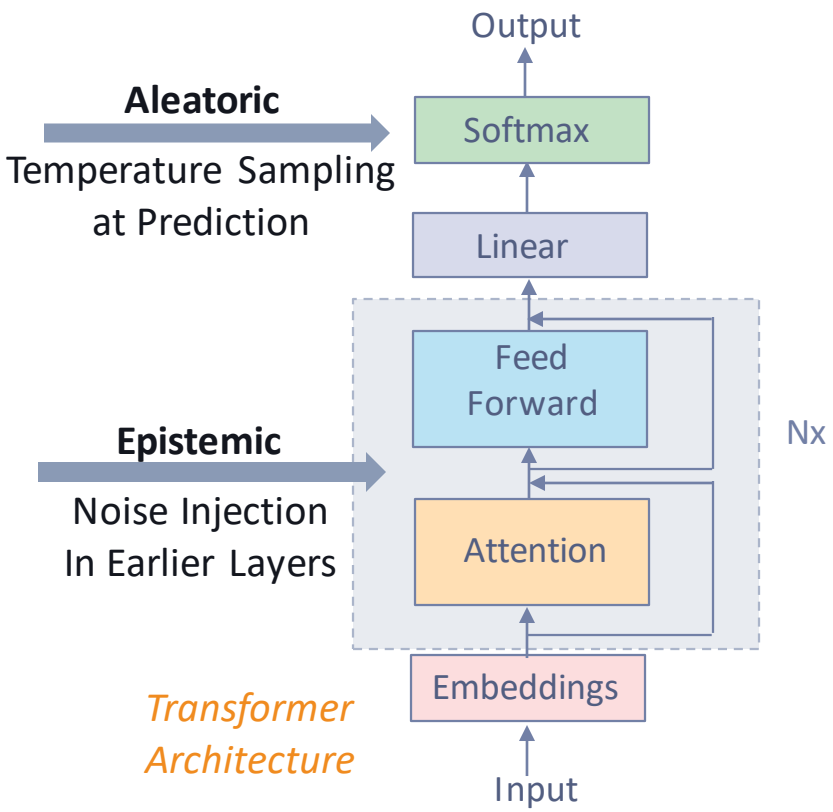
- Our work:

  → *How to sample to better capture uncertainty?*

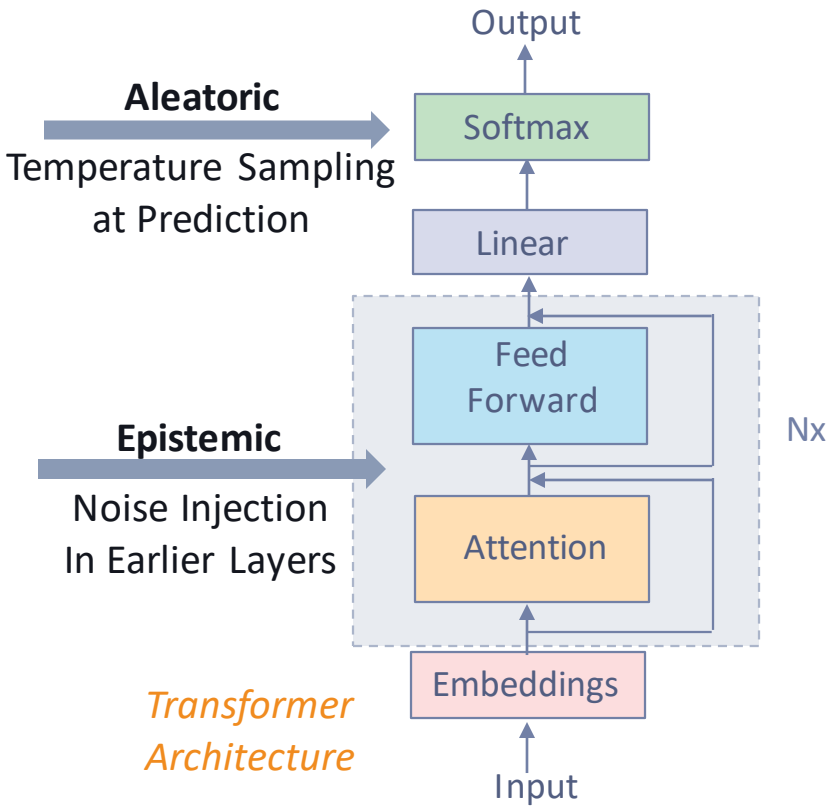
# View of Uncertainty: Aleatoric + Epistemic



# View of Uncertainty: Aleatoric + Epistemic



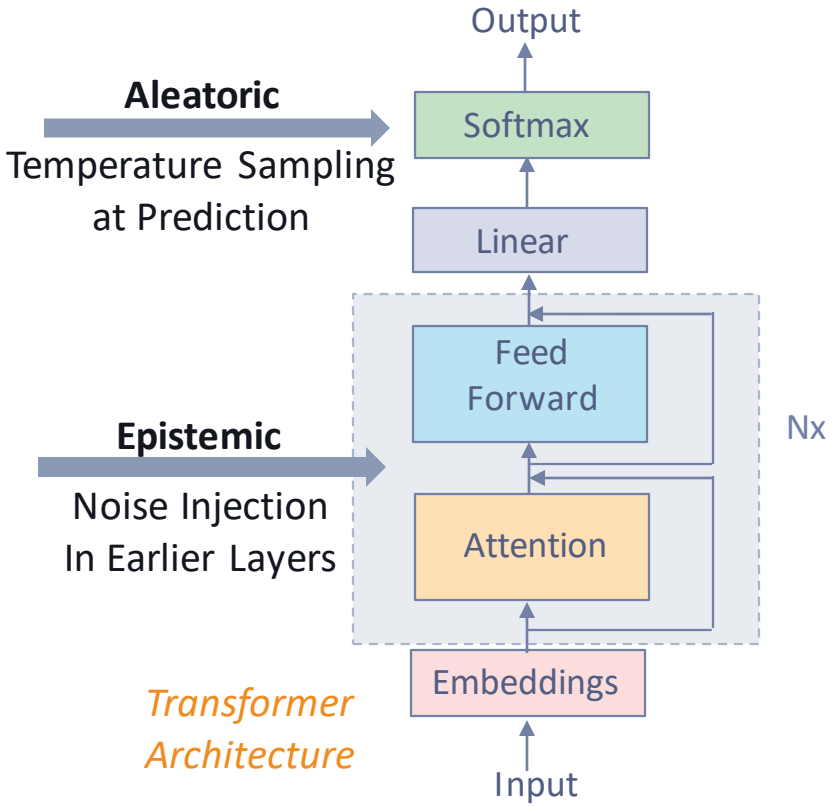
# View of Uncertainty: Aleatoric + Epistemic




- Noise injection enables a **training-free Bayesian ensemble** around pretrained model $\hat{\omega}$.

$$q(\omega) = \underbrace{\prod_{i \notin S} \delta(\omega_i - \hat{\omega}_i)}_{\text{unchanged}} \cdot \underbrace{\prod_{i \in S} q_i(\hat{\omega}_i, \alpha)}_{\text{noise perturbed}}$$

# View of Uncertainty: Aleatoric + Epistemic



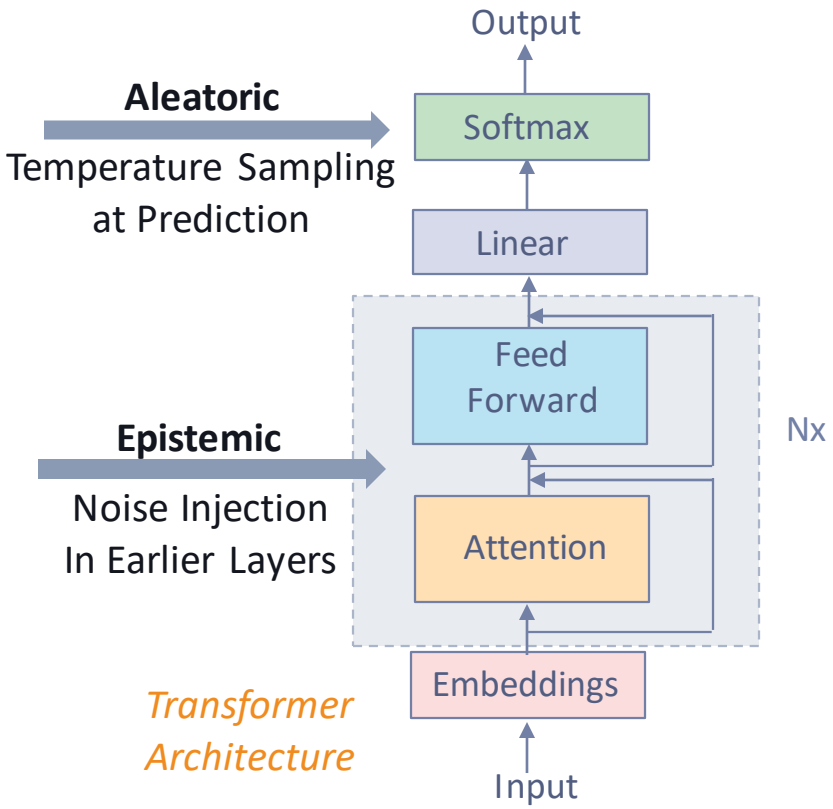

- Noise injection enables a **training-free Bayesian ensemble** around pretrained model $\hat{\omega}$.

$$q(\omega) = \underbrace{\prod_{i \notin S} \delta(\omega_i - \hat{\omega}_i)}_{\text{unchanged}} \cdot \underbrace{\prod_{i \in S} q_i(\hat{\omega}_i, \alpha)}_{\text{noise perturbed}}$$

- Sampling: combining aleatoric and epistemic uncertainty in a Bayesian manner.

$$p(\boldsymbol{y}|\boldsymbol{x}) = \int \Pi_t p(y_t|y_{<t}, \boldsymbol{x}, \omega) q(\omega) d\omega$$

# View of Uncertainty: Aleatoric + Epistemic



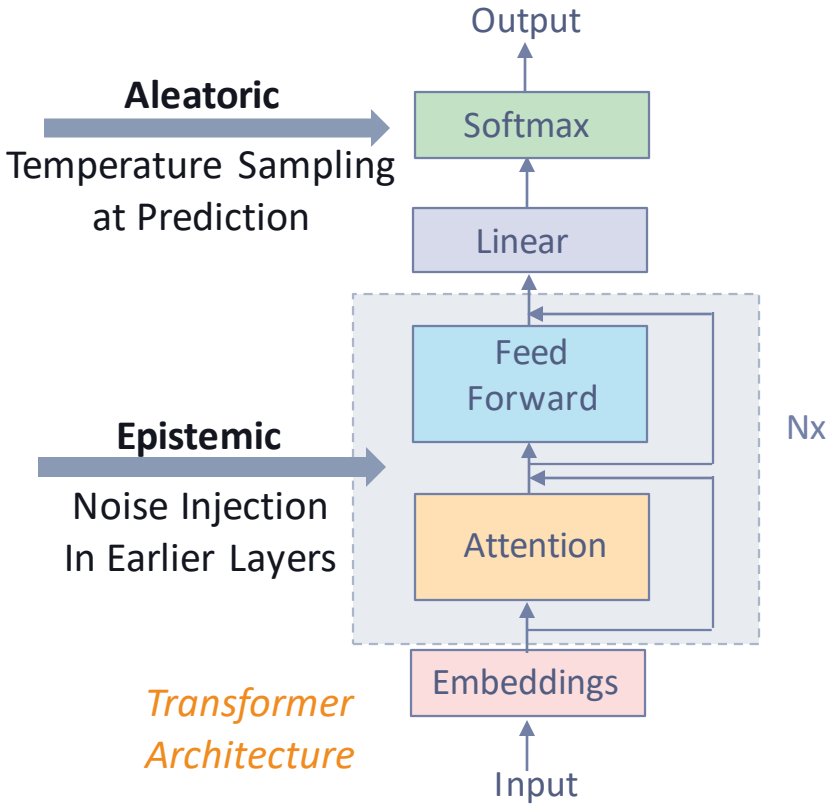

- Noise injection enables a **training-free Bayesian ensemble** around pretrained model $\hat{\omega}$.

$$q(\omega) = \underbrace{\prod_{i \notin S} \delta(\omega_i - \hat{\omega}_i)}_{\text{unchanged}} \cdot \underbrace{\prod_{i \in S} q_i(\hat{\omega}_i, \alpha)}_{\text{noise perturbed}}$$

- Sampling: combining aleatoric and epistemic uncertainty in a Bayesian manner.

$$p(\boldsymbol{y}|\boldsymbol{x}) = \int \Pi_t p(y_t|y_{<t}, \boldsymbol{x}, \omega) \underbrace{q(\omega) d\omega}_{\text{Noise Injection}}$$

# View of Uncertainty: Aleatoric + Epistemic




- Noise injection enables a **training-free Bayesian ensemble** around pretrained model $\hat{\omega}$.

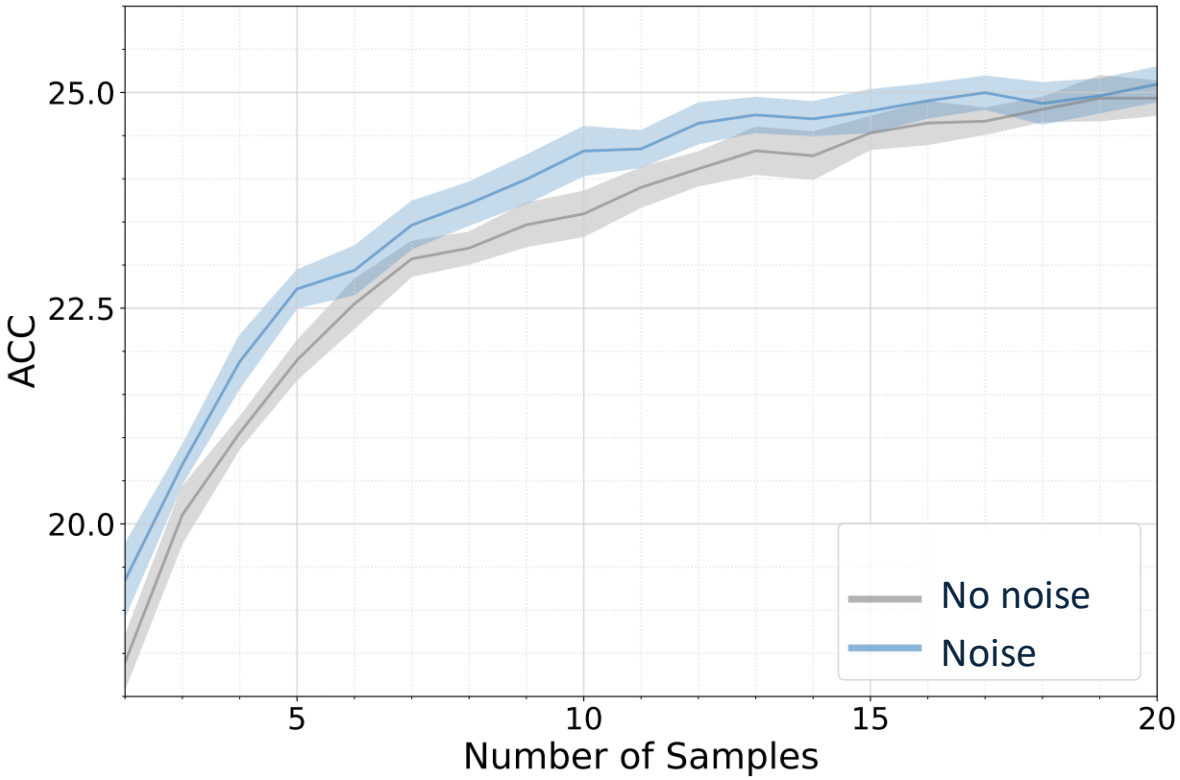
$$q(\omega) = \underbrace{\prod_{i \notin S} \delta(\omega_i - \hat{\omega}_i)}_{\text{unchanged}} \cdot \underbrace{\prod_{i \in S} q_i(\hat{\omega}_i, \alpha)}_{\text{noise perturbed}}$$

- Sampling: combining aleatoric and epistemic uncertainty in a Bayesian manner.

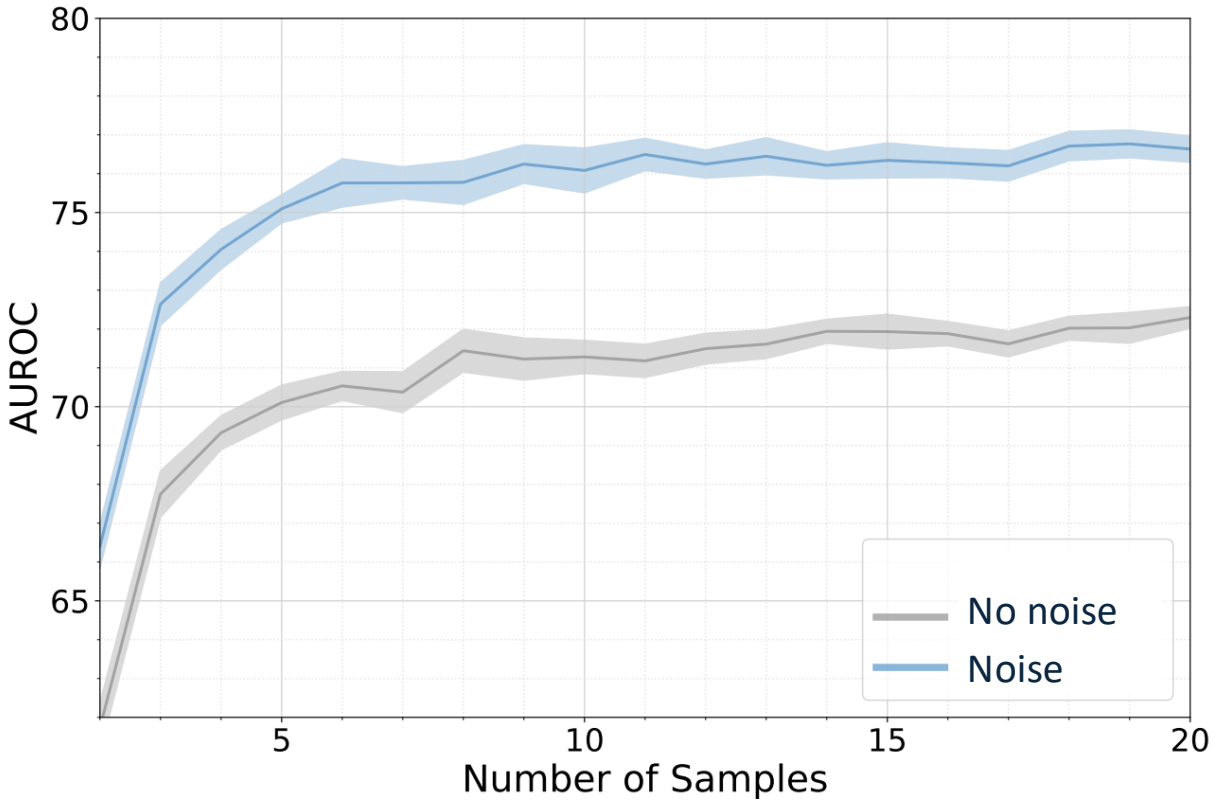
$$p(\boldsymbol{y}|\boldsymbol{x}) = \int \underbrace{\Pi_t p(y_t|y_{<t}, \boldsymbol{x}, \omega)}_{\text{Temperature Sampling}} \underbrace{q(\omega) d\omega}_{\text{Noise Injection}}$$

# Results

Noise Injection improves model **accuracy**.




Noise Injection improves **hallucination detection**.



# Results

- Noise Injection improves the effectiveness of hallucination detection (higher AUROC) across **models** and **datasets**.

| | GSM8K | CSQA | TriviaQA |
|---|---|---|---|
| Gemma-2B-it | 51.36 +/- 0.79 | 58.97 +/- 0.47 | 68.65 +/- 0.13 |
| Gemma-2B-it w/ Noise | 57.11 +/- 0.67 | 61.71 +/- 0.37 | 69.38 +/- 0.11 |
| Phi-3-mini-4k-instruct (3.8B) | 65.86 +/- 0.58 | 75.05 +/- 0.41 | 82.00 +/- 0.09 |
| Phi-3-mini-4k-instruct w/ Noise | 72.51 +/- 0.53 | 76.60 +/- 0.53 | 82.02 +/- 0.06 |
| Mistral-7B-Instruct | 78.06 +/- 0.30 | 72.96 +/- 0.45 | 77.59 +/- 0.08 |
| Mistral-7B-Instruct w/ Noise | 81.26 +/- 0.37 | 75.01 +/- 0.42 | 79.42 +/- 0.06 |
| Llama-2-7B-chat | 71.56 +/- 0.51 | 70.59 +/- 0.36 | 74.03 +/- 0.09 |
| Llama-2-7B-chat w/ Noise | 76.14 +/- 0.52 | 71.56 +/- 0.36 | 75.05 +/- 0.08 |
| Llama-2-13B-chat | 77.20 +/- 0.33 | 67.55 +/- 1.02 | 73.39 +/- 0.09 |
| Llama-2-13B-chat w/ Noise | 79.25 +/- 0.32 | 69.10 +/- 0.94 | 75.10 +/- 0.07 |

# Thank you

Qualcomm

Follow us on:

For more information, visit us at:

qualcomm.com & qualcomm.com/blog