DEAKIN UNIVERSITY | DEAKIN APPLIED ARTIFICIAL INTELLIGENCE INITIATIVE

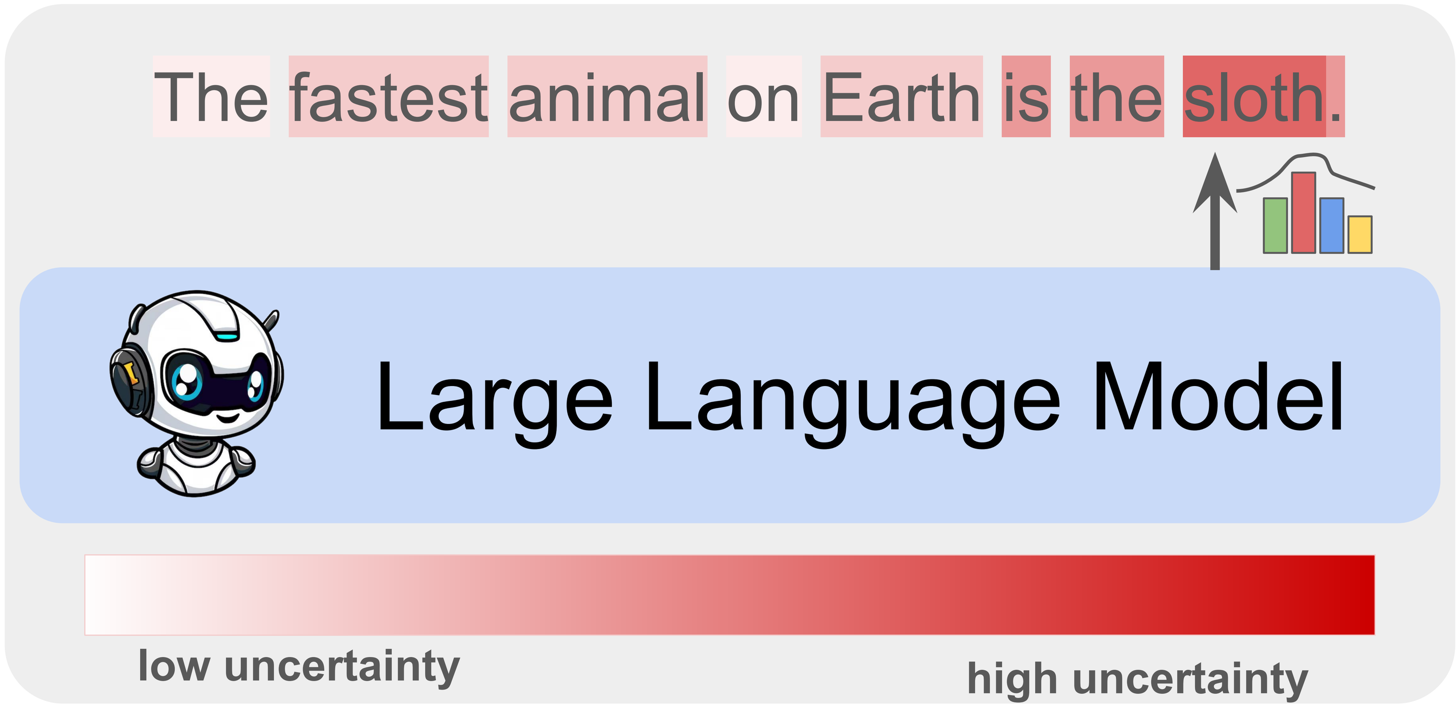
AAAI-26 / IAAI-26 / EAAI-26
JANUARY 20–27, 2026 | SINGAPORE

Probabilities Are All You Need: A Probability-Only Approach to Uncertainty Estimation in Large Language Models

Manh Nguyen, Sunil Gupta, Hung Le

Core idea: Approximate **Predictive Entropy** using only the response's **top-K probabilities**, where K is adaptively chosen by **thresholding α**

Question: What is the fastest animal on Earth?



Abstract

- Large Language Models (LLMs) perform well across NLP tasks but are prone to **hallucinations**—factually incorrect outputs that undermine reliability in real-world applications.
- Estimating uncertainty is a key strategy to detect hallucinations. However, existing methods often require sampling or extra computation to assess predictive entropy.
- We propose a **training-free, efficient method to estimate uncertainty** based on top-K output probabilities.

Background

In the context of LLMs, we can measure the uncertainty of a generation as:

$$U(x) = H(Y|x) = -\sum_{y} p(y|x) \log p(y|x)$$

The probability of generating sequence y given a prompt x:

$$p(y|x) = \prod_{t=1}^{T} p(y^t|y^{<t}, x)$$

where T is the length of the generated sequence, and y^t is the token at position t. Taking the logarithm, we get **Negative Log-Likelihood** (NLL):

$$\mathrm{NLL}(y|x) = -\sum_{t=1}^{T} \log p(y^t|y^{<t}, x).$$

$$p(y|x) = e^{-\mathrm{NLL}(y|x)}$$

However, NLL is **relying solely on a single generation** that can miss plausible alternatives, limiting the ability to capture response uncertainty in ambiguous or high-variance prompts.

Method

For simplicity, we use p_i^* to represent the probability of the top i-th generation $p(y_i^*|x)$. We introduce an approximation of **Predictive Entropy** as a **PR**obability-**O**nly uncertainty score (PRO):

$$\boxed{PRO(x) = -\log p_K^* - \sum_{i=1}^{K} p_i^* \log \frac{p_i^*}{p_K^*}}$$

Proposition 1. *Let $\mathbf{y}^* = (y_1^*, y_2^*, \ldots, y_K^*)$ be the top K generations of a LLM given prompt x. The predictive entropy approximation using the top K probabilities satisfies the following inequality:*

$$H(Y|x) \geq -\log p_K^* - \sum_{i=1}^{K} p_i^* \log \frac{p_i^*}{p_K^*}$$

Adaptive Top-K selection

Instead of using a fixed top-K, we propose an **adaptive constraint** that filters out low-probability generations, ensuring the uncertainty estimation focuses on the most confident and relevant responses.

$$\mathbf{p}_K = \{p_k \mid p_k \geq \alpha, 1 \leq k \leq N\}$$

Experiments

- **Baselines:** Semantic-based methods (SD, SE, Deg), Predictive Entropy (NE, PE), Negative Log-likelihood (ALL, NLL)

Dataset	Model	SD	SE	Deg	NE	PE	ALL	NLL	PRO (Ours)
TriviaQA	Gemma-2B	0.799	0.668	0.746	0.692	0.624	0.789	0.806	**0.819**
	Gemma-7B	0.831	0.690	0.715	0.702	0.652	0.833	0.812	**0.841**
	Llama2-13B	**0.862**	0.682	0.802	0.551	0.552	0.624	0.684	0.802
	Falcon-11B	0.706	0.592	**0.710**	0.555	0.604	0.577	0.668	0.668
	Falcon-40B	0.700	0.724	0.722	0.674	0.623	0.658	**0.765**	**0.765**
SciQ	Gemma-2B	0.719	0.570	0.725	0.601	0.605	0.719	0.728	**0.751**
	Gemma-7B	0.741	0.622	0.699	0.658	0.678	0.765	0.755	**0.787**
	Llama2-13B	0.706	0.574	**0.720**	0.481	0.543	0.515	0.600	0.716
	Falcon-11B	0.724	0.554	0.771	0.561	0.603	0.573	0.797	**0.799**
	Falcon-40B	0.668	0.613	0.626	0.592	0.577	0.660	**0.674**	**0.674**
NQ	Gemma-2B	0.618	0.599	0.620	0.600	0.613	0.607	0.694	**0.696**
	Gemma-7B	0.670	0.621	0.691	0.662	0.566	**0.698**	0.683	0.691
	Llama2-13B	0.627	0.562	0.713	0.540	0.649	0.691	0.737	**0.740**
	Falcon-11B	0.636	0.591	0.580	0.515	0.522	0.512	0.684	**0.685**
	Falcon-40B	0.632	0.603	0.579	0.544	0.585	0.475	0.638	**0.645**
Average AUC		0.709	0.618	0.695	0.595	0.600	0.646	0.715	**0.739**
Best Count		1	0	2	0	0	1	2	**11**

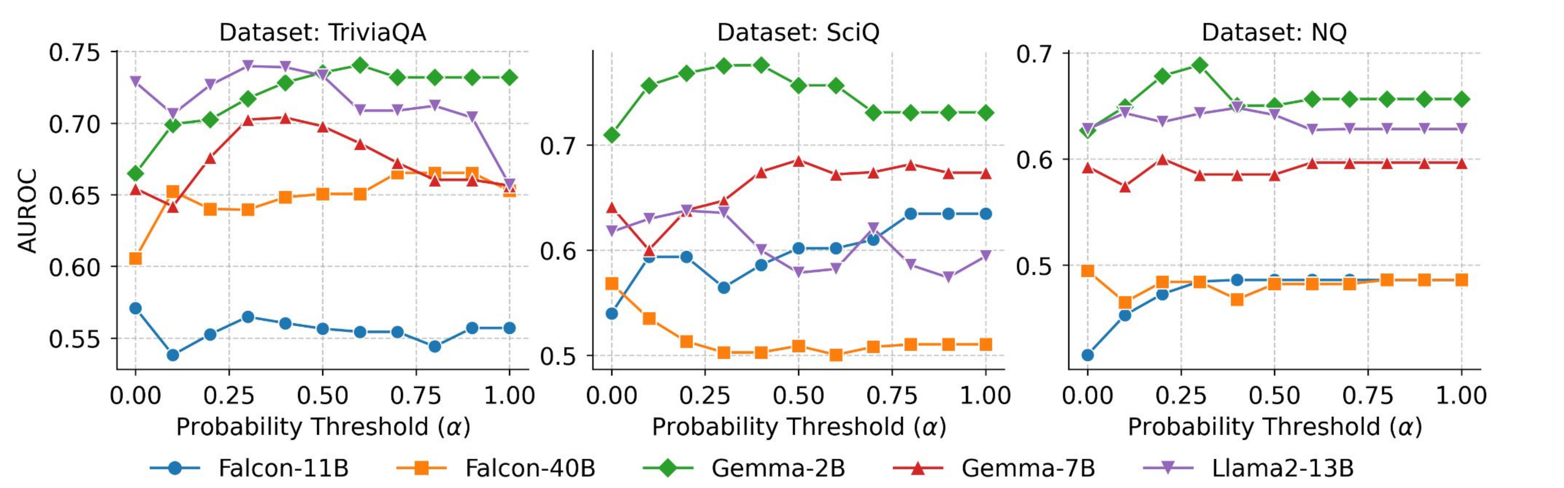


Figure 1: AUC performance when adjusting α

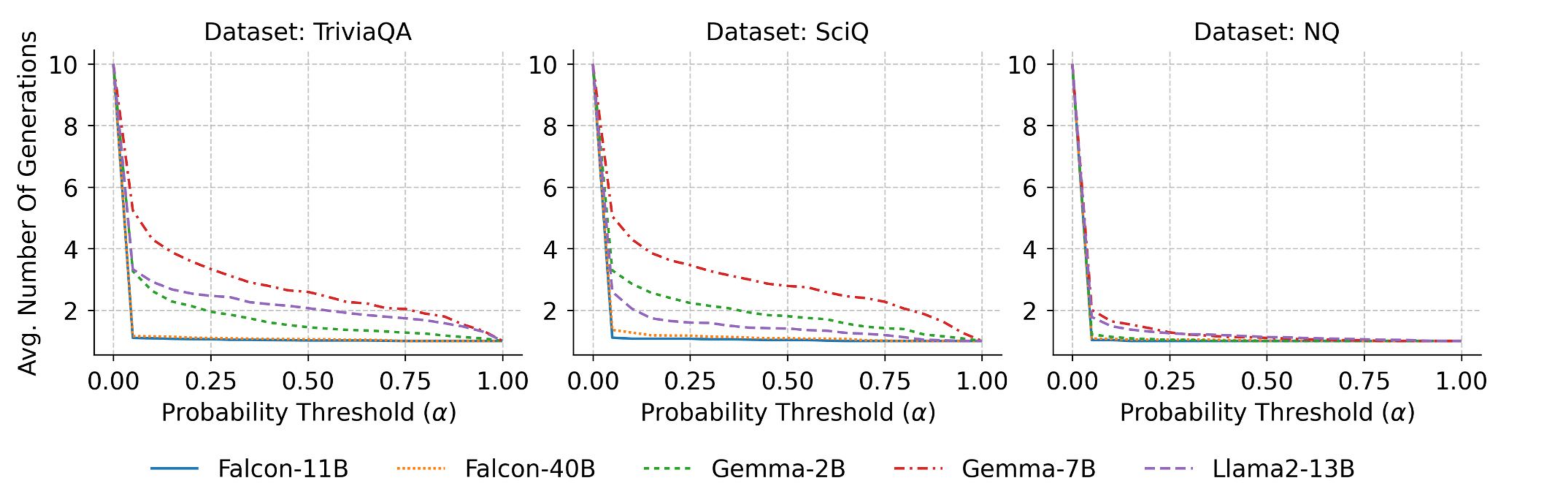


Figure 2: Relationship between number of selected generations and α

manh.nguyen@deakin.edu.au

Preprint: https://arxiv.org/pdf/2511.07694 | **Code:** https://github.com/manhitv/PRO