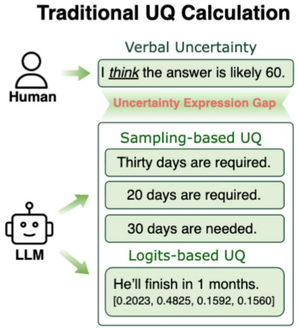
“very likely” Means “uncertain”? How LLMs Diverge from Humans in Linguistic Uncertainty Quantification

Jinhao Duan*[1], Zicheng Liu*[2], Zijie Liu*[1], Kaidi Xu[3], Tianlong Chen[1]
The University of North Carolina at Chapel Hill[1] The University of Hong Kong[2] City University of Hong Kong[3]

Introduction: Verbal Uncertainty Quantification



LLMs often express uncertainty with phrases like
“I think,” “likely,” or “possible.”

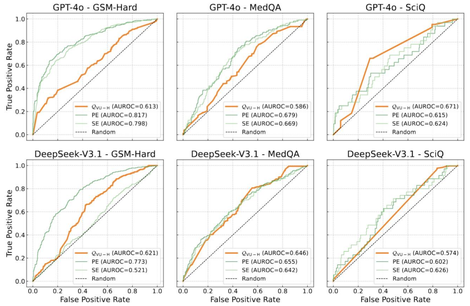
But traditional UQ relies on logits or sampling, which is costly and less human-interpretable.

Key Question:
Can verbal uncertainty markers reliably quantify LLM confidence?

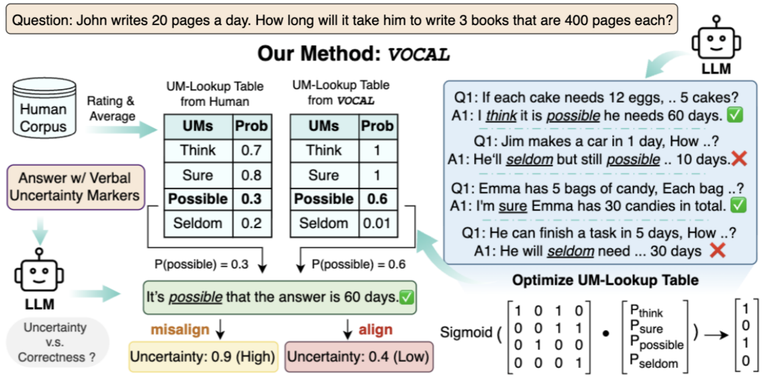
Motivation

1.Verbal markers carry useful UQ signals.
Human UM-Lookup achieves non-trivial AUROC, showing that linguistic uncertainty is informative.

2.LLMs diverge from humans in verbal UQ.
LLMs use human-like uncertainty phrases, but their implied confidence levels often differ from human interpretations.



The proposed *VOCAL*



VOCAL learns an LLM-specific UM-Lookup by fitting verbal uncertainty markers to empirical correctness. The overall optimization objective is defined as:

$$\mathcal{L}(\boldsymbol{c}) + \mathcal{L}_{\text{lap}}(\mathbf{c})$$

Correctness Alignment, a Binary Cross-Entropy (BCE) manner :

$$\mathcal{L}(\boldsymbol{c}) = \min_{\boldsymbol{c}} \mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})}\Big[-z \log \boldsymbol{c_y} - (1-z)\log(1-\boldsymbol{c_y})\Big],$$

The semantic smoothing regularizer is defined as:

$$\mathcal{L}_{\text{lap}}(\mathbf{c}) = \gamma\, \mathbf{c}^\top \boldsymbol{L}\mathbf{c} = \gamma \sum_{i,j} \boldsymbol{W}_{ij}(c_i - c_j)^2.$$

Experimental Results

- VOCAL Outperforms Single-sample UQ

Dataset	Model	G-NLL	PPL	VOCAL
Trivia QA	GPT-4o	0.538	**0.575**	0.573
	Qwen2.5-72B-Ins.	0.627	0.619	**0.645**
SciQ	GPT-4o	0.663	0.648	**0.700**
	Qwen2.5-72B-Ins.	0.568	0.555	**0.717**
GSM-Hard	DeepSeek-V3.1	0.520	0.567	**0.715**
	Qwen2.5-72B-Ins	0.507	0.580	**0.679**

Table 1: The comparison results between VOCAL and single-sample UQ baselines. It is shown that VOCAL is significantly better than these methods.

- VOCAL Learns LLM-tailored UM-Lookup

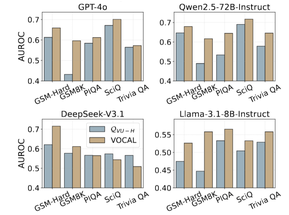


Figure 4. The evaluation results of VOCAL when comparing with human-sourced UM-Lookup, i.e., $Q_{\text{VU-H}}$. It demonstrates that VOCAL produces LLM-tailored UM-Lookup table.

- VOCAL Matches Multi-sample UQ

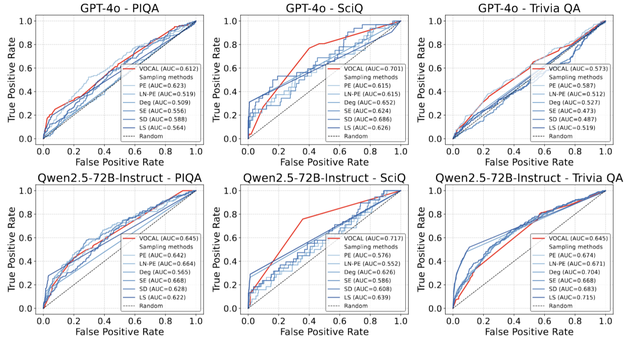


Figure 5. The evaluation results of VOCAL and multi-sample based UQ methods. It is worth noting that sampling-based methods rely on semantic consistency calculations, which are expensive and introduce latency in real-world deployment. It is shown that VOCAL achieves comparable performance to sampling-based UQ methods.

- Cross-LLM Transferability

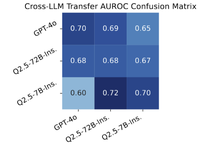


Figure 6. Cross-LLM transferability. LLMs share a substantial common structure in verbal uncertainty expression.

- LLMs Diverge from Humans in Verbal UQ

Phrase	GPT-4o Prob.	Human Prob.
absolutely certain	1.000	0.920
confident	0.839	0.900
positive	0.839	0.900
sure	0.839	0.830
i think	0.710	0.630
almost certain	0.677	0.920
think	0.645	0.490
can	0.355	0.570
reasonable to assume	0.355	0.605
very likely	0.355	0.853
likely	0.000	0.655