SIGIR 2025
Padova
ITALY

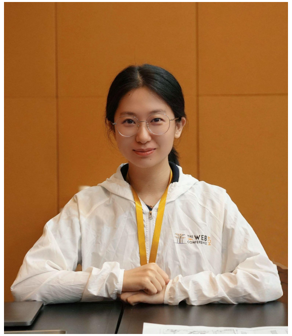
# Knowledge Boundary Identification

**Moxin Li**

National University of Singapore

NUS
National University of Singapore

# Speaker Info



Homepage

I am a fourth-year Ph.D. candidate at NExT Research Center, National University of Singapore (NUS).

**Current research interests**: trustworthy LLM, LLM safety, LLM evaluation.

**Experience**:
**2021 - Now**: Ph.D, NUS, School of Computing
**2016 - 2020**: B.S., Peking University, Yuanpei College

**Email**: limoxin@u.nus.edu
(**Actively looking for postdoc positions!**)

# Knowledge Boundary Identification

**Outline**

- **Uncertainty Estimation**
  - Quantifies the model's uncertainty about predictions.
  - High uncertainty tends to lie outside of boundary.
- **Confidence Calibration**
  - Aligns LLM's confidence with actual correctness of predictions.
  - High confidence tends to lie inside of boundary.
- **Internal State Probing**
  - Probes LLM internal states (e.g., attention heads, hidden layers, neurons) to assess factual accuracy.

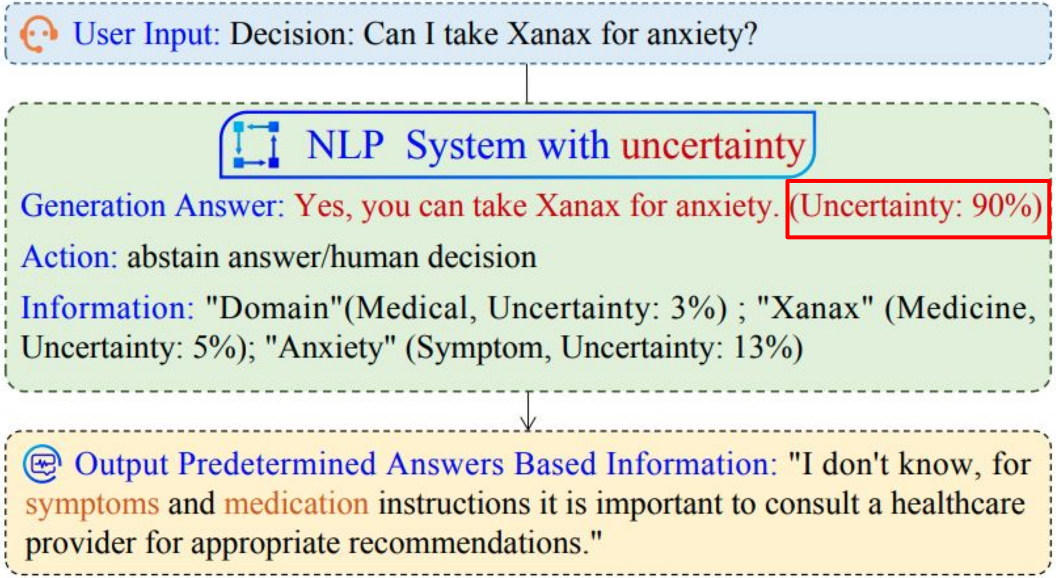
# Knowledge Boundary Identification

**Outline**

- ❑ Uncertainty Estimation
  - ❑ Token Probability-based
  - ❑ Semantic-based
  - ❑ Uncertainty Decomposition
  - ❑ Conformal Prediction
- ❑ Confidence Calibration
- ❑ Internal State Probing

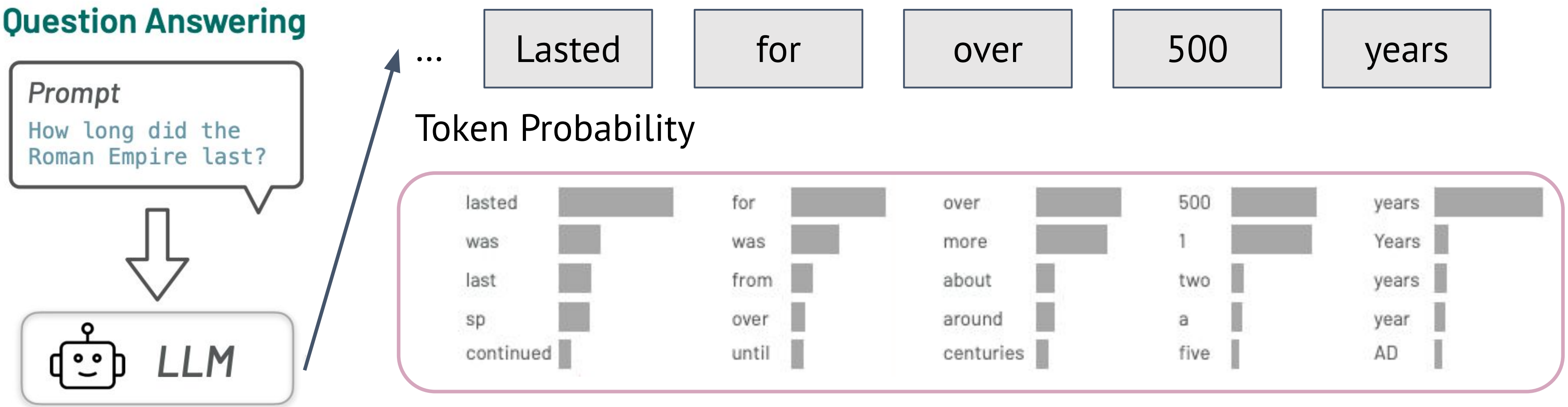
# Uncertainty Estimation

LLM, as a neural network, makes mistakes.
Estimating the reliability of LLM's output is important.




*Hu et al., 2023 "Uncertainty in Natural Language Processing: Sources, Quantification, and Applications" (CoRR'23)*

# Uncertainty Estimation (UE) – Token Probability-based




### Max/Avg Prob.

$$Max(-\log p)_i = \max_j(-\log p_{ij}),$$

$$Avg(-\log p)_i = -\frac{1}{J}\sum_j \log p_{ij},$$

### Max/Avg Entropy

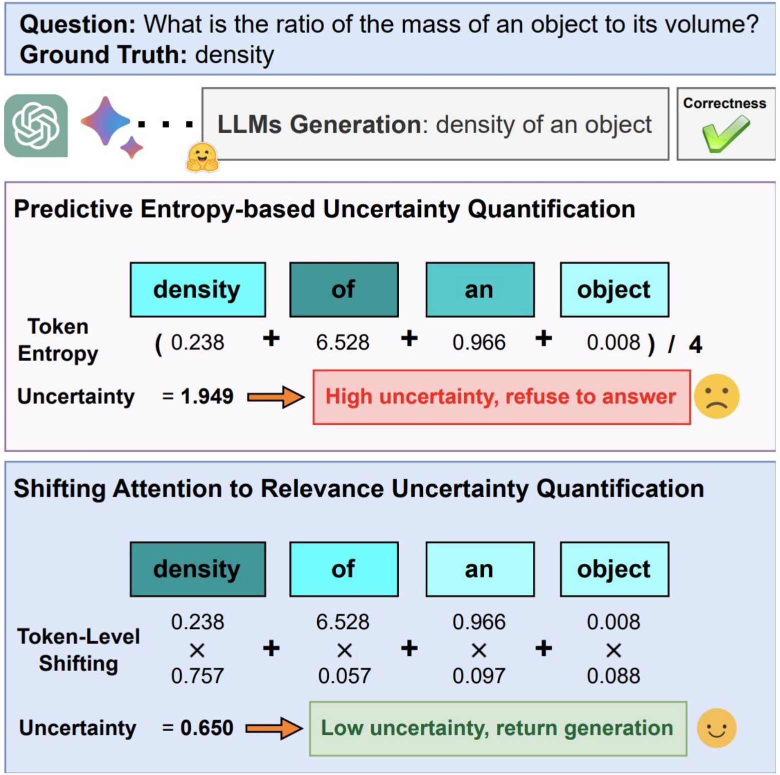
$$Max(\mathcal{H})_i = \max_j[\mathcal{H}_{ij}],$$

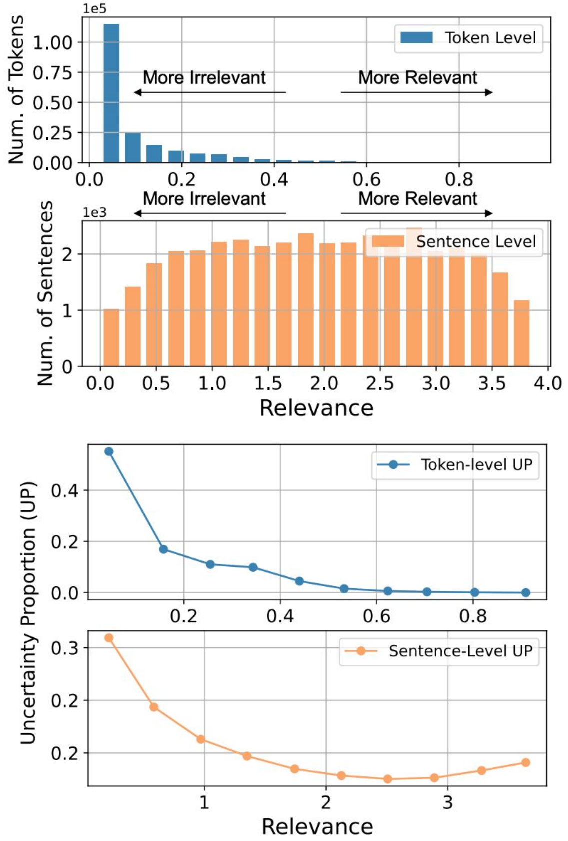
$$Avg(\mathcal{H})_i = \frac{1}{J}\sum_j \mathcal{H}_{ij},$$

*Huang et al., 2024 "Look Before You Leap: An Exploratory Study of Uncertainty Analysis for Large Language Models" (IEEE TRANSACTIONS ON SOFTWARE ENGINEERING'24)*

# Uncertainty Estimation (UE) – Token Probability-based

## Different token weights and granularity




*Duan et al., 2024 "Shifting Attention to Relevance: Towards the Predictive Uncertainty Quantification of Free-Form Large Language Models" (ACL'24)*

# Uncertainty Estimation (UE) – Semantic-based

## Semantic Entropy

1. **Generation:** Sample $M$ sequences $\{s^{(1)}, \ldots, s^{(M)}\}$ from the predictive distribution of a large language model given a context $x$.
2. **Clustering:** Cluster the sequences which mean the same thing using our bi-directional entailment algorithm.
3. **Entropy estimation:** Approximate semantic entropy by summing probabilities that share a meaning following Eq. (2) and compute resulting entropy. This is illustrated in Table 1.

$$p(c \mid x) = \sum_{\mathbf{s} \in c} p(\mathbf{s} \mid x) = \sum_{\mathbf{s} \in c} \prod_i p(s_i \mid s_{<i}, x).$$

(a) Scenario 1: No semantic equivalence

| Answer<br>$\mathbf{s}$ | Likelihood<br>$p(\mathbf{s} \mid x)$ | Semantic likelihood<br>$\sum_{\mathbf{s} \in c} p(\mathbf{s} \mid x)$ |
|---|---|---|
| Paris | 0.5 | 0.5 |
| Rome | 0.4 | 0.4 |
| London | 0.1 | 0.1 |
| Entropy | 0.94 | 0.94 |

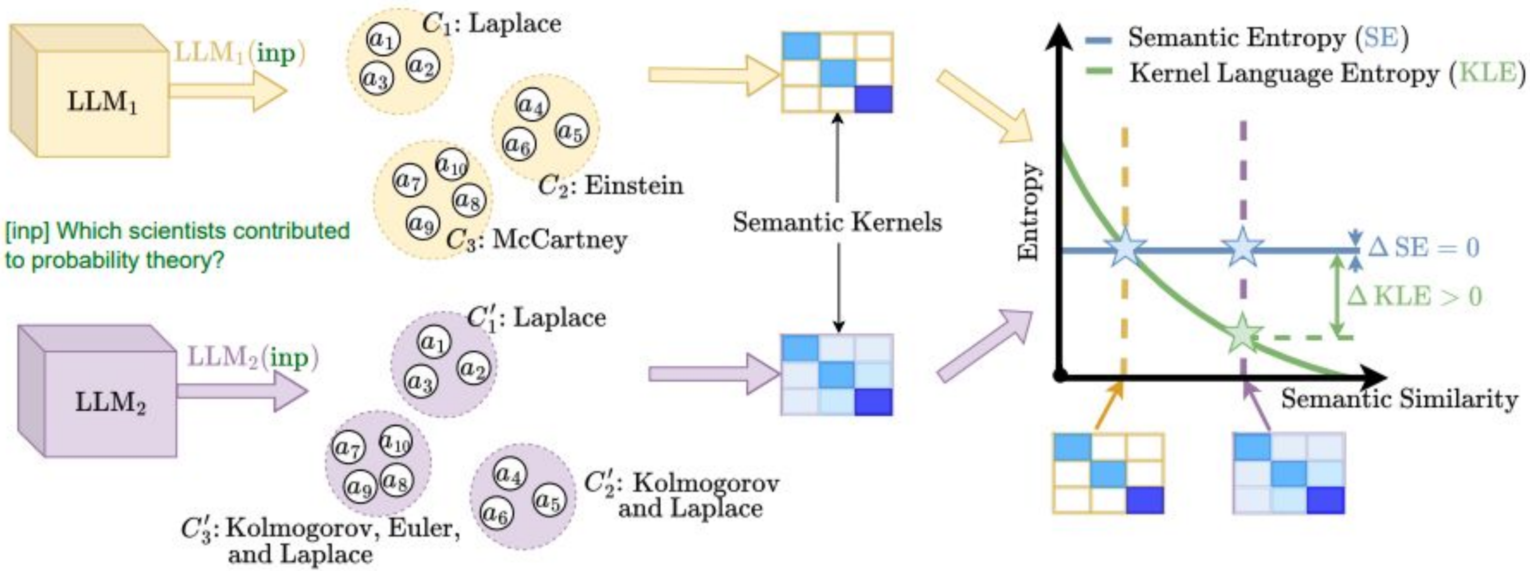
(b) Scenario 2: Some semantic equivalence

| Answer<br>$\mathbf{s}$ | Likelihood<br>$p(\mathbf{s} \mid x)$ | Semantic likelihood<br>$\sum_{\mathbf{s} \in c} p(\mathbf{s} \mid x)$ |
|---|---|---|
| **Paris** | 0.5 | 0.9 |
| **It's Paris** | 0.4 | |
| London | 0.1 | 0.1 |
| Entropy | 0.94 | 0.33 |

*Kuhn et al., 2023 "SEMANTIC UNCERTAINTY: LINGUISTIC INVARIANCES FOR UNCERTAINTY ESTIMATION IN NATURAL LANGUAGE GENERATION" (ICLR'23)*

# Uncertainty Estimation (UE) – Semantic-based

**Kernel Language Entropy**: considering inter-cluster similarity




Nikitin et al., 2024 *"Kernel Language Entropy: Fine-grained Uncertainty Quantification for LLMs from Semantic Similarities" (NeurIPS'24)*

# Uncertainty Estimation – Uncertainty Decomposition

**Uncertainty** = **Epistemic Uncertainty** + **Aleatoric Uncertainty**

Model uncertainty:
- Model lack of knowledge
- Suboptimal modeling
- Perturbation randomness
- Reducible with stronger model

**~ Parametric Knowledge Boundary**
**- Outward Knowledge Boundary**

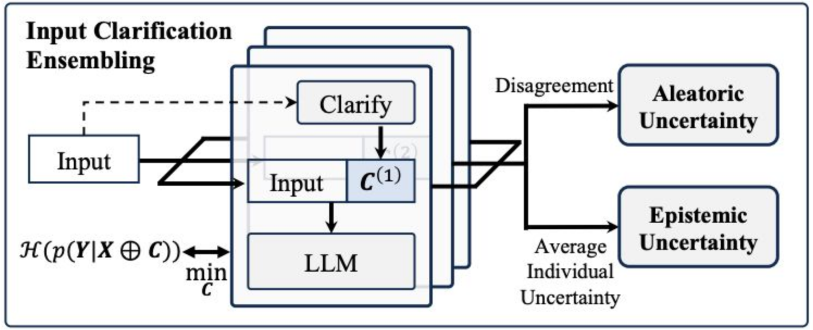
Data uncertainty:
- Question ambiguity
- Multiple answers
- Generation randomness (entropy)
- Less likely to be reducible

**~ Outward Knowledge boundary**

Most current works **do not distinguish** the two types of uncertainty and focus on the general identification of the **Outward Knowledge Boundary.**

*Bai et al., 2024 "Decomposing Uncertainty for Large Language Models through Input Clarification Ensembling" (ICML'24)*
*Gao et al., 2024 "SPUQ: Perturbation-Based Uncertainty Quantification for Large Language Models" (EACL'24)*
*Yadkori et al., 2024 "To Believe or Not to Believe Your LLM" (CoRR'24)*
*Ahdritz et al., 2024 "Distinguishing the Knowable from the Unknowable with Language Models" (ICML'24)*

# Uncertainty Estimation – Uncertainty Decomposition




$$\mathcal{H}(q(\boldsymbol{Y}|\boldsymbol{X})) = \underbrace{\mathcal{I}(\boldsymbol{Y};\boldsymbol{C}|\boldsymbol{X})}_{①'} + \underbrace{\mathbb{E}_{q(\boldsymbol{C}|\boldsymbol{X})}\mathcal{H}(q(\boldsymbol{Y}|\boldsymbol{X}\oplus\boldsymbol{C}))}_{②'}.$$

**Aleatoric uncertainty (1):**
mutual information between the model output distribution and the clarifications.
**Epistemic uncertainty (2):** average entropy of the output distribution given different clarifications.

*Bai et al., 2024 "Decomposing Uncertainty for Large Language Models through Input Clarification Ensembling" (ICML'24)*

# Uncertainty Estimation – Conformal Prediction

For classification task, conformal prediction produces a prediction set of labels $\mathcal{C}(X_t) \subset \mathcal{Y}$

$$p(Y_t \in \mathcal{C}(X_t)) \geq 1 - \alpha,$$

1. Identify a heuristic notion of uncertainty based on the model $f$;
2. Define a conformal score function $s(X, Y) \in \mathbb{R}$ with larger scores encoding worse agreement between $X$ and $Y$; **e.g., using softmax score corresponding to the true label** $s(X, Y) = 1 - f(X)_Y$
3. Compute conformal scores on the calibration set $s_1 = s(X_c^{(1)}, Y_c^{(1)}), \ldots, s_n = (X_c^{(n)}, Y_c^{(n)})$ and calculate a threshold $\hat{q}$ as the $\frac{\lceil (n+1)(1-\alpha) \rceil}{n}$ quantile of the calibration scores,

$$\hat{q} = \text{quant}\left(\{s_1, \ldots, s_n\}, \frac{\lceil (n+1)(1-\alpha) \rceil}{n}\right), \tag{2}$$

where $\lceil \cdot \rceil$ is the ceiling function;
4. Construct the prediction set for each test instance $X_t$ as

$$\mathcal{C}(X_t) = \{Y' \in \mathcal{Y} : s(X_t, Y') \leq \hat{q}\}. \tag{3}$$

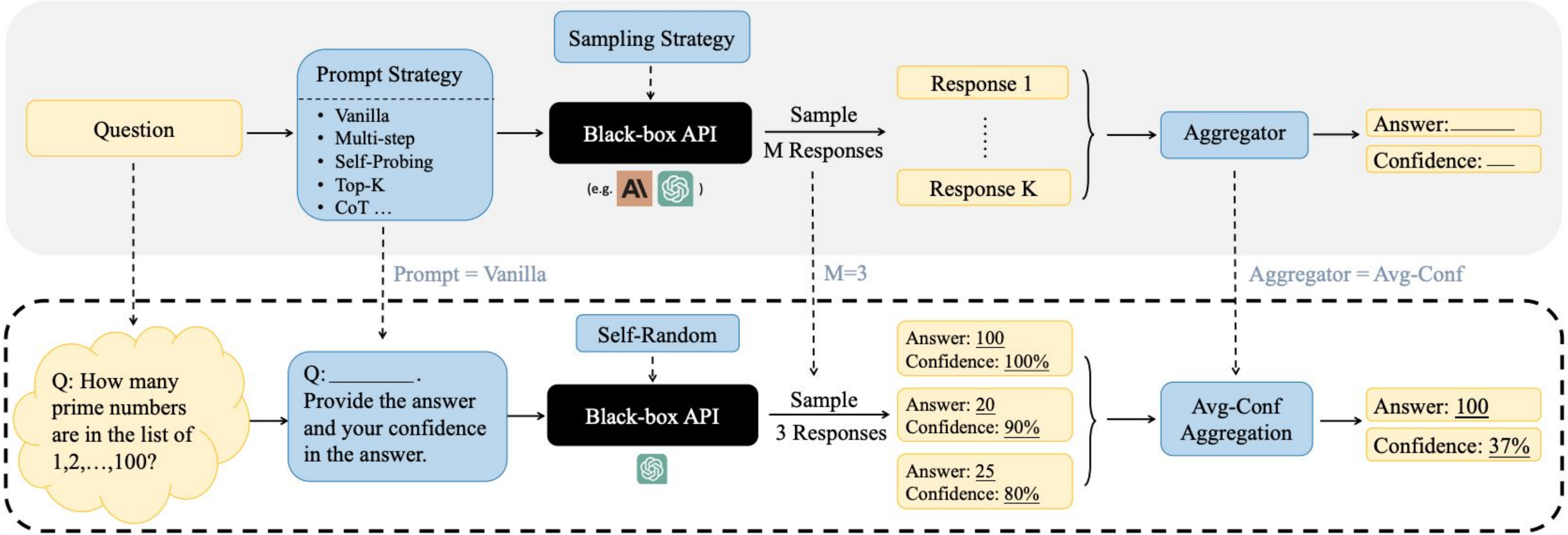
*Ye et al., 2024 "Benchmarking LLMs via Uncertainty Quantification" (NeurIPS'24)*

# Knowledge Boundary Identification

**Outline**

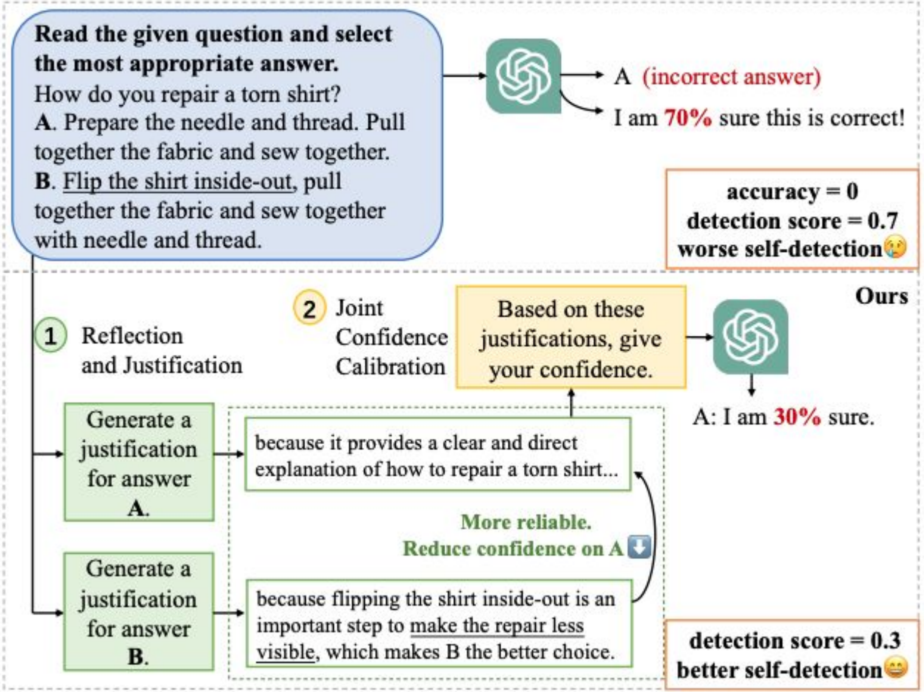
- ❑ Uncertainty Estimation (UE)
- ❑ Confidence Calibration
  - ❑ Prompt-based Calibration
  - ❑ Fine-tuning for Calibration
- ❑ Internal States Probing

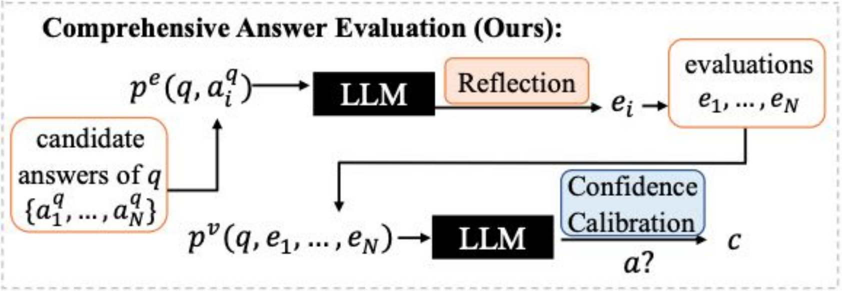
# Confidence Calibration – Prompt-based Calibration




*Xiong et al., 2024 "CAN LLMS EXPRESS THEIR UNCERTAINTY? AN EMPIRICAL EVALUATION OF CONFIDENCE ELICITATION IN LLMS" (ICLR'24)*

# Confidence Calibration – Prompt-based Calibration



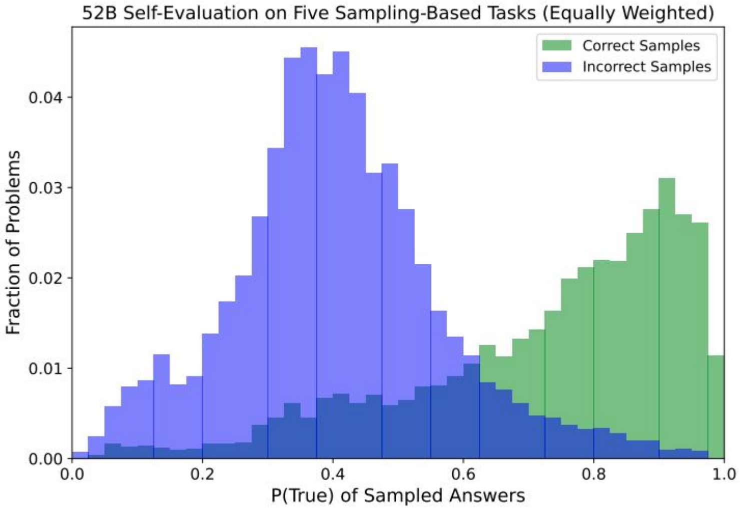



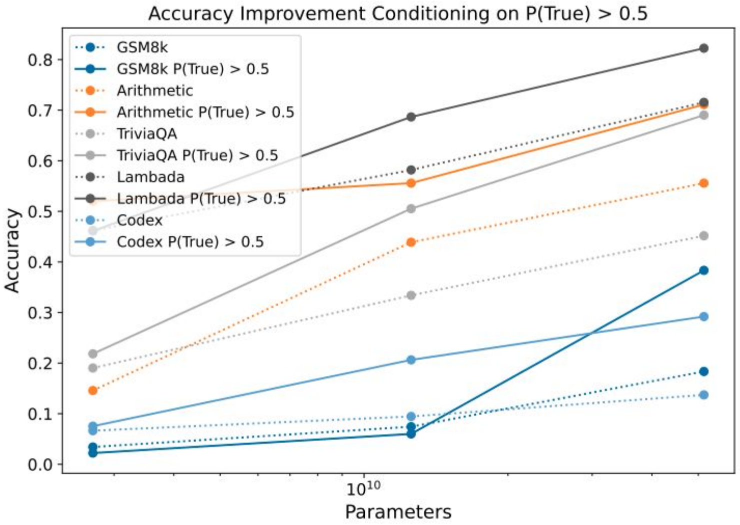

Li et al., 2024 "Think Twice Before Trusting: Self-Detection for Large Language Models through Comprehensive Answer Reflection" (EMNLP'24)

# Confidence Calibration – Prompt-based Calibration

***P(True)***

```
Question: Who was the first president of the United States?
Proposed Answer: George Washington
Is the proposed answer:
 (A) True
 (B) False
The proposed answer is:
```

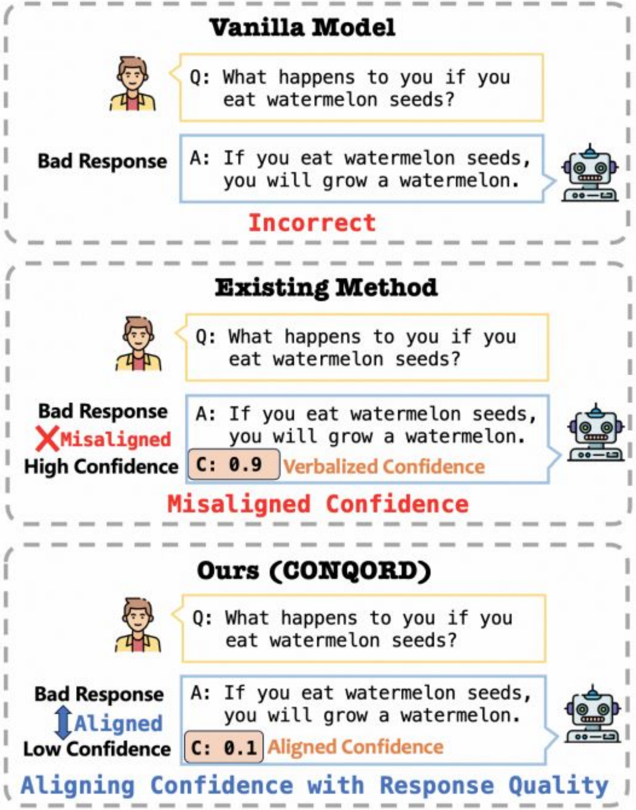



*Kadavath et al., 2022 "Language Models (Mostly) Know What They Know" (CoRR'22)*

# Confidence Calibration – Fine-tuning for Calibration

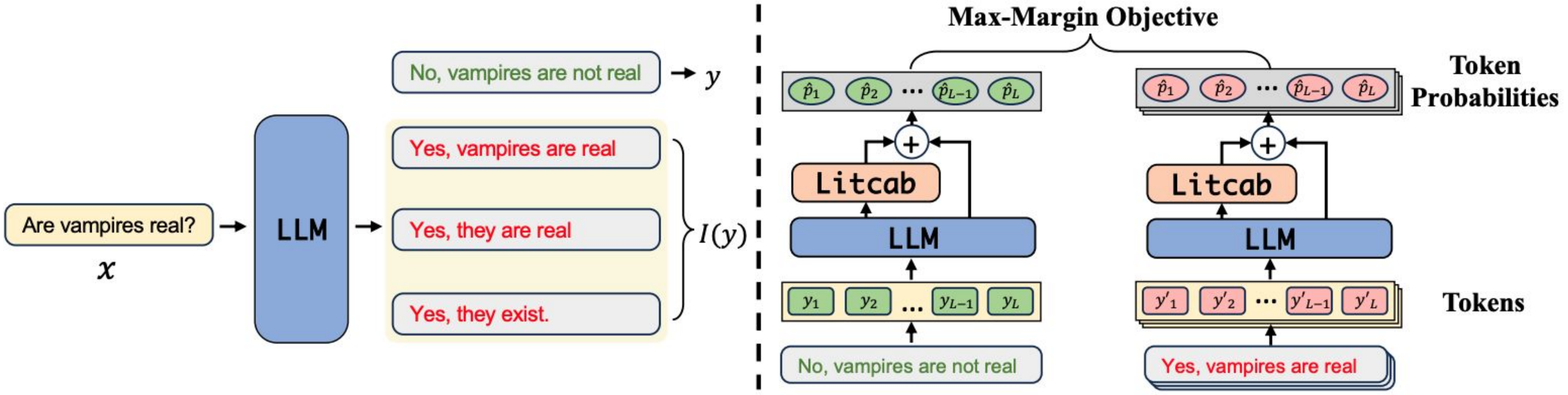
## Fine-tuning for verbalized confidence expression




*Tao et al., 2024 "When to Trust LLMs: Aligning Confidence with Response Quality" (ACL'24)*

# Confidence Calibration – Fine-tuning for Calibration

## Adjusting token probability




*Liu et al., 2024 "LITCAB: LIGHTWEIGHT LANGUAGE MODEL CALIBRATION OVER SHORT- AND LONG-FORM RESPONSES" (ICLR'24)*

# Knowledge Boundary Identification

**Outline**

- ❑ Uncertainty Estimation (UE)
- ❑ Confidence Calibration
- ❑ Internal State Probing

# Internal State Probing – Inference-Time Intervention (ITI)

**Motivation**

LLMs often know factual truth internally but still output falsehoods.

**Key Observation**

There's ~40% gap between what LLMs' hidden activations encode (via probe accuracy) vs. what they generate (output accuracy).

**Proposed Method (ITI)**

- During inference, identify a sparse subset of attention heads whose activations correlate with truth (measured by TruthfulQA).
- Shift activations along these "truthful directions" to nudge output toward truth.

**Result**

On Alpaca (instruction-tuned LLaMA), truthfulness jumped from 32.5% → 65.1% on TruthfulQA.

*Li et al., 2023 "Inference-Time Intervention: Eliciting Truthful Answers from a Language Model" (NeurIPS'23)*

# Internal State Probing – Universal Truthfulness Hyperplane

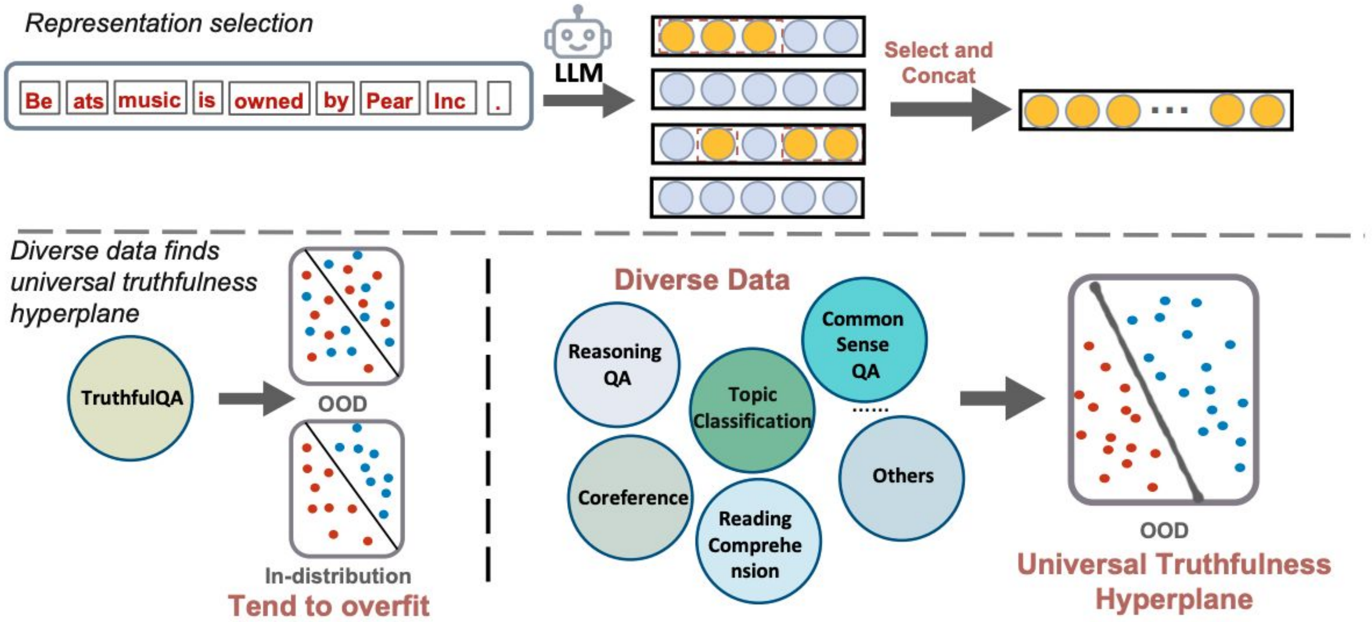



Figure 1: **Top**: we extract representations from the last token of the input sequence, then specific locations of the hidden states inside the LLM are selected and concatenated as input to train the probe. **Bottom**: Previous works mainly train the linear probe on one dataset which tends to overfit spurious features. Our work utilizes diverse datasets to examine whether a universal truthfulness hyperplane exists that can generalize to out-of-domain data.

*Liu et al., 2024 "On the universal truthfulness hyperplane inside llms" (EMNLP'24)*

# Knowledge Boundary Identification – Summary

**Uncertainty Estimation**

- Uncertainty Decomposition
- Token Probability-based
- Semantic-based
- Conformal Prediction

**Confidence Calibration**

- Prompt-based Calibration
- Fine-tuning for Calibration

**Internal State Probing**

Identification approaches should be designed for different knowledge boundaries, suiting different mitigation approaches.