Inv-Entropy: A Fully Probabilistic Framework for Uncertainty Quantification in Language Models

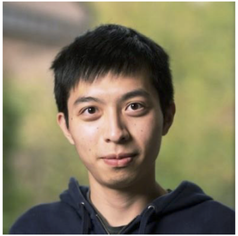
Haoyi Song

University of Michigan Ann Arbor

INFORMS 2025 Atlanta

Collaborators

Ruihan Ji
University of Minnesota



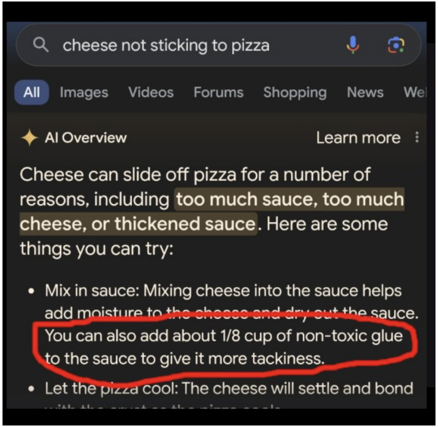
Naichen Shi
Northwestern University

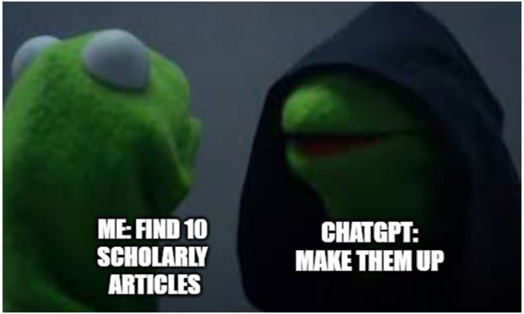
Lai Fan
University of Illinois

Raed Al Kontar
University of Michigan

Why Quantify Uncertainty?

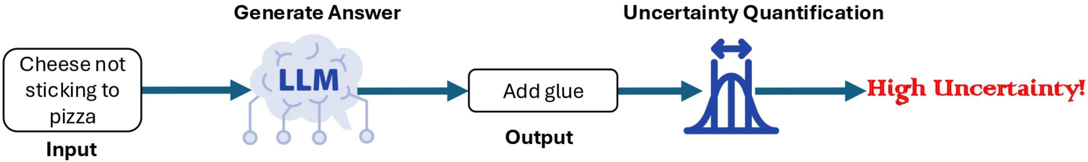
- LLMs can **make mistakes** and **hallucinate** (produce confidently wrong outputs).
- These undermine reliability in high-stakes applications (healthcare, legal and autonomous systems).





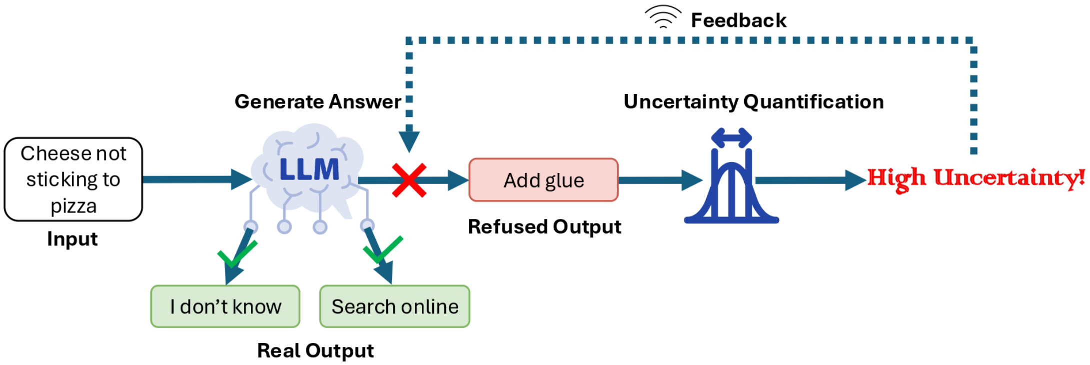
Why Quantify Uncertainty?

- Uncertainty Quantification enables LLMs to acknowledge their confidence in a generated output.

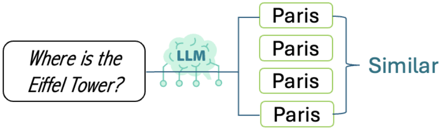
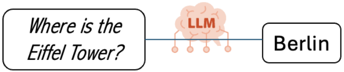
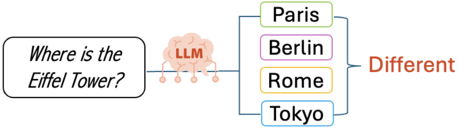


Why Quantify Uncertainty?

- Uncertainty Quantification enables LLMs to acknowledge their confidence in a generated output.
- It also allows downstream applications such as selective prediction.



What Exists?

	Token-probability methods [1-3]	Sampling-based methods [4-6]
Low Uncertainty	 P(Paris | Where is the Eiffel Tower?) = 0.9	
High Uncertainty	 P(Berlin | Where is the Eiffel Tower?) = 0.05	
Limitation	• Inapplicable to black-box LLMs • P(·) can be misspecified • Miss confidently wrong cases	• Heuristic • lack probabilistic foundation • Miss confidently wrong cases

Plot adapted from [6]

Our Goal

Develop a **fully probabilistic framework** to quantify uncertainty in LLMs:

Input → Random Variable

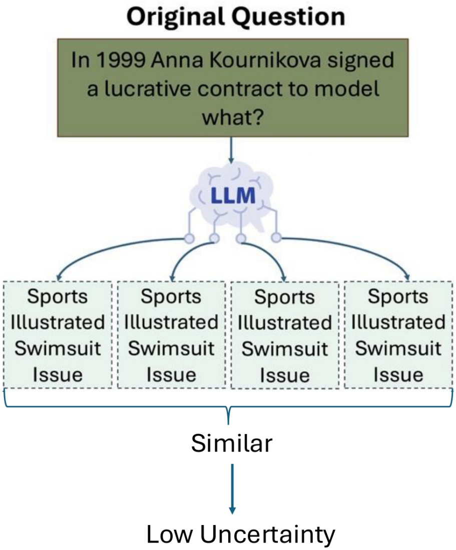
Output → Random Variable

→ Jointly Distributions → Uncertainty

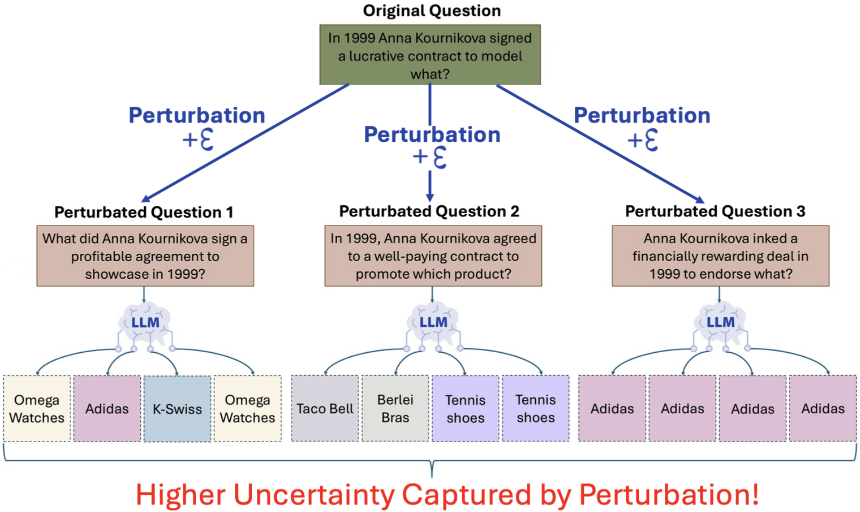
Our Tools

Our main tools are **Perturbation** and **Dual Random Walk**

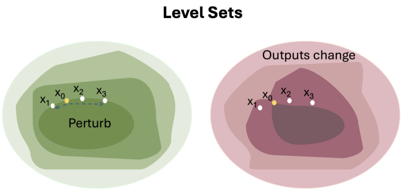
Why Perturb? An Example



Why Perturb? An Example



Why Perturb? A Theory



(a) Ground Truth (b) LLM

Var(perturbed outputs) $\sim$ Misalignment(Ground Truth, LLM)

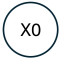
Theorem (Informal)

Assume $f^\star$ is the ground truth function, $\hat{f}$ is the LLM, x_0 is the original input and x' is the perturbed input with variance σ^2. Then,

$$\frac{1}{\sigma^2}\mathrm{Var}[\hat{f}(x')] = \|\nabla \hat{f}(x_0)\|^2 - \frac{\left(\nabla \hat{f}(x_0)^\top \nabla f^\star(x_0)\right)^2}{\|\nabla f^\star(x_0)\|^2} + \mathcal{O}(\sigma)$$

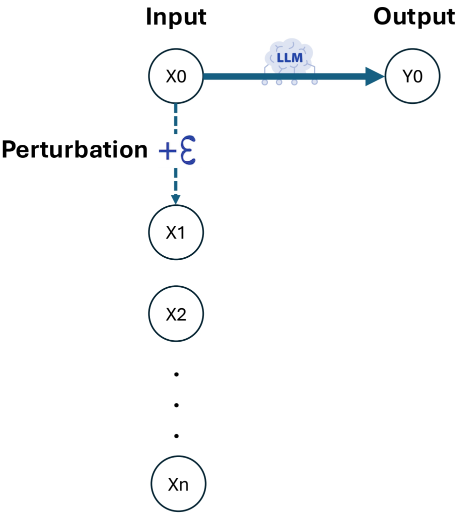
Setting

Input

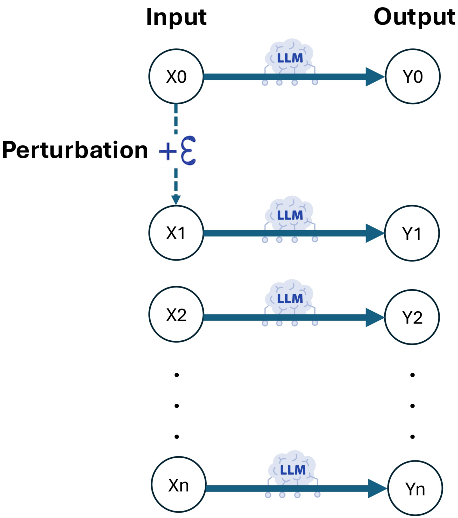


Setting

Setting



Setting


Input
Output
X0
LLM
Y0
Perturbation $+\varepsilon$
X1
LLM
Y1
X2
LLM
Y2
Xn
LLM
Yn

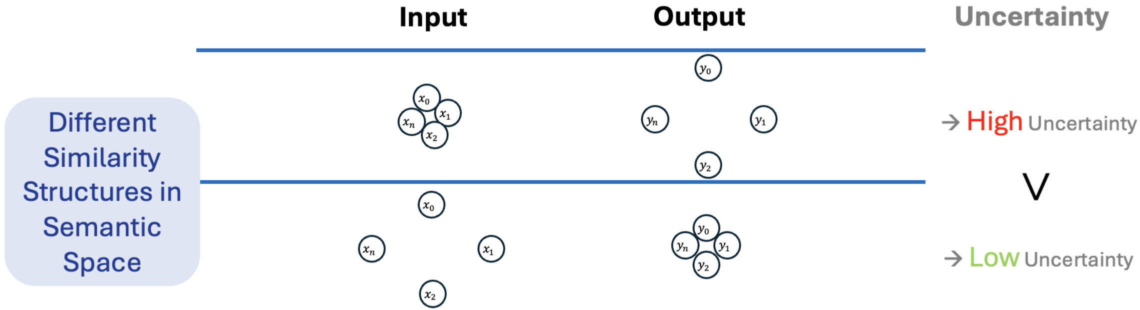
Our Goal: Fully Probabilistic Framework

$$\underbrace{x_0 \quad x_1 \quad x_2 \quad \cdots \quad x_n}_{\mathbb{P}(X)} \qquad\qquad \underbrace{y_0 \quad y_1 \quad y_2 \quad \cdots \quad y_n}_{\mathbb{P}(Y)}$$

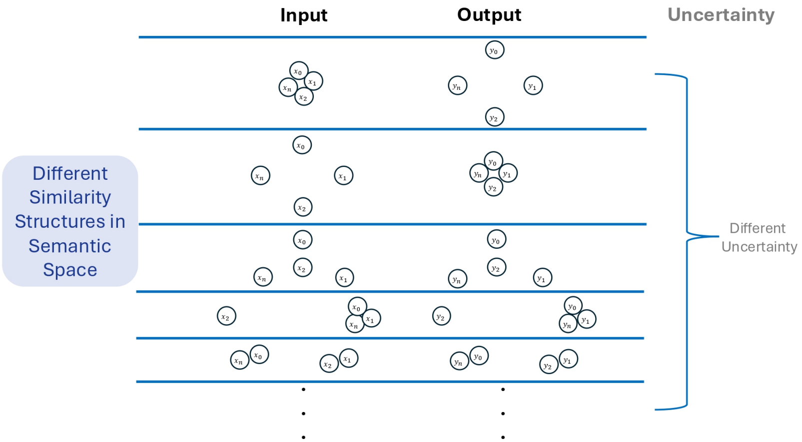
$$\underbrace{x_0 \quad x_1 \quad x_2 \quad \cdots \quad x_n}_{\mathbb{P}(X|Y=y_j)} \longleftarrow y_j \qquad\qquad \underbrace{y_0 \quad y_1 \quad y_2 \quad \cdots \quad y_n}_{\mathbb{P}(Y|X=x_i)} \longleftarrow x_i$$

$\mathbb{P}(X)$ $\mathbb{P}(Y)$ $\mathbb{P}(X|Y)$ $\mathbb{P}(Y|X)$ $\longrightarrow$ **Uncertainty** (e.g. Var(Y), H(Y|X), ...)

Similarity Structures in Semantic Space



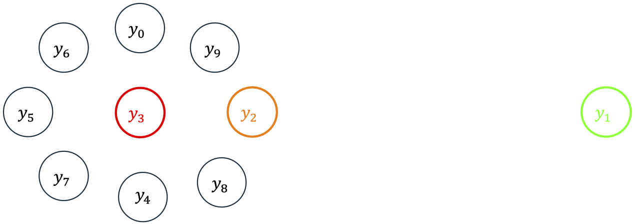
Similarity Structures in Semantic Space



Both inputs and outputs form their own **similarity** structures in the **semantic space**, causing varying degrees of **uncertainty**.

Principle of Distribution Design

$\mathbb{P}(Y)$ assigns higher probability to dense regions in the semantic space



$$\mathbb{P}(Y = y_3) > \mathbb{P}(Y = y_2) > \mathbb{P}(Y = y_1)$$

Constructing the distribution $\mathbb{P}(Y)$

$$\mathbb{P}(Y = y_j) = \frac{1}{n+1} \sum_{i=0}^{n} \underbrace{\frac{a_{\text{Similarity}}(y_i, y_j)}{\sum_k a_{\text{Similarity}}(y_i, y_k)}}_{\text{closeness of } y_j \text{ as a neighbor of } y_i} .$$

Random Walk: Interpretation of $\mathbb{P}(Y = y_j)$

closeness of y_j as a neighbor of y_i $\Leftrightarrow$ **Transition probability** $(y_i \to y_j)$

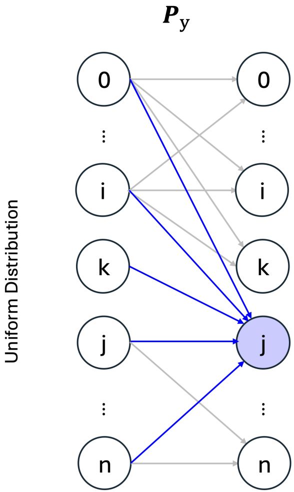
Model $\{y_0, y_1, \ldots, y_n\}$ as **a Random Walk**, with **transition probabilities** P_y defined by:

$$\mathbf{P}_y[i, j] = \frac{a_{\text{Similarity}}(y_i, y_j)}{\sum_k a_{\text{Similarity}}(y_i, y_k)}.$$

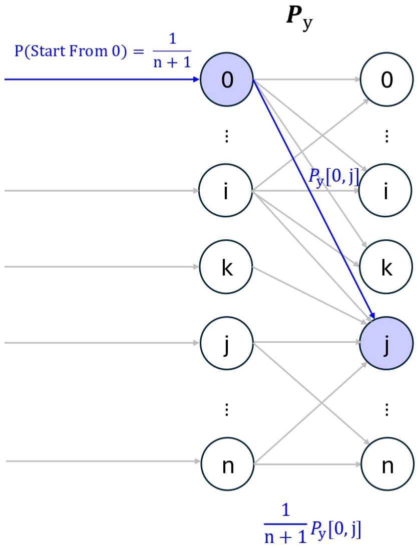
$$\pi_{\text{Uniform}} = [\frac{1}{n+1}, \frac{1}{n+1}, \ldots, \frac{1}{n+1}]$$

$$\mathbb{P}(Y = y_j) = \sum_{i=0}^{n} \frac{1}{n+1} \mathbf{P}_y[i, j] = (\pi_{\text{Uniform}} \mathbf{P}_y)[j].$$

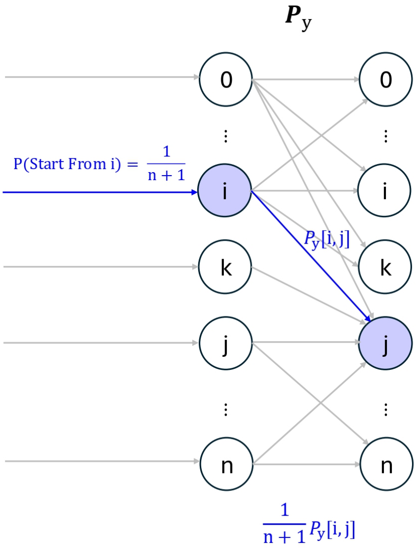
Random Walk: Interpretation of $\mathbb{P}(Y = y_j)$



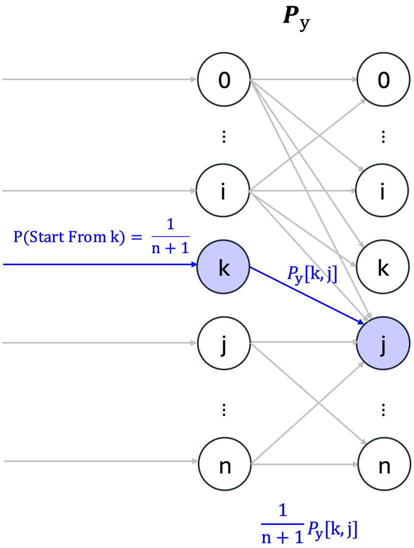
Random Walk: Interpretation of $\mathbb{P}(Y = y_j)$



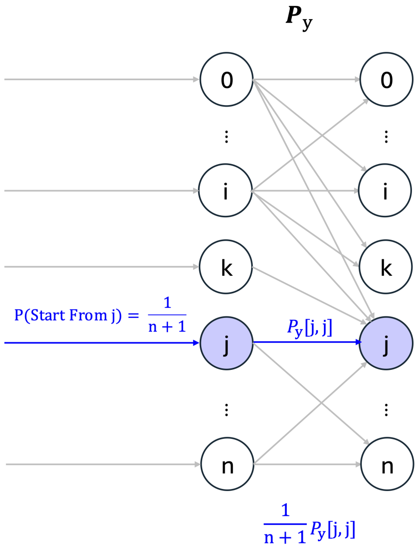
Random Walk: Interpretation of $\mathbb{P}(Y = y_j)$



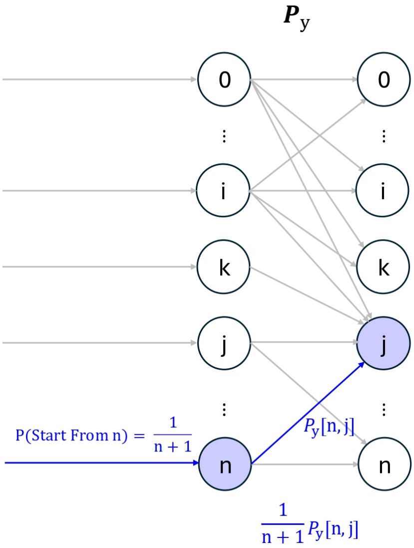
Random Walk: Interpretation of $\mathbb{P}(Y = y_j)$



Random Walk: Interpretation of $\mathbb{P}(Y = y_j)$



Random Walk: Interpretation of $\mathbb{P}(Y = y_j)$



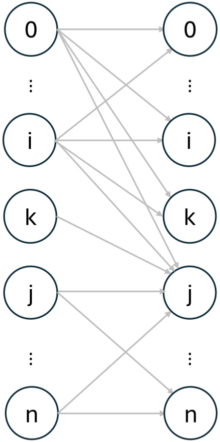
Random Walk on X

Model $\{x_0, x_1, \ldots, x_n\}$ as a Random Walk, with transition probabilities P_x defined by normalized similarity:

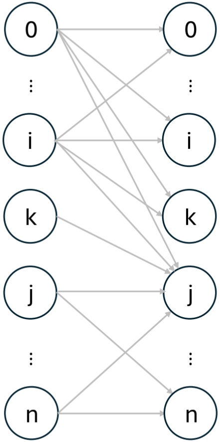
$$P_x[i,j] = \frac{a_{\text{Similarity}}(x_i, x_j)}{\sum_k a_{\text{Similarity}}(x_i, x_k)}.$$

Dual Random Walk on X and Y

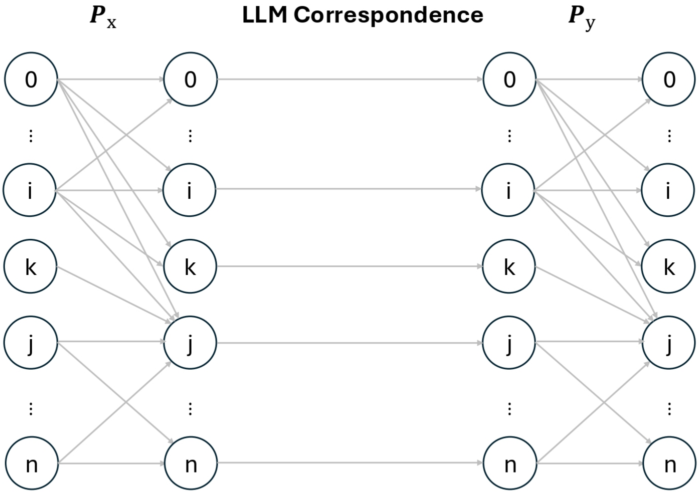
$\boldsymbol{P}_{\mathrm{x}}$



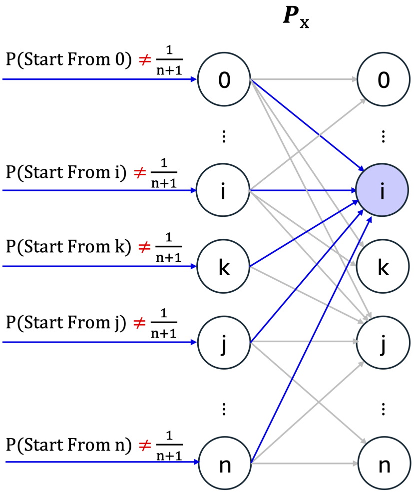
$\boldsymbol{P}_{\mathrm{y}}$



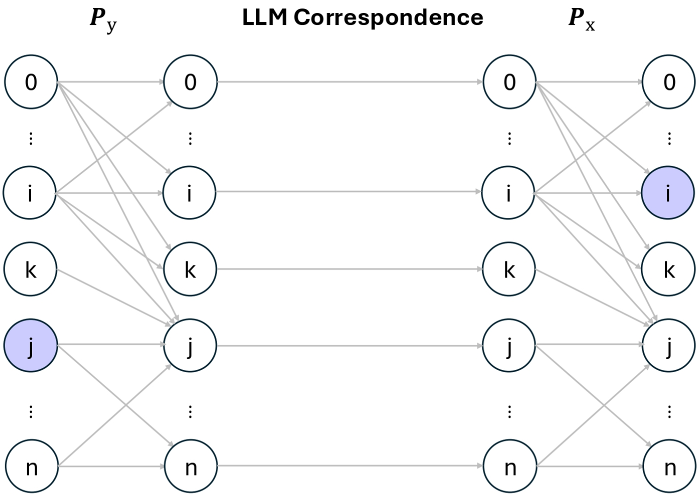
Connecting Input Space and Output Space: LLM Correspondence



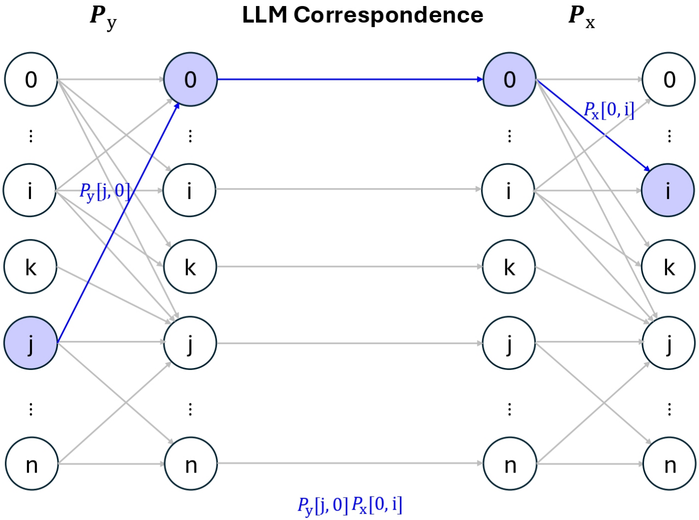
Constructing the Probability $\mathbb{P}(X = x_i \mid Y = y_j)$



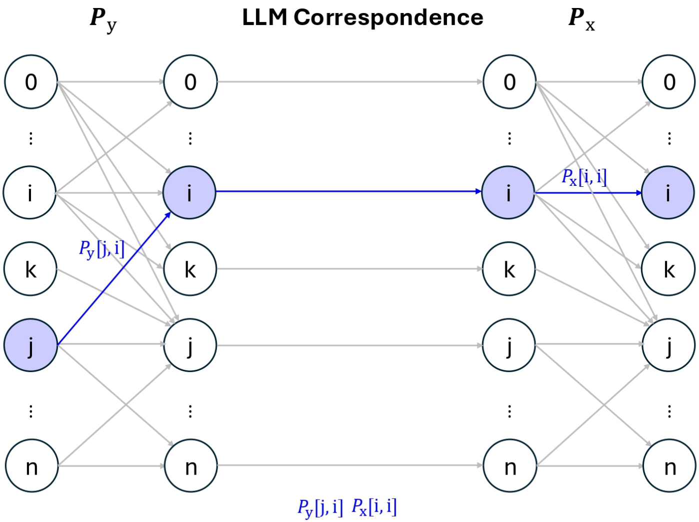
Constructing the Probability $\mathbb{P}(X = x_i \mid Y = y_j)$



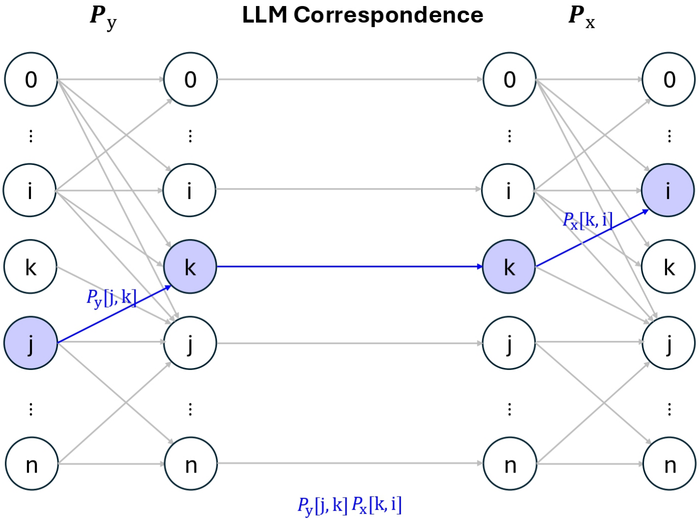
Constructing the Probability $\mathbb{P}(X = x_i \mid Y = y_j)$



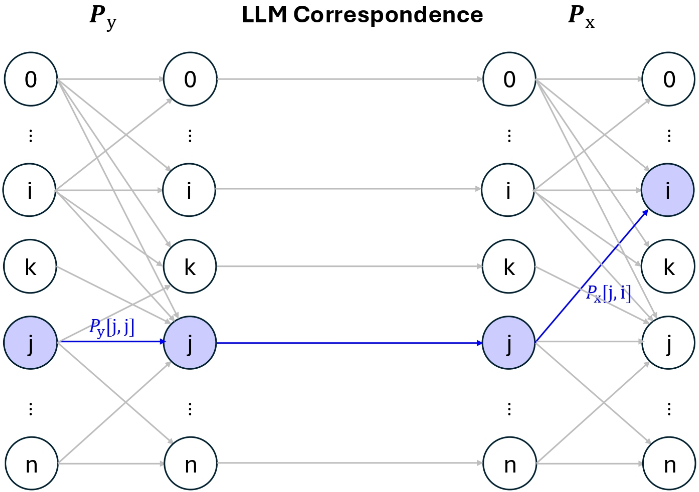
Constructing the Probability $\mathbb{P}(X = x_i \mid Y = y_j)$



Constructing the Probability $\mathbb{P}(X = x_i \mid Y = y_j)$

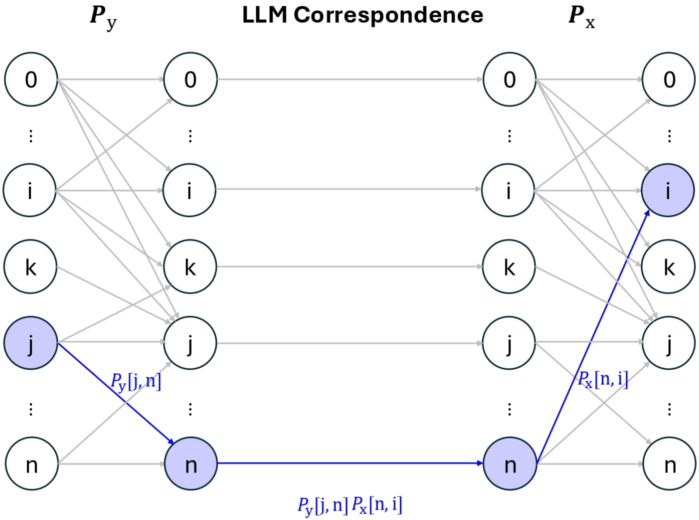


Constructing the Probability $\mathbb{P}(X = x_i \mid Y = y_j)$



$P_y[j, j]\ P_x[j, i]$

Constructing the Probability $\mathbb{P}(X = x_i \mid Y = y_j)$



Constructing the Distribution $P(X \mid Y)$

$$\mathbb{P}(X = x_i \mid Y = y_j) = (\mathbf{P}_y\mathbf{P}_x)[j, i] = \sum_k \underbrace{\frac{a_{\text{Similarity}}(y_j, y_k)}{\sum_\ell a_{\text{Similarity}}(y_j, y_\ell)}}_{\text{output transition } \mathbf{P}_y[j,k]} \underbrace{\frac{a_{\text{Similarity}}(x_i, x_k)}{\sum_m a_{\text{Similarity}}(x_m, x_k)}}_{\text{input transition } \mathbf{P}_x[k,i]}$$

Fully Probabilistic Framework

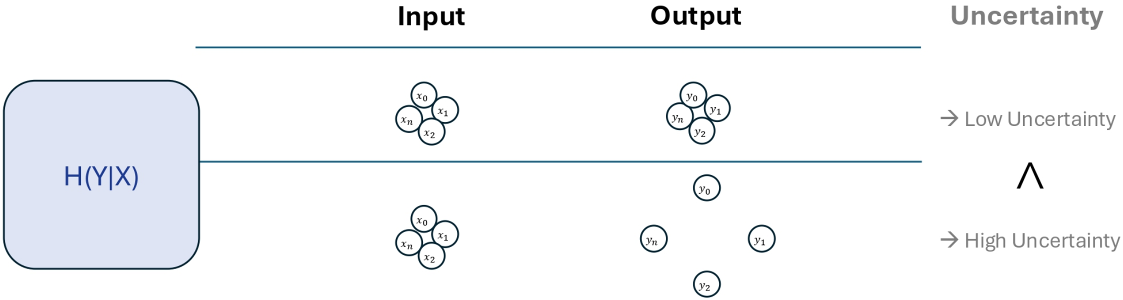
$$\mathbb{P}(Y), \mathbb{P}(X \mid Y) \Rightarrow \mathbb{P}(X), \mathbb{P}(Y \mid X)$$

Flexibility: Define any UQ measures on these distributions (entropy, divergences, distances)

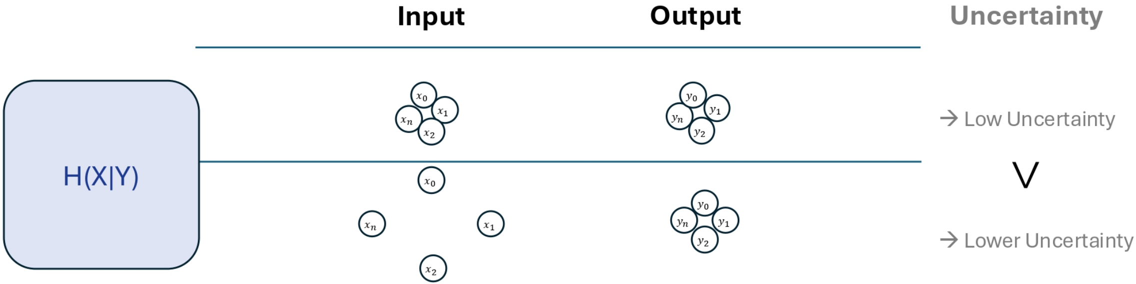
Inv-Entropy

$$H(X \mid Y) = -\sum_{i=0}^{n} \mathbb{P}(x_i \mid y_i) \log \mathbb{P}(x_i \mid y_i).$$

Uncertainty $H(Y \mid X)$



Uncertainty $H(X \mid Y)$



Why Inverse: From Y to X

Experiment Setting

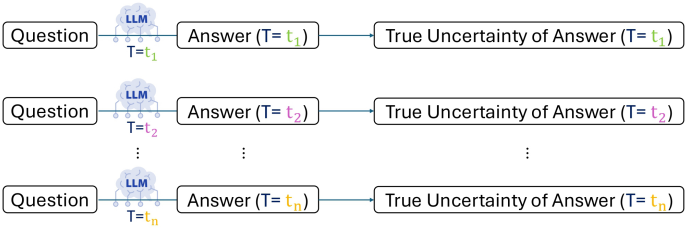
- Dataset: TriviaQA[7] (Question-Answering), SciQ[8] (Question-Answering), MMLU[9] (Multiple-choice)
- Baselines: Semantic Entropy[6], Verbalized Uncertainty (VU)[10], P(True)[11], Lexical Similarity (LexSim)[12], Degree Matrix (DegMat)[13], Long-text Uncertainty Quantification (LUQ)[14], Kernel Language Entropy (KLE)[15]
- Evaluation metrics: AUROC, PRR, Brier Score (uncertainty–correctness misalignment)

Experiments: Correctness-based Evaluation

Method	TriviaQA			SciQ			MMLU		
	AUROC↑	PRR↑	Brier↓	AUROC↑	PRR↑	Brier↓	AUROC↑	PRR↑	Brier↓
Semantic Entropy	0.579	0.517	0.166	0.679	0.763	0.173	0.518	0.690	0.208
VU	0.695	0.723	0.160	0.480	0.677	0.196	0.523	0.654	0.219
P(True)	0.604	0.797	0.172	0.522	0.679	0.215	0.474	0.671	0.215
LexSim	0.649	0.810	0.151	0.681	0.770	0.179	0.643	0.767	0.187
DegMat	0.734	0.882	0.140	0.672	0.802	0.164	0.608	0.771	0.191
LUQ	0.637	0.854	0.148	0.726	0.840	0.159	0.648	0.787	0.180
KLE	0.333	0.704	0.188	0.341	0.592	0.218	0.360	0.612	0.213
Inv-Entropy (ours)	**0.788**	**0.885**	**0.128**	**0.740**	**0.853**	**0.157**	**0.780**	**0.898**	**0.147**

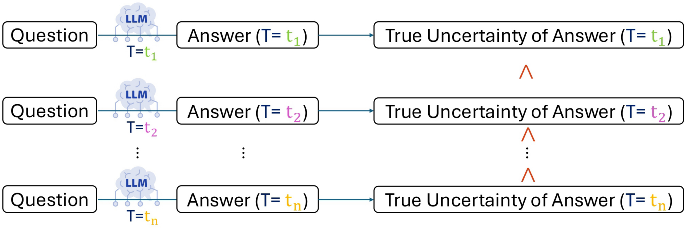
New Evaluation Metric: TSU

Temperature: $t_1 < t_2 < \cdots < t_n$



New Evaluation Metric: TSU

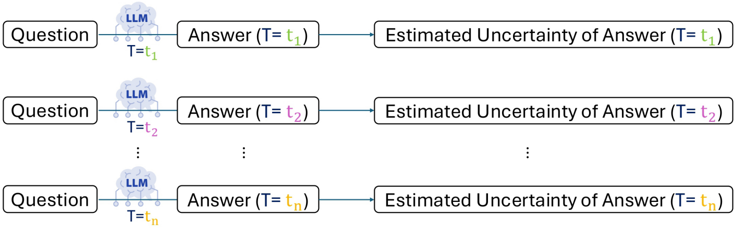
Temperature: $t_1 < t_2 < \cdots < t_n$



True Uncertainty of Answer (T= t_1) $<$ True Uncertainty of Answer (T= t_2) $< \cdots <$ True Uncertainty of Answer (T= t_n)

New Evaluation Metric: TSU

Temperature: $t_1 < t_2 < \cdots < t_n$



? Estimated Uncertainty of Answer (T= t_1) < Estimated Uncertainty of Answer (T= t_2) < ⋯ < Estimated Uncertainty of Answer (T= t_n)

TSU = the percentage of questions in a dataset that satisfy the rule above

New Evaluation Metric: TSU

$$\mathsf{TSU}(t_1, t_2 \dots, t_n) = \frac{1}{|\mathcal{D}|} \sum_{x \in \mathcal{D}} \mathbb{I}(\mathtt{UQ}(x, t_1) < \mathtt{UQ}(x, t_2) < \cdots < \mathtt{UQ}(x, t_n)) \,,$$

where $\mathcal{D}$ = dataset, $x \in \mathcal{D}$ = a question and $\mathtt{UQ}(x, t)$ = estimated uncertainty of x under temperature t.

Free of correctness labels!

Experiments: TSU Evaluation

Method	TriviaQA			MMLU		
	TSU(1.0–1.4)	TSU(0.3–1.0)	TSU(0.3–1.4)	TSU(1.0–1.4)	TSU(0.3–1.0)	TSU(0.3–1.4)
Semantic Entropy	17.35	5.18	3.94	33.20	7.14	2.09
VU	38.78	0.00	0.00	37.62	1.37	0.00
P(True)	3.85	0.00	0.00	5.87	0.00	0.00
LexSim	46.94	9.18	8.16	55.06	24.78	15.28
DegMat	45.37	18.37	13.27	69.39	21.46	14.34
LUQ	48.06	14.78	10.20	61.22	27.55	10.80
KLE	13.45	2.79	0.00	26.53	2.93	0.00
Inv-Entropy (ours)	**77.55**	**30.49**	**19.05**	**73.47**	**34.31**	**18.37**

Q & A

Contributions:

- Provide a **fully probabilistic framework** for Uncertainty Quantification in LLMs
- Provide a rigorous **theory for using perturbation**-based methods
- Introduce an **inverse perspective** that quantifies input diversity given an output
- Introduce **a new evaluation metric TSU** free of correctness labels

Haoyi Song