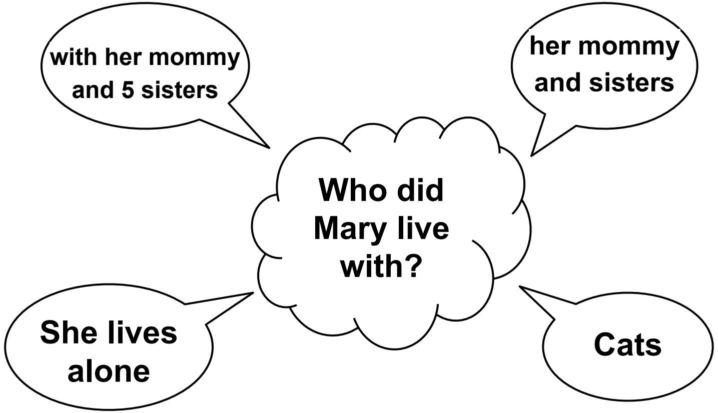
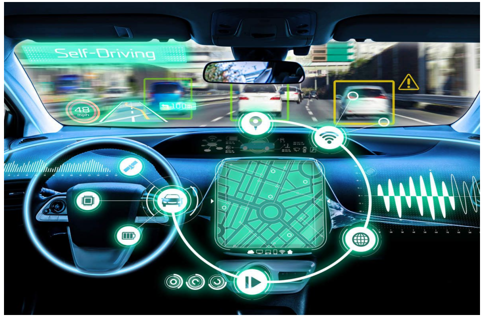
Semantic Density: Uncertainty Quantification for Large Language Models through Confidence Measurement in Semantic Space

Xin Qiu[1], Risto Miikkulainen[1,2]
[1]Cognizant AI Labs,
[2]The University of Texas at Austin
Neurips 2024

Motivation

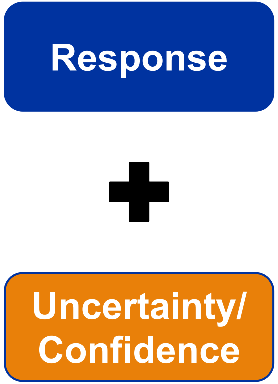
- **Trustworthiness of LLMs**
  - Applications in safety-critical domains
  - Unpredictable hallucinations/misinformation
  - Lack of uncertainty/confidence indicator





Motivation

- **Existing Solutions**
  - **Extending traditional UQ methods to LLMs**
    - Only work for classification tasks
  - **Train additional uncertainty/confidence predictors**
    - Not “off-the-shelf”, need extra data/training
  - **Lexical uncertainty/confidence metrics**
    - Ignoring semantic information
  - **Semantic Entropy**
    - Prompt-wise
    - No fine-grained semantic analysis

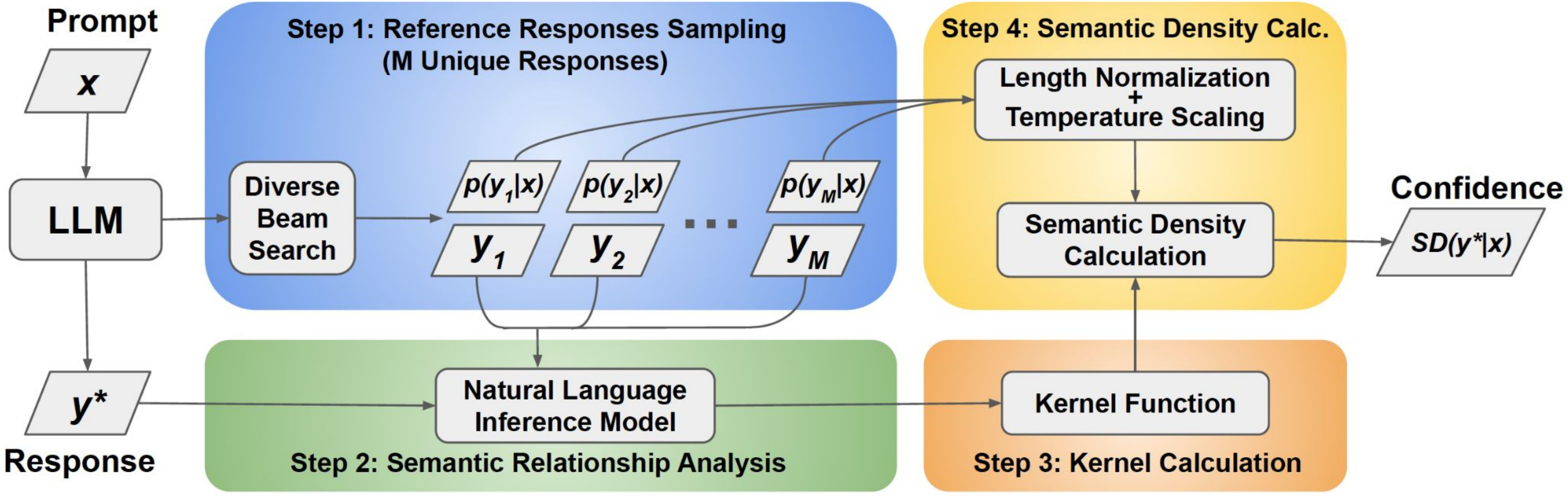


Proposed Solution

- **Semantic Density**
  - Measuring confidence for each response
  - Analyzing output distribution in semantic space
  - **Main advantages:**
    - Response-wise indicator
    - Considering fine-grained semantic relationships
    - “Off-the-shelf”:
      - directly applicable to pre-trained LLMs
      - no fine-tuning/re-training
    - Work for free-form generation tasks

Proposed Solution

- **Semantic Density**



Cognizant

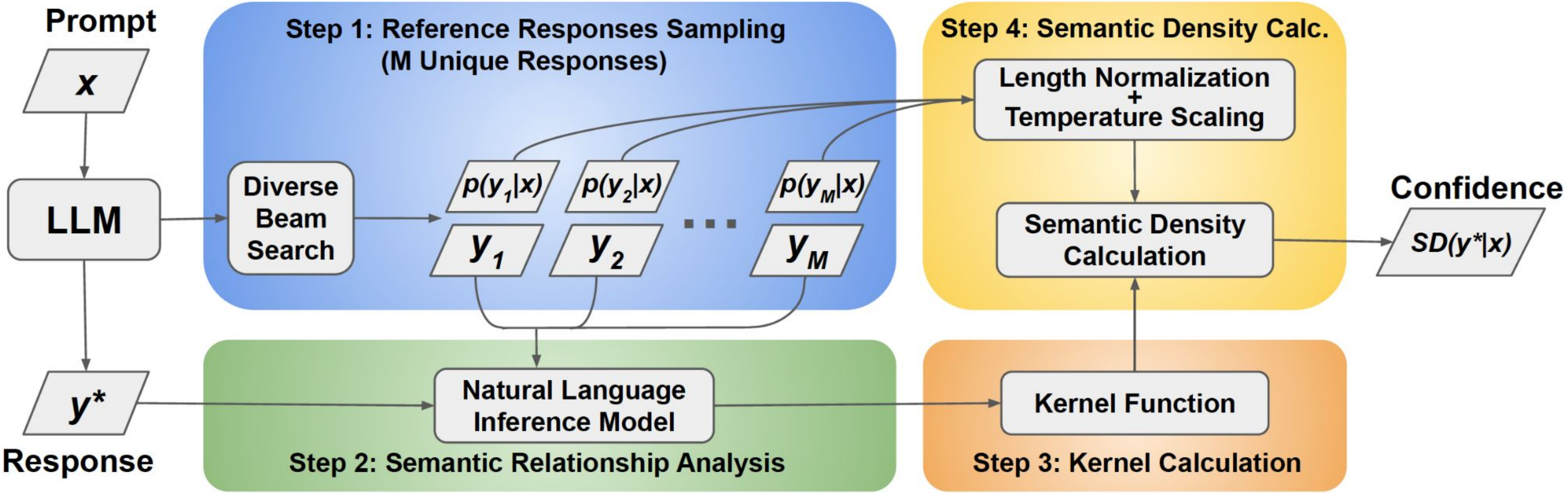
Proposed Solution

- **Semantic Density**

Step 2: Semantic Relationship Analysis

Contradictory Prob. | Neutral Prob. | Entailment Prob.

$$\begin{aligned}\mathbb{E}(||\boldsymbol{v}_* - \boldsymbol{v}_i||^2) &= 1^2 \cdot \boxed{p_{\mathrm{c}}(\boldsymbol{y}_*, \boldsymbol{y}_i|\boldsymbol{x})} + (\tfrac{\sqrt{2}}{2})^2 \cdot \boxed{p_{\mathrm{n}}(\boldsymbol{y}_*, \boldsymbol{y}_i|\boldsymbol{x})} + 0^2 \cdot \boxed{p_{\mathrm{e}}(\boldsymbol{y}_*, \boldsymbol{y}_i|\boldsymbol{x})} \\ &= p_{\mathrm{c}}(\boldsymbol{y}_*, \boldsymbol{y}_i|\boldsymbol{x}) + \tfrac{1}{2} \cdot p_{\mathrm{n}}(\boldsymbol{y}_*, \boldsymbol{y}_i|\boldsymbol{x})\end{aligned}$$



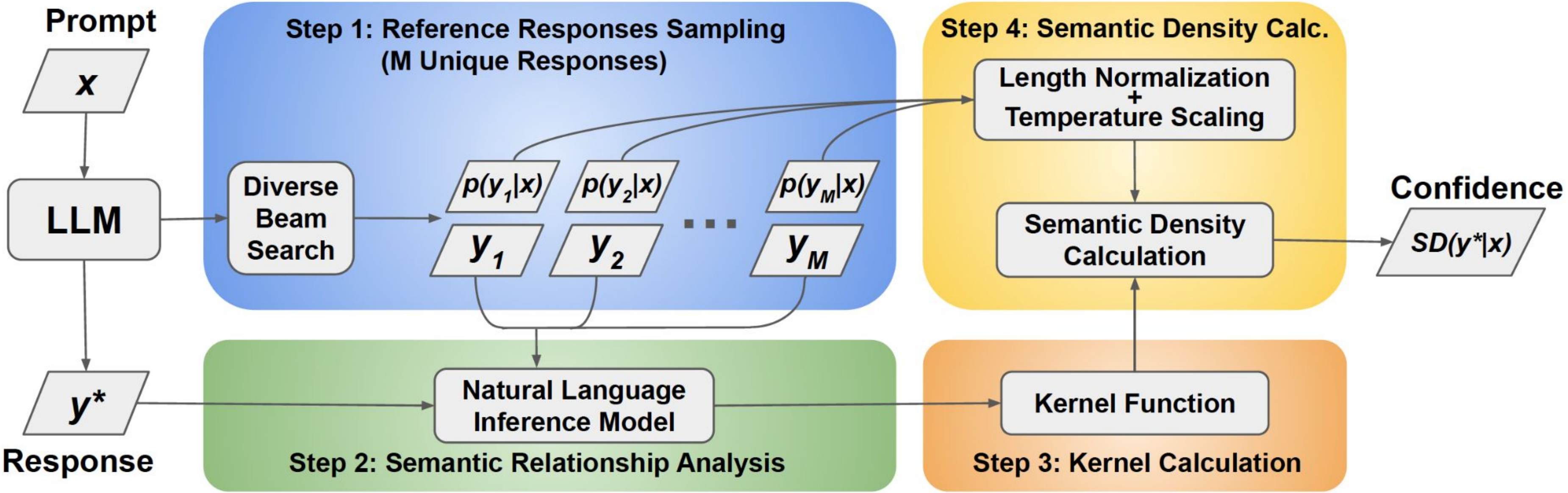
Proposed Solution

- **Semantic Density**

Step 3: Kernel Calculation

$$K(v_* - v_i) = (1 - \boxed{||v_* - v_i||^2})\mathbf{1}_{||v_* - v_i|| \leq 1}$$

Estimated by $\mathbb{E}(||v_* - v_i||^2)$



Cognizant

Proposed Solution

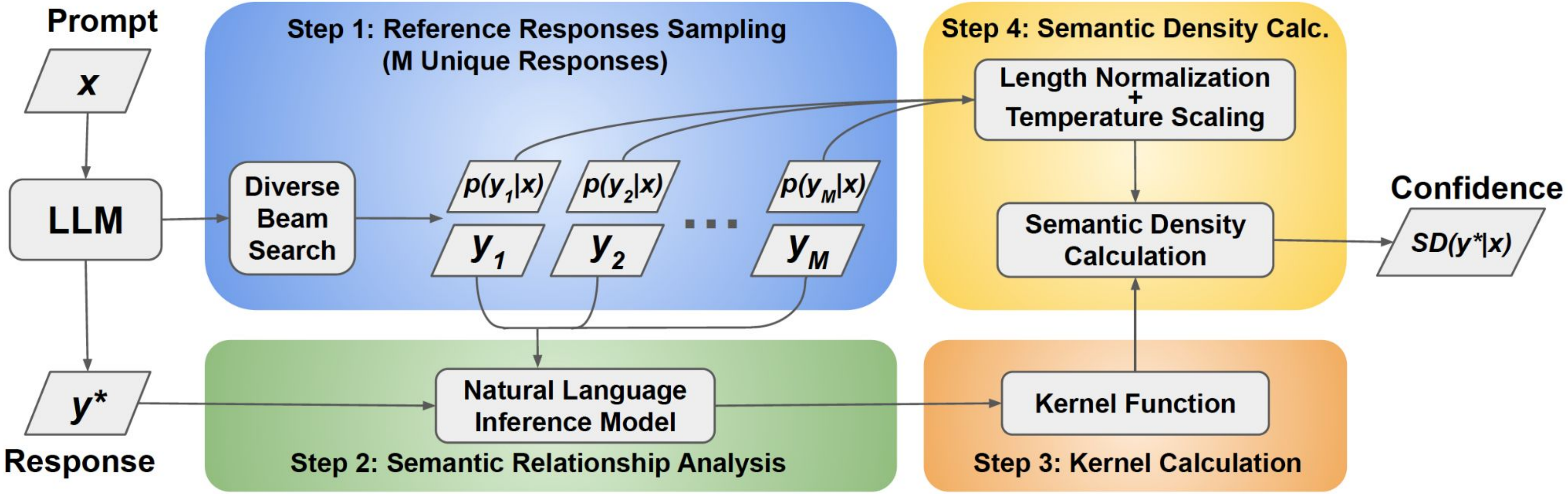
- **Semantic Density**

Step 4: Semantic Density Calc.

$$\mathrm{SD}(\boldsymbol{y}_*|\boldsymbol{x}) = \frac{1}{\sum_{i=1}^{M} p(\boldsymbol{y}_i|\boldsymbol{x})} \sum_{i=1}^{M} \boxed{p(\boldsymbol{y}_i|\boldsymbol{x})} K(\boldsymbol{v}_* - \boldsymbol{v}_i)$$

Post-processing by Length-normalization + Temperature Scaling

Sampling Probability by LLM



Semantic Density

- **Semantic Space**
  - **Oracle/idealized form:**
    - **Contextual embedding:** $\boldsymbol{v} = E(\boldsymbol{y}|\boldsymbol{x})$
    - **Fixed norm:** $||\boldsymbol{v}|| = \frac{1}{2}, \text{ for } \boldsymbol{v} = E(\boldsymbol{y}|\boldsymbol{x}), \ \forall \boldsymbol{x}, \boldsymbol{y}$
    - **Reflecting semantic relationship:**

$$||\boldsymbol{v}_i - \boldsymbol{v}_j|| = \begin{cases} 0, \text{ if } \boldsymbol{y}_i \text{ and } \boldsymbol{y}_j \text{ are semantically equivalent given context } \boldsymbol{x} \\ \frac{\sqrt{2}}{2}, \text{ if } \boldsymbol{y}_i \text{ and } \boldsymbol{y}_j \text{ are semantically irrelevant given context } \boldsymbol{x} \\ 1, \text{ if } \boldsymbol{y}_i \text{ and } \boldsymbol{y}_j \text{ are semantically contradictory given context } \boldsymbol{x} \end{cases}$$

$$||\boldsymbol{v}_i - \boldsymbol{v}_j|| < ||\boldsymbol{v}_i - \boldsymbol{v}_k||, \text{ if } \boldsymbol{y}_i \text{ is semantically closer to } \boldsymbol{y}_j \text{ than to } \boldsymbol{y}_k, \text{ given context } \boldsymbol{x}.$$

Semantic Density

- **Semantic**
  - **Implicit form via NLI model**
    - **Natural Language Inference (NLI) Model:**
      - Classifier for semantic relationships
    - **Estimating the expectation of** $||\boldsymbol{v}_* - \boldsymbol{v}_i||$ **:**

$$\mathbb{E}(||\boldsymbol{v}_* - \boldsymbol{v}_i||^2) = 1^2 \cdot \underbrace{\boxed{p_{\mathrm{c}}(\boldsymbol{y}_*, \boldsymbol{y}_i|\boldsymbol{x})}}_{\text{Contradictory Prob.}} + (\frac{\sqrt{2}}{2})^2 \cdot \underbrace{\boxed{p_{\mathrm{n}}(\boldsymbol{y}_*, \boldsymbol{y}_i|\boldsymbol{x})}}_{\text{Neutral Prob.}} + 0^2 \cdot \underbrace{\boxed{p_{\mathrm{e}}(\boldsymbol{y}_*, \boldsymbol{y}_i|\boldsymbol{x})}}_{\text{Entailment Prob.}}$$

$$= p_{\mathrm{c}}(\boldsymbol{y}_*, \boldsymbol{y}_i|\boldsymbol{x}) + \frac{1}{2} \cdot p_{\mathrm{n}}(\boldsymbol{y}_*, \boldsymbol{y}_i|\boldsymbol{x})$$

Cognizant

Semantic Density

- **Kernel Function Calculation**
  - **Traditional Epanechnikov kernel in standard KDE:**

Dimension Dependent

$$K(\boldsymbol{v}) = \boxed{\frac{\Gamma(2+\frac{D}{2})}{\pi^{\frac{D}{2}}}}(1-||\boldsymbol{v}||^2)\mathbf{1}_{||\boldsymbol{v}||\leq 1}$$

  - **Dimension-invariant kernel in semantic density:**

$$K(\boldsymbol{v}_* - \boldsymbol{v}_i) = (1 - \boxed{||\boldsymbol{v}_* - \boldsymbol{v}_i||^2})\mathbf{1}_{||\boldsymbol{v}_* - \boldsymbol{v}_i||\leq 1}$$

Estimated by $\mathbb{E}(||\boldsymbol{v}_* - \boldsymbol{v}_i||^2)$

    - Does not work for standard KDE
    - **Fits well in confidence/uncertainty estimation**
      - Making confidence scores comparable for embedding spaces with different dimensionalities

Semantic Density

- **Semantic Density Calculation**
  - **Expensive extension from standard KDE:**

$$\hat{p}(\boldsymbol{y}_*|\boldsymbol{x}) = \sum_{i=1}^{M} f_i K(\boldsymbol{v}_* - \boldsymbol{v}_i) = \frac{1}{\sum_{i=1}^{M} n_i} \sum_{i=1}^{M} \boxed{n_i} K(\boldsymbol{v}_* - \boldsymbol{v}_i)$$

Number of Occurrences

  - **Cost-effective realization in semantic density:**

$$\mathrm{SD}(\boldsymbol{y}_*|\boldsymbol{x}) = \frac{1}{\sum_{i=1}^{M} p(\boldsymbol{y}_i|\boldsymbol{x})} \sum_{i=1}^{M} \boxed{p(\boldsymbol{y}_i|\boldsymbol{x})} K(\boldsymbol{v}_* - \boldsymbol{v}_i)$$

Post-processing by Length-normalization + Temperature Scaling

Sampling Probability by LLM

Empirical Evaluations

- **Indicator for correctness of responses**
  - Area under the receiver operating characteristic curve (AUROC):

		CoQA					
AUROC	SD	SE[22]	P(True)[21]	Deg[26]	NL[31]	NE[29]	PE[21]
Llama-2-13B	**0.783**	0.633	0.594	0.734	0.709	0.629	0.647
Llama-2-70B	**0.783**	0.621	0.576	0.721	0.716	0.617	0.647
Llama-3-8B	0.738	0.599	0.593	**0.795**	0.676	0.608	0.604
Llama-3-70B	**0.789**	0.608	0.670	0.729	0.698	0.587	0.641
Mistral-7B	**0.788**	0.627	0.667	0.737	0.704	0.614	0.632
Mixtral-8x7B	**0.786**	0.626	0.589	0.728	0.708	0.617	0.651
Mixtral-8x22B	**0.791**	0.614	0.614	0.726	0.700	0.604	0.649

Cognizant

Empirical Evaluations

- **Indicator for correctness of responses**
  - Area under the receiver operating characteristic curve (AUROC):

			TriviaQA				
AUROC	SD	SE	P(True)	Deg	NL	NE	PE
Llama-2-13B	**0.848**	0.672	0.589	0.824	0.675	0.574	0.556
Llama-2-70B	**0.829**	0.677	0.556	0.787	0.714	0.582	0.566
Llama-3-8B	**0.866**	0.662	0.647	0.796	0.834	0.636	0.622
Llama-3-70B	**0.828**	0.663	0.654	0.764	**0.828**	0.611	0.596
Mistral-7B	**0.866**	0.690	0.589	0.828	0.745	0.615	0.536
Mixtral-8x7B	**0.846**	0.685	0.562	0.797	0.795	0.644	0.605
Mixtral-8x22B	**0.829**	0.686	0.604	0.762	0.801	0.644	0.607

Empirical Evaluations

- **Indicator for correctness of responses**
  - Area under the receiver operating characteristic curve (AUROC):

	SciQ						
AUROC	SD	SE	P(True)	Deg	NL	NE	PE
Llama-2-13B	**0.757**	0.570	0.572	0.727	0.693	0.513	0.574
Llama-2-70B	**0.746**	0.643	0.584	0.713	0.637	0.554	0.615
Llama-3-8B	**0.780**	0.611	0.564	0.731	0.686	0.597	0.651
Llama-3-70B	**0.771**	0.613	0.556	0.706	0.724	0.558	0.520
Mistral-7B	**0.771**	0.618	0.568	0.736	0.669	0.565	0.528
Mixtral-8x7B	**0.773**	0.612	0.585	0.716	0.726	0.612	0.658
Mixtral-8x22B	**0.775**	0.620	0.602	0.719	0.715	0.602	0.628

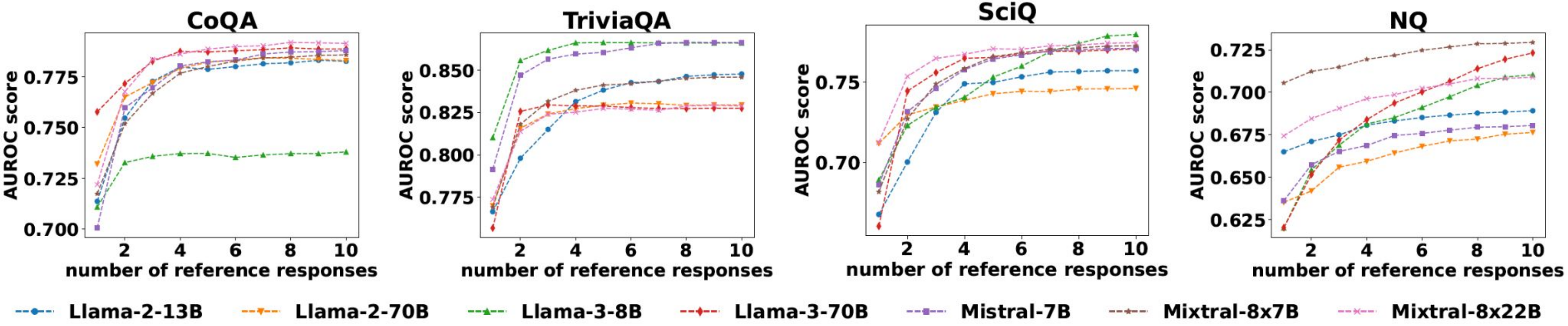
Empirical Evaluations

- **Indicator for correctness of responses**
  - Area under the receiver operating characteristic curve (AUROC):

			NQ				
AUROC	SD	SE	P(True)	Deg	NL	NE	PE
Llama-2-13B	**0.689**	0.581	0.592	0.686	0.588	0.571	0.640
Llama-2-70B	0.676	0.545	0.531	**0.691**	0.567	0.573	0.620
Llama-3-8B	**0.710**	0.583	0.517	0.706	0.601	0.603	0.615
Llama-3-70B	**0.723**	0.577	0.643	0.714	0.631	0.603	0.615
Mistral-7B	**0.680**	0.597	0.523	0.676	0.640	0.635	0.631
Mixtral-8x7B	**0.729**	0.599	0.576	0.720	0.654	0.603	0.608
Mixtral-8x22B	**0.709**	0.577	0.504	0.704	0.638	0.625	0.680

Empirical Evaluations

- **Sensitivity to number of reference responses:**

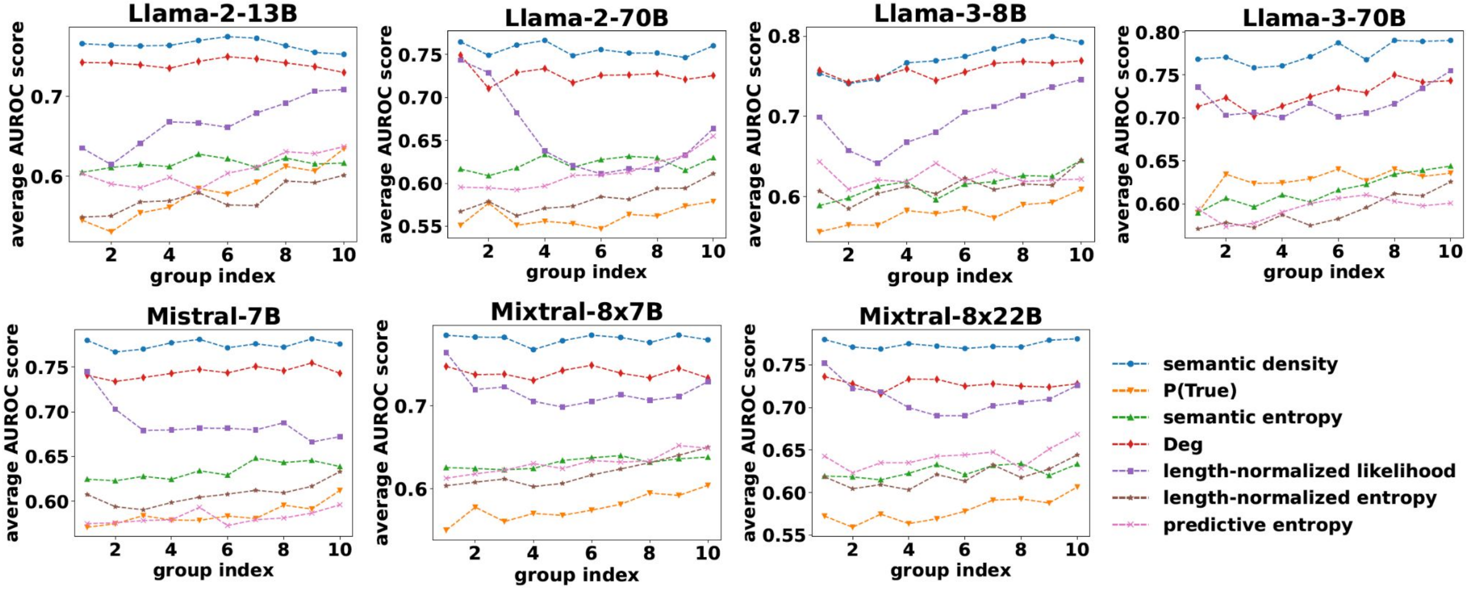


- Robust to reduced number of reference samples

Cognizant

Empirical Evaluations

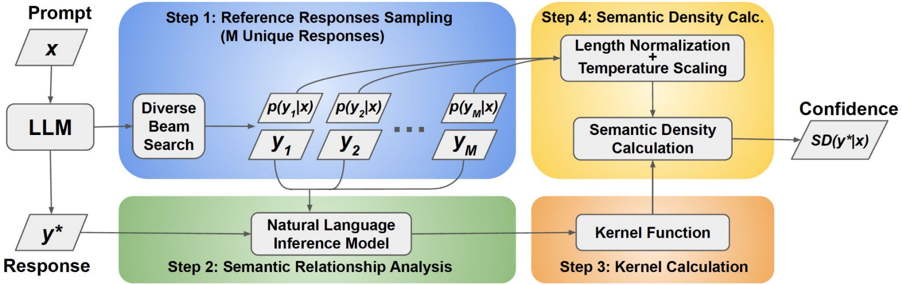
- **Performance changes for responses from different sampling strategies:**



- Robust for both greedy or diverse samplings

Summary

- **Conclusions**
  - Proposed semantic density as a confidence indicator for LLM responses
  - Response-wise, off-the-shelf for free-form generation tasks
- **Future work**
  - Improving sampling strategy for reference responses
  - Developing contextual embedding model
  - Dedicated kernel function
  - Better token probability calibration



- **Source code and contact**
  - https://github.com/cognizant-ai-labs/semantic-density-paper
  - qiuxin.nju@gmail.com, risto@cs.utexas.edu

Cognizant

Thank You