Uncertainty Quantification for Large Language Models

安浩 2024-9-20

Outline

1. Sample-based Uncertainty Quantification
2. Sample-free Uncertainty Quantification

Uncertainty Quantification

- To what extent should we trust LLM‘s answers?
  - Depends on uncertainty.
- How to quantify uncertainty? Some intuitive and simple Methods:
  - Predictive entropy

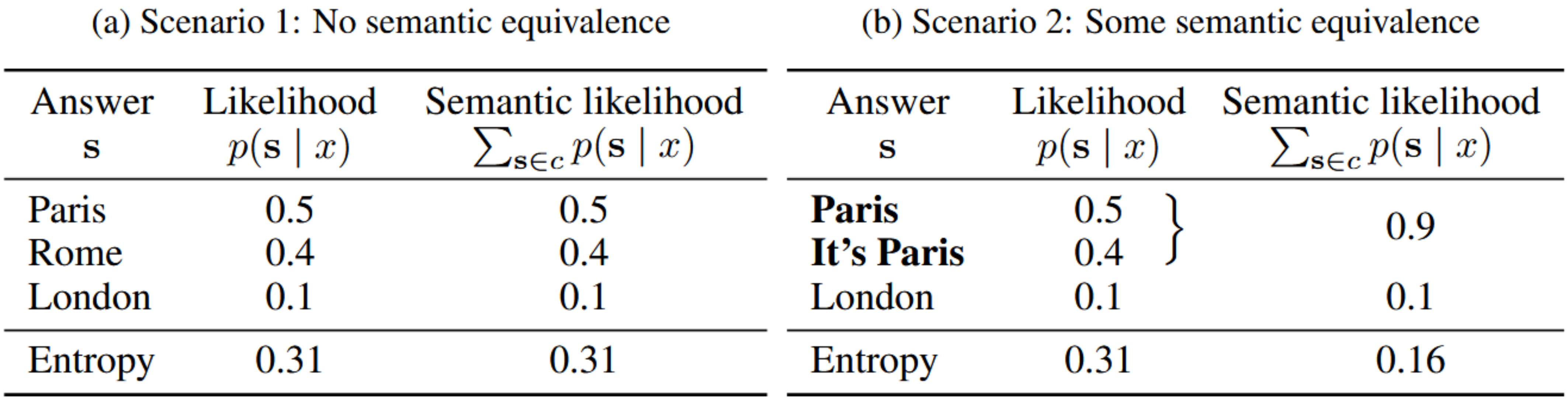
$$PE(x) = H(Y \mid x) = -\int p(y \mid x) \ln p(y \mid x) dy$$

  - Length-normalised predictive entropy.
  - P(true): ‘asking’ the model if its answer is correct.
  - Lexical similarity: $\frac{1}{C} \sum_{i=1}^{|\mathbb{A}|} \sum_{j=1}^{|\mathbb{A}|} \operatorname{sim}\left(s_i, s_j\right)$, sim is Rouge-L

Semantic Entropy

- Main idea

  When considering entropy, different answers that mean the same thing should be considered in the same cluster.



(a) Scenario 1: No semantic equivalence

Answer $\mathbf{s}$	Likelihood $p(\mathbf{s} \mid x)$	Semantic likelihood $\sum_{\mathbf{s} \in c} p(\mathbf{s} \mid x)$
Paris	0.5	0.5
Rome	0.4	0.4
London	0.1	0.1
Entropy	0.31	0.31

(b) Scenario 2: Some semantic equivalence

Answer $\mathbf{s}$	Likelihood $p(\mathbf{s} \mid x)$	Semantic likelihood $\sum_{\mathbf{s} \in c} p(\mathbf{s} \mid x)$
Paris	0.5 }	0.9
It's Paris	0.4 }	
London	0.1	0.1
Entropy	0.31	0.16

Semantic Entropy

- Method
  1. Generation: Sample **M** answers from the LLM given a context **x**.
  2. Clustering: Cluster the answers that mean the same thing using a bi-directional entailment algorithm.
  3. Entropy estimation: Approximate semantic entropy by summing probabilities that share a meaning and compute entropy.

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

$$SE(x) = -\sum_{c} p(c \mid x) \log p(c \mid x) = -\sum_{c} \left(\left(\sum_{\mathbf{s} \in c} p(\mathbf{s} \mid x) \right) \log \left[\sum_{\mathbf{s} \in c} p(\mathbf{s} \mid x) \right] \right)$$

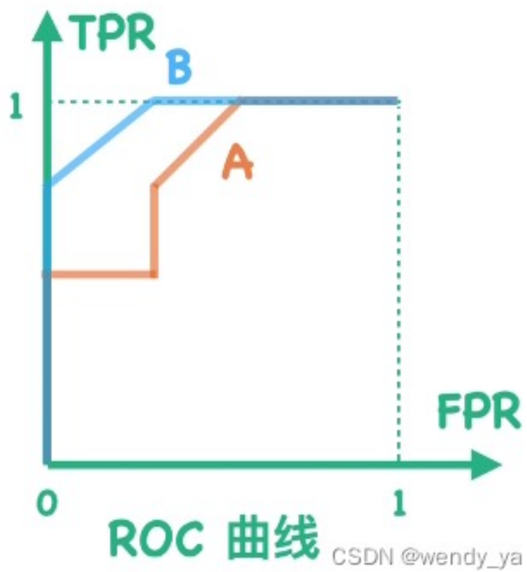
Semantic Entropy

- Natural Language Inference (NLI)
  - NLI determines the logical relationship between a pair of text sequences.
    - **Entailment**: the hypothesis can be inferred from the premise.
    - Contradiction: the negation of the hypothesis can be inferred from the premise.
    - Neutral: all the other cases.
  - Clustering based on bi-directional entailment
    - Using Deberta-large model fine-tuned on NLI and MNLI dataset.
    - Returns equivalent if and only if **both directions were entailment**.

The capital of France is Paris. ⇄ Paris is the capital of France.

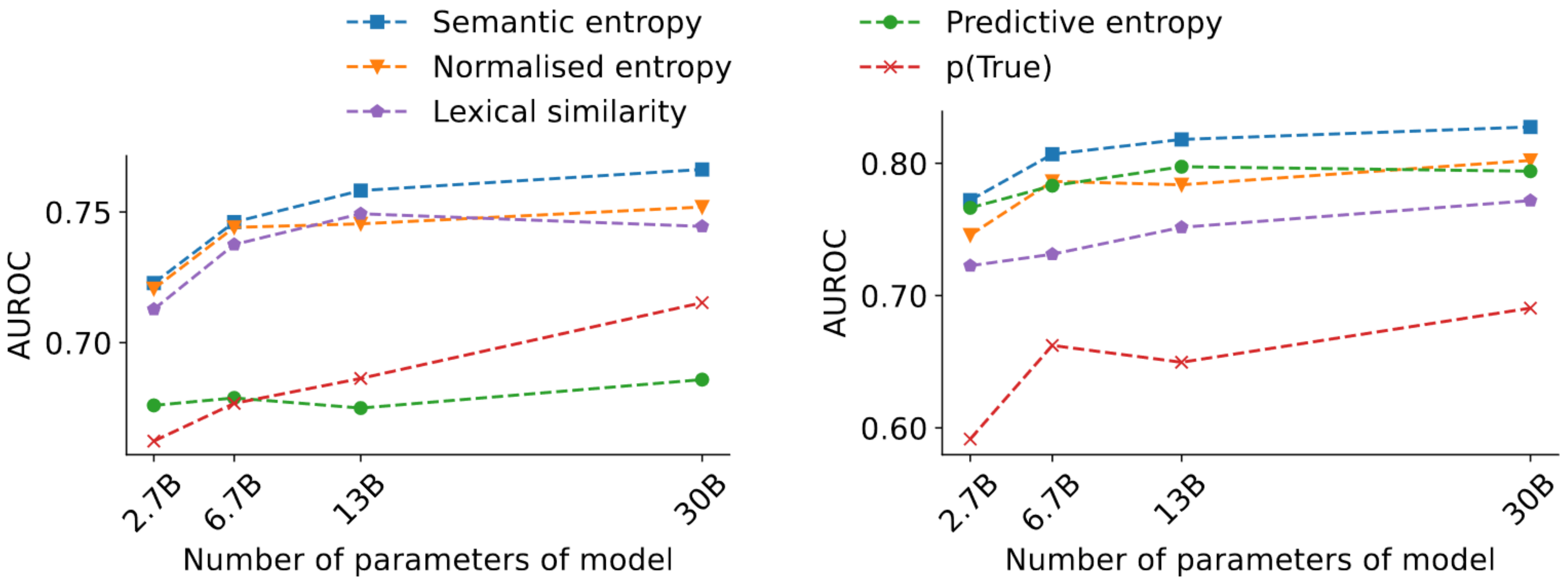
Semantic Entropy

- Evaluation
  - Very uncertain generations should be less likely to be correct
  - Metric: AUROC between -uncertainty score and Y-true.
  - Higher scores are better.
  - Y-true = 1 if the most likely answer is correct; otherwise, it is 0.
    - AUROC=1： Perfect classifier.
    - 0.5 < AUROC < 1: Better than random guessing.
    - AUROC=0.5： Random guessing.



Semantic Entropy

- Main Results

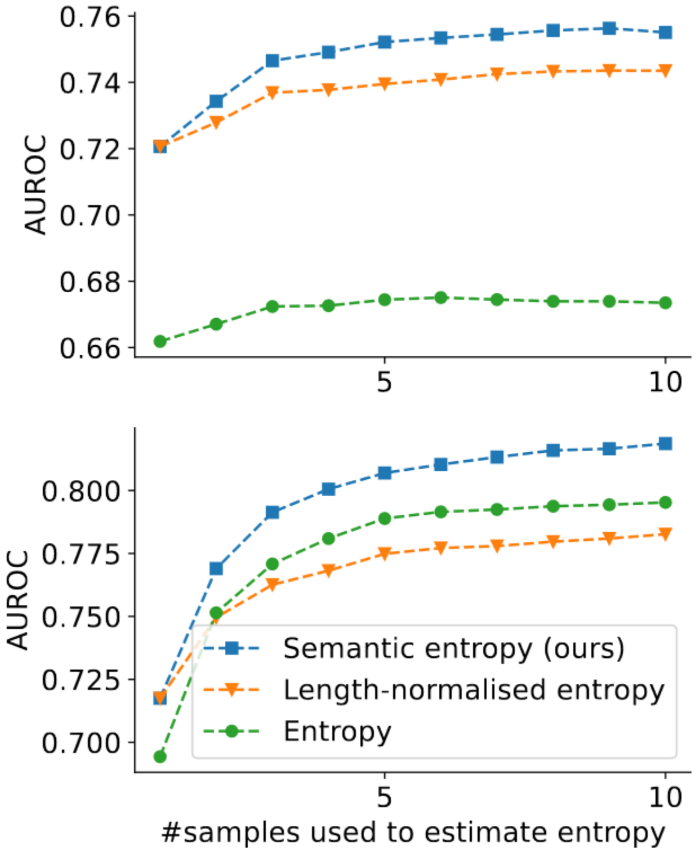


(a) CoQA

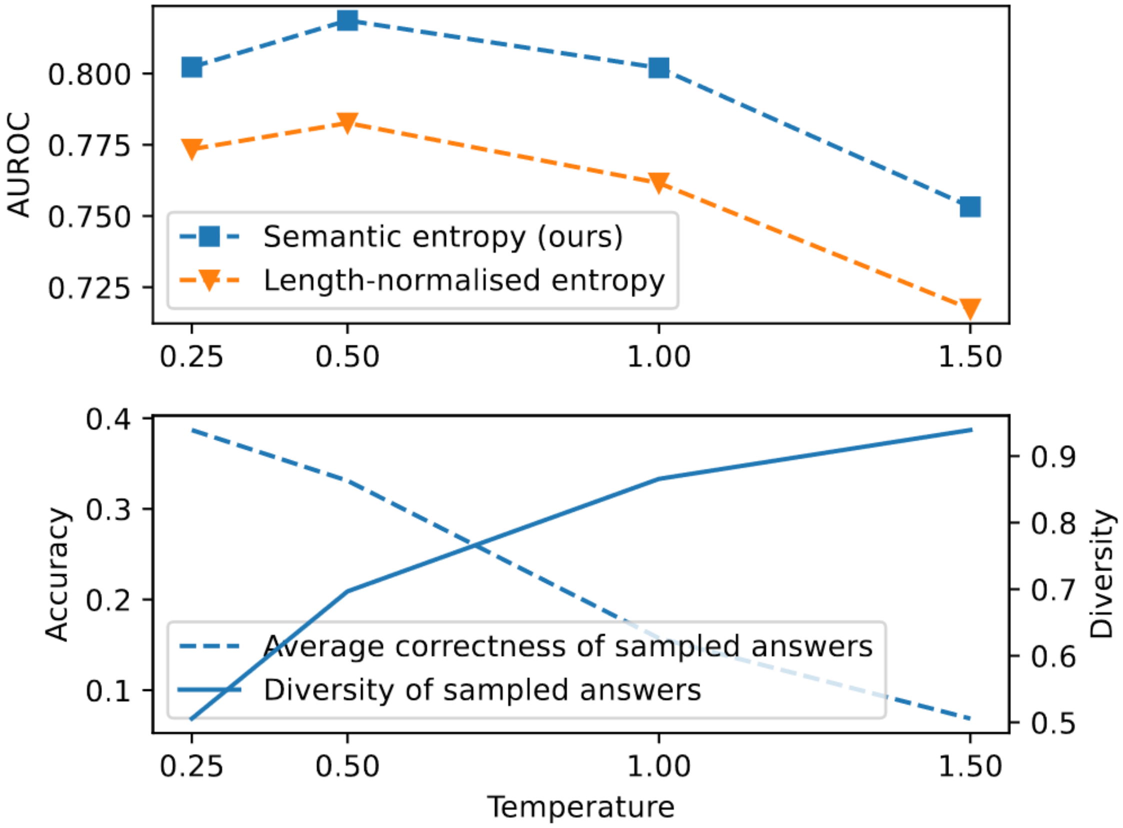
(b) TriviaQA

Semantic Entropy

- Results on sample number and temperature



(a) (top) CoQA, (bottom) TriviaQA



(b)

Mass Mean Probe

- Linear Probe
  - The truth concept of simple text is linear direction in hidden pace

- Consider the following prompt p:

The Spanish word ‘jirafa’ means ‘giraffe’. This statement is: TRUE [...]
The Spanish word ‘aire’ means ‘silver’. This statement is: FALSE
The Spanish word ‘uno’ means ‘floor’. This statement is:

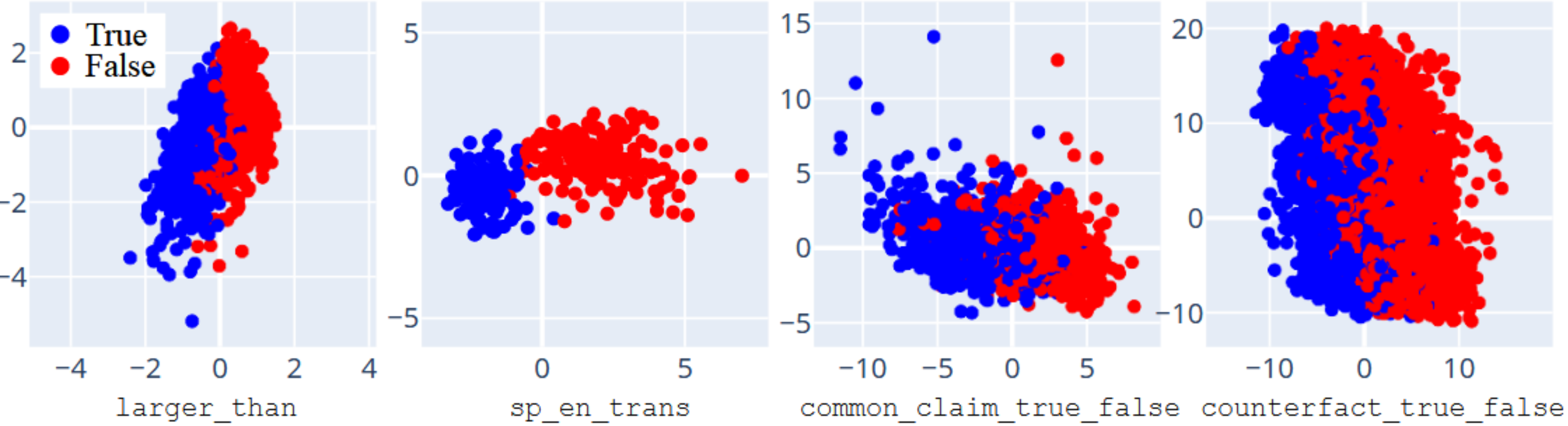
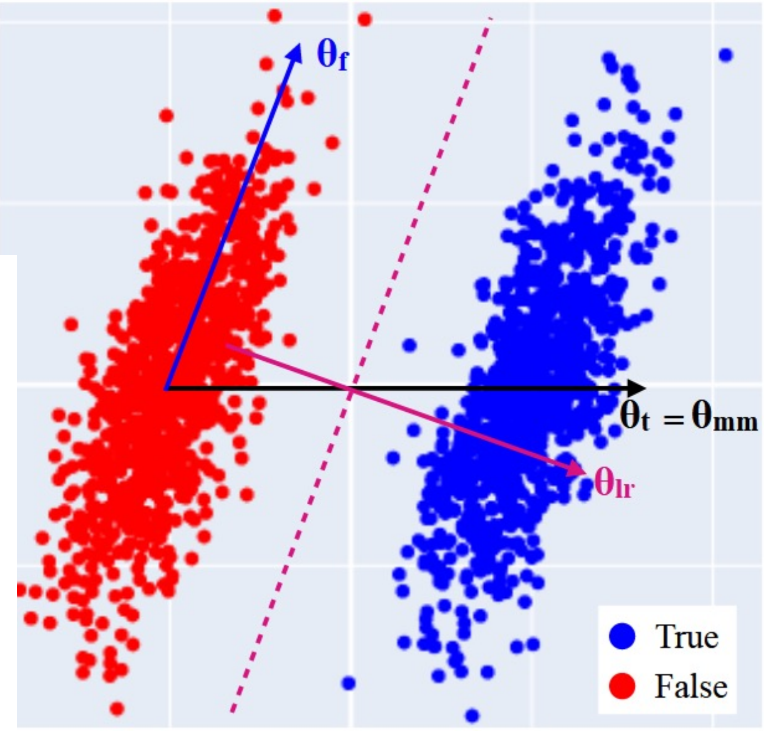


Figure 2: Projections of residual stream representations of datasets onto the top 2 PCs of cities.



igure 4: An illustration of a weakness of logistic regression.

Semantic Entropy Probe

- Semantic Entropy requires **sampling multiple answers**
- Sample-free method: Semantic Entropy Probe (SEP)
    - Training data
        - X: Hidden state
        - Y: Semantic entropy is high or low (binarized semantic entropy)
    - Model: logistic regression
    - Probing locations of hidden state:
        - The last token before generating (TBG)
        - The last token of the model response ——second last token (SLT)

Semantic Entropy Probe

- Predicting Semantic Entropy without Generation

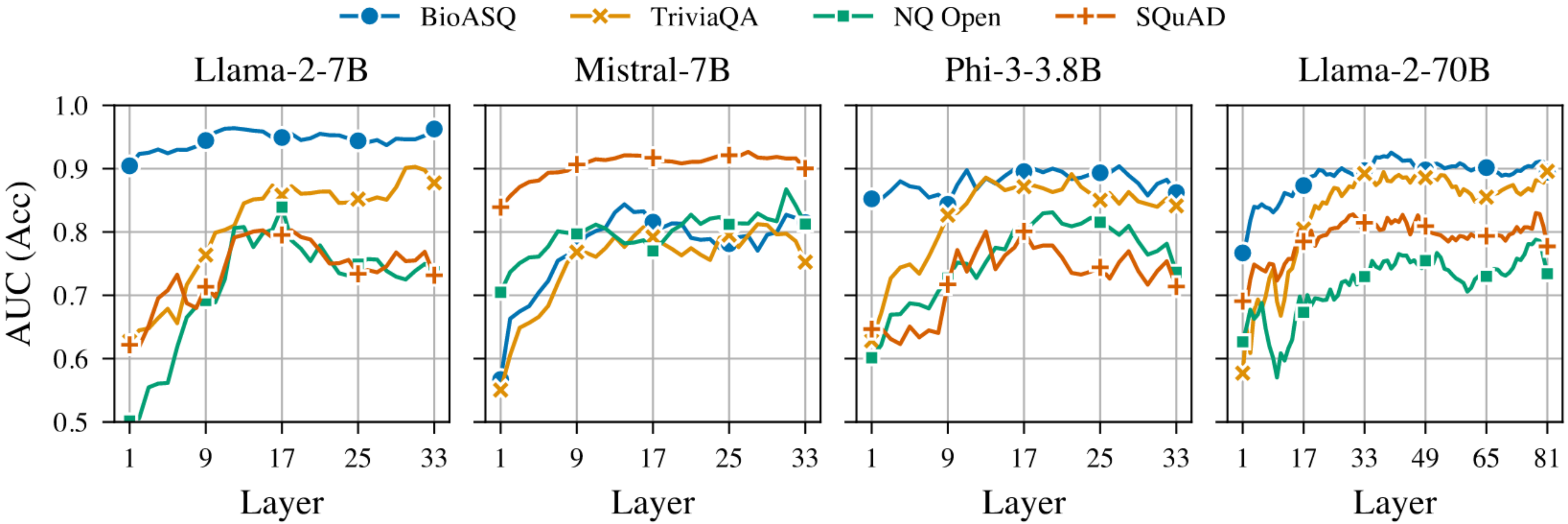


Figure 3: Semantic entropy can be predicted from the hidden states of the last input token, without generating any novel tokens. Short generations with SEPs trained on the token before generating (TBG).

Semantic Entropy Probe

- Predicting Hallucination on Unseen Task

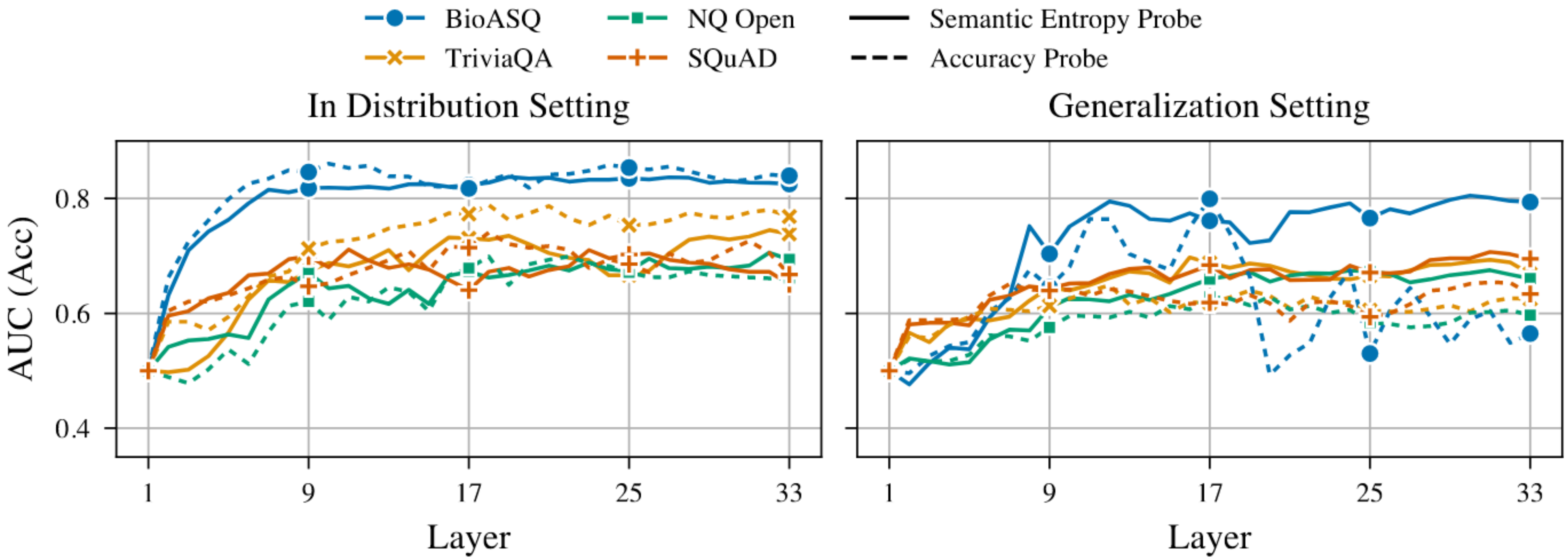


Figure 6: SEPs predict model hallucinations better than accuracy probes when generalizing to unseen tasks. In-distribution, accuracy probes perform better. Short generation setting with Llama-2-7B, SEPs trained on the second-last-token (SLT). For the generalization setting, probes are trained on all tasks except the one that we evaluate on.

Semantic Entropy Probe

- Predicting QA correctness

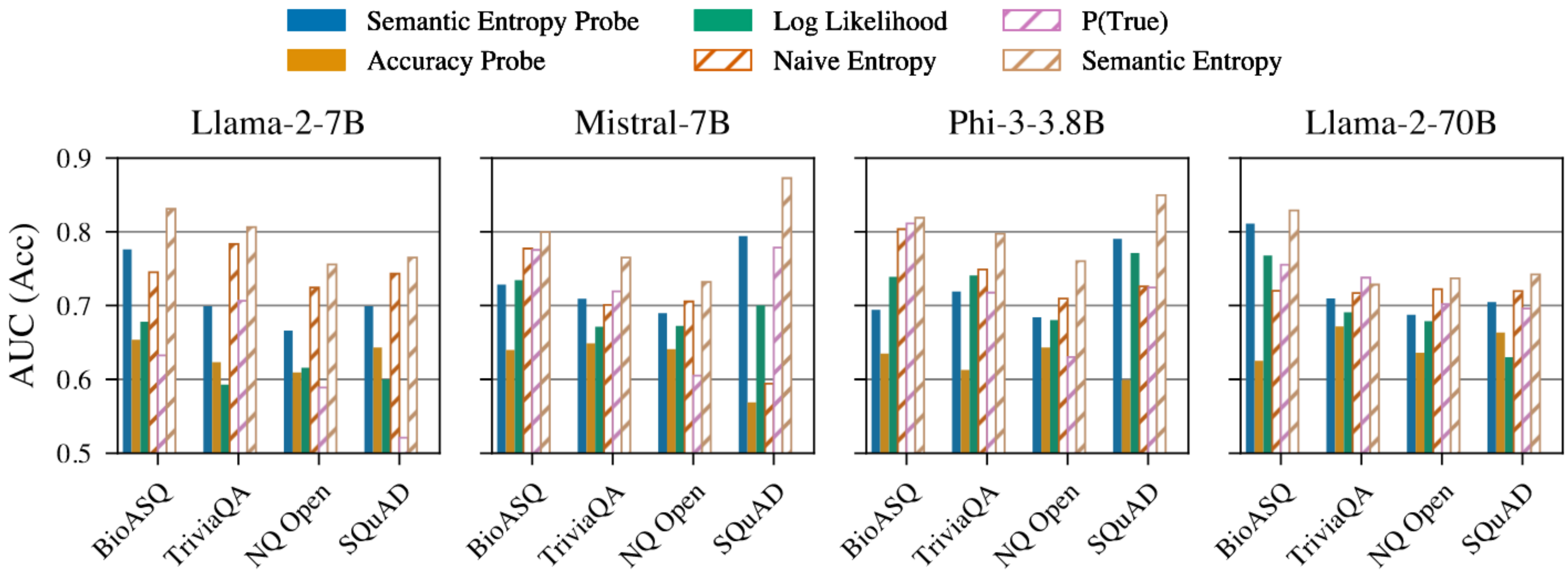
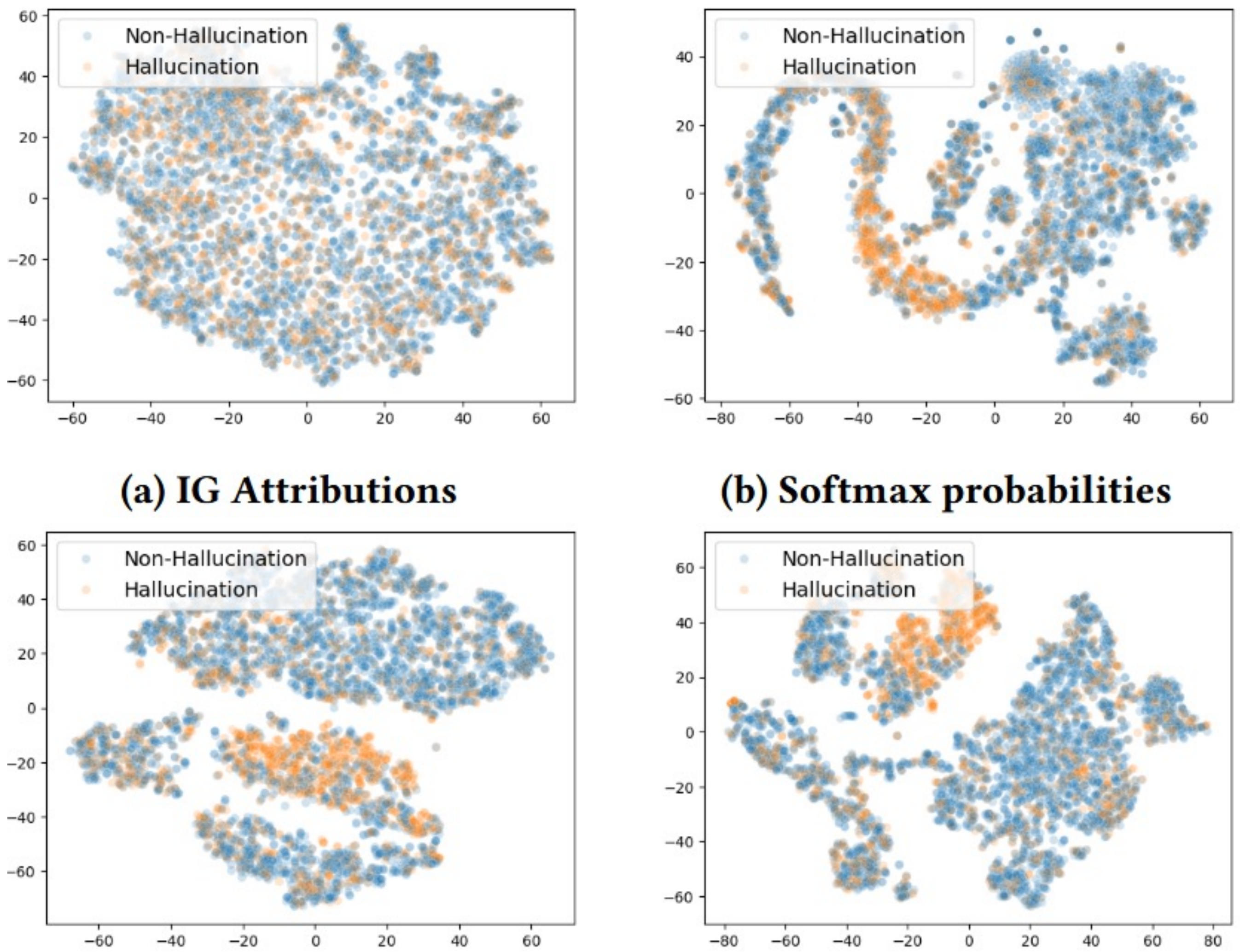


Figure 7: SEPs generalize better to new tasks than accuracy probes across models and tasks. They approach, but do not match, the performance of other, 10x costlier baselines (hatched bars). Short generation setting, SLT, performance for a selection of representative layers.

Accuracy probes

- Accuracy probes
  - Softmax probabilities.
  - Feature attributions.
  - Self-attention.
  - Hidden Activations.



(a) IG Attributions

(b) Softmax probabilities

(c) Self-attention scores

(d) Fully-connected activations

	LAM-13B	OPT-30B	FAL-40B
TriviaQA	0.60	0.57	0.48
Capitals	0.41	0.68	0.43
Founders	0.57	0.44	0.45
Birth Place	0.48	0.45	0.51
Combined	0.43	0.57	0.49

(a) IG attributions

	LAM-13B	OPT-30B	FAL-40B
TriviaQA	0.71	0.63	0.60
Capitals	0.68	0.69	0.63
Founders	0.61	0.67	0.66
Birth Place	0.66	0.62	0.66
Combined	0.69	0.67	0.62

(b) Softmax probabilities

	LAM-13B	OPT-30B	FAL-40B
TriviaQA	0.71	0.65	0.71
Capitals	0.72	0.72	0.72
Founders	0.73	0.68	0.71
Birth Place	0.81	0.61	0.81
Combined	0.78	0.71	0.79

(c) Self-attention scores

	LAM-13B	OPT-30B	FAL-40B
TriviaQA	0.72	0.64	0.72
Capitals	0.73	0.71	0.70
Founders	0.71	0.72	0.73
Birth Place	0.80	0.77	0.76
Combined	0.79	0.73	0.82

(d) Fully-connected activations

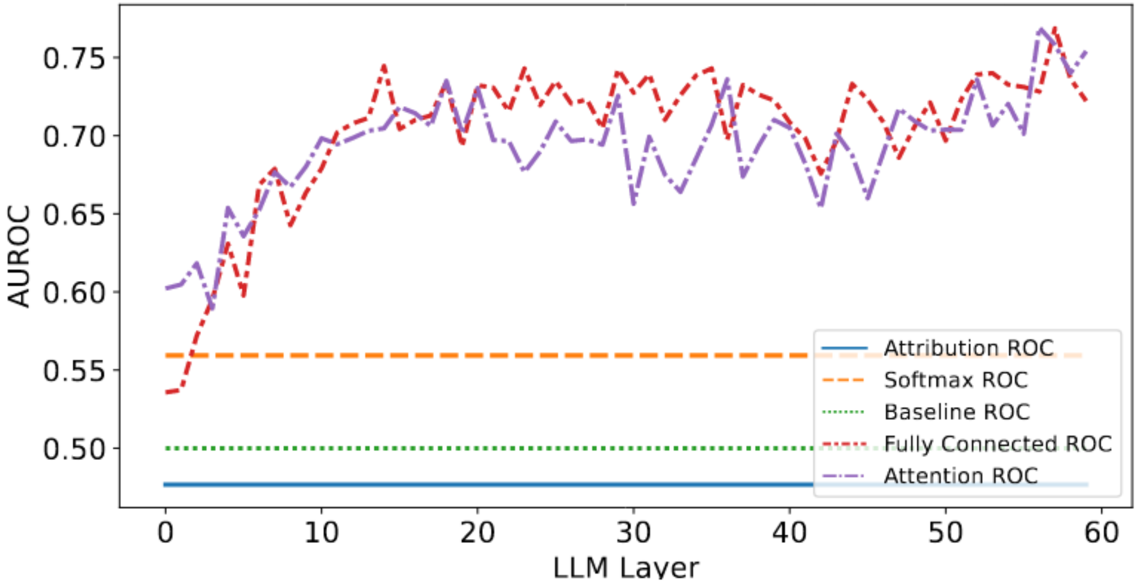
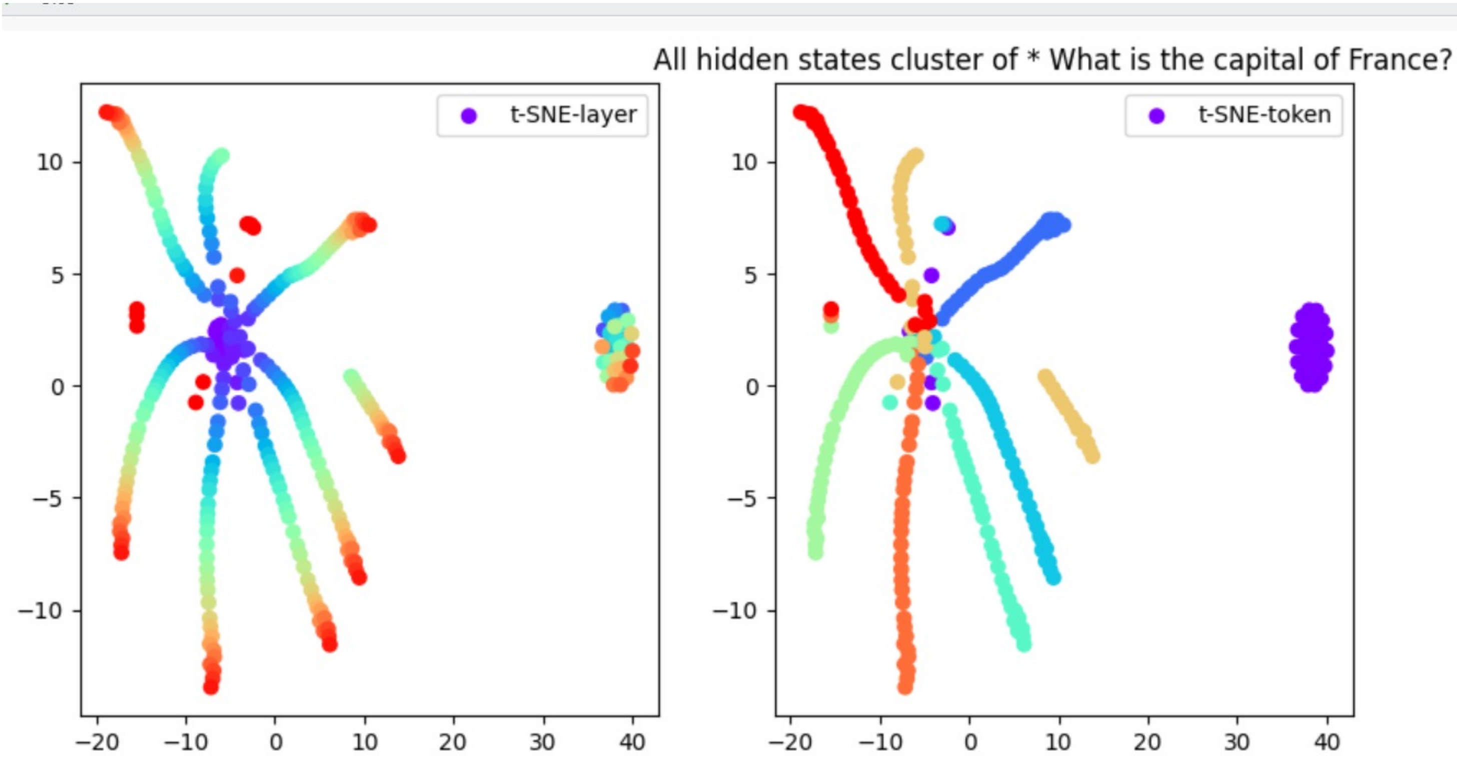


Figure 5: [FAL-40B on TriviaQA dataset] AUROC of the hallucination detectors using self-attention and fully-connected activations at different layers. The performance is better at later layers but has diminishing returns.

Distance as Uncertainty Score

- Distance-based Metrics
  - Pair-wise distance
    - L2 and Cosine
  - Layer-wise distance
    - L2 and Cosine
  - Linear Assignment distance
    - L2 and Cosine

- All of the above are round 0.5 AUROC



Reference

1. Semantic Uncertainty: Linguistic Invariances for Uncertainty Estimation in Natural Language Generation, University of Oxford, ICLR 2023
2. Semantic Entropy Probes: Robust and Cheap Hallucination Detection in LLMs, University of Oxford, ICML 2024 FM-Wild Workshop
3. On Early Detection of Hallucinations in Factual Question Answering, Amazon, KDD 2024
4. The Geometry of Truth: Emergent Linear Structure in Large Language Model Representations of True/False Datasets, COLM 2024, Northeastern University and MIT
5. Know Your Limits: A Survey of Abstention in Large Language Models, UW
6. Siren's Song in the AI Ocean: A Survey on Hallucination in Large Language Models
7. Internal Consistency and Self-Feedback in Large Language Models: A Survey