Detecting LLM Hallucination Through Layer-wise Information Deficiency

Analysis of Ambiguous Prompts and Unanswerable Questions

UNIVERSITY OF OXFORD

Hazel Kim, Tom A. Lamb, Adel Bibi, Philip Torr, Yarin Gal
hazel.kim@cs.ox.ac.uk

30th ANNIVERSARY EMNLP 2025
Suzhou, China | 中国苏州
November 4-9 | 11月4日–9日

Motivation:

- **Acknowledging uncertainty** is safer than giving confident but false answers under limited information.
- Existing approaches focus **mainly on final-layer outputs** or confidence scores.
- Training large LLMs to detect hallucinations is **costly and impractical** for most real-world systems.

Research Contributions:

1. A new, **test-time** metric to detect hallucinations (due to prompt ambiguity or question unanswerability) requiring **no architectural changes or additional training**.
2. LI effectively reflects model confidence under **varying instruction ambiguity** and **task difficulty.**
3. Tracking **information flow across all layers** gives **better insight** into internal model uncertainty than analyzing only the **final layer**.

Method:

Algorithm 1 Computing layer-wise usable information ($\mathcal{L}I$) without fine-tuning

Input: dataset $\mathcal{D} = \{(c_i, q_i)\}_{i=1}^{m}$, pretrained model with layers $\mathcal{L}$
Output: $\mathcal{L}I(C \to Q)$

$\varnothing \leftarrow$ empty context (null string)
for each example $(c_i, q_i) \in \mathcal{D}$ **do**
  for each layer $\ell \in \mathcal{L}$ **do**
    Compute token log-probs $p_\ell(q_t \mid q_{<t}, \varnothing)$
    Compute token log-probs $p_\ell(q_t \mid q_{<t}, c_i)$
    $H_\ell^{(i)}(Q \mid \varnothing) \leftarrow \frac{1}{T_i} \sum_{t=1}^{T_i} -\log_2 p_\ell(q_t \mid q_{<t}, \varnothing)$
    $H_\ell^{(i)}(Q \mid C) \leftarrow \frac{1}{T_i} \sum_{t=1}^{T_i} -\log_2 p_\ell(q_t \mid q_{<t}, c_i)$
    $I_\ell^{(i)} \leftarrow H_\ell^{(i)}(Q \mid \varnothing) - H_\ell^{(i)}(Q \mid C)$
  end for
end for
$\mathcal{L}I(C \to Q) \leftarrow \frac{1}{m} \sum_{i=1}^{m} \sum_{\ell \in \mathcal{L}} I_\ell^{(i)}$

1. Model (V) Usable Information (VI)

$$I_\mathcal{V}(X \to Y) = H_\mathcal{V}(Y|\varnothing) - H_\mathcal{V}(Y|X)$$

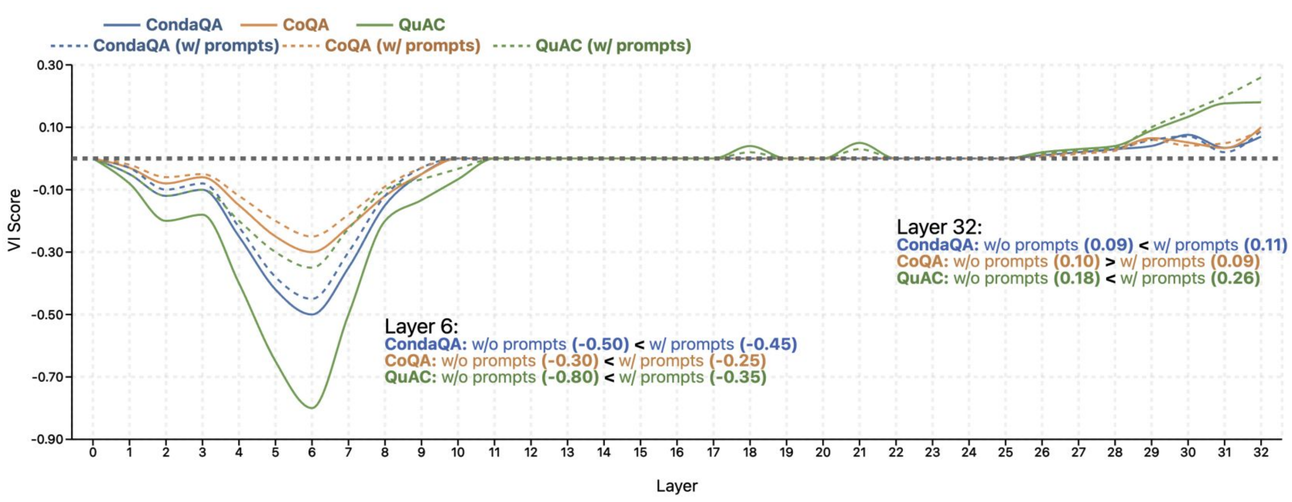
measures **how much information** a model V **can use** to predict outputs Y given inputs X.

2. Layer-wise Usable Information (LI)

$$\mathcal{L}I(c \to q) = \sum_{\ell \in \mathcal{L}} I_\ell(c \to q),$$
$$I_\ell(c \to q) = H_\ell(Q|\varnothing) - H_\ell(Q|C).$$

measures the **change in entropy** due to the presence of context **at layer ℓ**, and LI(c → q) **aggregates** these contributions across all layers.

Observation#1) Non-monotonic information flows (VI) across layers.



Final layer (32) shows inconsistent information quantities across prompt ambiguity over each dataset, *while premature layer (6) shows relatively consistent patterns.*

Observation #2) LI is more comprehensive to show the prompt ambiguity and question unanswerablity than VI on case level.

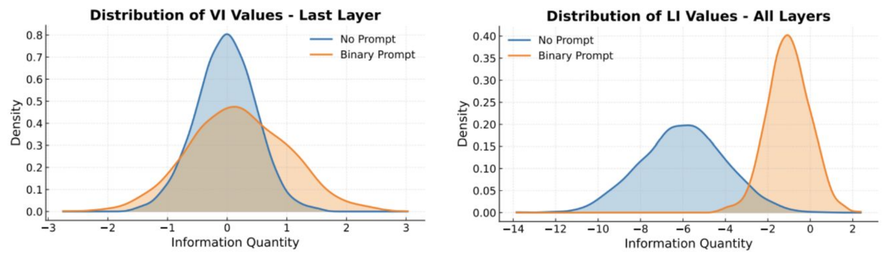
Context

Once there was a beautiful fish named Asta. Asta lived in the ocean. There were lots of other fish in the ocean where Asta lived. They played all day long. One day, a bottle floated by over the heads of Asta and his friends. They looked up and saw the bottle. ... They took the note to Asta's papa. "What does it say?" they asked. Asta's papa read the note. He told Asta and Sharkie, "This note is from a little girl. She wants to be your friend. If you want to be her friend, we can write a note to her. **But you have to find another bottle so we can send it to her." And that is what they did.**

Instruction Prompt	Question	Ground-Truth Label	Prediction	*VI*	*LI*
Binary: "Is this answerable?"	What was the name of the fish?	Yes.	✓	0.3	0.2
	Were they excited?	No.	✓	0.3	-0.4
Open-ended Prompt: "Answer the question or say dont́ know"	What was the name of the fish?	Asta.	✓	0.3	-0.7
	Were they excited?	Don't know.	✗	-0.3	-1.4
No prompts	What was the name of the fish?	Asta.	✓	1.2	-6.8
	Were they excited?	Unknown.	✗	0.2	-7.7

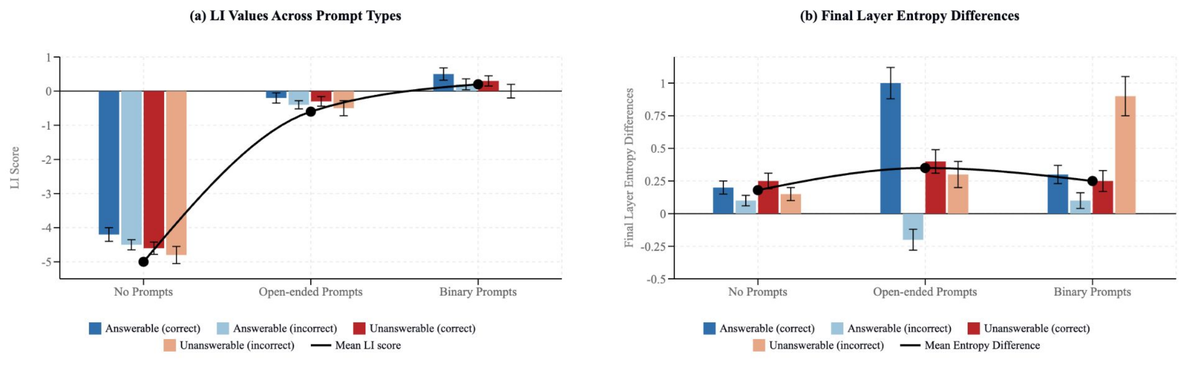
Table 1: The $\mathcal{L}I$ scores provide a more comprehensive overview of prompt ambiguity compared to $\mathcal{V}I$. ***Prediction***: prediction generated by a language model — ✓: correct, ✗: incorrect. $\mathcal{V}I$: $\mathcal{V}$-usable information only applied to the final layer (Xu et al., 2020). $\mathcal{L}I$: Layer-wise usable information accumulated across layers (our proposed method).

Observation #3) LI distributions show stronger signals of the prompt ambiguity than final layer VI.



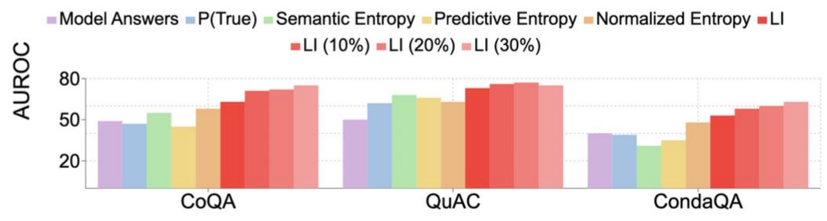
LI distribution shifts clearly toward higher usable information with additional instruction prompts, *while VI does not show clear separations. LI shows that the* ***explicit prompts help stabilize information flow,*** *while no-instruction prompts cause widespread uncertainty.*

Observation #4) Robustness of AURA in low-resource and OoD Settings



LI Scores *rise* as prompts become clearer *(no < open-ended < binary) as well as the prompt is easier* *(**incorrect**-unanswerable < **correct**-unanswerable < **incorrect**-answerable < **correct**-answerable)* **,** while VI on final layer shows inconsistent patterns.

Observation #5) LI distributions can detect unanswerable questions.



Performance of (un)answerability detection across datasets, comparing LI with other baselines. LI(10%), LI(20%), and LI(30%) shows the rejection rate based on low scores.

Observation #6) LI is computationally cheap.

Dataset	Context	Question	$\mathcal{L}I_{\text{context + question}}$	$\mathcal{L}I_{\text{question}}$	$\mathcal{L}I_{\text{total}}$ †
CoQA	271 words	5.5 words	1.00	0.02	1.02×
QuAC	401 tokens	6.5 tokens	1.00	0.01	1.01×
CondaQA	131 tokens	24.4 tokens	1.00	0.16	1.16×

LI requires only two forward passes, one with and one without context. Because the second pass covers only the short question sequence, ***the cost increase (LI total) is negligible.***