Hallucination Detection

서재형

Natural Language Processing
& Artificial Intelligence

Can Hallucination be defined as a single concept?

Survey of Hallucination in Natural Language Generation (2022)

Task	Sub-Task	Type	Source	Output
Abstractive Summarization [197]		Intrinsic	The first vaccine for Ebola was approved by the FDA in 2019 in the US, five years after the initial outbreak in 2014. To produce the vaccine, scientists had to sequence the DNA of Ebola, then identify possible vaccines, and finally show successful clinical trials. Scientists say a vaccine for COVID-19 is unlikely to be ready this year, although clinical trials have already started.	The first Ebola vaccine was approved in **2021**.
		Extrinsic		**China** has already started clinical trials of the COVID-19 vaccine.
Dialogue	Task-oriented [153]	Intrinsic	inform (NAME = pickwick hotel, PRICERANGE = moderate)	the hotel named pickwick hotel is in a **high** price range.
		Extrinsic		the pickwick hotel **in san diego** is a moderate price range
	Open-domain [226]	Intrinsic	Dialog History: Speaker 1: What do you think about Murray? Speaker 2: I think Murray is a great player he just needs to stay healthy in order to compete more. Who do you like best?	Speaker 1: I like **Roger Nadal** the most!
		Extrinsic	Knowledge: In tennis, the term Big Four refers to the quartet of men's singles players comprising Roger Federer, Rafael Nadal, Novak Djokovic, and Andy Murray.	Speaker 1: I like Djokovic. He has played **in the top ten singles players of the world**.
Generative QA [191]		Intrinsic	Question: dow jones industrial average please?	Answer: The Dow Jones Industrial Average(DJIA) is an index of **30 major U.S.stock indexes**. The DJIAwas created in 1896 by Charles Dow, a statistician and editor of The Wall Street Journal. The index was first published in the mid-1880s, and **was last updated on May 26,1896.**
		Extrinsic	Question: definition of sadduction	Answer: Sadducees. noun. **The definition of a sadducee is a person who acts in a deceitful or duplicitous manner. An example of a Sadducee is a politician who acts deceitfully in order to gain political power.**1 a member of a Jewish sect that was active during the Second Temple.

Can Hallucination be defined as a single concept?

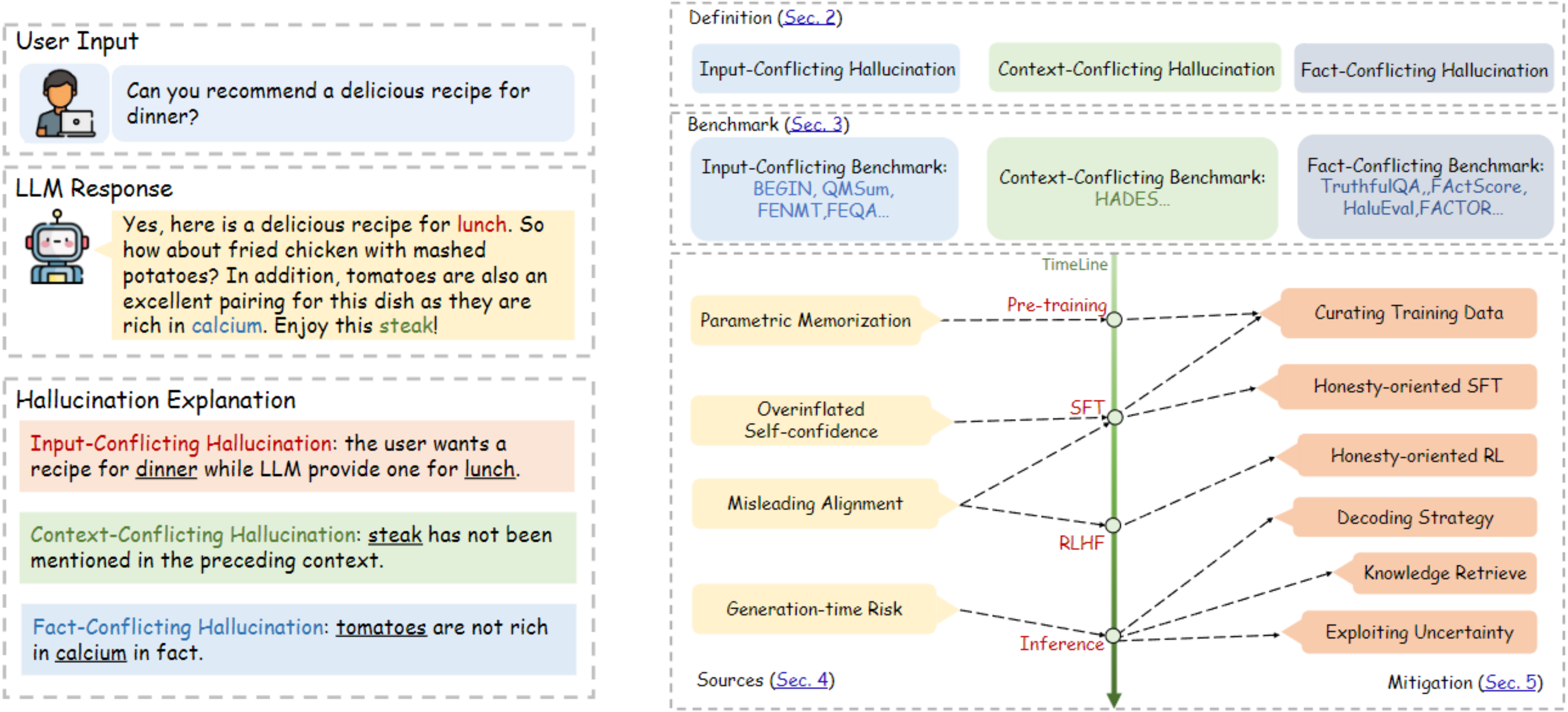
A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions (2023)

Hallucination Type	Sub-Type	User Input	Model Output
Factuality Hallucination	Factual Inconsistency	Tell me about the first person to land on the Moon.	Yuri Gagarin was the first person to land on the Moon.
	Factual Fabrication	Tell me about the historical origins of unicorns.	Unicorns were documented to have roamed the plains of Atlantis around 10,000 BC, where they were considered sacred creatures and were often associated with royalty.

Faithfulness Hallucination	Instruction Inconsistency	Translate the following English question into Spanish: "What is the capital of France?"	The capital of France is Paris.	The user explicitly provided an instruction to translate the English question into Spanish. **Instead of performing the translation as instructed, the LLM chose to answer the question.**
	Context Inconsistency	Summarize the following paragraph about the Nile: The Nile is a major north-flowing river in northeastern Africa. It's traditionally considered the longest river in the world. Originating in the Great Lakes region of central Africa, it flows through several countries before emptying into the Mediterranean Sea.	The Nile originates from the mountain ranges of central Africa and flows into the Mediterranean.	The LLM's summary incorrectly states that the Nile originates from the "mountain ranges of central Africa," while **the paragraph clearly mentions that the Nile originates in the Great Lakes region of central Africa.**
	Logical Inconsistency	Solve the equation 2x + 3 = 11 step by step.	Step 1: Subtract 3 from both sides to get: 2x = 8. Step 2: Divide both sides by 2 to get: x = 3. So the result of the equation is x=3.	The first step correctly isolates the term with x to get 2x = 8. However, **the next step inaccurately divides 8 by 2 to yield a result of x = 3**, which is inconsistent with the earlier reasoning.

Can Hallucination be defined as a single concept?

Siren's Song in the AI Ocean: A Survey on Hallucination in Large Language Models (2023)



Q: Can Hallucination be defined as a single concept?

Q: Can Hallucination be defined as a single concept?

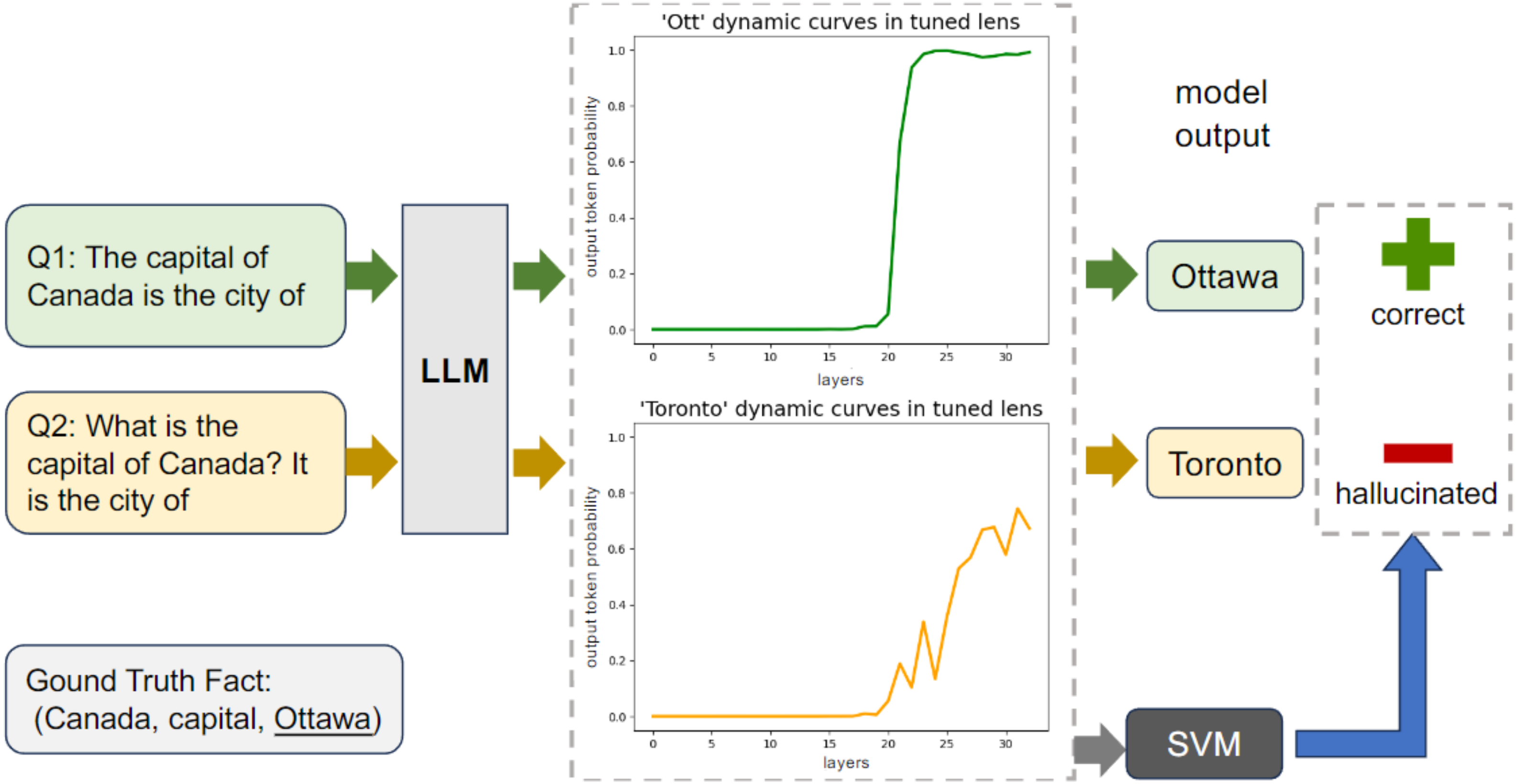
A: It depends on and is becoming more specialized

On Large Language Models' Hallucination with Regard to Known Facts

Che Jiang[1*], Biqing Qi[1], Xiangyu Hong[1], Dayuan Fu[1]
Yang Cheng[1], Fandong Meng[2], Mo Yu[2†], Bowen Zhou[1†], Jie Zhou[2]

NAACL 2024

Problem Statement



❑ **The mechanism behind the model's hallucination of previously memorized knowledge remains puzzling!!**

Known Fact Hallucination

Correct Answer → Memorized relevant information

- ❑ Challenging to ascertain what the model does not know (out of scope)

Failure in recalling parameterized knowledge

- ❑ Queried with **different prompt** for the **same knowledge triplet**
- ❑ Uncertain responses, irrelevant information, incorrect entities

→Investigate the **<u>dynamic inference characteristics of parameterized factual knowledge recall</u>** when LLM exhibits known fact hallucinations

Preliminary

*What differences are in the **dynamic change of hidden states** comparing **successful knowledge recalls and the failed ones**?*

1) Recall process of the object in triple knowledge

- ❑ (s, r, o)

2) COUNTERFACT (Meng et al., 2022a)

- ❑ 30K statement sentences or question-answer pairs
- ❑ s, r → input prompt

relation id	queries	prompts
P17	country of cities	"{*subject*}, which is located in the country of" "{*subject*} is located in the country of" "{*subject*} is situated in the country of"
P364	original language	"The original language of {*subject*} is" "What's the original language of {*subject*}? It is" "{*subject*} was originally filmed in the language of"

Preliminary

3) Model (Llama2-7B-chat)

- Model depth (L) = 32 layers, hidden state (d) = 4096, vocabulary size (V) = 32000
- Input T tokens $t_1, \ldots, t_T$, Embedding matrix $E \in R^{\{V*d\}}$
- Subsequently, they traverse through L transformer blocks, continuously evolving within the model space, generating a residual stream of shape $T \times L \times d$.
- Between layer $l - 1$ and l, the i^{th} token's hidden state x_i^{l-1} is updated by

→ $x_i^l = x_i^{l-1} + a_i^l + m_i^l$ (the outputs of attention and MLP layers, respectively)

- Tokens pass through an unembedding matrix $(d * V)$ → mapping vocabulary space before decoding
- First 10 tokens contain the answer w/o negation or multiple-choice format

Preliminary

4) Observation methods

❑ **Logit Lens**: Mapping from the model space to the vocabulary space at each position within the residual stream

M into two "halves," $\boldsymbol{M_{\leq \ell}}$ ***and*** $\boldsymbol{M_{>\ell}}$. The **function** $\boldsymbol{M_{\leq \ell}}$ consists of the layers of M up to and

including layer ℓ, and it **maps the input space to hidden states**.

Conversely, the **function** $\boldsymbol{M_{>\ell}}$ consists of the layers of M after ℓ, which **map hidden states to logits.**

(1) Layer ℓ updates the representation $$\boldsymbol{h}_{\ell+1} = \boldsymbol{h}_\ell + F_\ell(\boldsymbol{h}_\ell),$$

(2) $$\mathcal{M}_{>\ell}(\boldsymbol{h}_\ell) = \mathrm{LayerNorm}\Big[\boldsymbol{h}_\ell + \sum_{\ell'=\ell}^{L} \underbrace{F_{\ell'}(\boldsymbol{h}_{\ell'})}_{\text{residual update}}\Big] W_U.$$

(3) Residuals to zero $$\mathrm{LogitLens}(\boldsymbol{h}_\ell) = \mathrm{LayerNorm}[\boldsymbol{h}_\ell] W_U$$

Preliminary

4) Observation methods

❑ **Tuned Lens**: An advancement over Logit lens and involves training transformations at various layers within the model space

(1) Zero residuals to learnable b_l $$\mathrm{LogitLens}_{\ell}^{\mathrm{debiased}}(\boldsymbol{h}_{\ell}) = \mathrm{LogitLens}(\boldsymbol{h}_{\ell} + \mathbf{b}_{\ell})$$

(2) Affine Transformation $$\mathrm{TunedLens}_{\ell}(\boldsymbol{h}_{\ell}) = \mathrm{LogitLens}(A_{\ell}\boldsymbol{h}_{\ell} + \mathbf{b}_{\ell})$$

(3) Training (Distillation Loss) $$\operatorname{argmin} \mathop{\mathbb{E}}_{\boldsymbol{x}} \left[D_{KL}(f_{>\ell}(\boldsymbol{h}_{\ell}) \,||\, \mathrm{TunedLens}_{k}(\boldsymbol{h}_{\ell})) \right]$$

Experimental Setup

Observe the transformation of the hidden state x_T

→ corresponding to the **last token of the input** as the # of layers

(lens observation at positions $t < T$ concerning output tokens is minimal)

given knowledge triplet (s, r, o)

one **considers correct (p_r, a_r)** and **the other incorrect (p_w, a_w)**, $p_r = p_w$

(1) Successful Recall = $p_r \rightarrow a_r$ *ex) Canada's capital is → Ottawa*

(2) Failed Recall = $p_w \rightarrow a_r$ ex) *The capital of Canada is → Oranto*

(3) Hallucination Recall = $p_w \rightarrow a_w$ *ex) The capital of Canada is → Toronto*

Accuracy Statistics

Long tail knowledge

- Unpopular knowledge in Wikipedia pages based on browsing counts
- Can be memorized… but

Q. Does a subject's popularity significantly influence known fact hallucination?

Popularity	Incorrect	Uncertain	Irrelevant
$< 10^4$	28	12	11
$10^4 \sim 10^5$	26	8	16
$10^5 \sim 10^6$	27	9	16
$> 10^6$	28	9	14

Table 1: Statistic of hallucination categories across different popularity subjects.

Accuracy Statistics

Long tail knowledge

- ❑ Unpopular knowledge in Wikipedia pages based on browsing counts
- ❑ Can be memorized… but

Q. Does a subject's popularity significantly influence known fact hallucination?

A. No significant correlation between these error types and the popularity of the knowledge.

+) Less frequently accessed knowledge is weakly correlated with more knowledge extraction errors

➔ Invisible something???

Lens Observation

Q1. Did the model retrieve the correct knowledge when it hallucinated?

P_r = "The expertise of **Isaac Barrow** is in the field of,"

P_w = "What is **Isaac Barrow's** professional field? It is"

Erroneous output:

"not clear from the provided biographical information"

→ Failed to recall the memorized knowledge (low in graph)

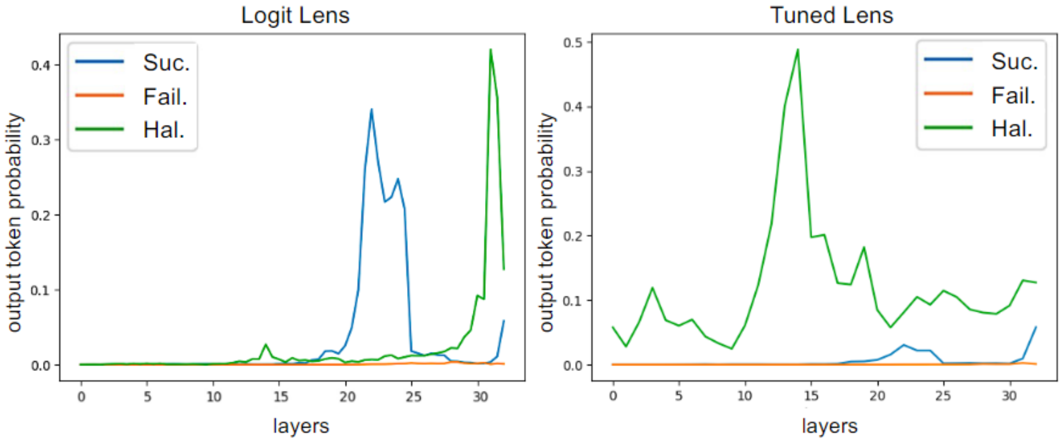


Figure 2: An example of the variation curves in the residual stream for three types of tokens under Logit Lens and Tuned Lens. The Fail. token is not extracted at all.

Lens Observation

Q1. Did the model retrieve the correct knowledge when it hallucinated?

[Logit Lens]

Suc. tokens establish output determination earlier

Hal. Tokens' decoding occurs almost at the final layer

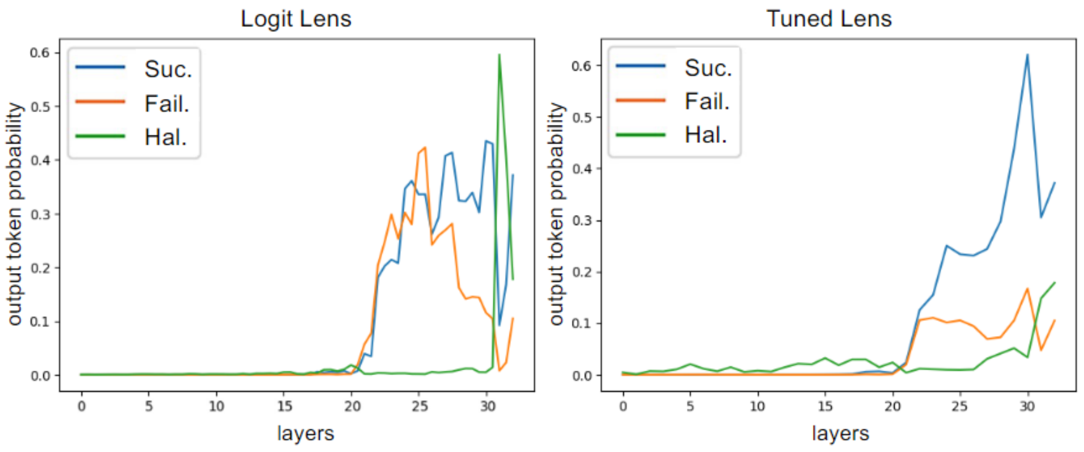


Figure 3: An example of the variation curves in the residual stream for three types of tokens under Logit Lens and Tuned Lens. The Fail. token is temporally recalled and is suppressed afterwards.

[Tuned Lens]

20th layer, model's confirmation of output information

→ Immediate switch to decoding model representation (correct)

= successful recall of knowledge indeed undergoes an 'information extraction point' → shifted to decoding mode

= failure recall of knowledge, the vast majority of knowledge remains unextracted

Lens Observation

Q1. Did the model retrieve the correct knowledge when it hallucinated?

Decoding Failure?

- Average occurrence frequency for the three types

Fail. Tokens: 31.28% (top 1), 56.71% (top 5) < Suc. & Hal.

➔ illusion occurs because knowledge is not successfully extracted in the intermediate steps

	Suc.	Fail.	Hal.
top1	77.57%	31.28%	68.04%
top5	93.21%	56.71%	92.70%

Table 2: Average occurrence frequency of three kinds of tokens in top1 and top5.

Lens Observation

Decoding Failure?

Fail. tokens have comparable probabilities to Suc. tokens at knowledge extraction positions but get suppressed in subsequent layers, resulting in decoding failure

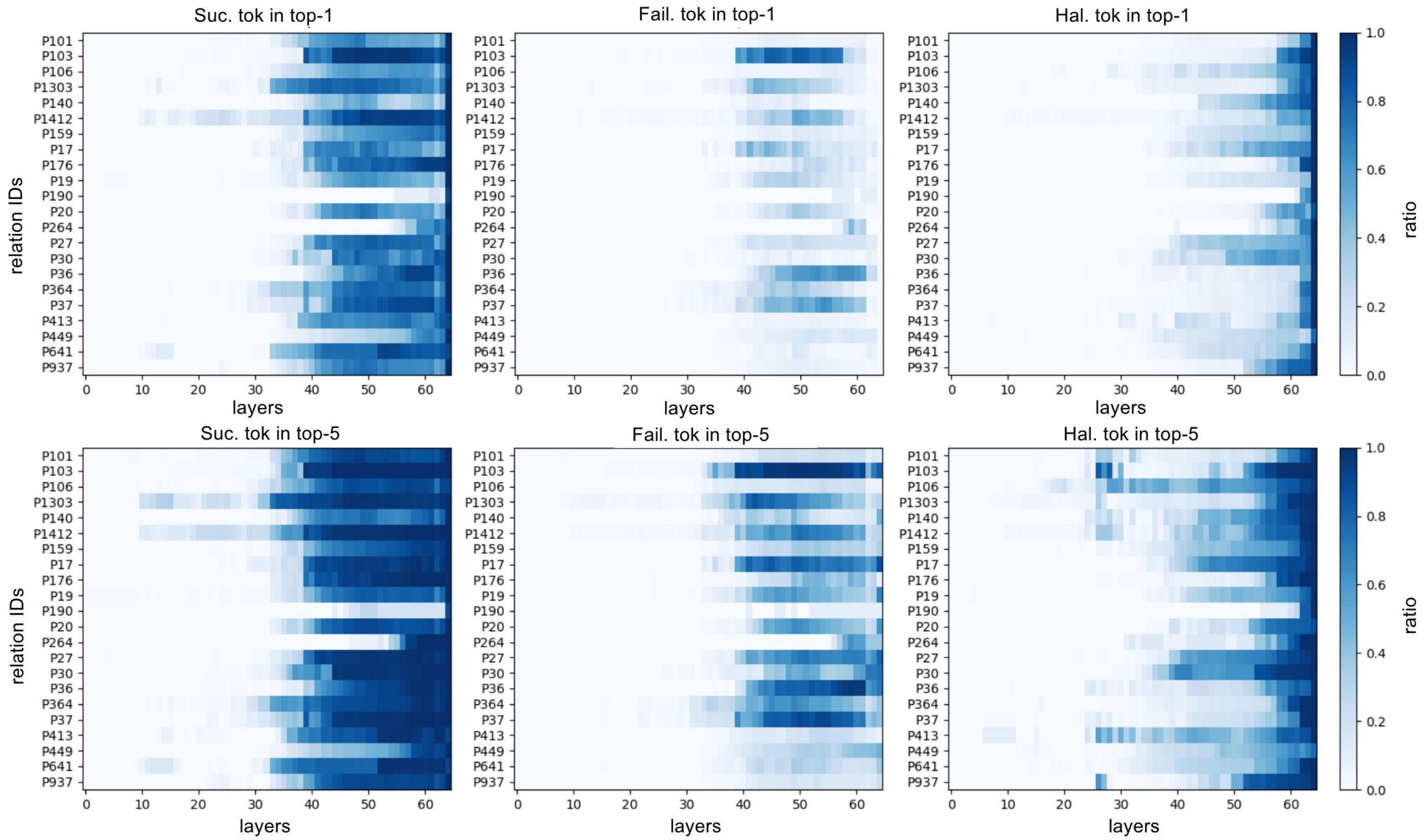


Figure 4: The ratio of the top-1 and top-5 appearances of three types of tokens in logits rankings varies across different relations as the number of layers changes.

Module contributions

Q2. Which module contributes more to hallucinations? What could be the potential process for this?

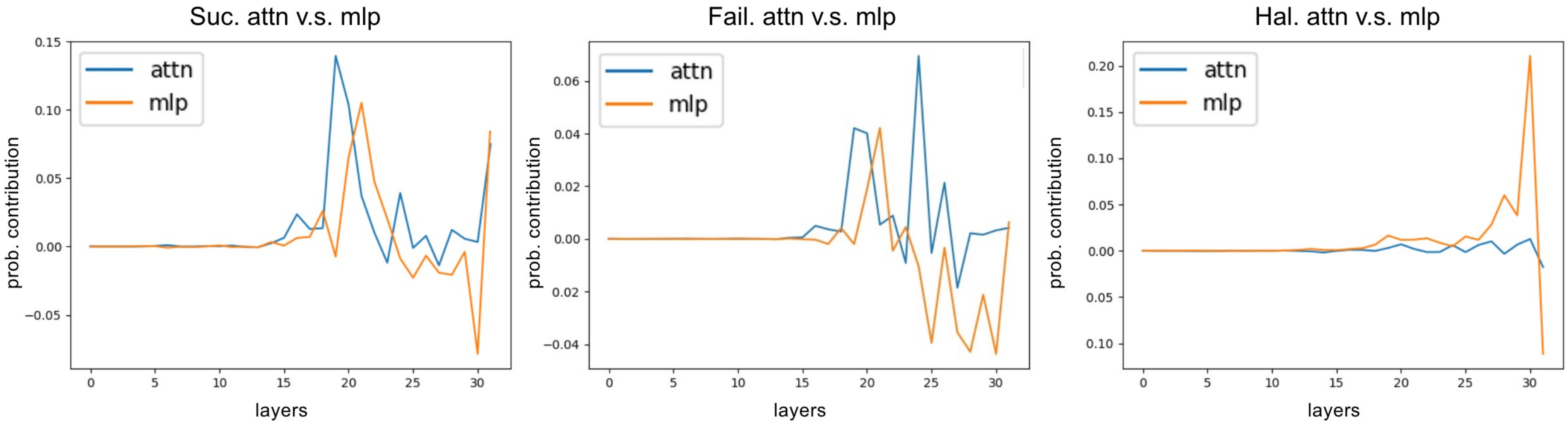


Figure 6: The average contributions of the attention module and the MLP module to the residual stream variations of three types of tokens.

- ❑ **MHSA and MLP demonstrate significant contributions to knowledge extraction, around the 20th layer**
- ❑ **MLP exerts a stronger inhibitory effect towards the erroneous output decoding**

Module contributions

Q2. Which module contributes more to hallucinations? What could be the potential process for this?

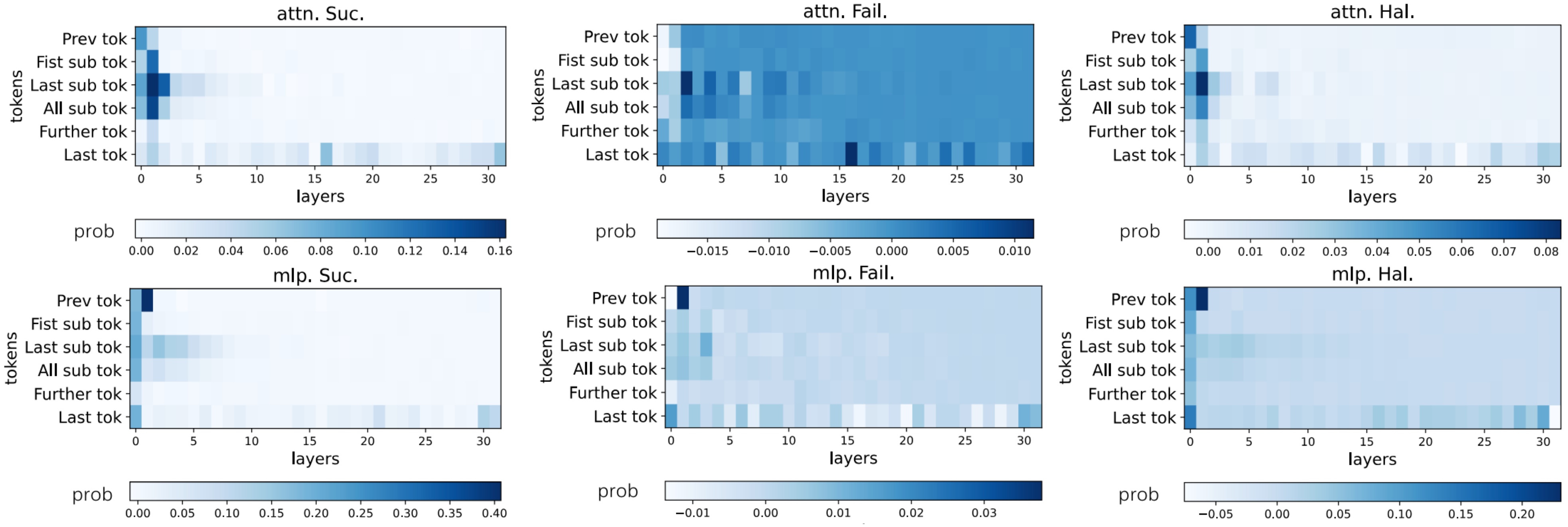


Figure 7: The ablation results of MHSA and MLP module of three types of tokens. The darker colors in the heatmap indicate a higher positive effect on the final output.

- ❑ The processing of output information mostly occurs at the position of the last token
- ❑ In the initial half of the model, the semantic parsing (knowledge extraction) of the query plays a crucial role

Logit evolution pattern

Q3. Are there any patterns in the inference dynamics of hallucination versus correct predictions?

- ❑ **Blend failed and successful samples**
- ❑ **Similar to previous experimental results**

→ early stages focus on query parsing and later stages on answer extraction and decoding

Hallucination outputs do not exhibit notable leaps at relevant positions; they often contain representations of the output token before semantic parsing completes

P36: country's capital

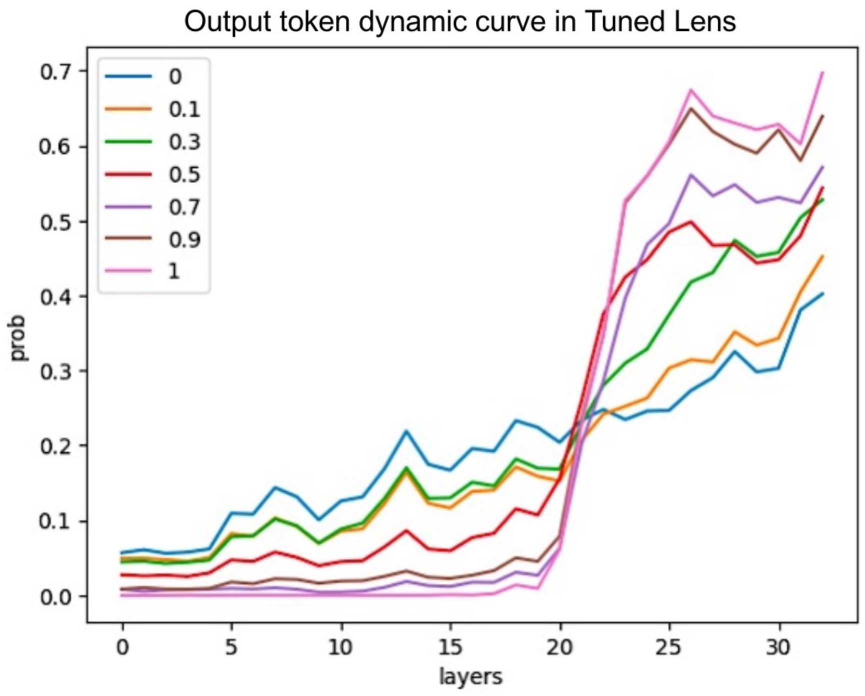


Figure 8: The average dynamic curve of output token under Tuned Lens mapping across various correct rate ratios for relation P36.

Logit evolution pattern

Q4. Can we benefit from the observed patterns for automatic hallucination detection?

linear SVM model using the probability variation curves after mapping with the two type of Lens

it only needs to backtrack the mapping pattern of the first token output (after the last input token)

Ex) [0.15, 0.05, … , 0.48] … sample1 → Correct
[0.10, 0.15, … , 0.85] … sample2 → Hallucination

Model	Logit	Tuned	Both
Llama-7B-chat	0.839	0.854	**0.879**
Llama-13B-chat	0.849	0.840	**0.878**
OPT-6.7B	0.856	0.858	**0.865**
Pythia-6.9B	**0.824**	0.764	0.822

Table 3: Hallucination classification accuracy using output token dynamics across different models.

INSIDE: LLMS' INTERNAL STATES RETAIN THE POWER OF HALLUCINATION DETECTION

Chao Chen[1], Kai Liu[2], Ze Chen[1], Yi Gu[1], Yue Wu[1], Mingyuan Tao[1]
Zhihang Fu[1*], Jieping Ye[1]
[1]Alibaba Cloud [2]Zhejiang University

ICLR 2024

Introduction

Hallucination = Unreliable generations

- Accurately **detecting and rejecting responses** when hallucinations occur in LLMs, has attracted more and more attention from the academic community

(1) Token-level uncertainty estimation (e.g., predictive confidence or entropy)

→ **How to drive sentence-level..?**

(2) Sentence-level uncertainty estimation (e.g., the output languages directly)

(3) Prompting LLMs to generate multiple responses (e.g., self-consistency)

However, such a post-hoc semantic measurement on decoded language sentences is **inferior to precisely modeling the logical consistency/divergence**

INSIDE (INternal States for hallucInation Detection)

Internal state of LLM's Hallucination

- LLMs preserve the **highly-concentrated semantic information** of the entire sentence **within their internal states** (Azaria & Mitchell, 2023), allowing for the direct detection of hallucinated responses in the sentence embedding space.

- First, **<u>skipping secondary semantic extraction via extra models,</u>** we **directly measure the self-consistency/divergence** of the output sentences using internal states of LLMs.

→ **EigenScore metric** regarding the eigenvalues of sentence embeddings' covariance matrix

- To handle the self-consistent (**<u>overconfident</u>**) hallucinations, we propose to **<u>rectify abnormal activations of the internal states</u>**

→ **Feature clipping** approach to truncate extreme features

Eigen Score

Logits & language space

- Neglect the dense semantic information that is retained within the internal states of LLMs
- To measure the semantic divergence in the sentence embedding space

Output token: y_t

Hidden states: h_t^l

Dimension: (d = 4096 for LLaMA-7B and d = 5120 for LLaMA-13B)

Sentence embedding: average of the token embedding $z = \frac{1}{T}\sum_{t=1}^{T} h_t$ **or last token embedding $z = h_T$ (Middle layer)**

K generated sequences: the covariance matrix of K sentence embeddings

Amos Azaria and Tom Mitchell. The internal state of an llm knows when its lying. arXiv preprint arXiv:2304.13734, 2023.

Eigen Score

$$\mathbf{\Sigma} = \mathbf{Z}^\top \cdot \mathbf{J}_d \cdot \mathbf{Z}$$

$\mathbf{\Sigma} \in \mathbb{R}^{K \times K}$ represents covariance matrix → captures the relation btw sentences in the embedding space

$\mathbf{Z} = [\boldsymbol{z}_1, \boldsymbol{z}_2, \cdots, \boldsymbol{z}_K] \in \mathbb{R}^{d \times K}$ represents the embedding matrix of K different sentences

$\mathbf{J}_d = \boldsymbol{I}_d - \frac{1}{d}\mathbf{1}_K\mathbf{1}_K^\top$ represents centering matrix

$$E(\mathcal{Y}|\boldsymbol{x}, \boldsymbol{\theta}) = \frac{1}{K}\log\det(\mathbf{\Sigma} + \alpha \cdot \mathbf{I}_K)$$ logarithm determinant (log det) of the covariance matrix

det (x) represents the **determinant of matrix X**, and a **small regularization term $\boldsymbol{\alpha} \cdot \boldsymbol{I_K}$** is added to the covariance matrix

$$E(\mathcal{Y}|\boldsymbol{x}, \boldsymbol{\theta}) = \frac{1}{K}\log(\prod_i \lambda_i) = \frac{1}{K}\sum_i^K \log(\lambda_i)$$

$\lambda = \{\lambda_1, \lambda_2, \cdots, \lambda_K\}$ denotes the eigenvalues of the regularized covariance matrix

Eigen Score

Remark 1. LogDet of covariance matrix represents the differential entropy in the sentence embedding space

$$H_e(X) = -\sum_X -p(x) \log p(x)$$ **Discrete Shannon Entropy**

$$H_{de}(X) = -\int_x f(x) \log f(x) dx$$ **Differential Entropy in continuous space with density function f(x)**

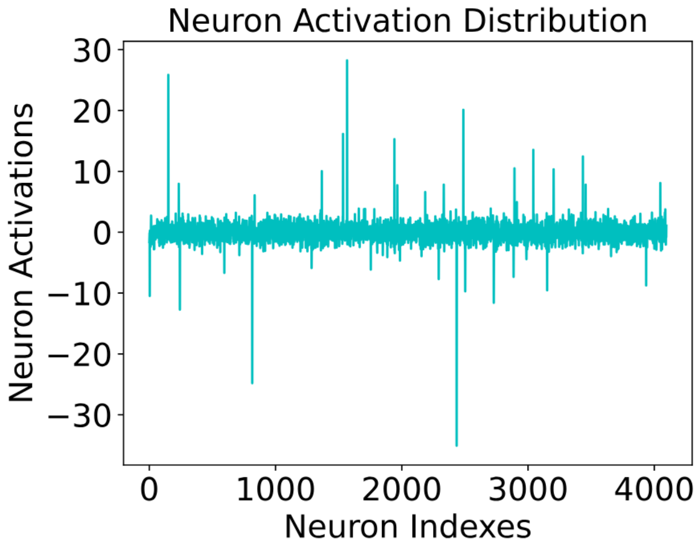
$$H_{de}(X) = \frac{1}{2} \log \det(\mathbf{\Sigma}) + \frac{d}{2}(\log 2\pi + 1) = \frac{1}{2} \sum_{i=1}^{d} \log \lambda_i + C$$

Multi-variant Gaussian Distribution $X \sim N(\boldsymbol{\mu}, \mathbf{\Sigma})$

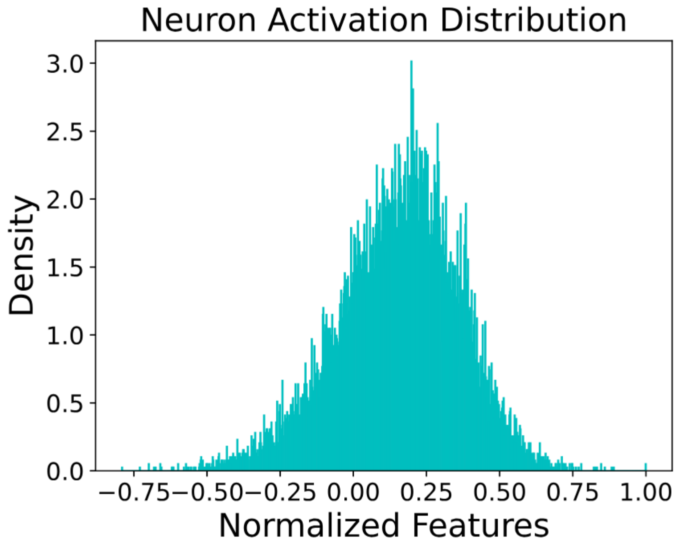
→ the differential entropy is determined by the eigenvalues (LogDet) of the covariance matrix

Test Time Feature Clipping

LLMs are subject to the risks of self-consistent (overconfident) hallucinations



(a) Neuron Activation



(b) Feature Distribution

Figure 2: Illustration of activation distributions in the penultimate layer of LLaMA-7B. (a) Activation distribution in the penultimate layer for a randomly sampled token. (b) Activation distribution for a randomly sampled neuron activation of numerous tokens.

Test Time Feature Clipping

Reduce overconfident prediction for Out-of-Distribution (OOD) detect with Piecewise function

$$FC(h) = \begin{cases} h_{min}, & h < h_{min} \\ h, & h_{min} \le h \le h_{max} \\ h_{max} & h > h_{max} \end{cases}$$

where h represents the feature of the hidden embeddings in the penultimate layer of the LLMs,
$\boldsymbol{h_{min}}$ **and** $\boldsymbol{h_{max}}$ are two thresholds for determining the minimum and maximum truncation activation

"Memory bank" which dynamically pushes and pops element in it to $\boldsymbol{N}$ **embedding tokens**
→ p-th percentiles of the features in the memory bank ($p = 0.2$)

Experimental Results

Table 1: Hallucination detection performance evaluation of different methods on four QA tasks. AUROC (AUC) and Pearson Correlation Coefficient (PCC) are utilized to measure the performance. AUC_s represents AUROC score with sentence similarity as correctness measure, and AUC_r represents AUROC score with ROUGE-L score as correctness measure. All numbers are percentages.

Models	Datasets / Methods	CoQA AUC_s	CoQA AUC_r	CoQA PCC	SQuAD AUC_s	SQuAD AUC_r	SQuAD PCC	NQ AUC_s	NQ AUC_r	NQ PCC	TriviaQA AUC_s	TriviaQA AUC_r	TriviaQA PCC
LLaMA-7B	Perplexity	64.1	68.3	20.4	57.5	60.0	10.2	74.0	74.7	30.1	**83.6**	83.6	54.4
	Energy	51.7	54.7	1.0	45.1	47.6	-10.7	64.3	64.8	18.2	66.8	67.1	29.1
	LN-Entropy	68.7	73.6	30.6	70.1	70.9	30.0	72.8	73.7	29.8	83.4	83.2	54.0
	Lexical Similarity	74.8	77.8	43.5	74.9	76.4	44.0	73.8	75.9	30.6	82.6	**84.0**	55.6
	EigenScore	**80.4**	**80.8**	**50.8**	**81.5**	**81.2**	**53.5**	**76.5**	**77.1**	**38.3**	82.7	82.9	**57.4**
LLaMA-13B	Perplexity	63.2	66.2	20.1	59.1	61.7	14.2	73.5	73.4	36.3	**84.7**	**84.5**	56.5
	Energy	47.5	49.2	-5.9	36.0	39.2	-20.2	59.1	59.8	14.7	71.3	71.5	36.7
	LN-Entropy	68.8	72.9	31.2	72.4	74.0	36.6	74.9	75.2	39.4	83.4	83.1	54.2
	Lexical Similarity	74.8	77.6	44.1	77.4	79.1	48.6	74.9	76.8	40.3	82.9	84.3	57.5
	EigenScore	**79.5**	**80.4**	**50.2**	**83.8**	**83.9**	**57.7**	**78.2**	**78.1**	**49.0**	83.0	83.0	**58.4**
OPT-6.7B	Perplexity	60.9	63.5	11.5	58.4	69.3	8.6	76.4	77.0	32.9	**82.6**	**82.0**	**50.0**
	Energy	45.6	45.9	-14.5	41.6	43.3	-16.4	60.3	58.6	25.6	70.6	68.8	37.3
	LN-Entropy	61.4	65.4	18.0	65.5	66.3	22.0	74.0	76.1	28.4	79.8	80.0	43.0
	Lexical Similarity	71.2	74.0	38.4	72.8	74.0	39.3	71.5	74.3	23.1	78.2	79.7	42.5
	EigenScore	**76.5**	**77.5**	**45.6**	**81.7**	**80.8**	**49.9**	**77.9**	**77.2**	**33.5**	80.3	80.4	0.485

Experimental Results

Table 2: Hallucination detection performance evaluation of different methods with and without (w/o) applying feature clipping (FC). "+FC" denotes applying feature clipping and EigenScore (w/o) denotes EigenScore without applying feature clipping. All numbers are percentages.

Model	LLaMA-7B				OPT-6.7B			
Datasets	CoQA		NQ		CoQA		NQ	
Methods	AUC_s	PCC	AUC_s	PCC	AUC_s	PCC	AUC_s	PCC
LN-Entropy	68.7	30.6	72.8	29.8	61.4	18.0	74.0	28.4
LN-Entropy + FC	70.0	33.4	73.4	31.1	62.6	21.4	74.8	30.3
Lexical Similarity	74.8	43.5	73.8	30.6	71.2	38.4	71.5	23.1
Lexical Similarity + FC	76.6	46.3	74.8	32.1	72.6	40.2	72.4	24.2
EigenScore (w/o)	79.3	48.9	75.9	38.3	75.3	43.1	77.1	32.2
EigenScore	80.4	50.8	76.5	38.3	76.5	45.6	77.9	33.5

Experimental Results

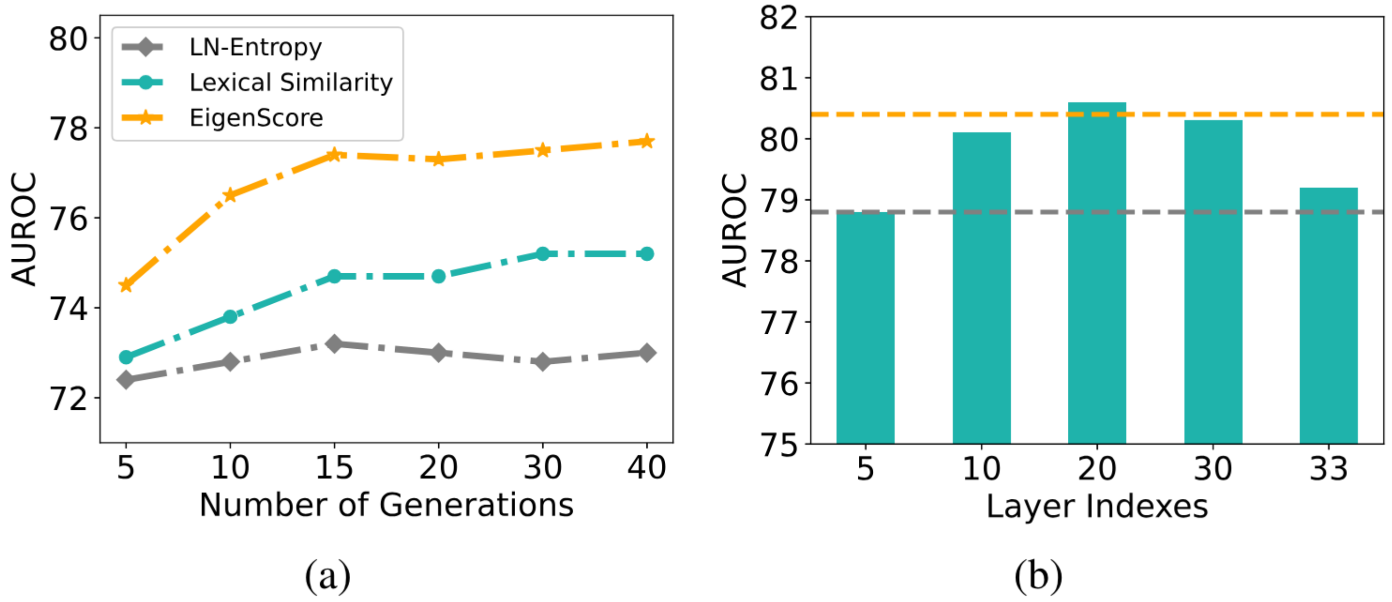


Figure 3: (a) Performance in LLaMA-7B and NQ dataset with different number of generations. (b) Performance in LLaMA-7B and CoQA dataset with sentence embedding in different layers. Orange line indicates using the last token's embedding in the middle layer (layer 17) as sentence embedding. Gray line indicates using the averaged token embedding in the last layer as sentence embedding. The performance is measured by AUROC_s.

Experimental Results

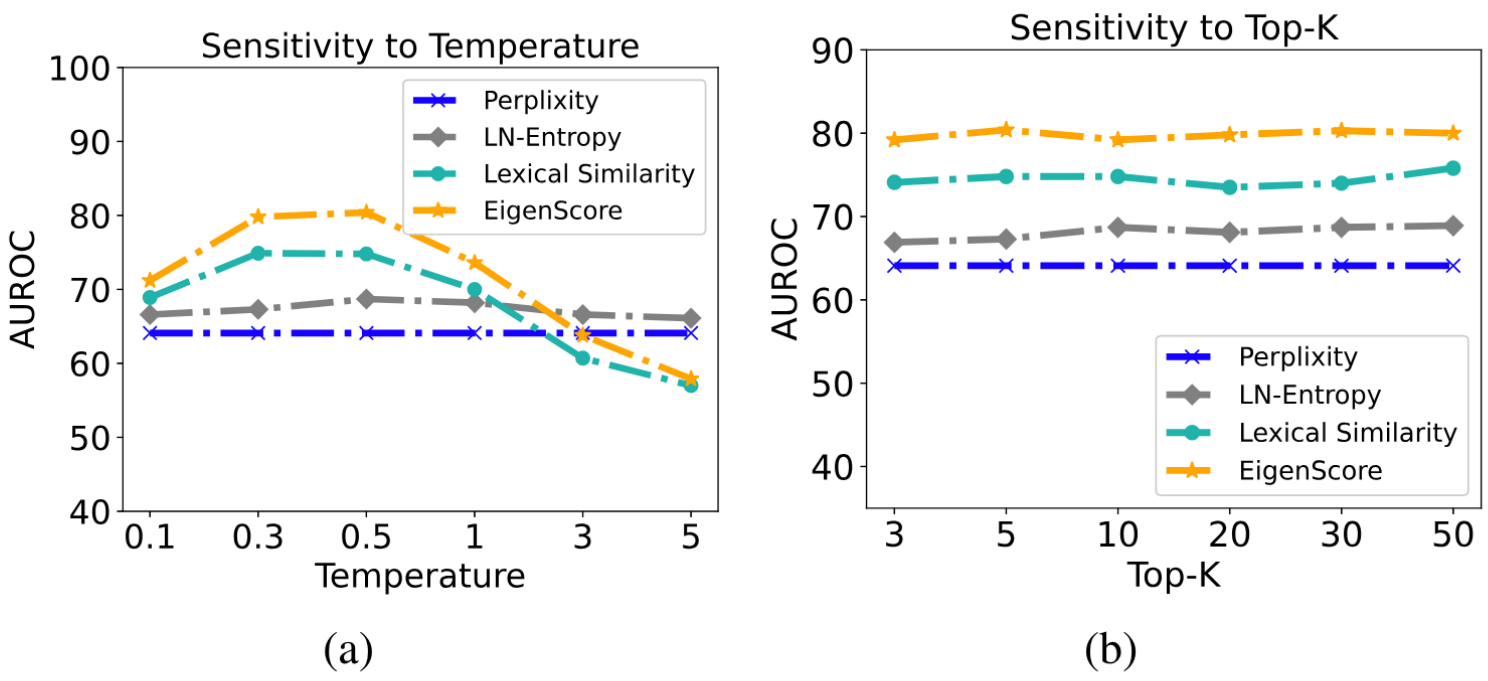


Figure 4: (a) Performance sensitivity to temperature. (b) Performance sensitivity to top-k. The performance is measured by AUROC_s.

Q & A