ICR Probe: Tracking Hidden State Dynamics for Reliable Hallucination Detection in LLMs

Zhenliang Zhang[1,2], **Xinyu Hu**[1], **Huixuan Zhang**[1]
Junzhe Zhang[1], **Xiaojun Wan**[1]
[1]Wangxuan Institute of Computer Technology, Peking University
[2]School of Software and Microelectronics, Peking University
{zhenliang,zhanghuixuan,junzhezhang}@stu.pku.edu.cn
{huxinyu,wanxiaojun}@pku.edu.cn

Abstract

Large language models (LLMs) excel at various natural language processing tasks, but their tendency to generate hallucinations undermines their reliability. Existing hallucination detection methods leveraging hidden states predominantly focus on static and isolated representations, overlooking their dynamic evolution across layers, which limits efficacy. To address this limitation, we shift the focus to the hidden state update process and introduce a novel metric, the ICR Score (**I**nformation **C**ontribution to **R**esidual Stream), which quantifies the contribution of modules to the hidden states' update. We empirically validate that the ICR Score is effective and reliable in distinguishing hallucinations. Building on these insights, we propose a hallucination detection method, the ICR Probe, which captures the cross-layer evolution of hidden states. Experimental results show that the ICR Probe achieves superior performance with significantly fewer parameters. Furthermore, ablation studies and case analyses offer deeper insights into the underlying mechanism of this method, improving its interpretability.

1 Introduction

Large language models (LLMs) demonstrate remarkable performance across various natural language processing tasks (Zhang et al., 2023). However, these models are still prone to generating hallucinations, which are nonsensical or irrelevant content that deviates from the intended output (Ji et al., 2023). This issue highlights the critical need for effective methods of hallucination detection.

Various methods for hallucination detection exist. Mainstream approaches analyze generated output through consistency checks or reference comparisons (Manakul et al., 2023), while probability-based methods focus on logit probability uncertainty (Ren et al., 2022). Another approach examines hidden states (e.g., embedding vectors) across LLM layers to detect hallucinations (Chen et al., 2024a,c). Methods based on output or logit probabilities often require ground truth references or multiple generations for consistency. In contrast, hidden state based detection offers the advantage of being reference-free, eliminating the need for external sources.

Current hallucination detection methods based on hidden states can be broadly categorized into training-based and training-free methods. Training-based methods often involve training individual or combination probes (Azaria and Mitchell, 2023; CH-Wang et al., 2024), or using semantic entropy probes (Kossen et al., 2024), whereas training-free methods calculate detection metrics directly from hidden states (Sriramanan et al., 2024). However, existing methods typically focus on static, high-dimensional hidden states (around 4000 dimensions), which limits feature extraction capabilities. These methods make it challenging to capture the updates of hidden states and the cross-layer evolution of the residual stream, ultimately restricting the effectiveness of hallucination detection.

To overcome these limitations, we introduce a novel approach that **shifts the focus from the hidden states themselves to their update process across layers**. Specifically, the **ICR** Score (**I**nformation **C**ontribution to **R**esidual Stream) quantifies the contribution of different modules (e.g., FFN or self-attention) to hidden state updates at each layer. Empirical results demonstrate that the ICR Score captures a stable and consistent pattern of residual stream updates, showing strong potential for distinguishing hallucinations.

We further introduce the **ICR Probe**, which aggregates ICR Scores across all layers to capture the comprehensive dynamics of the residual stream. Experimental results on three mainstream open-source LLMs demonstrate that the ICR Probe effectively detects hallucinations and outperforms previous methods across multiple datasets. Additionally, ablation studies are conducted to reveal

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 17986–18002

the underlying mechanisms. In summary, our core contributions are:

- **Novel Detection Signal**: We propose a hallucination detection method by focusing on the update patterns of hidden states across layers, introducing the ICR Score, a metric that captures the dynamic evolution of residual stream updates.
- **ICR Probe Development**: We introduce the ICR Probe, an effective and robust tool for hallucination detection, offering superior performance with fewer parameters.
- **Empirical Validation**: We conduct extensive empirical evaluations, showcasing the effectiveness, generalizability, and interpretability of our method across various datasets, reinforcing the understanding of its underlying mechanisms[1].

2 Background

Information Contribution of Different Modules In each layer of large language models (LLMs), multi-head self-attention (MHSA) and feed-forward networks (FFN) perform distinct but complementary functions in information processing and knowledge integration (Geva et al., 2021). MHSA functions as a **contextual signal router**, dynamically modulating token interactions based on query-key affinity scores. By attending to preceding tokens, MHSA redistributes existing contextual information and integrates it into the evolving residual stream of hidden states (Elhage et al., 2021). Importantly, MHSA does not directly extract new knowledge from the model parameters. Instead, it reallocates information already present in token representations, enhancing relevant associations within the context.

In contrast, FFNs act as **key-value memory banks**, facilitating sparse, content-based knowledge retrieval (Geva et al., 2023). Unlike MHSA, FFNs do not extract new information from the context but rather retrieve and integrate knowledge stored in the model's internal parameters, injecting learned knowledge into the residual stream.

Layer-wise Information Contribution Characteristics In LLMs, the contributions of FFNs and MHSA vary across layers, leading to distinct layer-specific characteristics. Stolfo et al. (2023) employ causal mediation analysis to investigate the information flow in LLMs during reasoning tasks. Their analysis reveals that FFNs predominantly influence the early and deeper layers, while MHSA plays a central role in the intermediate layers. Specifically, FFNs drive the processing of operands and operators, while MHSA manages the dynamic redistribution of information across tokens. In later stages, FFNs retrieve and integrate task-specific knowledge to generate results. Similar findings are reported by Elhage et al. (2021) and Geva et al. (2023), who observe that FFNs and MHSA dominate hidden state updates at different layers. These differences in information contribution provide valuable insights into the model's internal information flow patterns.

3 ICR Score

Previous research by Stolfo et al. (2023) suggests that the information contribution to the residual stream update process may be correlated with model performance, making it a potential signal for hallucination detection. Based on this insight, we formalize the residual stream and introduce the **ICR** Score (**I**nformation **C**ontribution to **R**esidual Stream), a metric that quantifies each module's contribution to the residual stream updates.

3.1 Residual Stream Update Process

We describe the residual stream update process, focusing on how the varying contributions of different modules lead to distinct characteristics across layers.

Residual stream update process in LLMs. LLMs process inputs through multiple sequential computational stages. First, the input text is tokenized into N discrete tokens $\{t_i\}_{i=1}^{N}$. Each token is then mapped to a d-dimensional hidden state. These hidden states evolve iteratively through the L layers of decoder architecture. At layer ℓ, the hidden states are updated as shown in Equation (1), where x_i^ℓ is the hidden state of token t_i in layer ℓ:

$$x_i^\ell = x_i^{\ell-1} + \underbrace{\text{MHSA}^\ell(x_i^{\ell-1})}_{a_i^\ell} + \underbrace{\text{FFN}^\ell(x_i^{\ell-1} + a_i^\ell)}_{m_i^\ell} \quad (1)$$

The hidden state is updated through two key modules: multi-head self-attention (MHSA) and feed-forward networks (FFN).

[1]The code is available in https://github.com/XavierZhang2002/ICR_Probe

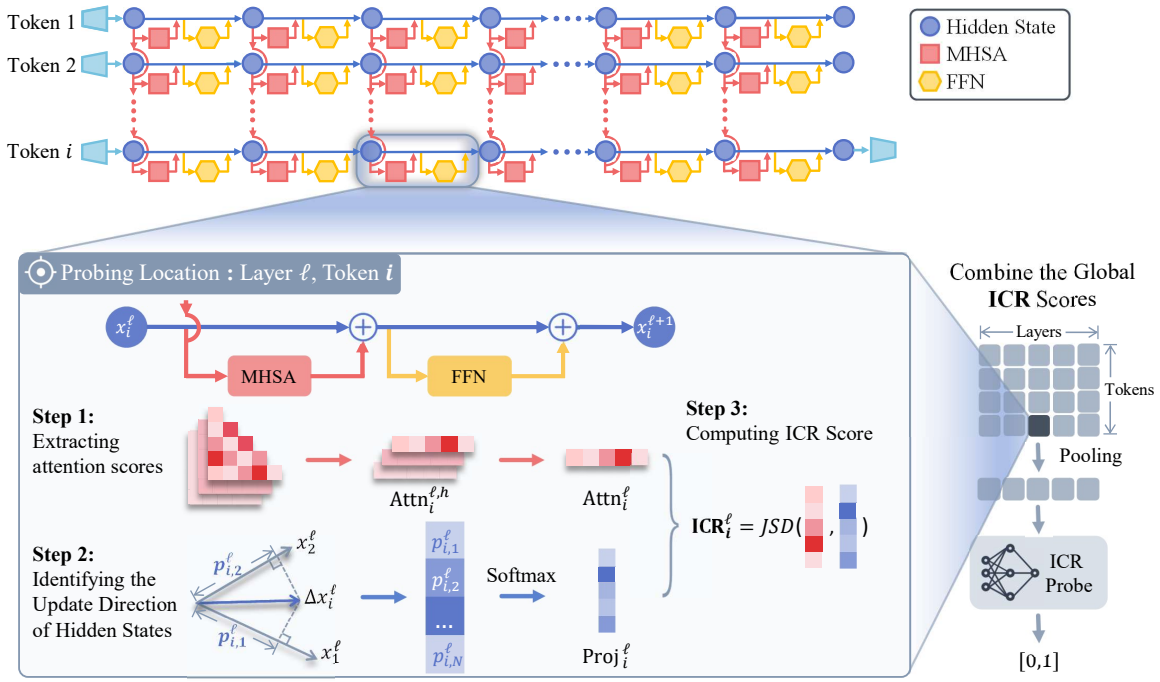


Figure 1: Overview of the ICR Score computation and ICR Probe detection process. We first probe each layer's residual stream to obtain the ICR Score, then aggregate the ICR Scores across all layers to capture global information. This aggregated signal serves as the input to the probe, which produces the final hallucination detection result.

a_i^ℓ represents the contribution from the MHSA module, and m_i^ℓ represents the contribution from the FFN module. As discussed in Section 2, contemporary research highlights the differing roles of these two modules in processing information.

Layer-wise Characteristics of Residual Stream Updates. As discussed in Section 2, the updates to the residual stream of hidden states differ across layers, with each layer contributing uniquely through the MHSA and FFN modules.

3.2 Constructing the ICR Score

To quantify the contribution of different modules to the updates of hidden states, we define the total update of hidden states at layer ℓ as:

$$\Delta x_i^\ell = x_i^\ell - x_i^{\ell-1} = a_i^\ell + m_i^\ell \tag{2}$$

Here, Δx_i^ℓ represents the update of the hidden state, which consists of two components: the contextual information a_i^ℓ derived from the MHSA module, and the parametric knowledge m_i^ℓ derived from the FFN module. Our objective is to quantify the specific contribution of each module.

A straightforward method for quantifying this contribution is to compute the ratio of the respective contributions in the residual stream, such as $\frac{a_i^\ell}{a_i^\ell + m_i^\ell}$. However, this ratio may not always be reliable, since m_i^ℓ is computed using a_i^ℓ, which complicates the distinction between the FFN's contribution of new parameterized knowledge and its reinforcement of information from the MHSA.

To address this issue, we propose an effective method for quantifying the contributions by measuring the consistency between the hidden state update Δx_i^ℓ and the attention score. As illustrated in Figure 1, the computation process consists of three steps:

Extracting Attention Scores For the i-th token in layer ℓ, the attention score with respect to the j-th token in the context for a given attention head h is computed as:

$$\text{Attn}_{i,j}^{\ell,h} = \frac{(Q_i^h)^T \cdot K_j^h}{\sqrt{d}} \tag{3}$$

where Q_i^h and K_j^h are the query and key vectors for the i-th and j-th tokens, and d is the dimension of the vectors.

The attention score vector for the i-th token across all N tokens is:

$$\text{Attn}_i^{\ell,h} = \left[\text{Attn}_{i,1}^{\ell,h}, \text{Attn}_{i,2}^{\ell,h}, \ldots, \text{Attn}_{i,N}^{\ell,h}\right] \in \mathbb{R}^N \tag{4}$$

To obtain a more robust attention score for the i-th token, we average the scores across all H attention heads:

$$\text{Attn}_i^\ell = \frac{1}{H}\sum_{h=1}^{H}\text{Attn}_i^{\ell,h} \in \mathbb{R}^N \tag{5}$$

Identifying the Update Direction of Hidden States The set $\{x_j^\ell\}_{j=1}^N$ represents the hidden states of all N input tokens. To quantify how much the update Δx_i^ℓ incorporates information from the context, we calculate the projection of Δx_i^ℓ onto the hidden states x_j^ℓ of all tokens $\{t_i\}_{i=1}^N$ in the input context. Specifically, we calculate the projection length $p_{i,j}^\ell$ between Δx_i^ℓ and x_j^ℓ, normalized by the norm of x_j^ℓ. This is given by Equation 6.

$$p_{i,j}^\ell = \frac{(\Delta x_i^\ell)^T \cdot x_j^\ell}{\|x_j^\ell\|} \tag{6}$$

where $\|x_j^\ell\|$ is the magnitude (norm) of the hidden state of the j-th token. The projection length $p_{i,j}^\ell$ measures the magnitude of the update Δx_i^ℓ in the direction of the hidden state x_j^ℓ.

We then construct the projection vector $\text{Proj}_i^\ell = [p_{i,j}^\ell]_{j=1}^N$ and normalize it using softmax.

Computing Consistency Between Hidden State Update Direction and Attention Scores We measure the consistency between the hidden state update direction Proj_i^ℓ and the attention scores Attn_i^ℓ using the Jensen-Shannon Divergence (JSD). The **ICR** Score (**I**nformation **C**ontribution to **R**esidual Stream) is defined in Equation 7, with the top k tokens selected based on the highest attention scores. Appendix A provides more details.

$$\text{ICR}_i^\ell = \text{JSD}(\text{Proj}_i^\ell, \text{Attn}_i^\ell) \tag{7}$$

The ICR score quantifies each module's contribution to the residual stream update at each layer, identifying the dominant module shaping the hidden state update Δx_i^ℓ. The interpretations of the ICR score are as follows:

- A small ICR score indicates that the hidden state update Δx_i^ℓ aligns closely with the attention scores Attn_i^ℓ. This implies that the **attention mechanism predominantly drives the hidden state update**, while the FFN primarily reinforces the information without adding significant new content.

- A large ICR score signifies a noticeable divergence between the hidden state update Δx_i^ℓ and the attention scores. **This divergence suggests that the FFN is more dominant in determining the update**, with the attention mechanism contributing less to the information stream.

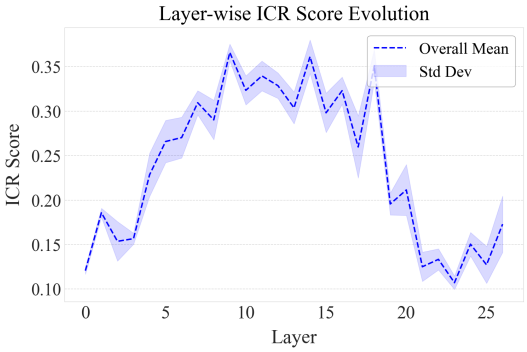


Figure 2: Mean ICR scores with standard deviation bands across four datasets. The narrow standard deviation bands indicate that the ICR score captures the residual stream patterns with strong cross-dataset consistency and stability.

The ICR score captures the relative contribution of each module across L layers and N tokens, revealing how attention and FFN modules shape the model's residual stream.

4 ICR Probe for Hallucination Detection

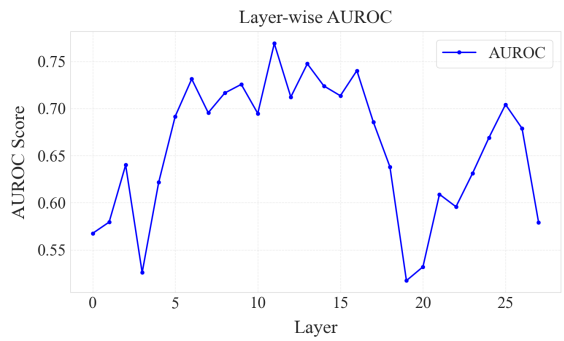
In this section, we analyze the layer-wise characteristics of the ICR Score and its ability to distinguish hallucinations, and design the ICR Probe for hallucination detection.

4.1 Empirical Study of the ICR Score

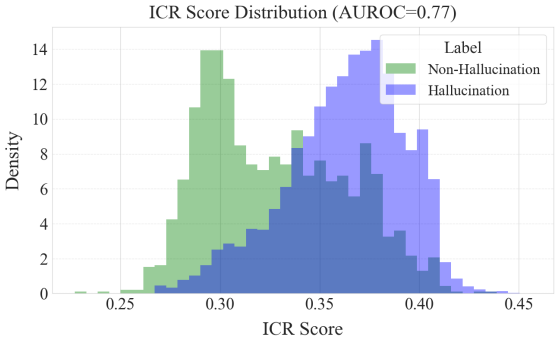
We investigate the potential of the ICR score for hallucination detection, addressing two requirements: (1) consistency/stability of layer-wise features, and (2) its ability to distinguish between hallucinated and non-hallucinated outputs.

Stability of Layer-wise Features in the ICR Score Using Qwen2.5 as an example, we compute the mean and variance of the ICR scores across layers on four different datasets (model and dataset details in Section 5.1) to examine the layer-wise characteristics and cross-dataset consistency.

As shown in Figure 2, the ICR score follows a clear progression across layers. In the early layers (0–3), low scores indicate that updates are primarily driven by MHSA, focusing on local contextual extraction. From layers 4–20, the ICR score increases, peaking around layers 10–15, suggesting a shift towards FFN-dominated updates, where parametric knowledge is integrated. Beyond layer 21,



(a) Layer-wise AUROC for hallucination detection **using direct ICR scores**, peaking at layer **11**, which indicates the layer most effective for hallucination detection.



(b) Probability density distribution of ICR Scores at layer 11, showing clear separability between hallucinated and non-hallucinated samples.

Figure 3: ICR Score for hallucination detection: (a) Layer-wise AUROC and (b) probability density distributions, highlighting ICR's effectiveness as a predictive metric.

the score declines, signaling a return to MHSA-driven updates for refining contextual integration. The narrow standard deviation bands in the figure further support the consistency of this pattern. This feature remains **consistent across datasets**, suggesting that the ICR score reflects an intrinsic property of the model rather than dataset-specific variations. It provides a structured representation of the information flow within LLMs.

Takeaway The ICR score systematically maps the model's intrinsic residual stream update pattern, providing a stable feature that confirms its potential as a diagnostic tool for understanding model performance.

ICR Score's Ability to Distinguish Hallucination To evaluate the detection capability of the ICR score, we assess its ability to distinguish hallucinated outputs from non-hallucinated ones. The ICR score is used as a direct feature for classification. Using the HaluEval dataset as an example, Figure 3a presents the AUROC values across layers, illustrating the inherent separability of ICR for hallucination detection. AUROC exceeds 0.7 in ten layers, peaking at 0.7690 at layer 11, which indicates that the ICR score effectively reflects the model's performance trends.
Figure 3b shows the probability density distribution of ICR scores at layer 11, with clear separation between hallucinated and non-hallucinated samples, confirming ICR's ability to distinguish reliable from erroneous generations.

Takeaway The ICR score effectively distinguishes between hallucinated and non-hallucinated outputs, confirming its potential as a predictive metric for hallucination detection.

4.2 ICR Probe

The previous analysis demonstrates that the ICR Score effectively captures the model's information contributions and serves as a reliable indicator for hallucination detection. However, utilizing the ICR score from a single layer provides only a partial representation of residual stream updates. To capture the global update pattern of the residual stream, we introduce the **ICR Probe**, a classifier trained on ICR Scores across all layers to detect hallucinations.

The ICR Probe takes the averaged ICR score as input, which is obtained by performing token-wise averaging of the original $N \times L$ matrix, resulting in a $1 \times L$ vector. The output is a scalar value between 0 and 1, representing the likelihood of a non-hallucinated generation.

Architecture Based on empirical evaluations, we adopt the model architecture $(L, 128, 64, 32, 1)$ for the ICR Probe, which optimally balances model complexity with computational efficiency. For models with $L < 42$, this configuration ensures that the total number of parameters remains below **16K**, maintaining computational efficiency. Detailed specifications are provided in Appendix B.4.

5 Experiments

We assess the performance of the ICR Probe for hallucination detection, comparing it with baselines. Additionally, we investigate its generalizability across different datasets, conduct ablation studies, and present a case study focused on token-level detection.

LLMs	Methods	HaluEval	SQuAD	HotpotQA	TriviaQA
Gemma-2	PPL	0.5546	0.5419	0.7196	0.7765
	LN-Entropy	0.7451	0.6531	0.7172	0.7200
	LLM-check	0.5683	0.5713	0.5600	0.5821
	SAPLMA	0.8101	0.7175	0.8193	0.7751
	SEP	0.6429	0.6417	0.6356	0.7834
	ICR Probe (Ours)	**0.8436**	**0.8142**	**0.8409**	**0.8001**
Qwen2.5	PPL	0.5439	0.5303	0.6125	0.6974
	LN-Entropy	0.7371	0.6497	0.6855	0.7004
	LLM-check	0.5292	0.5700	0.5445	0.5552
	SAPLMA	0.7799	0.6929	0.7750	**0.8225**
	SEP	0.6578	0.6363	0.6470	0.7520
	ICR Probe (Ours)	**0.8003**	**0.7456**	**0.7917**	0.7684
Llama-3	PPL	0.5920	0.6399	0.6721	0.7012
	LN-Entropy	0.6508	0.6306	0.6593	0.5860
	LLM-check	0.5206	0.5411	0.5491	0.5417
	SAPLMA	0.7238	0.7107	0.7701	**0.7650**
	SEP	0.7378	0.7201	0.6541	0.7116
	ICR Probe (Ours)	**0.7603**	**0.7634**	**0.7982**	0.7325

Table 1: Hallucination detection performance on four datasets. Higher AUROC values indicate better performance.

5.1 Settings

Models We evaluate the detection performance of ICR Probe on three mainstream open-source LLMs: Llama-3-8B-Instruct (AI@Meta, 2024), Qwen2.5-7B-Instruct (Team, 2024), and Gemma-2-9B-it (Team et al., 2024).

Datasets The models are evaluated on several benchmark datasets: HaluEval (Li et al., 2023), which focuses on hallucination detection; SQuAD (Rajpurkar et al., 2016), a reading comprehension dataset; TriviaQA (Joshi et al., 2017), which tests general knowledge; and HotpotQA (Yang et al., 2018), designed for multi-hop question answering.

Baselines We compare hallucination detection performance with several baselines. The training-free methods include **PPL** (Ren et al., 2022), **LN-Entropy** (Malinin and Gales, 2020), and **LLM-check** (Sriramanan et al., 2024). The training-based methods include **SAPLMA** (Azaria and Mitchell, 2023) and **SEP** (Kossen et al., 2024). Details of these baselines are in Appendix B.2.

Evaluation Metric Since hallucination detection is a binary classification task, we use the Area Under the Receiver Operating Characteristic Curve (AUROC), a threshold-independent metric that captures the trade-off between true and false positive rates.

ICR Probe Training The ICR Probe is trained on the pooled ICR Score to detect hallucinations. It is a lightweight multi-layer perceptron (MLP) classifier with four fully connected layers. The training follows a standard supervised learning setup with binary cross-entropy loss and Adam optimization.

Following the experimental setup outlined by Orgad et al. (2024), we partition each dataset into training and testing sets and report the results for each dataset accordingly. **The experimental setup for baselines matches ours exactly.** Additionally, we conduct a dedicated evaluation of generalization in Section 5.3.

5.2 Hallucination Detection Results

The performance of the ICR Probe is evaluated across four datasets and three mainstream open-source LLMs. Table 1 presents the key experimental results, showing that the ICR Probe outperforms both train-free and training-based baseline methods on most datasets.

Our approach offers several distinct advantages beyond its effectiveness in hallucination detection. First, the ICR Probe integrates global information from all layers, unlike methods that focus on isolated hidden states, which capture only local information. Our method also incorporates the contribution of modules, capturing the entire global residual stream update process. This provides a more comprehensive view of the model's behavior. Additionally, the ICR Probe eliminates the need to select specific layers or tokens, making it highly adaptable and scalable across different models and tasks. It also enables real-time hallucination detection from a single generation, eliminating the need for multiple samples. Furthermore, the ICR

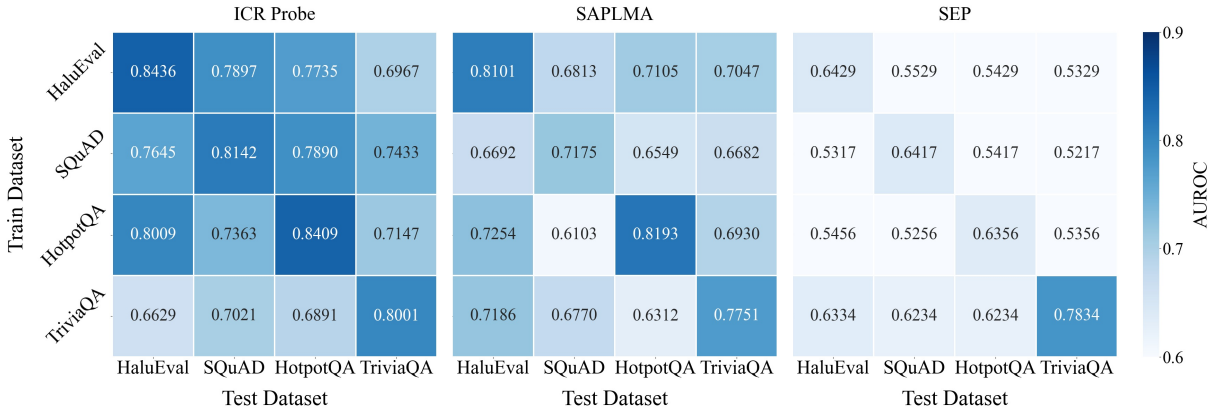


Figure 4: Cross-dataset generalization heatmaps for ICR Probe, SAPLMA, and SEP. Each subplot displays the AUROC when the model is trained on the row dataset and tested on the column dataset, with values annotated in each cell. A shared color scale highlights comparative generalization performance, showing that the ICR Probe not only achieves higher absolute AUROCs across unseen datasets but also exhibits the smallest average drop under domain shift.

Probe is parameter efficient, with only 16K parameters, significantly smaller than alternatives like SAPLMA, which has 110K parameters (Azaria and Mitchell, 2023). This ensures the approach remains both scalable and practical.

5.3 Generalization Analysis

We assess cross-dataset generalization by training each model on one dataset and evaluating on another, visualized as heatmaps in Figure 4. Each cell reports the AUROC of the model when trained on the row dataset and tested on the column dataset, using a shared color scale (0.6–0.9) for direct comparison.

Across unseen datasets, our ICR Probe maintains AUROC ≥ 0.66 and outperforms both SAPLMA and SEP. Furthermore, the average AUROC drop from in-domain to cross-domain is only 8.61 % for our method, versus 10.18 % for SAPLMA and 11.67 % for SEP, underscoring its superior accuracy and robustness under domain shift.

This advantage stems from leveraging layer-wise residual stream evolution patterns-capturing intrinsic LLM dynamics independent of dataset-specific artifacts (see Figures 2 and 5)—whereas SAPLMA and SEP rely on data—dependent representations and thus exhibit greater cross-domain sensitivity.

5.4 Ablation Experiments and Interpretability

We conduct ablation experiments on Gemma-2 to investigate two main aspects: (1) the contribution of different components of the ICR Score to hallucination detection, and (2) the impact of different layers on the Probe's performance.

Setting Attn_i^ℓ	Setting Proj_i^ℓ	HaluEval	SQuAD	HotpotQA	TriviaQA
×	×	0.5000	0.5000	0.5000	0.5000
×	✓	0.7993	0.7765	0.7829	0.7723
✓	✓	0.8436	0.8142	0.8409	0.8001

Table 2: Ablation study of the ICR score components, showing the impact of attention (Attn_i^ℓ) and hidden state projections (Proj_i^ℓ) on hallucination detection performance.

Ablation Study 1: Contribution of Each component. As defined in Equation 7, the ICR Score consists of two components: the update directions of hidden states (Proj_i^ℓ) and the attention scores (Attn_i^ℓ). To evaluate the contribution of each component, we design three experimental settings: In the NONE setting, both the hidden state (Proj_i^ℓ) and attention (Attn_i^ℓ) signals are excluded, resulting in random performance (AUROC = 0.5). In the HS ONLY setting, only the hidden state information (Proj_i^ℓ) is used, while the attention scores (Attn_i^ℓ) are set to a uniform distribution. At this stage, the calculation corresponds to the entropy of the Proj_i^ℓ. Finally, in the HS + ATTN setting, both the hidden state and attention signals (Proj_i^ℓ and Attn_i^ℓ) are integrated. Details of the ablation experiments are provided in Appendix B.5.

The results, shown in Table 2, demonstrate that even the hidden state component alone significantly improves hallucination detection. This suggests that the information entropy of Proj_i^ℓ is critical. Moreover, combining both components (hidden state and attention) improves performance, high-

Setting	HaluEval	SQuAD	HotpotQA	TriviaQA
Full Layers	0.8436	0.8142	0.8409	0.8001
w/o EL	0.7833	0.8045	0.8132	0.7885
w/o ML	0.7774	0.7582	0.7421	0.7197
w/o DL	0.7285	0.7924	0.7936	0.7657

Table 3: Layer-wise ablation study showing AUROC scores under different layer configurations.

lighting their complementary nature and validating the construction of the ICR Score.

Ablation Study 2: Layer Contribution Analysis. We perform a layer-wise ablation study to analyze the contributions of different layer groups in the LLM to hallucination detection. The model is divided into three stages: Early Layers (Layer 0-13), Middle Layers (Layer 15-28), and Deep Layers (Layer 29-41). In each setting, **remove** a specific layer group, e.g. Early Layers (EL), Middle Layers (ML), Deep Layers (DL).

The results in Table 3 show that removing Middle Layers significantly reduces performance, emphasizing their critical role in hallucination detection.

5.5 Token-level Detection: Case Study

Although our method is designed for sequence-level detection (as discussed in Section 4.2), we perform token-level analysis to explain the probe's detection capabilities and characteristics.

In token-level detection, each token is processed individually by the ICR Probe, with the prediction probability indicating its correctness. This method is effective only for detecting hallucinations in **key tokens** that are crucial to the answer. Table 4 shows that in hallucinated cases, key tokens like `hectares` and `Vietnam` have low probabilities (0.04 and 0.19), successfully identifying hallucinations, while common tokens like `The` and `.` have high probabilities (0.98, 0.99), causing misdetection. In non-hallucinated cases, the key token `Cree` is correctly detected, while `" "` is misidentified.

These results show that token-level detection is effective for key tokens but less reliable for frequent, less informative tokens. The ICR Probe primarily captures the residual stream pattern of key tokens, while irrelevant tokens act as noise, providing insights into its mechanism and guiding future optimizations.

Example
Question: Tiendesitas is a part of an interim business district with an area of more than what? **Response:** 18.5 hectares **Token-level Detection Result:** `"1"(0.34), "8"(0.99), "."(0.21), "5"(0.27), " hectares"(0.04), " "(0.96), "\n"(0.98)` **Correct Answer:** 100 ha
Question: What was the official name for the draft that Thomas E. O'Donnell was a public opponent of? **Response:** The Vietnam War draft. **Token-level Detection Result:** `"The"(0.98), " Vietnam"(0.19), " War"(0.28), " draft"(0.17), "."(0.99), " "(0.03), "\n"(0.99)` **Correct Answer:** Conscription in the United States
Question: Which First Nations group of over 200'000 members living in Canada signed an accord with the Quebec government, leading to the creation of Eeyou Istchee James Bay Territory? **Response:** The Cree. **Token-level Detection Result:** `"The"(0.54), " Cree"(0.55), "."(0.78), " "(0.06), "\n"(0.99)` **Correct Answer:** The Cree

Table 4: Token-level detection examples, with predicted probabilities displayed next to each token. Tokens are color-coded based on their probabilities: **Red** for hallucinations and **Green** for non-hallucinations.

6 Related Work

Hallucinations in LLMs Hallucinations are a prevalent phenomenon in large language models, typically defined as generated content that is nonsensical or deviates from the provided source content (Huang et al., 2025; Ji et al., 2023; Li et al., 2023). These hallucinations pose significant risks to the reliability and accuracy of LLMs. As a result, extensive research has been conducted to understand their underlying causes (Shi et al., 2023), improve detection methods (Azaria and Mitchell, 2023), and develop strategies for mitigating their effects (Chuang et al., 2024).

Hallucination Detection via Hidden States Several methods based on hidden states are proposed for hallucination detection. Azaria and Mitchell (2023) extract hidden states from specific layers and train probes, a method limited by its reliance on selected layers and tokens, which prevents it from capturing the dynamics across layers. CH-Wang et al. (2024) improve this approach by integrating probes, though this comes with high training costs and the need for extensive hyperparameter tuning. SEP (Kossen et al., 2024) utilizes indirect signals, such as semantic entropy from multiple outputs, to train probes for hallucination detection. Train-free

methods are also introduced, such as calculating the regularized covariance matrix of hidden states (Chen et al., 2024b) or the eigenvalues of attention kernels (Sriramanan et al., 2024). These methods eliminate the need for training, though they either require multiple generations or demonstrate limited detection performance. Additionally, Sun et al. (2024) propose the ReDeEP method, which focuses on RAG hallucinations by examining specific layers' external context and parameter knowledge. While this approach captures the contribution of different modules, it lacks a comprehensive view of the entire residual stream and is dependent on hyperparameter selection.

7 Conclusion

In this work, we introduce a novel approach to hallucination detection in LLMs by focusing on the dynamic updates of hidden states rather than their static representations. The ICR Score quantifies the contribution of model components to the residual stream updates, providing a more precise measure for detection. By aggregating these scores across layers, the ICR Probe captures the cross-layer evolution of the residual stream, enabling more accurate and comprehensive detection. Extensive experiments demonstrate that the ICR Probe outperforms baseline methods, with strong generalization across datasets and enhanced interpretability. Ablation studies and case studies further confirm its effectiveness and provide valuable insights into the underlying mechanisms of the approach.

Limitations

The proposed ICR Probe offers an effective approach to hallucination detection but has several limitations. First, it is applicable to open-source models, as it requires access to hidden states, which limits its use in proprietary models. Second, while this work focuses on hallucination detection, it does not directly address mitigating hallucinations in LLMs. We hope that future research, building on the insights and metrics proposed in this work, will explore interventions in the residual stream update process to reduce the tendency for hallucinations. Finally, although our experiments span diverse datasets (e.g., HaluEval, SQuAD), extending evaluations to more complex generation tasks would better validate the robustness of the approach.

Ethics Statement

In this research, we prioritize user privacy and data protection, as our methodology does not rely on personal or sensitive data. All experiments were conducted using publicly accessible datasets, ensuring adherence to privacy and ethical guidelines. While the goal of this work is to enhance the reliability and performance of large language models, we recognize the potential risks of misuse, including the creation of misleading or harmful AI-generated content. We are committed to promoting ethical AI practices, addressing concerns regarding misuse, and ensuring that our research contributes positively to society. Additionally, AI assistants were used solely for text polishing and were not involved in any aspect of the core research process.

Acknowledgements

This work was supported by Beijing Science and Technology Program (Z231100007423011) and Key Laboratory of Science, Technology and Standard in Press Industry (Key Laboratory of Intelligent Press Media Technology). We appreciate the anonymous reviewers for their helpful comments. Xiaojun Wan is the corresponding author.

References

AI@Meta. 2024. Llama 3 model card.

Amos Azaria and Tom Mitchell. 2023. The Internal State of an LLM Knows When It's Lying. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 967–976, Singapore. Association for Computational Linguistics.

Sky CH-Wang, Benjamin Van Durme, Jason Eisner, and Chris Kedzie. 2024. Do androids know they're only dreaming of electric sheep? In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 4401–4420, Bangkok, Thailand. Association for Computational Linguistics.

Chao Chen, Kai Liu, Ze Chen, Yi Gu, Yue Wu, Mingyuan Tao, Zhihang Fu, and Jieping Ye. 2024a. Inside: Llms' internal states retain the power of hallucination detection.

Chao Chen, Kai Liu, Ze Chen, Yi Gu, Yue Wu, Mingyuan Tao, Zhihang Fu, and Jieping Ye. 2024b. Inside: Llms' internal states retain the power of hallucination detection. *ArXiv*, abs/2402.03744.

Shiqi Chen, Miao Xiong, Junteng Liu, ZhengXuan Wu, Teng Xiao, Siyang Gao, and Junxian He. 2024c. In-context sharpness as alerts: An inner representation

perspective for hallucination mitigation. In *Proceedings of the 41st International Conference on Machine Learning*, Proceedings of Machine Learning Research, pages 7553–7567.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James R. Glass, and Pengcheng He. 2024. Dola: Decoding by contrasting layers improves factuality in large language models. In *The Twelfth International Conference on Learning Representations*.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. 2021. A mathematical framework for transformer circuits. *Transformer Circuits Thread*. Https://transformer-circuits.pub/2021/framework/index.html.

Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. 2023. Dissecting recall of factual associations in auto-regressive language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12216–12235, Singapore. Association for Computational Linguistics.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2021. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5484–5495, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Trans. Inf. Syst.*, 43(2).

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12).

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada. Association for Computational Linguistics.

Jannik Kossen, Jiatong Han, Muhammed Razzak, Lisa Schut, Shreshth A. Malik, and Yarin Gal. 2024. Semantic entropy probes: Robust and cheap hallucination detection in llms. *ArXiv*, abs/2406.15927.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. *arXiv preprint arXiv:2302.09664*.

Junyi Li, Xiaoxue Cheng, Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023. HaluEval: A large-scale hallucination evaluation benchmark for large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6449–6464, Singapore. Association for Computational Linguistics.

J. Lin. 1991. Divergence measures based on the shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151.

Andrey Malinin and Mark Gales. 2020. Uncertainty Estimation in Autoregressive Structured Prediction. In *International Conference on Learning Representations*.

Potsawee Manakul, Adian Liusie, and Mark Gales. 2023. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9004–9017, Singapore. Association for Computational Linguistics.

Hadas Orgad, Michael Toker, Zorik Gekhman, Roi Reichart, Idan Szpektor, Hadas Kotek, and Yonatan Belinkov. 2024. Llms know more than they show: On the intrinsic representation of llm hallucinations. *Preprint*, arXiv:2410.02707.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Jie Ren, Jiaming Luo, Yao Zhao, Kundan Krishna, Mohammad Saleh, Balaji Lakshminarayanan, and Peter J. Liu. 2022. Out-of-Distribution Detection and Selective Generation for Conditional Language Models. In *The Eleventh International Conference on Learning Representations*.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed Chi, Nathanael Schärli, and Denny Zhou. 2023. Large language models can be easily distracted by irrelevant context. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Gaurang Sriramanan, Siddhant Bharti, Vinu Sankar Sadasivan, Shoumik Saha, Priyatham Kattakinda, and Soheil Feizi. 2024. LLM-Check: Investigating Detection of Hallucinations in Large Language Models. In *Advances in Neural Information Processing Systems*, volume 37, pages 34188–34216. Curran Associates, Inc.

Alessandro Stolfo, Yonatan Belinkov, and Mrinmaya Sachan. 2023. A mechanistic interpretation of arithmetic reasoning in language models using causal mediation analysis. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7035–7052, Singapore. Association for Computational Linguistics.

ZhongXiang Sun, Xiaoxue Zang, Kai Zheng, Jun Xu, Xiao Zhang, Weijie Yu, Yang Song, and Han Li. 2024. ReDeEP: Detecting Hallucination in Retrieval-Augmented Generation via Mechanistic Interpretability. In *The Thirteenth International Conference on Learning Representations*.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, Anton Tsitsulin, Nino Vieillard, Piotr Stanczyk, Sertan Girgin, Nikola Momchev, Matt Hoffman, Shantanu Thakoor, Jean-Bastien Grill, Behnam Neyshabur, Olivier Bachem, Alanna Walton, Aliaksei Severyn, Alicia Parrish, Aliya Ahmad, Allen Hutchison, Alvin Abdagic, Amanda Carl, Amy Shen, Andy Brock, Andy Coenen, Anthony Laforge, Antonia Paterson, Ben Bastian, Bilal Piot, Bo Wu, Brandon Royal, Charlie Chen, Chintu Kumar, Chris Perry, Chris Welty, Christopher A. Choquette-Choo, Danila Sinopalnikov, David Weinberger, Dimple Vijaykumar, Dominika Rogozińska, Dustin Herbison, Elisa Bandy, Emma Wang, Eric Noland, Erica Moreira, Evan Senter, Evgenii Eltyshev, Francesco Visin, Gabriel Rasskin, Gary Wei, Glenn Cameron, Gus Martins, Hadi Hashemi, Hanna Klimczak-Plucińska, Harleen Batra, Harsh Dhand, Ivan Nardini, Jacinda Mein, Jack Zhou, James Svensson, Jeff Stanway, Jetha Chan, Jin Peng Zhou, Joana Carrasqueira, Joana Iljazi, Jocelyn Becker, Joe Fernandez, Joost van Amersfoort, Josh Gordon, Josh Lipschultz, Josh Newlan, Ju yeong Ji, Kareem Mohamed, Kartikeya Badola, Kat Black, Katie Millican, Keelin McDonell, Kelvin Nguyen, Kiranbir Sodhia, Kish Greene, Lars Lowe Sjoesund, Lauren Usui, Laurent Sifre, Lena Heuermann, Leticia Lago, Lilly McNealus, Livio Baldini Soares, Logan Kilpatrick, Lucas Dixon, Luciano Martins, Machel Reid, Manvinder Singh, Mark Iverson, Martin Görner, Mat Velloso, Mateo Wirth, Matt Davidow, Matt Miller, Matthew Rahtz, Matthew Watson, Meg Risdal, Mehran Kazemi, Michael Moynihan, Ming Zhang, Minsuk Kahng, Minwoo Park, Mofi Rahman, Mohit Khatwani, Natalie Dao, Nenshad Bardoliwalla, Nesh Devanathan, Neta Dumai, Nilay Chauhan, Oscar Wahltinez, Pankil Botarda, Parker Barnes, Paul Barham, Paul Michel, Pengchong Jin, Petko Georgiev, Phil Culliton, Pradeep Kuppala, Ramona Comanescu, Ramona Merhej, Reena Jana, Reza Ardeshir Rokni, Rishabh Agarwal, Ryan Mullins, Samaneh Saadat, Sara Mc Carthy, Sarah Perrin, Sébastien M. R. Arnold, Sebastian Krause, Shengyang Dai, Shruti Garg, Shruti Sheth, Sue Ronstrom, Susan Chan, Timothy Jordan, Ting Yu, Tom Eccles, Tom Hennigan, Tomas Kocisky, Tulsee Doshi, Vihan Jain, Vikas Yadav, Vilobh Meshram, Vishal Dharmadhikari, Warren Barkley, Wei Wei, Wenming Ye, Woohyun Han, Woosuk Kwon, Xiang Xu, Zhe Shen, Zhitao Gong, Zichuan Wei, Victor Cotruta, Phoebe Kirk, Anand Rao, Minh Giang, Ludovic Peran, Tris Warkentin, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, D. Sculley, Jeanine Banks, Anca Dragan, Slav Petrov, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Sebastian Borgeaud, Noah Fiedel, Armand Joulin, Kathleen Kenealy, Robert Dadashi, and Alek Andreev. 2024. Gemma 2: Improving open language models at a practical size. *Preprint*, arXiv:2408.00118.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.

Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. 2023. Alignscore: Evaluating factual consistency with a unified alignment function. *arXiv preprint arXiv:2305.16739*.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, et al. 2023. Siren's song in the ai ocean: a survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*.

A Calculation Details and Explanation of ICR Score

ICR Score is computed by Equation 7, where JSD quantifies the similarity between two distributions, and possesses the property of being symmetric and bounded.

A.1 Introduction to JSD

The Jensen-Shannon Divergence (JSD) (Lin, 1991) is calculated using the following formula:

$$\text{JSD}(P||Q) = \frac{1}{2}\left(D_{\text{KL}}(P||M) + D_{\text{KL}}(Q||M)\right)$$

where:

- P and Q are two probability distributions.
- $M = \frac{1}{2}(P + Q)$ is the average of the two distributions.

- $D_{\mathrm{KL}}(P||M)$ and $D_{\mathrm{KL}}(Q||M)$ are the Kullback-Leibler (KL) divergences between P and M, and Q and M, respectively.

The KL divergence is defined as:

$$D_{\mathrm{KL}}(P||Q) = \sum_i P(i) \log\left(\frac{P(i)}{Q(i)}\right)$$

where $P(i)$ and $Q(i)$ represent the probabilities of the i-th event in the distributions P and Q.

JSD quantifies the similarity between two probability distributions P and Q. It is symmetric and produces a value between 0 and 1. A JSD of 0 indicates identical distributions, while a value close to 1 indicates significant difference. The average distribution M is used to ensure symmetry and handle differing supports between P and Q.

Role and Impact of ICR Score The ICR Score quantifies the relative contribution of attention and FFN modules to the hidden state update in each layer. A small ICR score suggests the attention mechanism primarily drives the update, while a large score indicates the FFN plays a dominant role. By measuring these contributions, the ICR score helps reveal how attention and FFN modules influence the residual stream, offering insights into the model's internal dynamics and the distribution of responsibility between the two components.

A.2 ICR Score and Probe Workflow

We summarize the procedures for computing the ICR Score and deploying the ICR Probe; see Algorithm 1 and Algorithm 2, respectively.

B Details about Experiment

Computing Infrastructure Our experiments were conducted on a server equipped with 10 NVIDIA GeForce RTX 3090 GPUs (24 GB memory each), running CUDA 12.0 and Ubuntu 20.04.5 LTS. Across multiple rounds of experiments, the total computational budget amounted to 600-800 GPU hours.

B.1 Details about Datasets

For our experiments, we randomly sampled 10,000 instances from each dataset. Specifically, we used the QA subset of the HaluEval dataset and the rc.nocontext subset of TriviaQA. The datasets used in this study are publicly available and adhere to their respective licenses, which allow for academic research and non-commercial use. Each dataset is split into 80%-20% for training and testing. We train and test each dataset, reporting the corresponding results. **This experimental setup for baseline methods matches ours exactly.**

B.2 Baselines Methods

This section provides a brief overview of several baseline methods for hallucination detection, including metrics like PPL and LN-Entropy, as well as techniques such as LLM-check, SAPLMA, and SEP.

PPL: PPL assesses the likelihood of hallucinations in LLM-generated responses by calculating perplexity. A higher perplexity value reflects increased uncertainty in the model's output, as hallucinations are often associated with the model's lack of confidence or inaccurate knowledge.

Length-Normalized Entropy (LN-Entropy): LN-Entropy is designed to quantify sequence-level uncertainty across multiple generations. This metric normalizes entropy concerning sequence length, facilitating a more precise assessment of the stability and reliability of generated content. A high value of the normalized entropy may serve as an indicator of the presence of hallucinations or other forms of unreliable output.

LLM-Check: LLM-check uses the method of attention mechanism kernel similarity analysis to conduct hallucination detection. They found that in large language models, for an input sequence of length m with a self-attention heads per layer, the kernel similarity map Ker_i of each head is a lower-triangular (mm) matrix. By calculating $logdet(Ker_i) = \sum_{j=1}^{m} logKer_i^{jj}$ and aggregating the mean log - determinants of all heads to obtain the "Attention Score", it can effectively capture the characteristics related to hallucinations. This method is computationally efficient and performs well in hallucination detection across different datasets and settings. In black-box settings, an auxiliary LLM with teacher-forcing is used to acquire the scores for detection.

SAPLMA: SAPLMA trains a classifier to detect hallucinations using the hidden states of specific layers in LLMs. It captures internal signals from these layers, which the classifier uses to identify when the model is likely to generate hallucinated content.

Algorithm 1 ICR Score Computation

Require: Hidden states $\{x_i^{\ell-1}, x_i^{\ell}\}_{i=1}^{N}$ for $\ell = 1, \ldots, L$, attention logits $\{\mathrm{Attn}_i^{\ell,h}\}_{h=1}^{H}$, top-$k$
Ensure: ICR scores $\{\mathrm{ICR}_i^{\ell}\}$ for all tokens and layers

for $\ell = 1$ **to** L **do**
  for $i = 1$ **to** N **do**
    $a_i^{\ell} \leftarrow \mathrm{softmax}\Big(\frac{1}{H}\sum_{h=1}^{H} \mathrm{Attn}_i^{\ell,h}\Big)$ ▷ normalize multi-head attention
    $\Delta x_i^{\ell} \leftarrow x_i^{\ell} - x_i^{\ell-1}$ ▷ token update vector
    for $j = 1$ **to** N **do**
      $p_{i,j}^{\ell} \leftarrow \dfrac{(\Delta x_i^{\ell})^{\top} x_j^{\ell}}{\|x_j^{\ell}\|}$
    end for
    $\mathrm{proj}_i^{\ell} \leftarrow \mathrm{softmax}\big(\{p_{i,j}^{\ell}\}_{j=1}^{N}\big)$ ▷ projection distribution
    $S \leftarrow \mathrm{TopIndices}(a_i^{\ell}, k)$ ▷ indices of top-k attention weights
    $\mathrm{ICR}_i^{\ell} \leftarrow \mathrm{JSD}\big(\mathrm{proj}_i^{\ell}[S] \,\|\, a_i^{\ell}[S]\big)$
  end for
end for

SEP (Semantic Entropy Probe): SEP leverages linear probes trained on the hidden states of large language models. It detects hallucinations by analyzing the semantic entropy of tokens before generation. This approach assumes that the semantic entropy of tokens can provide insights into whether the subsequently generated content is hallucinatory.

B.3 Ablation Study on the Number of Selected Tokens (k)

To validate our choice of $k = 20$ when computing the ICR Score, we conduct an ablation study on the Gemma-2 model across four benchmark datasets. Table 5 summarizes detection performance for $k \in \{5, 20, 30, \mathrm{ALL}\}$.

Table 5: Detection performance (higher is better) under different values of k.

k	HaluEval	SQuAD	HotpotQA	TriviaQA
5	0.8117	0.7830	0.8172	0.7926
20	**0.8436**	**0.8142**	**0.8409**	0.8001
30	0.8431	0.8005	0.8360	**0.8005**
ALL	0.8289	0.7853	0.8327	0.7918

Why not use all tokens? Using only the top-k tokens yields consistently better detection accuracy than aggregating over all tokens. For example, on HaluEval the ICR Score with $k = 20$ outperforms the "ALL" baseline by +1.47%. Intuitively, selecting the highest-attention tokens filters out noisy or irrelevant positions in the attention map, leading to more robust attribution.

Sensitivity to k. The performance peak at $k = 20$ is stable: increasing to $k = 30$ changes the results by less than 0.5% across all datasets. Conversely, too small a k (e.g. $k = 5$) under-samples the important tokens and degrades detection accuracy by over 3% on some benchmarks.

Based on these results, we fix $k = 20$ for all further experiments in Table 1. Full ablation across other LLMs will be provided in the revised Appendix due to space constraints.

B.4 ICR Probe Training

Input and Output The ICR Probe is a multilayer perceptron (MLP) classifier trained to predict hallucinations using the pooled ICR score. Given a model generation, the ICR matrix $\mathbb{R}^{N \times L}$ is token-wise pooled to obtain a $1 \times L$ vector, which serves as the input to the probe. The label for each instance is derived from annotated hallucination datasets.

Model Architecture The probe consists of four fully connected layers, with batch normalization and dropout ($p = 0.3$) applied after each hidden layer. To strike a balance between performance and efficiency, we conducted preliminary experiments to select an appropriate probe structure. The results are shown in Table 6, taking Gemma - 2 and TriviaQA as examples. When the probe has four hidden layers, an optimal balance between parameter efficiency and performance is achieved.

The architecture is as follows:

$$L \to 128 \to 64 \to 32 \to 1$$

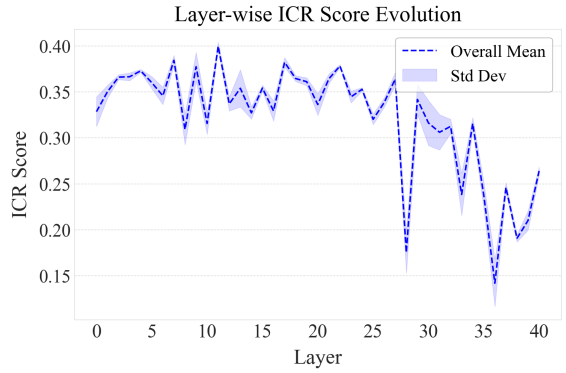
Algorithm 2 ICR Probe: Training and Inference

Require: LLM with L transformer layers, training corpus $\mathcal{D} = \{(q, a, y)\}$ with question q, generated answer a, label $y \in \{0, 1\}$ (*0: faithful, 1: hallucinated*); top-k for ICR Score; learning rate η

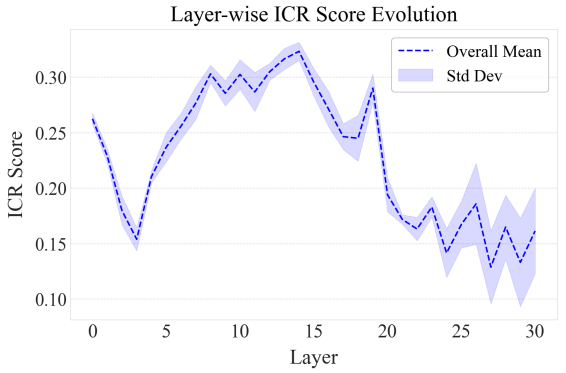
Ensure: Trained probe parameters θ and inference routine

procedure TRAIN($\mathcal{D}$)
  Initialize probe weights θ (logistic regression)
  for all $(q, a, y) \in \mathcal{D}$ **do**
    Run LLM once on (q, a), cache hidden states $\{x_i^\ell\}$ and attentions $\{\text{Attn}_i^{\ell,h}\}$
    for all answer tokens i **and** layers $\ell = 1..L$ **do**
      Compute ICR_i^ℓ by Algorithm 1 ▷ single forward-pass
    end for
    $f \leftarrow \left[\underset{i}{\text{mean}}\left(\text{ICR}_i^\ell\right) \right]_{\ell=1}^{L}$ ▷ layer-wise mean pool
    $\hat{y} \leftarrow \sigma(\theta^\top f)$ ▷ sigmoid
    $\theta \leftarrow \theta - \eta\,(\hat{y} - y)\,f$ ▷ SGD update
  end for
  return θ
end procedure

procedure INFER(q, a, θ)
  Run LLM on (q, a), cache $\{x_i^\ell\}, \{\text{Attn}_i^{\ell,h}\}$
  for all answer tokens i **and** layers ℓ **do**
    $\text{ICR}_i^\ell \leftarrow$ Algorithm 1
  end for
  $f \leftarrow [\text{mean}_i(\text{ICR}_i^\ell)]_{\ell=1}^{L}$
  $\hat{y} \leftarrow \sigma(\theta^\top f)$
  return $\hat{y}$ ▷ hallucination probability
end procedure



(a) Gemma-2: Layer-wise AUROC for hallucination detection **using direct ICR scores**, peaking at layer **21**.



(b) Llama-3: Layer-wise AUROC for hallucination detection **using direct ICR scores**, peaking at layer **10**.

Figure 5: Layer-wise mean ICR scores.

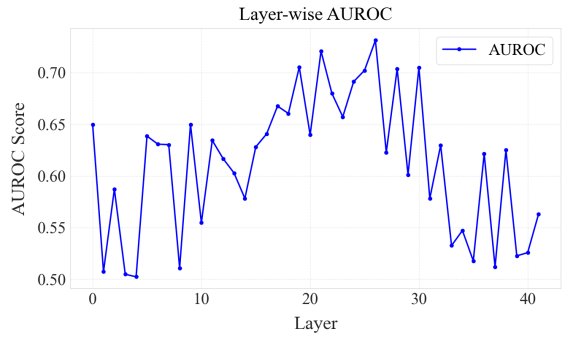
Hidden layers	**1**	**2**	**3**	**4**	**5**
Performance	0.6410	0.7398	0.7889	0.8000	0.8006

Table 6: Comparison results of performance of different probe structures.

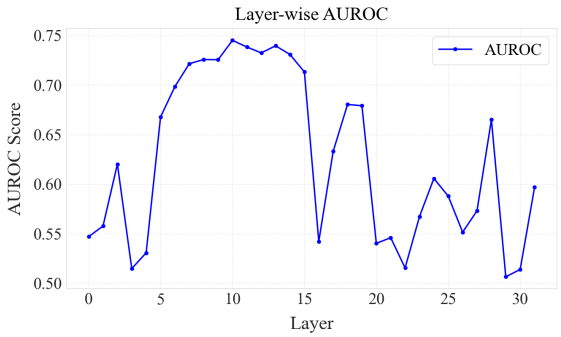
Each hidden layer employs the **Leaky ReLU** activation function ($\alpha = 0.01$), and the final output is passed through a **sigmoid** activation to obtain a probability score.

Details of Training Probe

- The model is trained using the binary cross-entropy loss with the Adam optimizer (lr $= 5 \times 10^{-4}$). We employ learning rate schedul-



(a) Gemma-2: Layer-wise AUROC for hallucination detection **using direct ICR scores**, peaking at layer **21**.



(b) Llama-3: Layer-wise AUROC for hallucination detection **using direct ICR scores**, peaking at layer **10**.

Figure 6: Layer-wise AUROC.

LLM	HaluEval	SQuAD	HotpotQA	TriviaQA
Gemma-2	0.7850	0.7321	0.7465	0.7883
Qwen2.5	0.7592	0.6903	0.7345	0.7582
Llama-3	0.7458	0.7331	0.7340	0.7206

Table 7: AUROC of the Semantic Entropy baseline.

ing using `ReduceLROnPlateau`), where the learning rate is reduced by a factor of 0.5 if the validation loss does not improve for 5 consecutive epochs.

- We use Kaiming initialization for linear layers and set batch normalization parameters to 1 (scale) and 0 (bias).
- Dropout ($p = 0.3$) is applied after each hidden layer to prevent overfitting.
- The model is trained for 50 epochs with a batch size of 32.
- To obtain stable and reliable results, we perform multiple runs and take the average.

B.5 Ablation Study Settings

In this appendix, we provide details on the three experimental settings used in the ablation study, including the calculation of ICR Score in each setting and the corresponding meanings.

1. NONE Setting In the **NONE** setting, both the hidden state (Proj_i^ℓ) and attention (Attn_i^ℓ) signals are excluded. This results in a random performance scenario, where the AUROC is equal to 0.5, indicating that the model's performance is no better than random guessing. In this setting, the ICR Score is not computed because no useful information is provided by either the hidden states or attention scores. This setup serves as a baseline to assess the importance of both components in hallucination detection.

2. HS ONLY Setting In the **HS ONLY** setting, only the hidden state information (Proj_i^ℓ) is used, while the attention scores (Attn_i^ℓ) are set to a uniform distribution. This effectively reduces the ICR Score to the entropy of the hidden state update directions. Specifically, the ICR Score in this setting is calculated as:

$$\text{ICR}^\ell_{i,\text{HS ONLY}} = \text{JSD}(\text{Proj}_i^\ell, \mathbf{p}_{\text{uniform}})$$

where $\mathbf{p}_{\text{uniform}}$ is a vector representing a uniform distribution, and JSD stands for the Jensen-Shannon Divergence.

JSD measures the similarity between two distributions. When we compare the hidden state updates (Proj_i^ℓ) to a uniform distribution, it is equivalent to measuring the entropy of the hidden state updates. This is because the entropy represents the amount of uncertainty or disorder within a distribution, and the uniform distribution can be considered as the maximum entropy distribution in the absence of any other information. Therefore, the JSD between the hidden state updates and a uniform distribution is a measure of how "disordered" or "uncertain" the update directions are, and this is equivalent to computing the entropy of Proj_i^ℓ.

Model	Method	HaluEval	SQuAD	HotpotQA	TriviaQA
Qwen2.5–14B	SAPLMA	0.7720	0.6984	0.7214	0.7759
	SEP	0.7016	0.6653	0.6874	0.7103
	ICR Probe	**0.8021**	**0.7632**	**0.7751**	**0.7846**
Qwen2.5–3B	SAPLMA	0.7538	0.6915	0.7747	0.7523
	SEP	0.6689	0.6560	0.6729	0.6945
	ICR Probe	**0.7917**	**0.7784**	**0.7905**	**0.7631**

Table 8: Performance of the ICR Probe on Qwen2.5 models of different scales.

3. HS + ATTN Setting In the **HS + ATTN** setting, both the hidden state (Proj_i^ℓ) and attention signals (Attn_i^ℓ) are integrated to compute the full ICR Score. The calculation of ICR Score in this setting is:

$$\text{ICR}^\ell_{i,\text{HS + ATTN}} = \text{JSD}(\text{Proj}_i^\ell, \text{Attn}_i^\ell)$$

Summary of Results The results from these three settings show that the hidden state alone (Proj_i^ℓ) is already sufficient for effective hallucination detection, as evidenced by the AUROC being significantly above 0.5. However, integrating both hidden state and attention signals leads to further improvement in detection performance, highlighting the complementary nature of these components. The entropy of the hidden state update directions captures meaningful information about the model's updates, while the inclusion of attention signals helps refine this information.

C Additional Experimental Results

C.1 Empirical Study Results

In this section, we provide further empirical analysis supporting the findings and conclusions presented in Section 4.1. Figure 5 illustrates the mean ICR scores, accompanied by standard deviation bands, across four datasets for both Llama-3 and Gemma-2. Narrow standard deviation bands across all models indicate the stability and consistency of the ICR scores across datasets.

Figure 6 presents the layer-wise AUROC for hallucination detection using direct ICR scores on Llama-3 and Gemma-2. Notably, the highest AUROC on Llama-3 occurs at layer 10 (0.7451), while on Gemma-2, the highest AUROC is observed at layer 21 (0.7313). These results further validate the discriminative power of the ICR approach for hallucination detection.

C.2 Additional Baseline

To provide a broader context, we report the cross-dataset AUROC of the Semantic Entropy baseline (Kuhn et al., 2023) alongside our ICR Probe. Note that Semantic Entropy relies on multiple generations and semantic clustering, whereas ICR Probe requires only a single forward pass, making the comparison computationally imbalanced. Table 7 summarizes the Semantic Entropy performance.

Discussion We deliberately exclude entailment-based methods (e.g., AlignScore, LLM prompting (Zha et al., 2023)) because they operate under a fundamentally different paradigm. Entailment-based approaches require an explicit premise-hypothesis setup, verifying consistency between a reference text and generated output, whereas our hallucination detection operates directly on hidden state representations without external references, enabling real-time single-pass inference. Consequently, entailment methods cannot be applied in scenarios lacking a well-defined premise (as in many open-ended QA or generation tasks; see Table 4), making them structurally incompatible with our evaluation protocol and task objectives.

C.3 Performance Across Model Scales

To evaluate the ICR Probe across different model sizes, we include both a smaller (3B) and a larger (14B) variant of Qwen2.5. Table 8 reports the AUROC of ICR Probe alongside the strongest hidden-state baselines, SAPLMA and SEP.

Discussion. Across both additional scales, ICR PROBE remains the top performer on every dataset. Compared with SAPLMA, it gains on average **3.9 %** AUROC at 14B and **3.8 %** at 3B; against SEP the average margin widens to **9.0 %** and **10.8 %**, respectively. These results confirm that the probe's layer-wise residual-stream cues are

Qwen2.5		Test			
		HaluEval	SQuAD	HotpotQA	TriviaQA
Train	HaluEval	**0.8003**	0.6835	0.7589	0.6626
	SQuAD	0.7360	**0.7456**	0.7414	0.6900
	HotpotQA	0.7822	0.7028	**0.7917**	0.6387
	TriviaQA	0.7576	0.6962	0.7222	**0.7684**

Table 9: Cross-dataset generalization evaluation for Qwen2.5. Each cell shows the AUROC when the probe is trained on the row dataset and evaluated on the column dataset.

Llama-3		Test			
		HaluEval	SQuAD	HotpotQA	TriviaQA
Train	HaluEval	**0.7603**	0.7210	0.7547	0.6289
	SQuAD	0.7342	**0.7634**	0.7294	0.5658
	HotpotQA	0.7775	0.7347	**0.7982**	0.5933
	TriviaQA	0.6972	0.6129	0.6749	**0.7325**

Table 10: Cross-dataset generalization evaluation for Llama-3. Each cell shows the AUROC when the probe is trained on the row dataset and evaluated on the column dataset.

largely scale-invariant, supporting its applicability to a broad spectrum of model sizes.

C.4 Generalization

In Section 5.3, we demonstrate the generalization ability of the ICR Probe on Gemma-2. Here, we provide additional generalization results. Tables 10 and 9 present the cross-dataset generalization evaluation results for Llama-3 and Gemma-2, respectively. The AUROC values for both models exceed 0.6, highlighting the strong generalization capability of our approach.