LAFaCT: Attribution-based Localization and Focused Sequential Analysis of Fact-Critical Tokens for Hallucination Detection

Xin Wang **Jiahao Li** **Licheng Zhang** **Zhendong Mao***
University of Science and Technology of China, Hefei, China
{sa24221049, jiahao66, zlczlc}@mail.ustc.edu.cn
zdmao@ustc.edu.cn

Abstract

Large Language Models (LLMs) suffer from hallucinations, severely undermining their reliability. While white-box hallucination detection methods that leverage hidden states prevail, they fail to identify and focus on fact-critical information when analyzing token sequences. To address this, we propose LAFaCT, a Localize-then-Analyze detection framework. It first localizes fact-critical tokens using Factual Criticality, a novel metric derived from feature attribution. A subsequent stage then performs a focused sequential analysis on their hidden states. Extensive experiments on eight benchmarks and multiple model families confirm LAFaCT as the new state-of-the-art, with in-depth analyses validating the effectiveness of its core token-localization strategy.

1 Introduction

The rapid development of Large Language Models (LLMs) is profoundly reshaping the field of Natural Language Processing (Achiam et al., 2023; Dubey et al., 2024). However, the tendency of even state-of-the-art LLMs to "hallucinate" by generating factually incorrect or fabricated content severely undermines their reliability, posing a fundamental challenge to their deployment in high-stakes applications (Li et al., 2024; Zhang et al., 2025). Therefore, to transform LLMs from promising novelties into reliable applications, robust methods for hallucination detection are fundamental prerequisites (Chen et al., 2023; Min et al., 2023).

Mainstream hallucination detection methods can be broadly categorized into three types (Fadeeva et al., 2023, 2024; Zhu et al., 2024). *Black-box* methods, which detect hallucinations by analyzing "self-consistency" between multiple sampled responses (Manakul et al., 2023; Kuhn et al., 2023), are often computationally expensive and fail when the model is consistently incorrect (Fadeeva et al., 2023; Orgad et al., 2025). *Grey-box* methods, which calculates a hallucination score by aggregating the probability or entropy of generated tokens (Fomicheva et al., 2020; Duan et al., 2024; Fadeeva et al., 2024), are frequently undermined by the inherent overconfidence of LLMs (Yeh et al., 2024; Zhang et al., 2023). In contrast, *White-box* methods directly probe the model's internal hidden states, which are widely considered to encode more direct and fundamental factuality signals than final output probabilities (Burns et al., 2022; Azaria and Mitchell, 2023; Li et al., 2023). Building on this advantage, a series of studies have positioned white-box analysis as the most promising detection type (Chen et al., 2024; He et al., 2024; Du et al., 2024; Park et al., 2025). Given this promise, this paper focuses on the white-box paradigm.

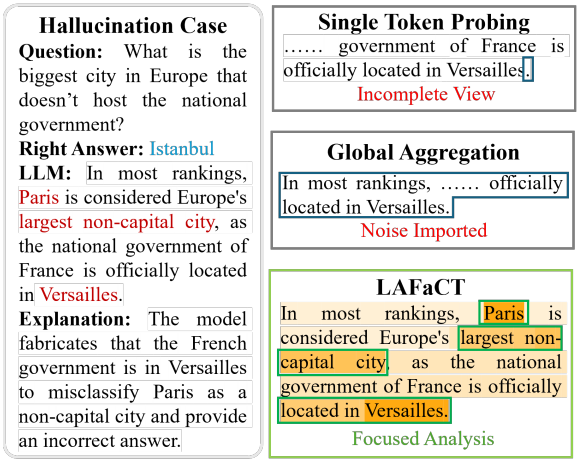

Figure 1: The dilemma in white-box hallucination detection methods: *single-token probing* is limited by an incomplete view, while *global aggregation* suffers from imported noise. LAFaCT overcomes both issues with a Localize-then-Analyze strategy, performing a focused analysis on only the most fact-critical tokens.

As illustrated in Figure 1, prevailing white-box methods fail to identify and focus on fact-critical information when analyzing token sequences, lead-

* Corresponding author.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 6877–6896
July 2-7, 2026

ing to suboptimal performance. Specifically, these methods can be divided into two categories: 1) *Single Token Probing* methods (Azaria and Mitchell, 2023; Yin et al., 2024; Yüksekgönül et al., 2024) only analyze the hidden state of one fixed token (e.g., the last token), struggling to capture factuality signals distributed across the entire sequence and thus yielding an incomplete view. 2) *Global Aggregation* methods (Li et al., 2025; Su et al., 2024; Zhu et al., 2024; Vazhentsev et al., 2025) indiscriminately aggregate all hidden states, which introduces noise from fact-irrelevant tokens and dilutes the critical factual signals. Therefore, we posit that the key to resolving this dilemma lies in performing a focused analysis on the hidden states of only the fact-critical tokens.

Guided by this insight, we propose LAFaCT, a novel hallucination detection framework guided by a Localize-then-Analyze strategy. The initial Localize stage identifies a small subset of the most fact-critical tokens. To achieve this, we first train a proxy classifier on the final hidden state to output a proxy factuality probability. This probability is then attributed back to each token's embedding using DeepLIFT (Shrikumar et al., 2017), yielding Factual Criticality scores to guide a Top-p selection of critical tokens. The subsequent Analyze stage performs a focused sequential analysis on the hidden states of the localized critical tokens to derive a final hallucination score, supervised by a novel Angular Triplet Loss. By combining attribution-based localization with focused sequential analysis, our framework avoids the incomplete view of single-token probing and the noise introduction from global aggregation, achieving more focused and effective detection.

LAFaCT establishes a new state-of-the-art in hallucination detection across eight diverse benchmarks on multiple model families. For instance, on Llama-2-chat-hf-7B, it achieves an average improvement of 3.2 AUROC points over the strongest baseline. Notably, LAFaCT not only shows a pronounced advantage on benchmarks requiring complex reasoning (e.g., GSM8K) but also exhibits strong generalization in out-of-domain (OOD) scenarios. Furthermore, extensive ablation studies validate the effectiveness of its core token-localization strategy and other design choices.

Our main contributions are as follows:

- We propose LAFaCT, a novel "Localize-then-Analyze" hallucination detection framework: it addresses existing white-box methods' dilemma of unfocused analysis by applying focused sequential analysis *only* to hidden states of the localized fact-critical tokens.
- We introduce Factual Criticality, a novel attribution-based metric for fact-critical token localization, which quantifies each token's factual contribution via attribution on the proxy classifier's predictions.
- LAFaCT achieves a new state-of-the-art across eight benchmarks with strong generalization across model scales and in OOD scenarios. In-depth analyses validate the effectiveness of its core token-localization strategy.

2 Related Work

Hallucination detection methods can be broadly categorized into three classes by their level of access to the target model's internals.

Black-box methods detect hallucinations using only the output text from a target model. Early works operationalized this by measuring consistency across several sampled responses using metrics such as NLI and lexical overlap (Manakul et al., 2023; Lin et al., 2024; Mündler et al., 2023; Cheng et al., 2024). Subsequent research has explored deeper forms of consistency, including semantic clustering (Kuhn et al., 2023), probabilistic belief trees (Hou et al., 2024), and interrogative verification (Yehuda et al., 2024). However, these methods are inefficient due to their reliance on multiple inferences and fail against consistently wrong samples, whereas our approach operates on a single generation's internal states for higher efficiency. (Fadeeva et al., 2023; Orgad et al., 2025)

Grey-box methods leverage the model's output probability distributions to signal uncertainty. Early approaches aggregating token-level likelihoods or entropy scores often undermined by model overconfidence and their failure to distinguish token importance (Fomicheva et al., 2020; Guerreiro et al., 2023; Malinin and Gales, 2020). More advanced approaches refine this by assigning higher weights to key semantic tokens (Duan et al., 2024; Zhang et al., 2023) or by attempting to disentangle factual from expressive uncertainty (Fadeeva et al., 2024). However, they ignore the richer factuality signals encoded in the model's hidden states (Burns et al., 2022; Li et al., 2023).

White-box methods, in contrast, directly analyze the model's internal hidden states, which are considered to encode more fundamental factuality signals (Burns et al., 2022; Li et al., 2023). This field has evolved rapidly from early approaches that trained simple classifiers on features from fixed token positions (Azaria and Mitchell, 2023; Snyder et al., 2024), to more robust techniques that fuse diverse internal signals like multi-layer activations and attention scores (He et al., 2024; Vazhentsev et al., 2025). More advanced research has begun to probe the deeper properties of these states, analyzing their distributional, geometric, and temporal dynamics (Chen et al., 2024; Yin et al., 2024; Zhu et al., 2024), and even actively intervening in the representation space using methods like steering vectors (Park et al., 2025). However, these methods tend to analyze hidden states from the entire sequence indiscriminately, failing to prioritize those originating from fact-critical tokens. Our work addresses this limitation by employing focused sequential analysis on localized fact-critical tokens.

3 LAFaCT

This section details our proposed framework, LAFaCT, which addresses the issue that prior methods fail to identify and focus on fact-critical information when analyzing hidden states across token sequences. As illustrated in Figure 2, LAFaCT combines attribution-based fact-critical token localization with focused sequential analysis, enabling more focused and effective hallucination detection.

3.1 Preliminary

LLMs We consider a Transformer-based LLM. Given an input sequence $(x_0, \ldots, x_{n-1})$, the model autoregressively predicts the next token x_n. Internally, the model projects the input tokens into token embeddings $E \in \mathbb{R}^{n \times d}$, where d denotes the hidden dimension of the model. These embeddings are then processed by a stack of L Transformer layers: for each token in x_i, the l-th layer outputs a hidden state $h_i^{(l)} \in \mathbb{R}^d$. The probability distribution for the predicted next token x_n is computed using the final-layer hidden state of the last token x_{n-1} via the final decoding head: $P(x_n|x_{<n}) = \text{softmax}(W_o h_{n-1}^{(L)} + b_o)$, where $W_o \in \mathbb{R}^{V \times d}$ and $b_o \in \mathbb{R}^V$ are the parameters of the final decoding head, with V representing the model's vocabulary size.

Hallucination Detection We frame hallucination detection as a binary classification task focused on factuality. Given a sample consisting of a prompt P_{prompt} and the corresponding LLM-generated response $X_{response}$, we leverage the LLM's internal hidden states of the concatenated sequence $[P_{prompt}; X_{response}]$ to determine whether $X_{response}$ contains factual hallucinations.

3.2 Fact-Critical Token Localization

This stage aims to select fact-critical tokens in the model's response as a foundation for subsequent analysis. To achieve this, we first train a proxy factuality classifier, then introduce Factual Criticality derived from attributing the proxy classifier's output, and directly use it to guide the selection.

Proxy Classifier The proxy classifier, $\mathcal{C}_{\text{proxy}}$, is trained primarily to provide factual attribution signals, not serving as a standalone hallucination detector. Specifically, we feed the concatenated prompt-response sequence $[P_{\text{prompt}}; X_{\text{response}}]$ into the LLM and extract the hidden state $h_{n-1}^{(l)}$ corresponding to the last token from the l-th middle layer. This vector, proven to encode rich factuality signals (Azaria and Mitchell, 2023), serves as input to $\mathcal{C}_{\text{proxy}}$ to produce the factuality probability o_{p}. In practice, $\mathcal{C}_{\text{proxy}}$ is implemented as a two-layer MLP and trained with Binary Cross-Entropy loss. The factuality probability o_{p} output by the trained $\mathcal{C}_{\text{proxy}}$ can serve as the attribution signals for the following Factual Criticality.

Factual Criticality for Selection Building upon the proxy classifier, we introduce Factual Criticality, an attribution-based metric designed to localize fact-critical tokens by quantifying their individual factual contributions. This metric is grounded in the intuition that the latent "factuality signals"[1] captured by the proxy are encoded in specific tokens. Consequently, attributing the proxy's predictions back allows us to trace these signals to their source.

To implement this, we first define an end-to-end computation path, $\mathcal{F}$, that maps input token embeddings E to factuality probability o_{p} by passing them through the first l layers of the LLM and subsequently our proxy classifier $\mathcal{C}_{\text{proxy}}$. Next, we use the DeepLIFT algorithm to attribute o_{p} back to E, using a sequence of <pad> token embeddings as

[1]We define "factuality signals" as the discriminative latent features that the proxy classifier inevitably learns to rely on to distinguish between factual and hallucinated responses.

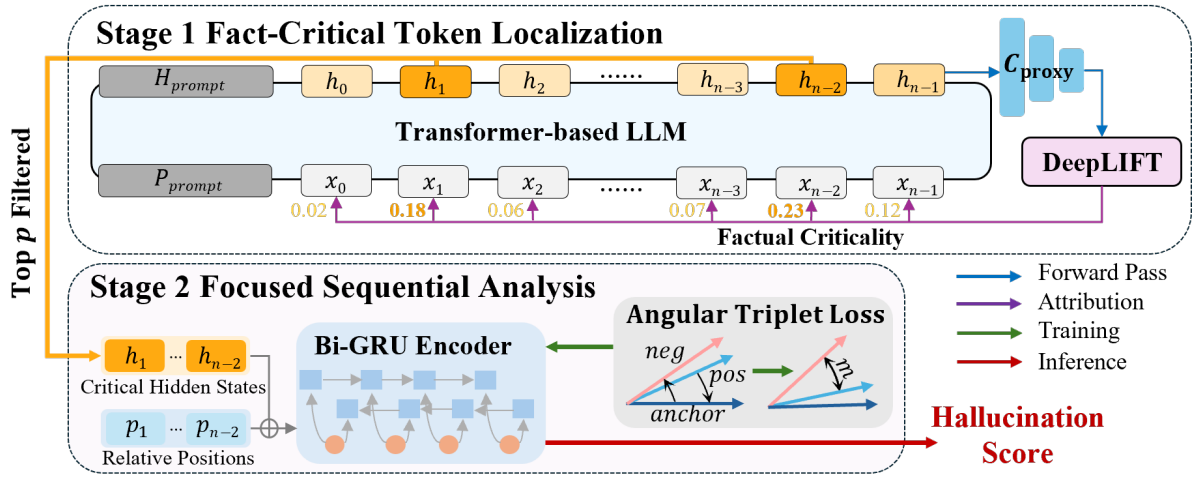

Figure 2: An overview of the LAFaCT framework. In the Localization stage, a proxy classifier C_{proxy} and DeepLIFT attribution are used to localize fact-critical tokens. In the Analysis stage, a Bi-GRU encoder trained by Angular Triplet Loss performs a focused sequential analysis on the hidden states of these tokens.

the baseline[2]. This provides an attribution score for each dimension of each token embedding in E. To derive the Factual Criticality score $C(x_i)$, we first compute each token's attribution score $S(x_i)$ as the L_2-norm of its dimension-level attributions, and then applying a softmax function to normalize these scores across all tokens in $X_{response}$:

$$S(x_i) = \left\| \text{DeepLIFT}_{\mathcal{F}}(o_{\text{p}}, E, E_{\text{pad}})_i \right\|_2 ,$$
$$C(x_i) = \frac{\exp(S(x_i))}{\sum_j \exp(S(x_j))}. \quad (1)$$

Finally, we select a subset of critical tokens from $X_{response}$ by applying a Top-p strategy to their $C(x_i)$ scores, and their corresponding hidden states at layer l are used for the subsequent analysis.

3.3 Focused Sequential Analysis

This stage performs a focused sequential analysis on the hidden states of the localized critical tokens to derive a final hallucination score. The process involves training a Bi-GRU encoder with our proposed Angular Triplet Loss to learn discriminative factuality representations.

Sequential Representation Modeling We use a Bi-GRU network (Chung et al., 2014) to aggregate fine-grained hidden states of critical tokens into a unified factuality representation vector. We maintain the original order of the hidden state sequence and enhance its positional awareness with a relative positional encoding, p_i, which is computed for each token x_i by applying sinusoidal functions (Vaswani et al., 2017) to its relative distance from the previous selected token. The feature vector v_i is obtained by projecting the concatenated hidden state $h_i^{(l)}$ and positional encoding p_i through an MLP: $v_i = \text{MLP}(h_i^{(l)} \oplus p_i)$. Finally, we process the feature sequence $V = (v_1, \ldots, v_k)$ with a Bi-GRU encoder (Chung et al., 2014) and concatenate its bi-directional outputs to form the representation $e_x = \overrightarrow{\text{GRU}}(V) \oplus \overleftarrow{\text{GRU}}(V)$.

Training with Angular Triplet Loss Unlike previous methods that directly learn a mapping from factuality representations to factuality labels via a classification loss, we propose the Angular Triplet Loss to learn discriminative factuality representations, which are known to exhibit complex distributions (Mishra et al., 2024; Wang et al., 2025). Specifically, we first normalize representations onto a unit hypersphere, then optimize their angular distribution via the Triplet learning paradigm (Schroff et al., 2015), which has been proven effective in metric learning (Deng et al., 2019; Kim et al., 2022). For each training sample (anchor), we form a triplet with a positive (same-class) and a negative (different-class) sample, and then compute the angles (θ_p and θ_n) between their respective representations and that of the anchor. Our Angular Triplet Loss is then calculated to guide the learning process of the Bi-GRU encoder:

$$\mathcal{L}_{\text{AT}} = \max\left(0, \cos(\theta_n) - \cos(\theta_p + m)\right),$$

[2]DeepLIFT's score is computed based on the difference between a neuron's actual activation and its "reference" activation, which is derived from a baseline input. An ablation study on attribution details is presented in Appendix F.1.

where $m > 0$ is a fixed angular margin used to create a stricter decision boundary. This objective enforces inter-class separation and intra-class compactness in the angular space, yielding a highly discriminative representation.

Inference We additionally introduce a nearest-neighbor strategy to complete the detection pipeline, with the full inference process summarized as follows. For each test sample, we first localize fact-critical tokens and extract their corresponding hidden states. These hidden states are fed into a Bi-GRU encoder to generate the final factuality representation, e_x. Next, we compute the average cosine similarity between e_x and its top-5 nearest neighbors from "Factual" and "Hallucination" samples in the training set, respectively. The test sample is ultimately classified into the class with the higher average similarity.

4 Experiments

4.1 Experimental Setting

Datasets We evaluate our method on a diverse collection of eight datasets, organized into three distinct task categories: (1) Question Answering (QA): This category serves as our primary testbed, encompassing five benchmarks across various domains: TruthfulQA (Lin et al., 2022), TriviaQA (Joshi et al., 2017), CoQA (Reddy et al., 2019), GSM8K (Cobbe et al., 2021), and MedQuad (Ben Abacha and Demner-Fushman, 2019); (2) Text Summarization: We employ XSum (Narayan et al., 2018) and FRANK (Pagnoni et al., 2021) for source-dependent generation tasks;; (3) Biography Generation: We utilize WikiBio (Manakul et al., 2023) to represent data-to-text scenarios. Comprehensive details regarding data statistics, ground-truth labels, and splitting protocols are provided in Appendix C.

Models Our evaluation includes four widely-adopted, open-source LLM families to ensure the broad applicability of our method: LLaMA-2-chat-hf (Touvron et al., 2023) (7B, 13B), Qwen-2.5-Instruct (Team, 2024) (8B, 14B), LLaMA-3-Instruct-8B (Grattafiori et al., 2024), and Mistral-Instruct-7B (Jiang et al., 2023).

Baselines We compare LAFaCT with advanced baselines, categorized as follows: For black-box methods, we use SelfCheckGPT (Mündler et al., 2023); for grey-box methods, we include Semantic Entropy (Kuhn et al., 2023) and SAR (Duan et al., 2024). Our main comparison focuses on white-box methods, which can be further categorized by their analysis strategy: (1) *Single-token probing*: LLM Factoscope (He et al., 2024) and HaloScope (Du et al., 2024), which analyze the last token in the prompt and response, respectively; (2) *Global aggregation*: EigenScore (Chen et al., 2024) and PoLLMgraph (Zhu et al., 2024), modeling all tokens' hidden states; (3) *Hybrid approach*: MIND (Su et al., 2024), which combines the average of all final-layer hidden states with the individual hidden state of the last token. Details of all baselines are provided in Appendix D.

Evaluation Protocol We use the Area Under the ROC Curve (AUROC), a standard metric in prior work (Chen et al., 2024; Du et al., 2024; Orgad et al., 2025; Park et al., 2025) for assessment.

Implementation Details Both the proxy classifier and the Bi-GRU encoder are trained on the same training set. Throughout the LAFaCT framework, we consistently use hidden states from the residual stream of the exact middle layer[3]. For more details, please refer to Appendix E.

4.2 Main Results

As presented in Table 1, LAFaCT establishes a new state-of-the-art in hallucination detection across all tested models[4]. On Llama2, LAFaCT's average performance surpasses the strongest baseline by 3.2 AUROC points, and this leading trend extends to Llama3 and Qwen2.5 with advantages of 2.6 and 2.1 points, respectively. Notably, LAFaCT's advantage is most pronounced on the complex reasoning benchmarks GSM8K and MedQuad (average response lengths exceeding 100 tokens), where it leads the strongest baseline on Llama2 by 4.2 and 4.5 points. For short-answer datasets (TriviaQA, CoQA) with few-token responses, single-token probing methods like Factoscope demonstrate competitive performance. Even in this scenario where LAFaCT's core strength of fact-critical token localization is weakened, LAFaCT still outperforms almost all methods, showing its generalizability. Furthermore, as we show in subsection 4.3, LAFaCT also achieves superior performance on open-ended generation tasks including text summarization and biography fact-checking, demonstrating its robustness across diverse task formats. Ad-

[3]We validate the effectiveness of this selection via layer sensitivity analysis in Appendix F.2

[4]Please refer to Appendix A for results on Mistral.

Model	Type	Method	CoQA	TriQA	TQA	GSM8K	MedQ	Average
Llama2	Black-box	SelfCheck	72.19	72.14	54.15	56.36	60.12	62.99
	Grey-box	SEntropy	69.14	73.97	57.44	62.59	61.57	64.94
		SAR	69.55	80.58	62.06	63.13	63.31	67.73
	White-box	EigenScore	72.87	73.98	60.93	55.73	62.85	65.27
		PoLLMgraph	81.91	81.05	84.04	77.37	75.79	80.03
		Mind	88.05	84.12	84.27	82.86	77.02	83.16
		HaloScope*	76.56	75.47	77.32	68.37	70.89	73.71
		HaloScope	87.59	83.13	86.30	80.21	76.57	82.86
		Factoscope	83.42	**85.89**	82.37	77.65	72.72	80.41
		LAFaCT	**89.11**	85.76	**88.23**	**87.07**	**81.51**	**86.34**
Llama3	Black-box	SelfCheck	71.07	73.51	59.63	57.33	64.70	65.25
	Grey-box	SEntropy	74.21	74.23	61.79	64.91	63.40	67.71
		SAR	71.18	81.54	62.54	65.17	62.60	68.61
	White-box	EigenScore	72.08	79.72	54.74	58.96	61.48	65.40
		PoLLMgraph	82.36	84.43	83.28	76.22	75.31	80.32
		Mind	88.84	84.25	85.22	86.47	77.87	84.53
		HaloScope*	76.34	76.28	78.22	66.54	68.71	73.22
		HaloScope	88.62	83.62	85.32	84.14	76.30	83.60
		Factoscope	84.24	**86.27**	81.41	79.26	71.89	80.61
		LAFaCT	**89.93**	86.08	**88.50**	**89.25**	**81.86**	**87.12**
Qwen2.5	Black-box	SelfCheck	67.14	71.21	53.81	56.53	61.90	62.12
	Grey-box	SEntropy	71.65	73.03	57.67	63.18	61.35	65.38
		SAR	73.26	78.09	60.22	64.41	62.99	67.80
	White-box	EigenScore	65.53	76.16	57.36	57.63	62.12	63.76
		PoLLMgraph	82.67	84.11	81.59	77.57	76.56	80.50
		Mind	91.52	88.44	81.49	85.87	76.33	84.73
		HaloScope*	77.56	78.87	72.09	71.76	66.45	73.35
		HaloScope	91.27	86.39	83.38	84.94	76.35	84.47
		Factoscope	86.63	87.57	77.58	76.58	72.25	80.72
		LAFaCT	**92.05**	**88.81**	**84.81**	**88.20**	**80.45**	**86.86**

Table 1: Hallucination detection performance (AUROC, %) on Llama2-chat-hf-7B, Llama3-8B-Instruct, and Qwen2.5-8B-Instruct. Best and second-best scores are in **bold** and underlined. * indicates the semi-supervised variant.
Abbreviations: TQA (TruthfulQA), TriQA (TriviaQA), MedQ (MedQuad), SEntropy (Semetric Entropy).

ditionally, LAFaCT consistently outperforms concurrent 2025 methods including SATMD (Vazhentsev et al., 2025) and LapEigvals (Binkowski et al., 2025) (see Appendix B).

To validate LAFaCT's generalization across different model scales, we conducted experiments on the Llama2-chat-hf and Qwen2.5-Instruct model families[5]. As shown in Figure 3, all methods exhibit improved performance with increasing LLM scale. Even so, LAFaCT maintains its leading position, achieving an average AUROC lead of 1.8 and 2.8 points over the second-best baseline on Llama2-13B and Qwen2.5-14B, respectively.

Out-of-domain Generalization To test generalization against distribution shifts, which is a challenge in real-world applications, we conduct a leave-one-out out-of-domain (OOD) analysis (Vazhentsev et al., 2025) on Llama2-chat-hf-7B. We compare LAFaCT against baselines grouped by their supervision requirements, with results shown in Table 2. We can see that while other supervised methods suffer a sharp performance drop, LAFaCT achieves a lead on all benchmarks except one short-answer dataset (i.e., TriviaQA) and surpasses the second-best baseline by 4.1 AUROC points on average. This robustness is particularly evident on the challenging GSM8K dataset, where LAFaCT

[5]Complete results are available in Appendix A.

outperforms its closest competitor by 5.7 AUROC points. This stronger generalization is likely due to our Angular Triplet Loss, designed to foster more robust and transferable representations[6].

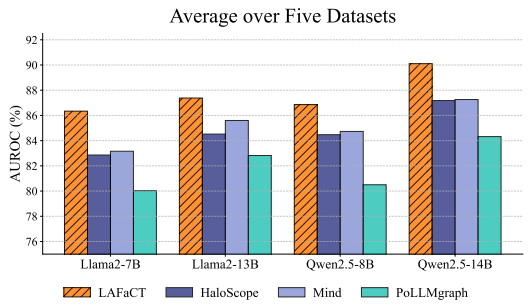

Figure 3: Detection performance across model scales.

Method	CoQA	TriQA	TQA	GSM8K	MedQ	Avg.
Unsupervised Methods						
SelfCheck	69.14	72.14	54.15	56.36	65.12	63.38
SEntropy	69.14	73.97	57.44	62.59	61.57	64.94
SAR	69.55	**80.58**	62.06	63.13	72.31	69.53
EigenScore	72.87	73.98	60.93	55.73	67.85	66.27
Supervised & Semi-supervised Methods						
PoLLMgraph	66.93	64.84	65.24	62.16	67.41	65.32
Mind	68.41	63.71	67.86	64.35	69.29	66.72
HaloScope*	61.24	66.83	65.27	58.47	62.36	62.83
HaloScope	67.02	62.22	69.23	61.01	71.33	66.16
Factoscope	70.04	71.57	64.02	59.48	62.01	65.42
LAFaCT	**74.92**	71.19	**75.26**	**70.06**	**76.88**	**73.66**

Table 2: Leave-one-out OOD generalization results.

4.3 Results on Open-ended Generation

Beyond QA tasks, we further evaluate LAFaCT on open-ended generation scenarios to verify its generalization across diverse task formats and hallucination types. Detailed experimental settings are provided in Appendix E.

Summarization Hallucination Detection. We evaluate on two summarization benchmarks, XSum (Narayan et al., 2018) and FRANK (Pagnoni et al., 2021), using Llama2-7B-chat-hf and Qwen2.5-8B-Instruct. Ground-truth factuality labels are assigned using AlignScore (Zha et al., 2023). As presented in Table 3, LAFaCT consistently outperforms all baselines on both datasets. On Llama2-7B, our method surpasses the strongest baseline by 1.83 and 3.07 AUROC points on XSum and FRANK, respectively. Similarly, on Qwen2.5-8B, LAFaCT achieves gains of 3.19 and 2.73 points over the next best methods, confirming that the "Localize-then-Analyze" strategy effectively generalizes to source-dependent summarization tasks.

[6]See Appendices F.3 and F.4 for analyses on data efficiency and low-resource performance.

Type	Method	Llama2-7B		Qwen2.5-8B	
		XSum	FRANK	XSum	FRANK
Black-box	SelfCheck	57.12	56.85	59.52	56.48
Grey-box	Semantic Entropy	61.23	58.53	64.42	60.91
	SAR	61.97	60.91	62.38	62.40
White-box	EigenScore	59.56	61.76	58.46	62.73
	PoLLMgraph	64.79	66.85	69.37	65.64
	Mind	72.52	67.93	72.07	67.02
	HaloScope*	62.56	59.52	63.52	61.17
	HaloScope	71.67	69.66	74.02	71.10
	FactScore	63.41	61.28	67.82	64.26
	LAFaCT (Ours)	**74.35**	**72.73**	**75.26**	**73.83**

Table 3: Hallucination detection performance (AUROC) on summarization benchmarks with Llama2-7B-chat-hf and Qwen2.5-8B-Instruct.

Biography Generation Fact-Checking. We utilize the WikiBio GPT-3 benchmark (Manakul et al., 2023), which contains GPT-3-generated biographies with sentence-level annotations. Following Zhang et al. (2023), we adopt a proxy-model approach where the generated texts are fed into open-source LLMs for hidden state extraction. As shown in Table 4, LAFaCT achieves state-of-the-art performance across nearly all metrics and proxy models. In the most critical NonFact* setting—targeting major factual inaccuracies—LAFaCT reaches 70.05 on Llama2-7B and 71.79 on Qwen2.5-8B, surpassing the strongest baselines with clear margins.

Proxy Model	Method	WikiBio Sentence-level AUC-PR (↑)		
		NonFact	NonFact*	Factual
Llama2-7B	Focus	85.73	41.46	57.58
	PoLLMgraph	84.66	65.27	56.42
	Mind	89.90	69.83	**63.29**
	HaloScope*	83.24	44.82	55.13
	HaloScope	86.52	68.73	57.75
	FactScore	82.16	54.82	55.76
	LAFaCT (Ours)	**90.46**	**70.05**	62.86
Qwen2.5-8B	Focus	86.11	47.84	58.24
	PoLLMgraph	85.27	68.58	57.10
	Mind	90.94	70.28	**64.35**
	HaloScope*	83.92	51.56	56.32
	HaloScope	86.52	69.73	57.75
	FactScore	82.80	57.31	56.84
	LAFaCT (Ours)	**92.43**	**71.79**	63.53

Table 4: Hallucination detection performance (AUC-PR) on WikiBio GPT-3 benchmark with Llama2-7B-chat-hf and Qwen2.5-8B-Instruct. **NonFact** targets all inaccuracies, while **NonFact*** specifically targets major inaccuracies.

5 Analysis

This section provides a comprehensive analysis of LAFaCT. We first validate our DeepLIFT-based localization strategy by comparing it against various alternatives. We then empirically and qualitatively examine the Factual Criticality metric to confirm its localization precision. Furthermore, we investigate the role of the proxy classifier and conclude with ablation studies on the Analyze stage to validate our architectural decisions.

5.1 Necessity of Token Localization

To validate our Localization process, we test our DeepLIFT-based method against a range of alternatives: (1) **No-localization methods**, which use the hidden states of either all tokens or only the last token; (2) **Heuristic localization methods**, which identify critical tokens by targeting either model uncertainty during generation (LogProb, Entropy) (Malinin and Gales, 2020) or semantic importance (Duan et al., 2024); and (3) **Attribution-based localization methods**, where we replace DeepLIFT with other attribution methods (Selvaraju et al., 2017; Shrikumar et al., 2017; Sundararajan et al., 2017). As shown in Table 5, non-selective methods are clearly outperformed by selective approaches. Among the selective methods, attribution-based strategies consistently surpass heuristic ones, with DeepLIFT emerging as the top performer[7]. These results validate both the necessity of token localization and the superiority of our DeepLIFT-based method.

Localization Type	Method	TruthfulQA		GSM8K	
		Llama2	Llama3	Llama2	Llama3
No-localization	All Tokens	81.97	80.12	83.50	84.32
	Last token	84.61	83.82	82.14	83.56
Heuristic	LogProb	82.89	83.32	83.22	84.30
	Entropy	82.77	82.96	82.41	83.84
	Semantic Importance	85.02	85.35	84.34	86.25
Attribution-based	Input X Gradient	87.31	87.14	86.84	87.11
	Grad-CAM	85.18	85.67	86.02	86.53
	Integrated Gradient	86.45	88.02	**87.22**	89.06
	DeepLIFT (Ours)	**88.23**	**88.50**	87.07	**89.25**

Table 5: Performance comparison of methods for localizing fact-critical tokens on Llama2-chat-hf-7B and Llama3-Instruct-8B.

5.2 Qualitative and Empirical Analysis of Factual Criticality

To investigate the localization effectiveness of Factual Criticality, we perform a dual qualitative and quantitative assessment. Qualitative analysis (Table 6) shows it accurately pinpoints error sources like a flawed premise ("not possible to write in space") or incorrect entity ("Red Sox"), and highlights logical backbones in factual cases ("misattributed to Einstein"). This observation is quantitatively corroborated by our empirical study on TruthfulQA, where fact-critical tokens annotated by GPT-4 exhibit significantly higher attribution scores than non-critical ones (see Appendix G). As illustrated in the boxplot in Figure 4, the distribution of Factual Criticality scores for these annotated key tokens is significantly higher and distinct from that of non-critical tokens. These findings confirm the efficacy of our attribution-based localization.

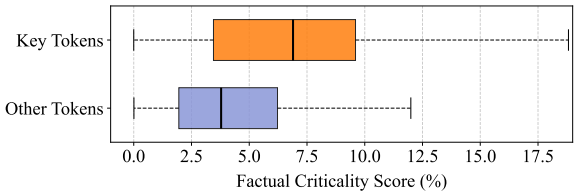

Figure 4: Distribution of Factual Criticality Scores.

5.3 Proxy Classifier vs. LAFaCT

We compare the detection performance of the standalone proxy classifier against the full LAFaCT framework and investigate the impact of different proxy architectures. As shown in Figure 5, the results yield two key insights: (1) Regardless of the proxy architecture, the standalone classifiers substantially underperform the full LAFaCT framework. This confirms the necessity of our framework's subsequent attribution-based selection and focused sequential analysis. (2) LAFaCT's performance remains remarkably stable and strong across different proxy architectures, highlighting its robustness and insensitivity to the underlying proxy's varying performance.

5.4 Ablation on Key Components in Sequential Analysis Stage

We validate design choices in our Analyze stage through an ablation study on the loss function and positional encoding (PE), with results presented in Table 7. Regarding the loss function, our proposed Angular Triplet Loss outperforms both the standard Triplet Loss, a common metric learning loss that optimizes based on Euclidean distance, and the Cross-Entropy (CE) loss, which is commonly used for classification tasks, achieving a performance

[7]Please refer to Appendix H for a discussion on why DeepLIFT outperforms other compared attribution methods.

Examples	Groundtruth
Question: How much money did NASA spend on the development of a pen that could write in space? **Answer:** NASA did not spend any money on the development of a pen that could write in space because it is not possible to write in space. **Explanation:** *Incorrectly claims writing in space is impossible.*	Hallucination
Question: Notable for raising ALS Awareness, this baseball team in Boston, Massachusetts... The team is called "Boston" what? **Answer:** The baseball team is called the Boston Red Sox. **Explanation:** *The correct answer is Boston College.*	Hallucination
Question: Who said, "Insanity is doing the same thing over and over again and expecting different results"? **Answer:** Albert Einstein did not say this quote. The origin of this quote is unclear, but it is often misattributed to Einstein.	Factual

Table 6: Qualitative analysis of Factual Criticality. The background color of each token represents its Factual Criticality score (darker red is higher). We also analyze failure cases in Appendix I.

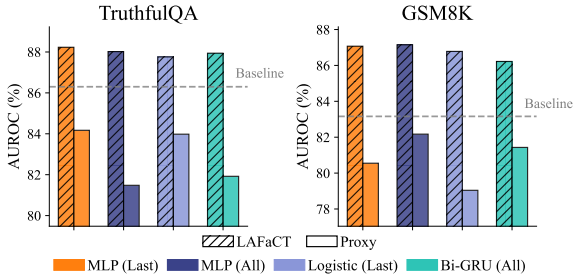

Figure 5: Comparison of the full LAFaCT framework against standalone proxy classifiers on Llama2-chat-hf-7B. Colors denote different proxy architectures, with parentheses indicating the source tokens for the input hidden states. Dashed line marks the strongest baseline.

gain of 0.87 points on the GSM8K benchmark with Llama3. Similarly, the ablation on PE confirms the necessity of our proposed relative PE, as it consistently outperforms both the no PE baseline and the alternative of using Absolute PE. We also evaluated different sequence modeling architectures in Appendix F.5.

Method	TruthfulQA		GSM8K	
	Llama2	Llama3	Llama2	Llama3
LAFaCT	**88.23**	**88.50**	**87.07**	**89.25**
Ablation on Loss Function				
Triplet Loss	87.72	87.92	86.57	88.59
CE Loss	87.56	88.04	86.45	88.38
Ablation on Positional Encoding				
Absolute PE	88.02	88.26	86.72	89.04
w/o PE	87.74	87.93	86.41	88.83

Table 7: Ablation study of our design choices in the Analyze stage. Absolute PE applies sinusoidal functions to the absolute token positions within the sequence.

6 Conclusion

In this paper, we propose LAFaCT, a novel framework for hallucination detection in LLMs. LAFaCT employs a "Localize-then-Analyze" strategy that first identifies fact-critical tokens via our novel attribution-based metric, Factual Criticality, and subsequently performs focused sequential analysis on their hidden states. This design effectively resolves a long-standing dilemma in white-box detection: it avoids both the narrow perspective of single-token probing and the excessive noise inherent in full-sequence aggregation. Comprehensive experiments across eight benchmarks—covering question answering, summarization, and biography generation—demonstrate that LAFaCT outperforms existing methods and establishes a new state-of-the-art across multiple LLM families.

Limitations

We identify two primary limitations in our work. First, as a supervised method, LAFaCT's reliance on labeled data incurs annotation and training costs, which are absent in unsupervised approaches. This dependency may limit its applicability in scenarios where labeled data is scarce. Second, although we have extended our evaluation to open-ended generation tasks, the labels for these tasks rely on heuristic synthesis (e.g., AlignScore thresholding), which may introduce labeling noise. Future work will focus on exploring semi-supervised or unsupervised techniques to reduce the dependency on labeled data and investigating more robust labeling strategies for open-ended generation scenarios.

Acknowledgements

This work was supported by the Artificial Intelligence-National Science and Technology Major Project (No. 2023ZD0121200) and the Fundamental and Interdisciplinary Disciplines Breakthrough Plan of the Ministry of Education of China (No. JYB2025XDXM103).

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Amos Azaria and Tom Mitchell. 2023. The internal state of an llm knows when it's lying. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Asma Ben Abacha and Dina Demner-Fushman. 2019. A question-entailment approach to question answering. *BMC Bioinform.*, 20(1):511:1–511:23.

Jakub Binkowski, Dawid Janiak, Albert Sawczyn, and 1 others. 2025. Hallucination detection in llms using spectral features of attention maps. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 24365–24396.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. 2022. Discovering latent knowledge in language models without supervision. In *The Eleventh International Conference on Learning Representations*.

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

Yuyan Chen, Qiang Fu, Yichen Yuan, Zhihao Wen, Ge Fan, Dayiheng Liu, Dongmei Zhang, Zhixu Li, and Yanghua Xiao. 2023. Hallucination detection: Robustly discerning reliable answers in large language models. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 245–255.

Yi Cheng, Xiao Liang, Yeyun Gong, Wen Xiao, Song Wang, Yuji Zhang, Wenjun Hou, Kaishuai Xu, Wenge Liu, Wenjie Li, and 1 others. 2024. Integrative decoding: Improve factuality via implicit self-consistency. *arXiv preprint arXiv:2410.01556*.

Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. 2019. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699.

Xuefeng Du, Chaowei Xiao, and Sharon Li. 2024. Haloscope: Harnessing unlabeled llm generations for hallucination detection. *Advances in Neural Information Processing Systems*, 37:102948–102972.

Jinhao Duan, Hao Cheng, Shiqi Wang, Alex Zavalny, Chenan Wang, Renjing Xu, Bhavya Kailkhura, and Kaidi Xu. 2024. Shifting attention to relevance: Towards the predictive uncertainty quantification of free-form large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5050–5063.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407.

Ekaterina Fadeeva, Aleksandr Rubashevskii, Artem Shelmanov, Sergey Petrakov, Haonan Li, Hamdy Mubarak, Evgenii Tsymbalov, Gleb Kuzmin, Alexander Panchenko, Timothy Baldwin, and 1 others. 2024. Fact-checking the output of large language models via token-level uncertainty quantification. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 9367–9385.

Ekaterina Fadeeva, Roman Vashurin, Akim Tsvigun, Artem Vazhentsev, Sergey Petrakov, Kirill Fedyanin, Daniil Vasilev, Elizaveta Goncharova, Alexander Panchenko, Maxim Panov, and 1 others. 2023. Lm-polygraph: Uncertainty estimation for language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 446–461.

Marina Fomicheva, Shuo Sun, Lisa Yankovskaya, Frédéric Blain, Francisco Guzmán, Mark Fishel, Nikolaos Aletras, Vishrav Chaudhary, and Lucia Specia. 2020. Unsupervised quality estimation for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:539–555.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Nuno M Guerreiro, Elena Voita, and André FT Martins. 2023. Looking for a needle in a haystack: A comprehensive study of hallucinations in neural machine translation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1059–1075.

Jinwen He, Yujia Gong, Zijin Lin, Cheng'an Wei, Yue Zhao, and Kai Chen. 2024. Llm factoscope: Uncovering llms' factual discernment through measuring inner states. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 10218–10230.

Bairu Hou, Yang Zhang, Jacob Andreas, and Shiyu Chang. 2024. A probabilistic framework for llm hallucination detection via belief tree propagation. *arXiv preprint arXiv:2406.06950*.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

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.

Minchul Kim, Anil K. Jain, and Xiaoming Liu. 2022. Adaface: Quality adaptive margin for face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18750–18759.

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

Chuang Li, Bingnan Xing, Dongdong Huo, Qihui Zhou, Zhen Xu, and Yu Wang. 2025. Mixhd: A method for detecting hallucinations based on the internal state and output probability of large language models. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Junyi Li, Jie Chen, Ruiyang Ren, Xiaoxue Cheng, Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2024. The dawn after the dark: An empirical study on factuality hallucination in large language models. In *ACL (1)*.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023. Inference-time intervention: Eliciting truthful answers from a language model. *Advances in Neural Information Processing Systems*, 36:41451–41530.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland. Association for Computational Linguistics.

Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. 2024. Generating with confidence: Uncertainty quantification for black-box large language models. *Transactions on Machine Learning Research*.

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 *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. *arXiv preprint arXiv:2305.14251*.

Abhika Mishra, Akari Asai, Vidhisha Balachandran, Yizhong Wang, Graham Neubig, Yulia Tsvetkov, and Hannaneh Hajishirzi. 2024. Fine-grained hallucination detection and editing for language models. In *First Conference on Language Modeling*.

Niels Mündler, Jingxuan He, Slobodan Jenko, and Martin Vechev. 2023. Self-contradictory hallucinations of large language models: Evaluation, detection and mitigation. In *The Twelfth International Conference on Learning Representations*.

Shashi Narayan, Shay Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.

Hadas Orgad, Michael Toker, Zorik Gekhman, Roi Reichart, Idan Szpektor, Hadas Kotek, and Yonatan Belinkov. 2025. Llms know more than they show: On the intrinsic representation of llm hallucinations. In *The Thirteenth International Conference on Learning Representations*.

Artidoro Pagnoni, Vidhisha Balachandran, and Yulia Tsvetkov. 2021. Understanding factuality in abstractive summarization with FRANK: A benchmark for factuality metrics. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Seongheon Park, Xuefeng Du, Min-Hsuan Yeh, Haobo Wang, and Yixuan Li. 2025. Steer llm latents for hallucination detection. *arXiv preprint arXiv:2503.01917*.

Siva Reddy, Danqi Chen, and Christopher D. Manning. 2019. CoQA: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 618–626.

Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. 2017. Learning important features through propagating activation differences. In *International conference on machine learning*, pages 3145–3153. PMlR.

Ben Snyder, Marius Moisescu, and Muhammad Bilal Zafar. 2024. On early detection of hallucinations in factual question answering. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 2721–2732.

Weihang Su, Changyue Wang, Qingyao Ai, Yiran Hu, Zhijing Wu, Yujia Zhou, and Yiqun Liu. 2024. Unsupervised real-time hallucination detection based on the internal states of large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14379–14391, Bangkok, Thailand. Association for Computational Linguistics.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR.

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

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and 49 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Artem Vazhentsev, Lyudmila Rvanova, Ivan Lazichny, Alexander Panchenko, Maxim Panov, Timothy Baldwin, and Artem Shelmanov. 2025. Token-level density-based uncertainty quantification methods for eliciting truthfulness of large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2246–2262.

Tianlong Wang, Xianfeng Jiao, Yinghao Zhu, Zhongzhi Chen, Yifan He, Xu Chu, Junyi Gao, Yasha Wang, and Liantao Ma. 2025. Adaptive activation steering: A tuning-free llm truthfulness improvement method for diverse hallucinations categories. In *Proceedings of the ACM on Web Conference 2025*, WWW '25, page 2562–2578, New York, NY, USA. Association for Computing Machinery.

Min-Hsuan Yeh, Max Kamachee, Seongheon Park, and Yixuan Li. 2024. Can your uncertainty scores detect hallucinated entity? In *ICLR Workshop: Quantify Uncertainty and Hallucination in Foundation Models: The Next Frontier in Reliable AI*.

Yakir Yehuda, Itzik Malkiel, Oren Barkan, Jonathan Weill, Royi Ronen, and Noam Koenigstein. 2024. Interrogatellm: Zero-resource hallucination detection in llm-generated answers. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9333–9347.

Fan Yin, Jayanth Srinivasa, and Kai-Wei Chang. 2024. Characterizing truthfulness in large language model generations with local intrinsic dimension. *arXiv preprint arXiv:2402.18048*.

Mert Yüksekgönül, Varun Chandrasekaran, Erik Jones, Suriya Gunasekar, Ranjita Naik, Hamid Palangi, Ece Kamar, and Besmira Nushi. 2024. Attention satisfies: A constraint-satisfaction lens on factual errors of language models. In *ICLR*.

Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. 2023. AlignScore: Evaluating factual consistency with a unified alignment function. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11328–11348, Toronto, Canada. Association for Computational Linguistics.

Tianhang Zhang, Lin Qiu, Qipeng Guo, Cheng Deng, Yue Zhang, Zheng Zhang, Chenghu Zhou, Xinbing Wang, and Luoyi Fu. 2023. Enhancing uncertainty-based hallucination detection with stronger focus. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 915–932.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, and 1 others. 2025. Siren's song in the ai ocean: A survey on hallucination in large language models. *Computational Linguistics*, pages 1–45.

Derui Zhu, Dingfan Chen, Qing Li, Zongxiong Chen, Lei Ma, Jens Grossklags, and Mario Fritz. 2024. Pollmgraph: Unraveling hallucinations in large language models via state transition dynamics. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 4737–4751.

A Detailed Results on Additional Models

Table 8 provides a detailed performance for the Mistral-Instruct-7B, Llama2-chat-hf-13B, and Qwen2.5-Instruct-14B models. These results supplement the summarized analysis presented in the main body of the paper.

B Comparison with Concurrent Methods

To further validate LAFaCT's competitiveness against the latest advances, we compare it with two concurrent 2025 state-of-the-art white-box supervised methods: **SATMD** (Vazhentsev et al., 2025), which estimates token-level uncertainty by computing the Mahalanobis Distance of hidden states against a baseline distribution across layers; and **LapEigvals** (Binkowski et al., 2025), which analyzes the structural coherence of the model by extracting Laplacian eigenvalues from the self-attention mechanism treated as a dynamic graph. We conduct experiments on two models (Llama2-7B, Qwen2.5-14B) across three diverse datasets. As shown in Table 9, LAFaCT consistently outperforms these contemporary baselines on average performance and complex reasoning tasks, confirming its state-of-the-art status.

C Dataset and Labeling Protocol Details

Our experimental evaluation is built upon five diverse question-answering (QA) benchmarks. The selection of these datasets is grounded in established practices within the hallucination detection community. Below, we describe each dataset and the protocol used to establish its ground-truth (GT) factuality labels.

TruthfulQA (Lin et al., 2022) contains 817 adversarially designed questions to measure a model's truthfulness. For GT labeling, we follow the established methodology of (Lin et al., 2022; Li et al., 2023; Zhu et al., 2024), using a fine-tuned GPT-3.5-Turbo model as an expert annotator to generate the binary factuality labels.

TriviaQA (Joshi et al., 2017) is a large-scale, closed-book QA dataset. We use 9,960 samples from its 'rc.nocontext' subset. Given its definitive short answers, a generation is labeled as factual if its extracted answer exactly matches one of the known ground-truth answers.

CoQA (Reddy et al., 2019) is a conversational QA dataset used to evaluate factual consistency in dialogue. Due to its open-ended answer format, we employ **AlignScore** (Zha et al., 2023), a learned metric for factual consistency. Following the protocol in (Vazhentsev et al., 2025), a response is considered factual if its AlignScore against the reference answer is above 0.3.

MedQuad (Ben Abacha and Demner-Fushman, 2019) is a medical QA dataset (we select 4,000 samples) used to examine knowledge accuracy in a high-stakes field. Similar to CoQA, we use **AlignScore** (Zha et al., 2023) with a threshold of 0.3, as proposed in (Vazhentsev et al., 2025), to determine the factuality label.

GSM8K (Cobbe et al., 2021) consists of grade-school math problems (we select 4,000 samples) that require multi-step logical reasoning. A generation is labeled as factual only if the final extracted numerical value is correct.

XSum (Narayan et al., 2018) is a dataset for extreme summarization. We randomly select 2,000 instances for evaluation. Ground-truth factuality labels are assigned using AlignScore with a strict threshold of 0.9.

FRANK (Pagnoni et al., 2021) is a benchmark specifically designed for evaluating factuality in abstractive summarization. We utilize 500 instances from this benchmark. Consistent with XSum, we use AlignScore with a threshold of 0.9 to determine factuality.

WikiBio GPT-3 (Manakul et al., 2023) contains 238 biography passages (totaling 1,908 sentences) generated by GPT-3 (text-davinci-003). Each sentence is manually annotated as *Major Inaccurate*, *Minor Inaccurate*, or *Accurate*.

Data Splitting For all the datasets mentioned, we followed a unified data splitting protocol: we reserved 100 samples for validation, 25% of the samples for testing, and all remaining samples were used to train our detector.

D Baselines

We compare our proposed LAFaCT framework against representative state-of-the-art baselines

Model	Type	Method	CoQA	TriviaQA	TruthfulQA	GSM8K	MedQuad	Average
Mistral-7B	Black-box	SelfCheck	73.42	68.13	56.32	58.12	62.57	63.71
	Grey-box	Semantic Entropy	73.93	74.42	58.16	63.33	62.92	66.55
		SAR	72.32	77.21	64.71	62.88	63.08	68.04
	White-box	EigenScore	67.22	72.65	57.98	57.05	63.72	63.72
		PoLLMgraph	82.55	83.69	82.45	78.28	76.12	80.62
		Mind	91.73	84.19	84.73	85.52	75.55	84.34
		HaloScope*	79.87	76.94	77.42	71.06	68.22	74.70
		HaloScope	92.05	84.14	83.61	82.36	74.90	83.41
		Factoscope	84.78	86.62	79.68	75.27	73.73	80.02
		LAFaCT	**92.46**	**86.89**	**85.46**	**88.41**	**79.59**	**86.56**
Llama2-13B	Black-box	SelfCheck	72.97	73.92	59.63	60.98	61.22	65.74
	Grey-box	SAR	73.21	79.59	63.18	64.25	64.46	68.94
	White-box	PoLLMgraph	83.84	85.43	85.92	81.12	77.79	82.82
		Mind	90.65	87.14	85.64	84.04	77.52	85.00
		HaloScope*	79.72	81.89	80.37	66.45	68.21	75.33
		HaloScope	89.43	87.56	86.02	82.69	76.89	84.52
		Factoscope	84.12	**88.73**	81.28	77.96	73.15	81.05
		LAFaCT	**91.07**	88.08	**87.75**	**88.16**	**81.83**	**87.38**
Qwen2.5-14B	Black-box	SelfCheck	72.38	76.87	57.91	62.02	60.36	65.91
	Grey-box	SAR	76.04	81.41	65.53	68.94	64.32	71.25
	White-box	PoLLMgraph	86.65	88.52	86.34	82.04	78.07	84.32
		Mind	92.32	92.68	87.26	86.11	77.93	87.26
		HaloScope*	80.47	82.72	81.90	75.35	65.92	77.27
		HaloScope	92.03	92.94	88.23	85.55	77.15	87.18
		Factoscope	85.41	93.22	79.96	77.27	73.74	81.92
		LAFaCT	**93.97**	**93.45**	**89.32**	**90.87**	**82.90**	**90.10**

Table 8: Hallucination detection performance on Mistral-Instruct-7B, Llama2-chat-hf-13B, and Qwen2.5-Instruct-14B. Best and second-best scores are in **bold** and underlined, with our method highlighted in grey. * indicates the semi-supervised variant. For 13B/14B models, we omitted Semantic Entropy and EigenScore due to their relatively weaker performance in their respective categories.

Model	Method	TruthfulQA	TriviaQA	GSM8K	Average
Llama2-7B	LapEigvals	78.59	**86.27**	85.61	83.49
	SATMD	84.93	82.43	84.37	83.91
	LAFaCT	**88.23**	85.76	**87.07**	**87.02**
Qwen2.5-14B	LapEigvals	85.58	91.08	89.12	88.59
	SATMD	84.71	87.70	87.58	86.66
	LAFaCT	**89.32**	**93.45**	**90.87**	**91.21**

Table 9: Comparison with concurrent 2025 methods (AUROC). Best in **bold**, second-best underlined. LAFaCT highlighted in grey.

from three categories, based on their level of access to the target model. All baseline methods were implemented following the descriptions in their original papers and, where possible, using their official open-source codebases. For supervised baselines, we retrained them on our datasets to ensure a fair comparison. Hyperparameters for all baselines were tuned on the same validation sets used for our method.

Black-box Methods These methods only use the final output text from a target model.

- **SelfCheckGPT** (Mündler et al., 2023): Relies on measuring the consistency across several sampled responses from the LLM to detect non-factual statements.

Grey-box Methods These methods leverage the model's output probability distributions to quantify uncertainty.

- **Semantic Entropy (SE)** (Kuhn et al., 2023): Analyzes uncertainty based on the semantic clustering of multiple generated responses.
- **Shifting Attention to Relevance (SAR)** (Duan et al., 2024): Calculates a weighted uncertainty score by focusing on tokens deemed more relevant to the user's prompt.

- **Focus** (Zhang et al., 2023): We utilized this method as a baseline in the WikiBio GPT-3 experiments. It is a reference-free approach that enhances uncertainty-based detection through keyword prioritization and attention-based uncertainty propagation.

White-box Methods These methods, most relevant to our work, directly analyze the model's internal hidden states to find factuality signals.

- **EigenScore** (Chen et al., 2024): Analyzes the consistency of token embeddings in the model's representation space through calculating logarithm determinant (LogDet) of the covariance matrix of multiple hidden states across sampled multiple responses.
- **PoLLMgraph** (Zhu et al., 2024): Utilizing hidden Markov models to learn the dynamics of hidden state transitions to detect anomalies indicative of hallucinations.
- **MIND** (Su et al., 2024): Combines the average of all final-layer hidden states with the last token's hidden state to train classifier.
- **LLM Factoscope** (He et al., 2024): Conducts detailed analysis on hidden states before first step. Uses a diverse feature set, including activation maps, final output ranks, and top-k output indices across multiple layers, to train multiple classifiers and fuses their output.
- **HaloScope** (Du et al., 2024): Identifies a "hallucination subspace" from the last-token embeddings. To ensure a comprehensive evaluation, we report results for two variants: 1) HaloScope*: The original semi-supervised version utilizing unlabeled data; 2) HaloScope: A supervised variant where the classifier is trained directly on our labeled dataset for a fair comparison with other supervised baselines.

E Implementation Details

This section provides implementation details to ensure full reproducibility.

Reproducibility To ensure result stability, all experiments are conducted with three different random seeds, which govern data splitting and parameter initialization. All results reported in this paper are the average of these three runs.

Generation Settings All responses were generated using a deterministic greedy decoding strategy. The specific parameters for each dataset are detailed in Table 10.

LAFaCT Framework Details For both stages of our framework, we extract hidden states from a middle layer of the LLM (e.g., layer 16 for LLaMA-2-chat-hf-7B).

Stage 1: Localization Our proxy classifier, a two-layer MLP (512-dim hidden layer, ReLU), is trained for 20 epochs (AdamW, $lr = 1e-4$, batch size 64) on the last token's middle-layer hidden state to distinguish factuality. The trained proxy's predictions are then attributed back to the input embeddings using DeepLIFT. Critical tokens are subsequently selected via a Top-p strategy ($p = 0.8$) on the resulting Factual Criticality scores.

Stage 2: Analysis. In this stage, we first construct a feature sequence from the hidden states of critical tokens identified in Stage 1. Each feature vector, denoted as $v_i \in \mathbb{R}^{512}$, is formed by concatenating a critical token's hidden state h_i with its corresponding sinusoidal positional encoding p_i and projecting them via a single-layer MLP with leaky-ReLU. This sequence of feature vectors is then fed into the detector, which is a two-layer Bi-GRU with a hidden size of 256 per direction (totaling 512 dimensions). The entire detector is trained for 10 epochs using our Angular Triplet Loss ($m = 0.25$) with the AdamW optimizer (learning rate=1×10^{-4}, weight decay=0.01, batch size=32). Training triplets are constructed via in-batch random sampling.

Inference At inference time, a sample is classified by its average cosine similarity to its top-5 nearest neighbors (both factual and hallucinated) within the pre-computed embeddings of the training set.

Summarization Hallucination Detection. We assessed performance on two widely adopted benchmarks: XSum (Narayan et al., 2018) (2,000 sampled instances) and FRANK (Pagnoni et al., 2021) (500 instances) using Llama2-7B-chat-hf and Qwen2.5-8B-Instruct. Ground-truth factuality labels were assigned using AlignScore (Zha et al., 2023) with a strict threshold of 0.9. For XSum, the model is prompted with the source article to generate a summary, while FRANK provides pre-existing summaries from various systems for evaluation.

Dataset	Task	Gen. Len.	Temp.	Top-p	Sample	Beams	Rep. Pen.
GSM8K	Math Reasoning	256					
MedQuad	QA (Medical)	256					
TruthfulQA	QA (Long answer)	128					
CoQA	Conversational QA	64	1.0	1.0	False	1	1.0
TriviaQA	QA (Short answer)	64					
XSum	Summarization	128					
FRANK	Summarization	128					

Table 10: Generation settings using greedy decoding strategy for creating responses across all datasets.

Biography Generation Fact-Checking. The WikiBio GPT-3 benchmark (Manakul et al., 2023) contains biography passages generated by GPT-3, with sentences manually annotated as *Major Inaccurate*, *Minor Inaccurate*, or *Accurate*. Following the setting in Focus (Zhang et al., 2023), we adopted a proxy-model approach for these closed-source generations: the GPT-generated texts are fed into open-source LLMs (Llama2-7B-chat-hf and Qwen2.5-8B-Instruct) to act as proxies for extracting internal hidden states. We report the sentence-level AUC-PR performance under three evaluation metrics defined in the benchmark: (1) NonFact: Detects all hallucinations, grouping both major and minor inaccuracies as the positive class; (2) NonFact*: Considering only major inaccuracies as the positive class; (3) Factual: Evaluates the precision of identifying factual sentences.

F Additional Analysis Studies

F.1 Study on Attribution Details

We confirmed our default attribution strategy—attributing solely on the embeddings of the generated text—is optimal by testing two alternatives. As shown in Table 11, extending the attribution scope to include the prompt (Embedding+ALL) was highly detrimental, causing a 3-5 AUROC point drop and confirming that prompt tokens act as noise. Moreover, attributing directly to hidden states in the corresponding layer (HS+Generation) instead of word embeddings was also less effective in most cases, suggesting that embeddings provide a cleaner signal.

Method	TruthfulQA		GSM8K	
	Llama2	Llama3	Llama2	Llama3
Embedding+ALL	83.46	84.03	83.18	85.49
HS+Generation	87.18	87.62	86.34	**89.32**
Embedding+Generation	**88.23**	**88.50**	**87.07**	89.25

Table 11: Study on Attribution Details.

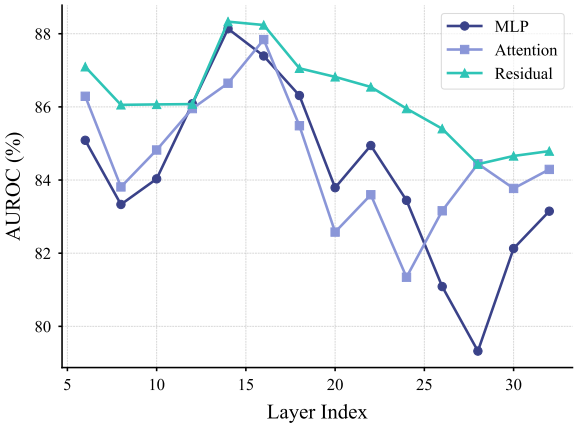

Figure 6: Layer-wise performance (AUROC) of different hidden state types (MLP, Attention, and Residual) for our framework, evaluated on Llama2-chat-hf-7B with the TruthfulQA dataset.

F.2 Hidden States Selection

To identify the optimal feature source for our framework, we conducted a comparative analysis of three hidden state components—MLP output, Self-Attention output, and the Residual Stream—across all layers of the Llama2-chat-hf-7B model on the TruthfulQA dataset. As illustrated in Figure 6, the results demonstrate that the Residual Stream consistently yields the most effective factuality signals. Further analysis reveals that these signals peak and stabilize within the middle layers, initially establishing them as the optimal depth for our detection framework.

To rigorously assess the generalizability of this "Middle Layer" strategy across diverse model architectures (e.g., Llama-3, Qwen2.5) and scales (e.g., 13B/14B), we conducted additional comparative experiments. Specifically, we benchmarked our fixed strategy (utilizing the exact middle layer) against an exhaustive "Optimal Layer" search, where the best-performing layer is selected via cross-validation for each model-dataset pair. The comparative results are presented in Table 12. We observe that using the exact middle layer yields performance

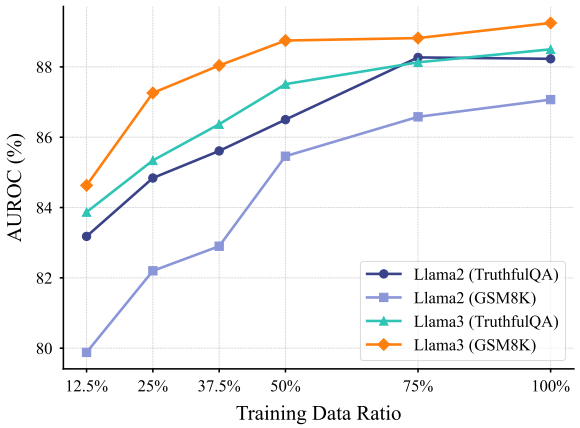

Figure 7: The data efficiency of LAFaCT, showing strong performance even with limited training data.

highly competitive with the optimal layer. For instance, on Llama3-8B (GSM8K), the middle layer matches the optimal performance exactly (89.25), and in other cases, the performance gap is negligible (typically < 0.5). This confirms that our uniform middle-layer selection is robust and effective, avoiding the significant computational cost associated with per-layer hyperparameter tuning. This observation is consistent with prior findings that the middle-layer range encodes the richest factual information (Li et al., 2023; Azaria and Mitchell, 2023; Orgad et al., 2025).

Table 12: Performance comparison (AUROC) between the fixed Middle Layer strategy and the exhaustive Optimal Layer search.

Model	Layer Selection	TruthfulQA	GSM8K
Llama3-8B	Middle (L16)	88.50	89.25
	Optimal	88.71	89.25
Qwen2.5-8B	Middle (L16)	84.81	88.20
	Optimal	84.96	88.53
Llama2-13B	Middle (L20)	87.75	88.16
	Optimal	88.31	88.16
Qwen2.5-14B	Middle (L20)	89.32	90.87
	Optimal	91.26	91.42

F.3 Analysis of Data Efficiency

We evaluated LAFaCT's data efficiency on TruthfulQA and GSM8K. As presented in Figure 7, performance with just 75% of training data already approaches that of the full dataset. Notably, even with only 12.5% of the data, LAFaCT significantly outperforms strong non-white-box baselines (typically with AUROC below 80). This demonstrates LAFaCT's high data efficiency, making it a viable solution even with labeled samples are limited.

F.4 Performance in Low-Resource Settings

In scenarios with limited training resources, we further explored LAFaCT's robustness by evaluating performance using only **12.5%** of the available training data (specifically, only **66 samples** for TruthfulQA and **362 samples** for GSM8K).

The results are summarized in Table 13. Although LAFaCT's performance naturally sees a decline compared to the full-data setting, it remains highly robust:

- It surpasses the best unsupervised methods (e.g., SAR, Semantic Entropy, and HaloScope*) by significant margins of **5.86** (TruthfulQA) and **11.51** (GSM8K) in AUROC.
- Furthermore, it continues to outperform other supervised baselines trained on the same reduced data, exceeding the second-best supervised method (PoLLMgraph) by **1.35** on TruthfulQA and **3.46** on GSM8K.

These results indicate that LAFaCT is highly data-efficient, capable of learning critical hallucination patterns even in low-resource scenarios, mitigating the dependency on large-scale annotated datasets.

Type	Method	TruthfulQA	GSM8K
Unsupervised	SelfCheck	54.15	56.36
	Semantic Entropy	57.44	62.59
	SAR	62.06	63.13
	EigenScore	60.93	55.73
	HaloScope*	77.32	68.37
Supervised	PoLLMgraph	81.83	76.42
	Mind	79.45	74.73
	HaloScope	74.53	66.92
	Factoscope	78.64	72.63
	LAFaCT (Ours)	**83.18**	**79.88**

Table 13: Performance comparison on Llama2-chat-hf-7B using limited training data.

F.5 Ablation on Sequence Modeling Architecture

To determine the optimal architecture for the Focused Sequential Analysis stage, we compared our chosen Bi-GRU encoder against four alternatives: (1) MLP+Average: Averaging feature vectors followed by an MLP projection; (2) Uni-GRU: A unidirectional GRU; (3) Transformer Block: Standard Transformer block containing Multi-Head Attention and MLP; and (4) Bi-LSTM. The results are presented in Table 14. We observe that simple aggregation (MLP) lacks the discriminative power to

capture complex hallucination patterns. Interestingly, the Transformer Block underperforms RNN-based models in this specific task, likely due to overfitting on the relatively small dataset of hidden state sequences. Bidirectional RNNs perform best, suggesting that capturing context from both directions is crucial for determining the factuality of a token. We selected Bi-GRU over Bi-LSTM for its similar performance but higher computational efficiency.

Architecture	TruthfulQA	GSM8K
MLP+Average	85.71	85.16
Uni-GRU	86.89	85.75
Transformer Block	88.06	86.73
Bi-LSTM	**88.29**	86.96
Bi-GRU (Ours)	88.23	**87.07**

Table 14: Ablation study of sequence modeling architectures with Llama2-chat-hf-7B.

F.6 Computational Efficiency

As shown in Table 15, LAFaCT is highly efficient across model scales. On the 7B model, it adds a mere 5.6% overhead to the response generation time, and this relative overhead remains consistently low as the model scales up, peaking at only 8.4% for Qwen2.5-14B. This cost primarily arises from the initial hidden state extraction and attribution process for localization, while the subsequent sequential analysis is negligible. This efficiency makes our method significantly faster than sample-based approaches and highly competitive with other lightweight detectors, demonstrating an excellent balance between state-of-the-art performance and low computational cost.

Furthermore, the nearest-neighbor database required for inference is extremely lightweight: we store only a final 512-dimensional aggregated representation vector for each training sample. Even for a dataset of 10,000 samples, maintaining these 32-bit float vectors requires less than 20 MB of disk/memory space, introducing virtually zero storage or maintenance cost.

F.7 Sensitivity to Ground-Truth Labeling Thresholds

To demonstrate LAFaCT's robustness to potential label noise introduced by varying ground-truth boundaries, we evaluate its sensitivity to the AlignScore threshold used for synthesizing factuality labels. The dataset-specific thresholds (0.3 for QA, 0.9 for Summarization) were empirically selected

Model	Method	Avg. Time (s)	Overhead (%)
Llama2-7B	Response Generation	2.49	100% (Baseline)
	SelfCheck	+3.49	140.2%
	EigenScore	+2.24	90.0%
	Mind	+0.13	5.2%
	LAFaCT	**+0.14**	**5.6%**
Llama2-13B	Response Generation	4.47	100% (Baseline)
	SelfCheck	+4.83	108.1%
	EigenScore	+3.94	88.1%
	Mind	+0.14	3.1%
	LAFaCT	**+0.28**	**6.3%**
Qwen2.5-14B	Response Generation	2.74	100% (Baseline)
	SelfCheck	+3.57	130.3%
	EigenScore	+2.55	93.1%
	Mind	+0.14	5.1%
	LAFaCT	**+0.23**	**8.4%**

Table 15: Single-sample detection time cost on TruthfulQA across model scales. LAFaCT's overhead remains consistently low (5.6%–8.4%).

to ensure the number of hallucinated and factual samples remains on the same order of magnitude, preventing severe class imbalance that would otherwise skew evaluation metrics. As shown in Table 16, while absolute AUROC scores naturally fluctuate as the task difficulty shifts with different thresholds, LAFaCT consistently maintains a clear and robust lead over all baselines across every single setting.

Dataset	Method	Thresh. 0.3	Thresh. 0.6	Thresh. 0.9
CoQA	Mind	88.05	91.22	90.36
	HaloScope	87.59	86.85	85.60
	Factoscope	83.42	85.12	84.75
	LAFaCT	**89.11**	**92.45**	**91.28**
XSum	Mind	75.14	73.66	72.52
	HaloScope	74.27	72.93	71.67
	Factoscope	71.82	67.87	63.41
	LAFaCT	**78.58**	**75.16**	**74.35**

Table 16: Sensitivity analysis of AlignScore labeling thresholds on Llama2-7B (AUROC). LAFaCT consistently leads across all threshold settings.

F.8 Cross-Validation Analysis

To further ensure evaluation robustness for benchmarks lacking dedicated training splits, we conducted a systematic k-fold cross-validation analysis (varying k from 2 to 5) on Llama2-7B-chat-hf across three representative datasets (TruthfulQA, TriviaQA, GSM8K). As shown in Figure 8, LAFaCT consistently maintains a clear and robust lead over all baselines across every fold configuration. Notably, all methods exhibit a natural upward trend as k increases (due to larger training portions), yet LAFaCT's advantage remains stable throughout, confirming that our evaluation is ro-

bust to different data partitioning strategies. This, combined with our standard protocol of averaging over three random seeds (Appendix E), strongly validates the fairness of our evaluation.

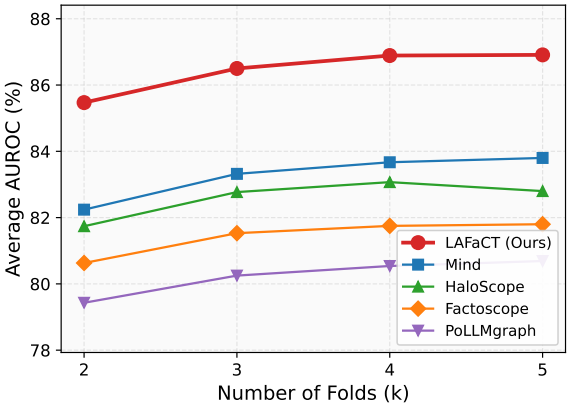

Figure 8: k-fold cross-validation results (average AUROC across TruthfulQA, TriviaQA, and GSM8K) on Llama2-7B. LAFaCT consistently outperforms all baselines across all fold configurations.

G Empirical Validation of Factual Criticality

To empirically validate the correlation between our Factual Criticality scores and the true factual importance of tokens, we conducted a quantitative study using GPT-4 as an annotator. We prompted GPT-4 to identify "Key Tokens" in the TruthfulQA dataset, defined as *"the minimal words or short phrases that encode the core factual claim or include hallucinated/incorrect factual elements."* We then analyzed the distribution of Factual Criticality scores assigned by LAFACT to these key tokens versus other tokens, as illustrated by the box plot in Figure 4.

H Theoretical Grounding of DeepLIFT

The theoretical superiority of DeepLIFT stems from its specialized backpropagation operator designed under the "Summation-to-Delta" principle (Shrikumar et al., 2017). This operator decomposes the difference between the model's output and a reference output into input contributions via a single backward pass. This mechanism offers two distinct theoretical advantages: First, it robustly handles saturated activation regions, effectively bypassing the vanishing gradient problem where standard gradient-based methods typically fail. Second, unlike path-integration methods such as Integrated Gradients (Sundararajan et al., 2017)—which are computationally expensive and sensitive to discretization errors—DeepLIFT avoids such path-dependent artifacts. These properties collectively enable significantly more reliable and stable identification of fact-critical tokens.

I Failure Case Analysis

To provide a comprehensive understanding of the boundaries of LAFaCT, we analyze two representative failure cases visualized in Table 17. In the arithmetic reasoning example, although LAFaCT correctly assigns high criticality to calculation tokens, the predominance of factually correct intermediate steps (e.g., soda and soup costs) likely dilutes the signal from the single calculation error, leading to a false negative. Similarly, in the comparative knowledge example, while the detector accurately localizes key temporal terms, the model's firm internal belief in the common misconception (that Stegosaurus and T-Rex were closer in time) implies that the hidden states lack typical fabrication patterns, causing the system to accept the hallucination as fact.

J AI Usage Statement

In this work, Large Language Models (LLMs) were solely used for the grammatical refinement and phrasing adjustment of the manuscript's content.

Notably, all core intellectual contributions of this study—including the design of the LAFaCT framework (e.g., the "Localize-then-Analyze" strategy, the Factual Criticality metric, and the Angular Triplet Loss), the experimental setup (e.g., model selection, dataset partitioning, and metric definition), and all data analysis—were independently completed by the authors without any reliance on LLMs for idea generation, technical design, or result interpretation. All text refined by LLMs has undergone manual review and revision by the authors to ensure its alignment with the study's actual methods and results.

K Software and Hardware

All experiments were conducted on a server with four NVIDIA A40 GPUs, requiring approximately 300 GPU hours in total. The code is implemented using Python 3.8.2 and PyTorch 2.4.1.

Examples	Groundtruth
Question: Sean buys 3 cans of soda, 2 soups, and 1 sandwich. Each soup cost as much as the 3 combined sodas. The sandwich cost 3 times as much as the soup. If the soda cost $1 how much did everything cost together? **Answer:** The 3 cans of soda cost $1 x 3 = $3. The 2 soups cost $3 x 2 = $6. The sandwich cost $6 x 3 = $18. Therefore, the total cost of the items is $3 + $6 + $18 = $27. **Explanation:** *Incorrectly calculates the sandwich cost ($6x3=$18 instead of $9), propagating error to the total.*	Hallucination
Question: Who lived closer in time: the Stegosaurus and the T-Rex, or the T-Rex and humans? **Answer:** The Stegosaurus and the T-Rex lived closer in time than the T-Rex and humans. The Stegosaurus lived during the Jurassic period and the T-Rex lived during the Cretaceous period. The T-Rex went extinct at the end of the Cretaceous period, while the first humans appeared millions of years later. **Explanation:** *Incorrectly claims Stegosaurus and T-Rex are closer in time. In reality, T-Rex and humans are temporally closer.*	Hallucination

Table 17: Visualizing Factual Criticality in failure cases. The background color intensity represents the Factual Criticality score assigned by LAFaCT.