

Tracking Truth Through Layers: Temporal Dynamics of Fact Representations for Hallucination Detection in LLMs

Yanqiang Zhang, Dawei Feng[†], Bo Ding, Yuanzhao Zhai, Huaimin Wang

College of Computer Science and Technology, National University of Defense Technology,
State Key Laboratory of Complex & Critical Software Environment,
National Key Laboratory of Parallel and Distributed Computing

[†]Correspondence: davyfeng.c@qq.com

Abstract

While large language models (LLMs) have exhibited remarkable performance across diverse natural language processing tasks, their inherent propensity to generate hallucinations poses a critical challenge to reliable deployment in high-stakes scenarios. Existing hallucination detection methods often rely on static single-layer representations or aggregated features across all layers, which often fail to capture fine-grained factual inconsistencies and thus exhibit limited generalization capability. To address this limitation, this paper proposes **TemHal** (**Temporal Hallucination Detector**), a novel detection framework that models the cross-layer evolutionary trajectory of internal representations.¹ Specifically, TemHal tracks the dynamic representational progression of key factual tokens across selected intermediate MLP activations and employs a bidirectional temporal modeling network to identify representational patterns that differentiate factually consistent outputs from hallucinated ones. Extensive experimental results demonstrate that the proposed method achieves consistent performance improvements across multiple mainstream model architectures and diverse benchmark datasets, while also exhibiting robust cross-dataset generalization across multiple LLM backbones. Furthermore, ablation studies validate the contribution of each architectural component and confirm the importance of temporal modeling and intermediate-layer focusing for detection performance.

1 Introduction

Large language models (LLMs) have demonstrated substantial proficiency across a broad spectrum of natural language processing tasks (Brown et al., 2020; Hoffmann et al., 2022; Chowdhery et al., 2023). Nevertheless, these models remain susceptible to *hallucinations*—instances where LLMs gen-

erate responses that appear plausible but are factually incorrect or contextually disconnected (Li et al., 2025). This inherent limitation impedes their deployment in high-stakes domains, such as scientific research (Moores and Zuin Zeidler, 2025) and medical diagnosis (Kim et al., 2025), where output reliability is paramount. Consequently, developing robust hallucination detection mechanisms is imperative to facilitate rigorous confidence assessments of model outputs.

Various hallucination detection methods have been proposed, which can be broadly categorized into three paradigms: (1) **consistency-based approaches** (Kuhn et al., 2023), which assess reliability by quantifying agreement across multiple stochastic generations; (2) **external-knowledge-based approaches** (Dipta and Ferraro, 2025; Wanner et al., 2025; Upravitelev et al., 2025), which verify factual correctness by corroborating model outputs against retrieved external evidence; and (3) **internal-representation-based approaches** (Belinkov, 2022), which detect hallucinations by analyzing internal representations within LLMs (e.g., hidden states, MLP activations, and output logits). While consistency-based methods require multiple generations and external-knowledge approaches depend on retrieval infrastructure, internal-representation methods operate in a single forward pass without additional resources, making them particularly attractive for real-time deployment.

Existing internal-representation methods can be stratified into trained and training-free paradigms. Trained approaches typically employ probes (e.g., linear classifiers) operating on internal representations from either individual layers (Orgad et al., 2025) or naively aggregated combinations across all layers (Zhang et al., 2025c; Kadavath et al., 2022). Conversely, training-free techniques compute heuristic metrics directly from raw activations within single layers (Ren et al., 2022; Sriramanan

¹Code and evaluation results are available at: <https://github.com/Origamizyq/TemHal>

et al., 2024).

Despite these distinctions, both paradigms share a fundamental limitation: suboptimal utilization of layer-wise information. Recent mechanistic studies (Skean et al., 2025; Chen et al., 2025) indicate that factual knowledge in LLMs is non-uniformly distributed across layers, with discriminative signals concentrated in specific regions (particularly intermediate layers), rather than being uniformly distributed. This discrepancy renders current strategies inherently suboptimal: isolated single-layer analysis lacks sufficient representational capacity, while comprehensive aggregation across all layers introduces substantial noise from layers irrelevant to factual encoding. Moreover, existing methods treat layer-wise representations as static or independently pooled, neglecting the evolutionary dynamics of factual signals as they propagate through the transformer hierarchy. Consequently, such approaches constrain the overall efficacy of hallucination detection.

To address these limitations, we propose **TemHal** (**Temporal Hallucination Detector**), a novel framework that explicitly models the layer-wise activation trajectory of tokens corresponding to precise factual claims. Our method is grounded in two key insights: (1) hallucination signals are embedded in the trajectory of representations across layers rather than in any single static state, and (2) a focused window of intermediate layers (approximately one-third of total depth) provides an optimal signal-to-noise ratio for detection. Specifically, TemHal first identifies factual key tokens in the model’s response using an LLM-guided extraction strategy. For each such token, we construct a temporal sequence by stacking MLP outputs from a contiguous block of intermediate layers. This sequence captures the progressive refinement of factual signals throughout the inference process. We then leverage a streamlined temporal modeling network (a bidirectional LSTM with linear attention) to classify whether this activation trajectory corresponds to a hallucinated or factually grounded response.

Extensive empirical evaluations across three mainstream open-source LLMs (Llama-3.1, Mistral-v0.3, Qwen-3) and five benchmark datasets demonstrate that TemHal consistently surpasses prior methods, achieving state-of-the-art performance in 11 out of 15 evaluation settings. Furthermore, ablation studies validate our architectural design choices and elucidate the underlying

mechanisms, confirming that ordered cross-layer dependencies contribute beyond multi-layer feature availability and model capacity.

We summarize the core contributions of this work as follows:

- **Novel Formulation:** We propose treating activation trajectories across contiguous intermediate layers as temporal sequences, explicitly capturing cross-layer representational dynamics that prior static or pooled approaches overlook.
- **Methodological Framework:** We devise TemHal, a hallucination detection framework informed by mechanistic insights, leveraging bidirectional temporal modeling to trace the layer-wise evolution of factual tokens within a focused one-third layer window.
- **Empirical Validation:** We substantiate the effectiveness and generalizability of our approach through extensive evaluations across diverse model architectures and five benchmark datasets spanning knowledge-intensive QA and commonsense reasoning tasks.

2 Related Work

2.1 Hallucination

Hallucinations, defined as factually incorrect or contextually disconnected content (Huang et al., 2021; Ji et al., 2023), remain a critical challenge despite significant advancements in LLMs across reasoning (Wang et al., 2024; Ferrag et al., 2025), visual question answering (VQA) (Liu et al., 2024; Carolan et al., 2024), and summarization (Ravaut et al., 2024; Zhang et al., 2025a) (Huang et al., 2025; Li et al., 2025). Hallucinations pose significant risks in safety-critical domains such as medical diagnosis (Kim et al., 2025) and finance (Roychowdhury, 2024). Their origins are multifaceted, stemming from training data noise (Shao et al., 2025), error accumulation via maximum likelihood estimation (Aksitov et al., 2023), and instruction tuning that may prioritize adherence over accuracy (Han et al., 2025). Consequently, accurate detection remains a critical prerequisite for reliable deployment.

2.2 Hallucination Detection via Internal Representations

Detection methods have evolved from analyzing output logits to examining **internal repre-**

sentations. Early approaches utilized confidence scores (Kadavath et al., 2022; Lin et al., 2022) and uncertainty metrics (Ren et al., 2022; Sriraman et al., 2024), but these effectively compress multi-layer dynamics into single probability distributions, constraining detection granularity. Recent work shifts focus to internal states (Li et al., 2023), leveraging the functional specialization of Multi-Head Self-Attention (MHSA) for context integration (Elhage et al., 2021) and Feed-Forward Networks (FFN) for knowledge retrieval (Geva et al., 2023). Methodologies range from single-layer probes (Azaria and Mitchell, 2023; Orgad et al., 2025) to cross-layer aggregations (Sky et al., 2024; Zhang et al., 2025c,b; Suresh et al., 2025) and training-free metrics (Chen et al., 2024; Sriraman et al., 2024). However, critical limitations persist: single-layer analysis lacks sufficient representational capacity, while full-layer aggregation introduces substantial noise (Chen et al., 2025), hindering zero-shot generalization. This trade-off motivates our focused temporal modeling approach.

3 Preliminaries

Given a large language model F , we denote the dataset as $\mathcal{D} = \{(q_i, a_i)\}_{i=1}^N$, where q_i represents an input question and a_i denotes the corresponding ground-truth answer.

When interacting with the model, each question q_i elicits a generated response:

$$\hat{y}_i = F(q_i),$$

where $\hat{y}_i$ is the open-ended textual output produced by the model. We denote the token sequence of this response as $\{t_i^{(j)}\}_{j=1}^{T_i}$, where T_i is the length of the generated answer for the i -th query, and $t_i^{(j)}$ represents the j -th token.

To assess factual consistency, we assign a binary label $z_i \in \{0, 1\}$ to each instance, indicating whether $\hat{y}_i$ is factually correct ($z_i = 1$) or hallucinated ($z_i = 0$). Labels can be obtained through reference-based metrics (e.g., lexical overlap measures) or via LLM-as-a-judge methods that evaluate factual consistency between the generated answer and the ground truth (Zheng et al., 2023).

The core challenge in hallucination detection lies in predicting the label z_i *without access to the ground-truth answer* a_i during inference. Under this reference-free setting, the detector’s input consists solely of information extracted from the

model-question interaction—specifically, the textual pair $(q_i, \hat{y}_i)$ and/or the model’s internal representations. Formally, the detector implements a function:

$$\hat{z}_i = f(x_i; \theta),$$

where x_i denotes the input features (e.g., textual Q-A pairs, internal representations), θ represents learnable parameters, and $\hat{z}_i$ is the predicted label aiming to approximate the true label z_i . The central objective is thus to learn a function f that effectively captures the relationship between the model’s internal representations and the factual correctness of its response.

Within this paradigm, a prominent approach (known as *Probe*) trains a classifier (typically linear) on the model’s hidden states or activations to distinguish between hallucinated and factual outputs. This method has been shown to be among the most effective strategies for internal-signal-based hallucination detection (Belinkov, 2022).

4 Method

To address the limited generalization and coarse-grained nature of existing faithfulness detection approaches in knowledge-intensive question answering, we propose a novel method that leverages the *layer-wise temporal evolution* of internal representations across transformer layers. Instead of relying on static snapshots from a single layer or aggregated output embeddings, our approach tracks how fact-relevant semantic units are progressively refined through the model’s deep architecture. This enables the construction of a fine-grained, distribution-robust signal for factual consistency.

The proposed framework consists of three key components, as illustrated in Figure 1.

4.1 Fact-Critical Token Identification

Not all tokens in a generated sequence contribute equally to factual correctness. Discriminative signals for hallucination detection are concentrated in *exact answer tokens* (those encoding core semantic information, e.g., entities, attributes, numerical values) rather than auxiliary tokens serving grammatical or discourse functions. This distinction aligns with human fact-checking behavior, which prioritizes core factual claims over contextual elaboration.

Empirical evidence supports this focus. Orgad et al. (Orgad et al., 2025) demonstrate that probing ex-

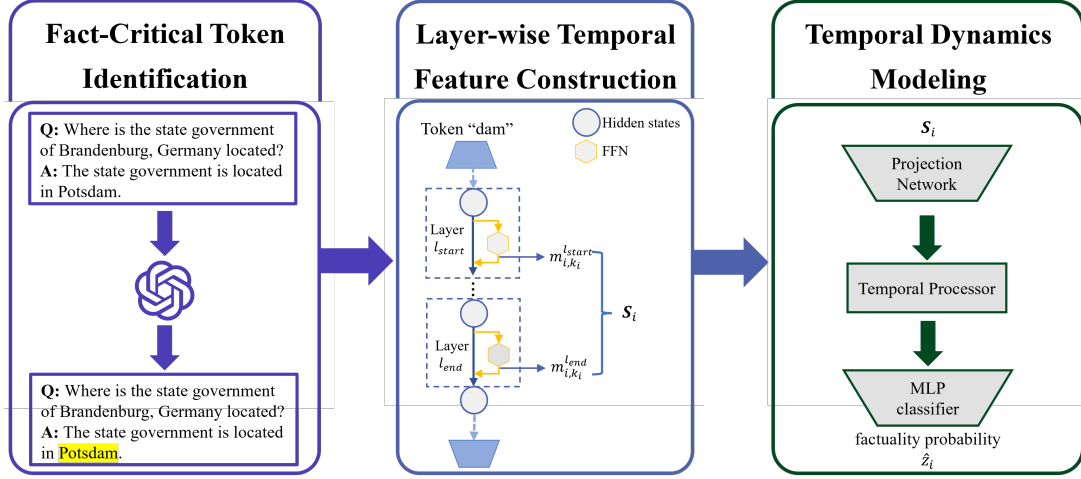


Figure 1: Overview of our framework. We extract exact answer tokens from the generated response and probe their MLP activations across selected transformer layers to capture the layer-wise evolution of factuality signals.

act answer tokens (defined as generated tokens semantically aligning with the ground-truth answer) yields superior detection signals compared to full-text or heuristic sampling strategies. This suggests that factual content induces distinguishable activation patterns in the model’s internal representation space.

Following this paradigm, we identify the core factual keywords through an LLM-guided extraction strategy. Our approach is primarily designed for scenarios with single-keyword answers. Given a question q_i and generated response $\hat{y}_i = F(q_i)$, our primary benchmark protocol follows Probe-Exact and uses the reference answer a_i to locate the corresponding factual keyword for correct generations. For incorrect generations or cases without a valid reference-aligned span, Mistral-7B-Instruct-v0.2 extracts the factual keyword using only q_i and $\hat{y}_i$. Following Probe-Exact, examples for which the extractor returns NO ANSWER or no valid span can be mapped to the generated response are excluded from exact-token probing. We further evaluate a strictly reference-free setting in which no gold answer is supplied for any generation, as detailed in Appendix B.1.

Crucially, for each extracted keyword, we **exclusively utilize its last token** for probing. We explicitly avoid using any intermediate or preceding tokens. In our primary scenario, the keyword is a single token, which naturally serves as its own last token. In the minority of cases where a keyword is segmented into multiple tokens by the tokenizer, we still strictly select only the terminal token, as it inherently aggregates the semantic representation

of the entire keyword (Luo et al., 2026).

Formally, let k_i denote the index of the last token corresponding to the extracted factual keyword in sample i . Our analysis centers strictly on this single target token t_{i,k_i}^{exact} , probing its MLP activations across a contiguous block of intermediate layers to characterize representational distinctions between factual and non-factual generation.

4.2 Layer-wise Temporal Feature Construction

Experiments in previous mechanistic analyses of Transformer-based models, ranging from BERT-base-based (Devlin et al., 2019) to the latest architectures (Dai et al., 2022; Skean et al., 2025; Chen et al., 2025), have shown that model knowledge is predominantly associated with intermediate layers, a finding further corroborated in Section 5.4. To capture the dynamic refinement of factual representations, we construct a sequence of MLP outputs corresponding to the single target token t_{i,k_i}^{exact} across a contiguous block of intermediate layers.

Specifically, let L be the total number of transformer layers in model F , where $1 \leq l_{\text{start}} \leq l_{\text{end}} \leq L$. Let $\mathbf{m}_{i,k_i}^{(l)} \in \mathbb{R}^d$ denote the MLP output at the target token position k_i in layer l . For the selected layer range $[l_{\text{start}}, l_{\text{end}}]$, we construct

$$\mathbf{S}_i = \text{Stack} \left(\mathbf{m}_{i,k_i}^{(l_{\text{start}})}, \dots, \mathbf{m}_{i,k_i}^{(l_{\text{end}})} \right) \in \mathbb{R}^{w \times d},$$

where $w = l_{\text{end}} - l_{\text{start}} + 1$ is the layer-window size and d is the hidden dimension.

We focus on MLP outputs based on the findings of Geva et al. (Geva et al., 2021), who demonstrated

through systematic probing that Feed-Forward Network (FFN) layers functionally operate as key-value memory units: the “keys” correspond to specific patterns in the input context, while the “values” encode the associated knowledge representations. Consequently, modeling MLP outputs offers a more direct lens for capturing the dynamic evolution of factual knowledge. This sequence encodes how the representation of the critical token evolves as it propagates through deeper processing stages, capturing stages of activation, integration with contextual knowledge, and potential distortion due to hallucination.

4.3 Temporal Dynamics Modeling

Given the layer-wise MLP activation sequence $\mathbf{S}_i \in \mathbb{R}^{w \times d}$, we design a temporal discriminator to predict the **instance-level factuality** of each response from the representation of its single target token. As illustrated in Figure 1, the discriminator processes $\mathbf{S}_i$ through three stages: (1) dimensionality projection, (2) temporal encoding via a Temporal Processor, and (3) classification.

The input sequence is first projected into a lower-dimensional space $\bar{\mathbf{S}}_i \in \mathbb{R}^{w \times d'}$ via a learnable projection network. The Temporal Processor then models the layer-wise dynamics:

$$\mathbf{e}_{\text{temp},i} = \text{TP}(\bar{\mathbf{S}}_i), \quad (1)$$

where $\mathbf{e}_{\text{temp},i} \in \mathbb{R}^{2d_{\text{hidden}}}$ is the aggregated temporal embedding. Finally, the classifier produces the instance-level factuality probability:

$$\hat{z}_i = \sigma(\text{MLP}(\mathbf{e}_{\text{temp},i})), \quad (2)$$

where $\sigma(\cdot)$ is the sigmoid function and $\hat{z}_i \in [0, 1]$ estimates the probability that the generated response is factually correct.

The discriminator is trained with binary cross-entropy loss using the instance-level ground-truth label $z_i \in \{0, 1\}$, where $z_i = 1$ indicates factual consistency and $z_i = 0$ denotes hallucination:

$$\mathcal{L}_{\text{BCE}} = -\frac{1}{N} \sum_{i=1}^N [z_i \log \hat{z}_i + (1 - z_i) \log(1 - \hat{z}_i)]. \quad (3)$$

No token-level labels are introduced. Since our primary scenario focuses on single-keyword answers, each sample contains exactly one fact-critical token. Consequently, the instance-level prediction $\hat{z}_i$ directly serves as the final response-level factuality score, eliminating the need for aggregation across multiple tokens.

4.4 Architecture of the Temporal Processor

The Temporal Processor (TP) is the core component that captures the evolutionary dynamics of factual representations across transformer layers. We employ a stacked Bidirectional LSTM (Bi-LSTM) (Graves, 2012) followed by an attention pooling mechanism, as illustrated in Figure 2. This architectural design is motivated by our empirical findings that hallucination signals are better captured through layer-wise temporal dynamics than static representations.

Bidirectional Temporal Encoding. The Bi-LSTM processes the ordered layer sequence in both the forward (lower-to-higher layer) and backward (higher-to-lower layer) directions, allowing the representation at each position to incorporate information from both sides of the selected layer sequence. Given the projected sequence

$$\bar{\mathbf{S}}_i = \left\{ \bar{\mathbf{m}}_{i,k_i}^{(l)} \right\}_{l=l_{\text{start}}}^{l_{\text{end}}},$$

the Bi-LSTM computes forward and backward hidden states $\{\vec{\mathbf{h}}_{i,l}\}$ and $\{\overleftarrow{\mathbf{h}}_{i,l}\}$, which are concatenated as

$$\mathbf{h}_{i,l} = \left[\vec{\mathbf{h}}_{i,l}; \overleftarrow{\mathbf{h}}_{i,l} \right] \in \mathbb{R}^{2d_{\text{hidden}}}.$$

Our ablations show that Bi-LSTM, RNN, and Transformer processors achieve comparable performance, suggesting that capturing ordered cross-layer dependencies is more important than the specific processor architecture. We select Bi-LSTM because it achieves the highest numerical AUROC while using substantially fewer parameters than the Transformer variant.

Attention Pooling. To adaptively focus on critical layers where factual signals stabilize or distort, we apply an attention pooling mechanism instead of uniform averaging. The attention weights $\alpha_{i,l}$ are computed via a linear projection followed by softmax normalization over the sequence dimension:

$$\alpha_{i,l} = \frac{\exp(\mathbf{w}_a^T \mathbf{h}_{i,l})}{\sum_{j=l_{\text{start}}}^{l_{\text{end}}} \exp(\mathbf{w}_a^T \mathbf{h}_{i,j})}, \quad (4)$$

where $\mathbf{w}_a \in \mathbb{R}^{2d_{\text{hidden}}}$ is a learnable attention vector. The final temporal embedding is obtained via weighted sum:

$$\mathbf{e}_{\text{temp},i} = \sum_{l=l_{\text{start}}}^{l_{\text{end}}} \alpha_{i,l} \mathbf{h}_{i,l}. \quad (5)$$

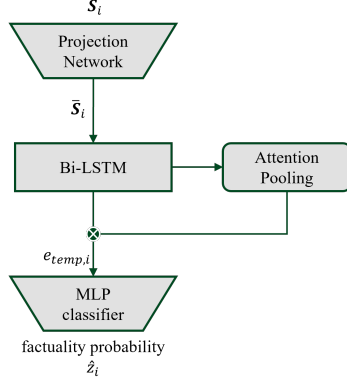


Figure 2: Architecture of the Factuality Discriminator. The model processes layer-wise MLP activation sequences through a projection network, Bi-LSTM encoder, attention pooling mechanism, and MLP classifier to predict factuality probability.

5 Experiment

5.1 Settings

Models. We evaluate our method on three instruction-tuned LLM families: **Llama-3.1-8B-Instruct** (AI@Meta, 2024) (hereafter *Llama-3.1*), **Mistral-7B-Instruct-v0.3** (Jiang et al., 2023) (hereafter *Mistral-v0.3*), and **Qwen-3-8B-Instruct** (Yang et al., 2025) (hereafter *Qwen-3*). Detailed specifications are provided in Appendix A.1.

Datasets. We evaluate our approach across five datasets spanning diverse domains and reasoning tasks: **HotpotQA** and its with-context variant **HotpotQA-WC** (Yang et al., 2018), **Natural Questions** (Kwiatkowski et al., 2019), **TriviaQA** (Joshi et al., 2017), and **Winogrande** (Sakaguchi et al., 2021). Following established practices (Orgad et al., 2025), we employ dataset-appropriate labeling strategies based on domain characteristics; details on labeling protocols, dataset statistics, prompt templates, and hallucination rates are provided in Appendix A.2.

Baselines. We compare TemHal against representative hallucination detection baselines spanning three categories: (1) **classic generation probability-based methods**, including **Perplexity** (Ren et al., 2022) and **P(True)** (Kadavath et al., 2022), which estimate factual consistency directly from output logits or generation probabilities; (2) **training-free internal representation methods**, primarily **LLM-Check** (Sriramanan et al., 2024), for which we adopt the attention-based modeling

variant in our main comparisons due to its superior preliminary performance (results for other variants are provided in Table 7); and (3) **trainable internal representation methods**, including **Probe-Exact** (Orgad et al., 2025), which probes single-layer internal representations, **ICR** (Zhang et al., 2025c), which jointly models representations across all transformer layers, **MHAD** (Zhang et al., 2025b), which selects hallucination-aware neurons and layers, and **CLAP** (Suresh et al., 2025), which models cross-layer representations from the residual stream. Implementation details for all baselines are provided in Appendix A.3.

Evaluation Metric. We adopt the area under the receiver operating characteristic curve (AUROC) as our primary metric, a threshold-independent measure robust across varying decision thresholds. Each experiment is run with 5 random seeds (0, 5, 26, 42, 63).

5.2 Hallucination Detection Results

Table 1 summarizes the AUROC results of TemHal against all evaluated baselines across five benchmark datasets and three mainstream open-source LLMs. TemHal achieves the best performance in 11 out of 15 settings and obtains the highest average AUROC across all three model families, outperforming the strongest competing method by 0.0137–0.0444 AUROC in terms of model-level average performance.

Beyond detection accuracy, TemHal maintains strong performance across the three evaluated model families (though the detector is trained separately for each backbone), extracts more discriminative signals than single-layer probing, static feature selection, or alternative cross-layer aggregations, and employs a fixed intermediate one-third layer configuration determined through ablation studies (Section 5.4), avoiding exhaustive layer-wise search.

TemHal also remains effective when fact-critical tokens are extracted without gold answers, outperforming Probe-Exact by 0.0565 AUROC on average under the strictly reference-free protocol (Appendix B.1).

5.3 Generalization Analysis

We assess cross-dataset generalization by training on one dataset and evaluating on held-out benchmarks within each LLM family. Figure 3 illustrates AUROC differences between TemHal and Probe-

Table 1: Performance comparison of hallucination detection methods across different LLMs and benchmark datasets. Bold indicates the best result, underlined indicates the second-best.

LLMs	Methods	HotpotQA	HotpotQA-WC	Natural Questions	TriviaQA	Winogrande
Llama-3.1	PPL	0.5061	0.6286	0.6154	0.6803	0.5910
	P(True)	0.8215	0.5602	0.7341	0.7760	0.5985
	LLM-check	0.7596	0.5416	0.6499	0.5749	0.5894
	ICR	0.8609	0.5819	0.7510	0.7809	0.6181
	MHAD	0.8083	0.7059	<u>0.8224</u>	<u>0.8082</u>	0.7604
	CLAP	0.8173	0.7058	0.8226	0.8071	0.7642
	Probe-exact	0.8259	<u>0.7441</u>	0.7762	0.7996	<u>0.8715</u>
	TemHal (Ours)	<u>0.8375</u>	0.7679	0.8131	0.8349	0.8869
Mistral-v0.3	PPL	0.5747	0.5393	0.6188	0.5613	0.4833
	P(True)	0.6280	0.6411	0.6852	0.7019	0.5964
	LLM-check	0.5950	0.5490	0.5075	0.5934	0.6081
	ICR	0.6505	0.5758	0.6637	0.6422	0.6039
	MHAD	0.7923	0.6477	0.7624	0.7992	0.7228
	CLAP	0.7951	0.6517	<u>0.7761</u>	0.8146	0.7393
	Probe-exact	<u>0.8396</u>	<u>0.7674</u>	0.7719	<u>0.8537</u>	0.7202
	TemHal (Ours)	0.8506	0.7869	0.7901	0.8680	<u>0.7255</u>
Qwen-3	PPL	0.6211	0.6369	0.5843	0.6336	0.5873
	P(True)	0.5673	0.5857	0.4828	0.5479	0.6035
	LLM-check	0.5246	0.5756	0.6402	0.5635	0.6201
	ICR	0.6765	0.5549	0.8038	0.6708	0.6450
	MHAD	<u>0.8298</u>	0.6909	0.8733	<u>0.8366</u>	0.7633
	CLAP	0.8157	0.6772	<u>0.8571</u>	0.8266	0.7511
	Probe-exact	0.8222	<u>0.7315</u>	0.8116	0.8252	<u>0.7727</u>
	TemHal (Ours)	0.8812	0.7684	0.8462	0.8969	0.8234

Exact, where positive values favor our method; complete raw scores are provided in Appendix B.2.

TemHal outperforms Probe-Exact in most transfer settings on Llama-3.1 and Mistral-v0.3. Generalization is particularly effective within similar task categories (e.g., factual-to-factual and reasoning-to-reasoning), suggesting that temporal dynamics transfer more effectively between tasks with related cognitive demands.

For Qwen-3, certain cross-category transfers degrade: under substantial domain shift, the baseline’s single-layer features tend to remain near chance level, whereas TemHal’s temporal signals, being more sensitive to task-specific dynamics, may occasionally invert below chance. This highlights a trade-off: multi-layer temporal modeling provides stronger discriminative signals within related task categories but can be more sensitive to domain shift than static single-layer probes. Overall, TemHal demonstrates robust cross-dataset generalization, particularly when source and target tasks share similar characteristics.

5.4 Ablation Study

We ablate TemHal on Qwen-3-8B-Instruct, focusing on layer-range sensitivity and core temporal modeling components.

Sensitivity to Layer Range. Previous work indicates non-uniform knowledge distribution across transformer layers (Chen et al., 2025). Table 2 compares early, middle, and late layer segments: intermediate layers generally yield superior performance, with our selected window (layers 13–24) achieving the best results on 4 out of 5 benchmarks. This advantage is particularly pronounced on reasoning-intensive tasks, where the intermediate window substantially outperforms late layers, suggesting that mid-level layers capture syntactic-to-semantic transitions.

For knowledge-intensive tasks (e.g., Natural Questions and TriviaQA), intermediate layers remain optimal, although late layers are competitive but slightly lower. Using all layers consistently performs worse than the focused window, suggesting that less informative layers dilute the factuality signals. We select a one-third layer window to optimize signal discriminability. Fine-grained tuning confirms stability within the 10–16 layer range around the middle, indicating that the exact boundary is not overly sensitive; although a 16-layer window shows marginal gains on HotpotQA-WC, the 12-layer configuration offers the best overall balance and is adopted as the default.

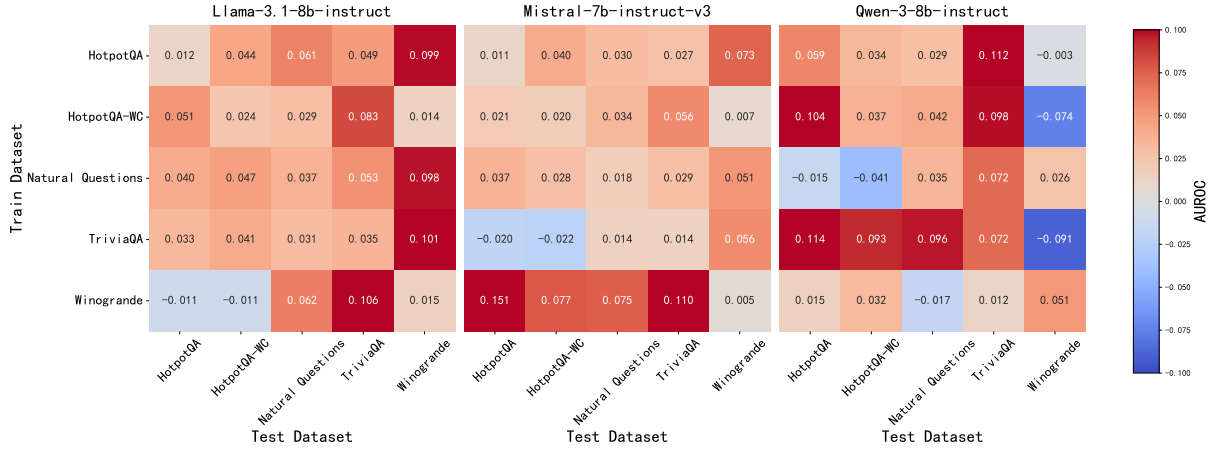


Figure 3: Cross-dataset generalization performance across three LLM families. Each cell shows the AUROC difference (ours – Probe-Exact) for the corresponding train–test pair. Red indicates improvement, while blue indicates degradation relative to the baseline.

Table 2: Ablation on layer range selection (AUROC).

Layer Range	HotpotQA	HotpotQA-WC	Natural Questions	TriviaQA	Winogrande
ours (12 layers [13-24])	0.8812	0.7684	0.8462	0.8969	0.8234
All layers [1-36]	0.8659	0.7520	0.8383	0.8838	0.8002
<i>Position Selection (1/3 Split)</i>					
Early [1-12]	0.8216	0.6477	0.7595	0.7923	0.6461
Late [25-36]	0.8546	0.7537	0.8258	0.8601	0.7371
<i>Range Size Tuning (Around Mid Layers)</i>					
8 layers [15-22]	0.8791	0.7684	0.8424	0.8957	0.8172
10 layers [14-23]	<u>0.8796</u>	0.7685	0.8452	<u>0.8966</u>	<u>0.8162</u>
14 layers [12-25]	0.8767	<u>0.7704</u>	<u>0.8460</u>	0.8939	0.8118
16 layers [11-26]	0.8753	0.7705	0.8454	0.8938	0.8129

Contribution of Core Components. The original single-layer ablation changes both the number of input layers and the model capacity, making it insufficient to isolate the effect of temporal ordering. We therefore introduce four controlled baselines that use the same 12 intermediate-layer representations (layers 13–24) as TemHal. Mean pooling, attention pooling, and concatenation aggregate these representations without explicitly modeling layer-to-layer transitions. Random Permutation retains the same layer features and identical Bi-LSTM architecture as TemHal but disrupts their original order. All variants use capacity-controlled embedding, projection, and classification heads and are trained independently under the same protocol.

As shown in Table 3, TemHal numerically outperforms all four controls by 0.0095–0.0176 AUROC. In particular, Random Permutation uses the same 12 representations and the same Bi-LSTM architecture as TemHal, differing only in whether the original layer order is preserved. Its performance

Table 3: Controlled comparison with non-sequential or order-disrupted aggregation methods on Winogrande. All methods use representations from layers 13–24. Values are mean AUROC scores over five random seeds.

Method	AUROC	Δ vs. TemHal
Mean Pooling	0.8115	−0.0119
Attention Pooling	0.8063	−0.0171
Random Permutation	0.8058	−0.0176
Concatenation	0.8139	−0.0095
TemHal	0.8234	–

drop provides direct evidence that ordered cross-layer dependencies contribute beyond multi-layer feature availability and model capacity.

Table 4 reports the remaining component and architecture ablations; the single-layer MLP is retained as a compact reference rather than primary evidence, since it changes both input features and model capacity.

Bi-LSTM, RNN, and Transformer processors achieve relatively similar AUROC, suggesting that

Table 4: Ablation study on core components on Winogrande. Parameter counts are reported for the temporal-processor comparison. Values are mean AUROC scores over five random seeds.

Variant	AUROC	Δ vs. Full	Params.
TemHal	0.8234	–	$\sim 11.3\text{M}$
Transformer-based TP	0.8183	–0.0051	$\sim 47.5\text{M}$
RNN-based TP	0.8197	–0.0037	$\sim 9.2\text{M}$
w/o Attention	0.8124	–0.0110	–
Single-layer MLP	0.7727	–0.0507	–
Uni-LSTM TP	0.7901	–0.0333	–

modeling ordered cross-layer dependencies is more important than the specific processor architecture. Performance is also not determined by parameter count alone: the Transformer-based processor contains approximately $4.2\times$ as many parameters as the Bi-LSTM but achieves a slightly lower AUROC. The RNN provides a competitive lightweight alternative, while we select Bi-LSTM as the default for its highest numerical AUROC with substantially fewer parameters.

Removing bidirectionality reduces AUROC by 0.0333, suggesting that information from both lower and higher layers is useful for modeling the selected layer sequence. Removing attention leads to a decrease of 0.0110 on Winogrande, indicating that adaptive layer weighting provides an additional benefit in this setting. Overall, these results confirm the effectiveness of ordered cross-layer modeling and support Bi-LSTM as an efficient default architecture.

6 Conclusion

We presented TemHal, a novel hallucination detection framework that explicitly models the temporal activation trajectories of factual tokens across intermediate transformer layers. Motivated by mechanistic insights that factual knowledge evolves non-uniformly across the transformer hierarchy, our approach captures the dynamic refinement of representations during inference, contrasting with prior static or single-layer probing methods that obscure discriminative signals embedded in representational evolution. Extensive evaluations across three LLM families and five benchmarks demonstrate TemHal achieves state-of-the-art performance in the vast majority of settings, with particular strengths in reasoning tasks and cross-dataset generalization within task categories. By leveraging a focused intermediate-layer window, TemHal maintains robust performance across di-

verse tasks and evaluated model families while avoiding exhaustive layer-wise search. Our findings underscore temporal dynamics in internal representations as a discriminative signal for factuality, suggesting that future detection frameworks may benefit from explicitly modeling information-flow dynamics and aligning strategies with the characteristics of target tasks.

Limitations

Despite its strong performance, TemHal has several limitations. First, it relies on accessing layer-wise internal representations, restricting deployment to models with exposed internals; developing black-box-compatible variants is left to future work. Second, the default layer window was empirically optimized; models with substantially different depths may require proportional adjustment, motivating automated layer selection. Third, although TemHal generalizes effectively within task categories, cross-category transfer under extreme domain shift can be sensitive, as temporal signals may adapt to task-specific dynamics; task-invariant temporal representations or adaptive calibration mechanisms are potential solutions. Finally, complex reasoning scenarios (e.g., multi-step mathematical proofs, chain-of-thought verification, or logical deduction with intermediate fact-checking) remain underexplored; adapting temporal dynamics modeling to reasoning-specific evolutionary patterns is a promising direction.

Ethical Considerations

Our methodology does not rely on personal or sensitive data: all experiments use publicly available benchmarks (HotpotQA, Natural Questions, TriviaQA, Winogrande) curated to exclude personally identifiable information and offensive content, and we follow original dataset licenses without additional data collection. AI assistants were used solely for text polishing; all core research contributions were independently conducted by the authors. We recognize potential misuse risks and commit to responsible AI deployment.

Acknowledgments

This work was supported by the Open Fund of the National Key Laboratory of Parallel and Distributed Computing (PDL) under Grant No. 2024KJWPDL-02.

References

- AI@Meta. 2024. [Llama 3 model card](#).
- Renat Aksitov, Chung-Ching Chang, David Reitter, Siamak Shakeri, and Yunhsuan Sung. 2023. Characterizing attribution and fluency tradeoffs for retrieval-augmented large language models. *arXiv preprint arXiv:2302.05578*.
- Amos Azaria and Tom Mitchell. 2023. The internal state of an llm knows when it’s lying. *arXiv preprint arXiv:2304.13734*.
- Yonatan Belinkov. 2022. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.
- Kilian Carolan, Laura Fennelly, and Alan F Smeaton. 2024. A review of multi-modal large language and vision models. *arXiv preprint arXiv:2404.01322*.
- 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, Zehao Li, Shuangjie You, Zhengyu Chen, Jingwen Chang, Yi Zhang, Weinan Dai, Qingpei Guo, and Yanghua Xiao. 2025. Attributive reasoning for hallucination diagnosis of large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 23660–23668.
- Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.
- Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. 2022. Knowledge neurons in pretrained transformers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8493–8502.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.
- Shubhashis Roy Dipta and Francis Ferraro. 2025. If we may de-presuppose: Robustly verifying claims through presupposition-free question decomposition. In *Proceedings of the 14th Joint Conference on Lexical and Computational Semantics (* SEM 2025)*, pages 253–266.
- Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, et al. 2021. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 1(1):12.
- Mohamed Amine Ferrag, Norbert Tihanyi, and Merouane Debbah. 2025. From llm reasoning to autonomous ai agents: A comprehensive review. *arXiv preprint arXiv:2504.19678*.
- Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. 2023. Dissecting recall of factual associations in auto-regressive language models. *arXiv preprint arXiv:2304.14767*.
- 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.
- Alex Graves. 2012. Long short-term memory. *Supervised sequence labelling with recurrent neural networks*, pages 37–45.
- Kyubeen Han, Junseo Jang, Hongjin Kim, Geunyeong Jeong, and Harksoo Kim. 2025. Exploring the impact of instruction-tuning on llm’s susceptibility to misinformation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 26711–26731.
- Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. 2022. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*.
- Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55.
- Yichong Huang, Xiachong Feng, Xiaocheng Feng, and Bing Qin. 2021. The factual inconsistency problem in abstractive text summarization: A survey. *arXiv preprint arXiv:2104.14839*.
- 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 computing surveys*, 55(12):1–38.
- Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego

- de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L  lio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timoth  e Lacroix, and William El Sayed. 2023. *Mistral 7b*. *Preprint*, arXiv:2310.06825.
- Mandar Joshi, Eunsol Choi, Daniel S 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.
- Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.
- Yubin Kim, Hyewon Jeong, Shan Chen, Shuyue Stella Li, Chanwoo Park, Mingyu Lu, Kumail Alhamoud, Jimin Mun, Cristina Grau, Minseok Jung, et al. 2025. Medical hallucinations in foundation models and their impact on healthcare. *arXiv preprint arXiv:2503.05777*.
- Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. *arXiv preprint arXiv:2302.09664*.
- Tom Kwi  tkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.
- Chaozhuo Li, Pengbo Wang, Chenxu Wang, Litian Zhang, Zheng Liu, Qiwei Ye, Yuanbo Xu, Feiran Huang, Xi Zhang, and Philip S Yu. 2025. Loki’s dance of illusions: A comprehensive survey of hallucination in large language models. *arXiv preprint arXiv:2507.02870*.
- Kenneth Li, Oam Patel, Fernanda Vi  gas, 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. Teaching models to express their uncertainty in words. *arXiv preprint arXiv:2205.14334*.
- Yixin Liu, Kai Zhang, Yuan Li, Zhiling Yan, Chujie Gao, Ruoxi Chen, Zhengqing Yuan, Yue Huang, Hanchi Sun, Jianfeng Gao, et al. 2024. Sora: A review on background, technology, limitations, and opportunities of large vision models. *arXiv preprint arXiv:2402.17177*.
- Wen Luo, Guangyue Peng, Wei Li, Shaohang Wei, Feifan Song, Liang Wang, Nan Yang, Xingxing Zhang, Jing Jin, Furu Wei, et al. 2026. Two pathways to truthfulness: On the intrinsic encoding of llm hallucinations. *arXiv preprint arXiv:2601.07422*.
- Audrey Moores and V  nia G Zuin Zeidler. 2025. Don’t let generative ai shape how we see chemistry. *Nature Reviews Chemistry*, 9(10):649–650.
- 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 *International Conference on Learning Representations*, volume 2025, pages 66880–66913.
- Mathieu Ravaut, Aixin Sun, Nancy Chen, and Shafiq Joty. 2024. On context utilization in summarization with large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2764–2781.
- 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. *arXiv preprint arXiv:2209.15558*.
- Sohini Roychowdhury. 2024. Journey of hallucination-minimized generative ai solutions for financial decision makers. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, pages 1180–1181.
- Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. Winogrande: An adversarial winograd schema challenge at scale. *Communications of the ACM*, 64(9):99–106.
- Jiandong Shao, Yao Lu, and Jianfei Yang. 2025. Benford’s curse: Tracing digit bias to numerical hallucination in llms. *arXiv preprint arXiv:2506.01734*.
- Oscar Skea, Md Rifat Arefin, Dan Zhao, Niket Patel, Jalal Naghiyev, Yann LeCun, and Ravid Shwartz-Ziv. 2025. Layer by layer: Uncovering hidden representations in language models. *arXiv preprint arXiv:2502.02013*.
- CH-Wang Sky, 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.
- 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. *Advances in neural information processing systems*, 37:34188–34216.

- Malavika Suresh, Rahaf Aljundi, Ikechukwu Nkisi-Orji, and Nirmalie Wiratunga. 2025. Cross-layer attention probing for fine-grained hallucination detection. *arXiv preprint arXiv:2509.09700*.
- Max Upravitelev, Premtim Sahitaj, Arthur Hilbert, Veronika Solopova, Jing Yang, Nils Feldhus, Tatiana Anikina, Simon Ostermann, and Vera Schmitt. 2025. Exploring semantic filtering heuristics for efficient claim verification. In *Proceedings of the Eighth Fact Extraction and VERification Workshop (FEVER)*, pages 229–237.
- Qineng Wang, Zihao Wang, Ying Su, Hanghang Tong, and Yangqiu Song. 2024. Rethinking the bounds of llm reasoning: Are multi-agent discussions the key? *arXiv preprint arXiv:2402.18272*.
- Miriam Wanner, Benjamin Van Durme, and Mark Dredze. 2025. Dndscore: Decontextualization and decomposition for factuality verification in long-form text generation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 23620–23637.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.
- 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.
- Haopeng Zhang, Philip S Yu, and Jiawei Zhang. 2025a. A systematic survey of text summarization: From statistical methods to large language models. *ACM Computing Surveys*, 57(11):1–41.
- Luan Zhang, Dandan Song, Zhijing Wu, Yuhang Tian, Changzhi Zhou, Jing Xu, Ziyi Yang, and Shuhao Zhang. 2025b. Detecting hallucination in large language models through deep internal representation analysis. In *IJCAI*, pages 8357–8365.
- Zhenliang Zhang, Xinyu Hu, Huixuan Zhang, Junzhe Zhang, and Xiaojun Wan. 2025c. Icr probe: Tracking hidden state dynamics for reliable hallucination detection in llms. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 17986–18002.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhaghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.

A Experimental Details

A.1 Model Specifications

Table 5 summarizes the three LLMs evaluated in this work. All models are decoder-only transformers with instruction-tuned checkpoints publicly available via Hugging Face.

Inference Configuration. For all models, we use standard sampling with Temperature = 1.0 and Top-P = 1.0 to maintain generation diversity. To ensure reproducibility, we fix the random seed for all experiments. The maximum generation length is set to 1024 tokens. We use the Hugging Face transformers library with output_hidden_states=True to obtain residual-stream hidden states for baselines that require them. For TemHal, following Probe-Exact, we register forward hooks on the MLP submodule of each transformer block and directly record its output at the selected fact-critical token position.

A.2 Dataset Details

Table 6 provides detailed statistics for the five evaluation datasets used in our experiments, including domain, task type, and sample counts for both training and testing splits. Below we describe each dataset in detail.

HotpotQA (Yang et al., 2018). HotpotQA is a benchmark designed for diverse multi-hop question answering, originally comprising Wikipedia documents to facilitate reasoning. In our experiments, we employ two distinct settings to evaluate factual consistency under varying conditions: (1) **HotpotQA** (without-context variant), where questions are presented directly without accompanying documents, requiring the model to leverage both parametric factual knowledge and multi-step reasoning capabilities—distinguishing it from datasets like TriviaQA that primarily involve knowledge retrieval; and (2) **HotpotQA-WC** (with-context variant), where original Wikipedia documents are provided alongside questions. This latter setting preserves the dataset’s intended retrieval-augmented structure and evaluates the model’s ability to adhere to and utilize provided contextual information when external grounding is available.

Natural Questions (Kwiatkowski et al., 2019). Natural Questions (NQ) is a large-scale corpus designed for training and evaluating open-domain question answering systems. It is the first dataset to replicate the end-to-end process of humans seeking

Table 5: Specifications of evaluated LLMs.

Model	Organization	Hugging Face Repository
Llama-3.1-8B-Instruct	Meta	meta-llama/Llama-3.1-8B-Instruct
Mistral-7B-Instruct-v0.3	Mistral AI	mistralai/Mistral-7B-Instruct-v0.3
Qwen3-8B	Qwen	Qwen/Qwen3-8B

answers to questions: real, anonymized queries submitted by users to Google Search, with answers extracted from corresponding Wikipedia pages. Unlike artificially curated QA benchmarks, NQ reflects authentic information-seeking behaviors and natural language patterns encountered in practical deployment, offering higher ecological validity.

TriviaQA (Joshi et al., 2017). TriviaQA is a collection of trivia question-answer pairs. Questions are presented without any supporting context, requiring the model to rely solely on its internal parametric knowledge. The dataset includes multiple acceptable answer variations, enabling robust automatic evaluation.

Winogrande (Sakaguchi et al., 2021). Winogrande is a large-scale commonsense reasoning dataset designed to explore errors in commonsense reasoning. It consists of Winograd-style coreference challenges, where each example presents a sentence containing two entities and a pronoun. The objective is to determine which entity the pronoun refers to, relying on commonsense reasoning. For instance, in the sentence "The trophy doesn't fit into the suitcase because it's too large," the pronoun "it" refers to the trophy, not the suitcase.

Train/Test Split. Each dataset is randomly partitioned into 10,000 training samples and 10,000 testing samples via simple random sampling. As shown in Table 6, hallucination rates exhibit notable patterns across datasets and models.

Prompt Templates. We adopt dataset-specific prompt templates to elicit free-form answers. All prompts follow a unified structure: [Instruction] + [Question] + [Optional Context] + [Answer Prefix]. For reproducibility, the complete prompt templates for all datasets, along with implementation details, are publicly available in our code repository.

Labeling Strategy. Following Probe-Exact (Orad et al., 2025), we assign an instance-level binary label $z_i \in \{0, 1\}$ to each generated response, where $z_i = 1$ denotes a factually correct response

and $z_i = 0$ denotes a hallucinated response. The selected fact-critical token inherits the label of its corresponding response.

For HotpotQA, HotpotQA-WC, TriviaQA, and WinoGrande, a generation is labeled as factual if its normalized answer span matches an accepted reference answer; otherwise, it is labeled as hallucinated. For Natural Questions, we use Mistral-7B-Instruct-v0.2 as an automatic verifier to assess factual consistency given the question, generated response, and reference answer.

A.3 Baseline Implementation Details

Perplexity (Ren et al., 2022). Computed as the exponential of the average negative log-likelihood over all tokens in the generated sequence. Lower perplexity indicates higher confidence in factual consistency.

P(True) (Kadavath et al., 2022). We query the model with the following prompt:

```
Question: {question}
Possible answer: {answer}
Is the possible answer:
(A) False
(B) True
The possible answer is:
```

Following the original implementation, we extract the logits of the option tokens "A" and "B" at the first generation position, then apply a softmax normalization over these two logits to obtain a calibrated probability distribution. The probability of the "True" option (i.e., token "B") serves as the factual confidence score. This approach leverages the model's self-assessment capability while mitigating token-frequency bias through pairwise normalization.

LLM-Check (Sriramanan et al., 2024). The original work introduces three training-free detection methods operating on single-layer representations: *Window Entropy*, *Hidden Score (HS)*, and *Attention Score (AS)*. We evaluate HS and AS across all layers of each assessed LLM and report their optimal validation-selected performance in Table 7, with the corresponding layer indices indicated in parentheses. Window Entropy is evaluated follow-

Table 6: Hallucination rates (%) across different LLM architectures. For each dataset, we randomly sample **10,000 training** and **10,000 testing** instances from the original benchmark.

Dataset	Task	Qwen-3		Llama-3.1		Mistral-v0.3	
		Train	Test	Train	Test	Train	Test
HotpotQA	Generation	64.58	67.69	73.51	77.62	58.70	63.47
HotpotQA-WC	Generation	13.12	32.60	17.23	35.90	19.94	36.20
Natural Questions	Generation	34.47	34.15	35.25	34.84	39.37	38.49
TriviaQA	Generation	41.79	42.65	35.55	35.18	32.51	32.54
Winogrande	Multiple-choice	35.55	35.55	28.40	28.71	28.25	26.84

ing its original configuration. Although no single variant consistently dominates across all model–dataset settings, AS achieves the highest average AUROC across the 15 settings (0.5928), compared with 0.5864 for HS and 0.5670 for Window Entropy. We therefore use AS as the fixed representative of LLM-Check in the main experiments. All implementations follow the official repository provided by the original authors².

Probe-Exact (Orgad et al., 2025). Following the original work, we train a linear probe on the MLP output at the fact-critical token position. We conduct a layer-wise search across the full depth of each backbone and select the optimal layer based on validation AUROC. Our feature-extraction and token-selection procedures follow the public code-base³. To maintain a unified optimization protocol across trainable baselines, we implement the linear classifier in PyTorch and optimize it using AdamW, with early stopping on a held-out validation set comprising 20% of the training data.

ICR (Zhang et al., 2025c). We follow ICR Probe, which computes an Information Contribution to Residual Stream (ICR) score at each token and layer. Specifically, the score is defined as the Jensen–Shannon divergence between the pooled attention distribution over selected context positions and a distribution derived from the alignment of the residual hidden-state update with the corresponding attended hidden states. The resulting token-by-layer score matrix is averaged over tokens to form an L -dimensional cross-layer feature vector, which is processed by an MLP with dimensions $(L, 128, 64, 32, 1)$. Following the original implementation, we train the probe using binary cross-entropy loss and Adam. We refer readers to the official code repository⁴ for complete architec-

tural details and implementation.

MHAD (Zhang et al., 2025b). The original work introduces three variants based on different types of internal representations: *MHAD-AO*, *MHAD-FO*, and *MHAD-LO*, which respectively use attention outputs, feed-forward network outputs, and layer outputs. Their complete results are reported in Table 9. Although the three variants achieve comparable performance overall, MHAD-LO exhibits a modest but more consistent advantage, achieving the highest AUROC among the MHAD variants in 8 out of 15 evaluation settings. It also obtains the highest average AUROC of 0.7749, compared with 0.7716 for MHAD-FO and 0.7664 for MHAD-AO. We therefore use MHAD-LO as the fixed representative of MHAD in our main experiments. Because the project repository accompanying MHAD currently does not provide implementation code, we independently reimplement the three variants following the architecture, neuron-selection, layer-selection, and classification procedures described in the original paper. Our reimplementation will be released together with TemHal in our public GitHub repository.

CLAP (Suresh et al., 2025). CLAP processes residual-stream representations across the entire model depth as an ordered joint sequence and applies a Transformer-based probe for hallucination detection. We use the implementation released by the authors⁵ and adapt it to our evaluated model backbones and datasets. CLAP is evaluated using the same data splits and evaluation protocol as TemHal.

²https://github.com/GaurangSriraman/LLM_Check_Hallucination_Detection

³<https://github.com/technion-cs-nlp/LLMsKnow>

⁴<https://github.com/XavierZhang2002/ICR-Probe>

⁵<https://github.com/itsmemala/CLAP>

Table 7: Comparison of LLM-Check variants (Window Entropy, HS, AS) across LLMs and benchmarks (AUROC). For HS and AS, numbers in parentheses indicate the optimal layer index (1-based), whereas Window Entropy reports its best performance without layer annotation. The * symbol denotes values originally below 0.5 due to negative correlation and inverted using $1 - \text{AUROC}$. Bold indicates the best result for each LLM–dataset setting. All results are computed on 10K test samples.

LLMs	Method	HotpotQA	HotpotQA-WC	Natural Questions	TriviaQA	Winogrande
Llama-3.1	Window Entropy	0.5150*	0.6267	0.5340	0.7240	0.6156
	Hidden Score(HS)	0.5895* (32)	0.5321 (7)	0.6008* (32)	0.6639 (14)	0.6037 (21)
	Attention Score(AS)	0.7596* (31)	0.5416 (29)	0.6499* (31)	0.5749* (31)	0.5894 (28)
	TemHal	0.8375	0.7679	0.8131	0.8349	0.8869
Mistral-v0.3	Window Entropy	0.5284	0.5089	0.5527	0.5261	0.5112
	Hidden Score(HS)	0.5301* (32)	0.5409 (32)	0.5851* (32)	0.5630 (4)	0.6213 (31)
	Attention Score(AS)	0.5950 (31)	0.5490 (31)	0.5075* (31)	0.5934 (31)	0.6081 (31)
	TemHal	0.8506	0.7869	0.7901	0.8680	0.7255
Qwen-3	Window Entropy	0.6070	0.5898	0.5651	0.5946	0.5059
	Hidden Score(HS)	0.5121* (36)	0.5657 (36)	0.7014* (36)	0.5535 (34)	0.6330 (36)
	Attention Score(AS)	0.5246 (32)	0.5756 (32)	0.6402* (35)	0.5635 (4)	0.6201 (2)
	TemHal	0.8812	0.7684	0.8462	0.8969	0.8234

Table 8: Optimal layer indices for Probe-Exact across models and datasets.

Model (L layers)	HotpotQA	HotpotQA-WC	Natural Questions	TriviaQA	Winogrande
Llama-3.1 (32)	12	12	12	13	13
Mistral-v0.3 (32)	15	17	14	14	16
Qwen-3 (36)	35	19	35	20	22

B Additional Experiments

B.1 Strictly Reference-Free Token Identification

To evaluate TemHal without access to gold answers, we introduce a strictly reference-free token-identification protocol. Unlike the primary Probe-Exact compatible protocol, the reference answer a_i is not used to locate fact-critical tokens for any example, including correct generations. Instead, Qwen-3-8B-Instruct receives only the question q_i and generated response $\hat{y}_i$ and identifies the shortest span expressing the central factual answer. Following the primary protocol, examples for which no valid extracted span can be mapped to the generated response are excluded from exact-token probing.

TemHal and Probe-Exact use the same extracted token location for each example, ensuring that their comparison is not affected by different token-selection information. Neither the token extractor nor the hallucination detector receives the gold answer or correctness annotation. Gold annotations are used only after inference to compute AUROC.

As shown in Table 10, TemHal outperforms Probe-Exact on all five datasets, with improvements ranging from 0.0399 to 0.0833 AUROC and an average improvement of 0.0565. These

results demonstrate that TemHal remains effective when gold answers are unavailable to the inference pipeline and that its observed advantage is not solely attributable to gold-assisted token localization.

B.2 Raw Cross-Dataset Transfer Results

We report the complete AUROC scores underlying the performance-difference heatmaps in Figure 3. For each experiment, the detector is trained on the source dataset indicated by the row and evaluated on the target dataset indicated by the column. All transfer experiments are conducted separately within the same LLM backbone and therefore evaluate cross-dataset, rather than cross-model, generalization. In Table 11, P and T denote Probe-Exact and TemHal, respectively, and bold indicates the better result for each train–test pair.

TemHal outperforms Probe-Exact in AUROC for 65 of the 75 source–target pairs. The gains are most consistent for Llama-3.1 and Mistral-v0.3, whereas Qwen-3 exhibits several degradations under cross-category transfer. These results support the conclusion that TemHal provides effective cross-dataset transfer, particularly when the source and target tasks have related characteristics.

Table 9: Comparison of MHAD variants with TemHal across three LLMs and five benchmarks (AUROC). Bold indicates the best result for each model and dataset.

LLM	Method	HotpotQA	HotpotQA-WC	Natural Questions	TriviaQA	Winogrande
Llama-3.1	MHAD-AO	0.7966	0.6969	0.8152	0.8005	0.7422
	MHAD-FO	0.8092	0.7081	0.8213	0.8065	0.7579
	MHAD-LO	0.8083	0.7059	0.8224	0.8082	0.7604
	TemHal	0.8375	0.7679	0.8131	0.8349	0.8869
Mistral-v0.3	MHAD-AO	0.7939	0.6325	0.7453	0.7821	0.6856
	MHAD-FO	0.7907	0.6474	0.7526	0.7898	0.7029
	MHAD-LO	0.7923	0.6477	0.7624	0.7992	0.7228
	TemHal	0.8506	0.7869	0.7901	0.8680	0.7255
Qwen-3	MHAD-AO	0.8341	0.6807	0.8781	0.8473	0.7646
	MHAD-FO	0.8322	0.6839	0.8744	0.8403	0.7574
	MHAD-LO	0.8298	0.6909	0.8733	0.8366	0.7633
	TemHal	0.8812	0.7684	0.8462	0.8969	0.8234

Table 10: AUROC under strictly reference-free fact-critical token identification. No gold answer is provided to the token extractor or hallucination detector. Results are reported on Qwen3-8B.

Dataset	TemHal	Probe-Exact	Δ
HotpotQA	0.8318	0.7773	+0.0545
HotpotQA-WC	0.7283	0.6833	+0.0450
Natural Questions	0.8285	0.7886	+0.0399
TriviaQA	0.8433	0.7600	+0.0833
Winogrande	0.8099	0.7502	+0.0597
Average	0.8084	0.7519	+0.0565

Table 11: Raw cross-dataset transfer AUROC scores across three LLM families. P and T denote Probe-Exact and TemHal, respectively. NQ denotes Natural Questions. Bold indicates the better result for each source–target pair.

Source	Target									
	HotpotQA		HotpotQA-WC		NQ		TriviaQA		Winogrande	
	P	T	P	T	P	T	P	T	P	T
Llama-3.1										
HotpotQA	0.8259	0.8375	0.6369	0.6805	0.6650	0.7256	0.7289	0.7778	0.5543	0.6531
HotpotQA-WC	0.6491	0.7002	0.7441	0.7679	0.5965	0.6251	0.6481	0.7314	0.5697	0.5834
Natural Questions	0.6003	0.6405	0.5291	0.5757	0.7762	0.8131	0.7014	0.7548	0.5708	0.6684
TriviaQA	0.6631	0.6959	0.5912	0.6320	0.6800	0.7108	0.7996	0.8349	0.5827	0.6834
Winogrande	0.6632	0.6526	0.5698	0.5592	0.5932	0.6555	0.5558	0.6620	0.8715	0.8869
Mistral-v0.3										
HotpotQA	0.8396	0.8506	0.6629	0.7034	0.6825	0.7129	0.7824	0.8097	0.4891	0.5622
HotpotQA-WC	0.7418	0.7625	0.7674	0.7869	0.5622	0.5961	0.6773	0.7336	0.4736	0.4810
Natural Questions	0.6227	0.6592	0.5400	0.5679	0.7719	0.7901	0.7656	0.7945	0.6143	0.6657
TriviaQA	0.7162	0.6964	0.6339	0.6118	0.7117	0.7258	0.8537	0.8680	0.5003	0.5559
Winogrande	0.5729	0.7238	0.5361	0.6135	0.5554	0.6308	0.5918	0.7017	0.7202	0.7255
Qwen-3										
HotpotQA	0.8222	0.8812	0.6808	0.7147	0.7310	0.7597	0.7476	0.8598	0.4482	0.4452
HotpotQA-WC	0.7134	0.8171	0.7315	0.7684	0.5904	0.6323	0.6515	0.7499	0.5173	0.4429
Natural Questions	0.6664	0.6510	0.6222	0.5812	0.8116	0.8462	0.7191	0.7909	0.4630	0.4895
TriviaQA	0.6554	0.7698	0.5915	0.6841	0.6414	0.7378	0.8252	0.8969	0.4858	0.3952
Winogrande	0.4334	0.4488	0.4898	0.5216	0.5225	0.5051	0.5845	0.5963	0.7727	0.8234