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

Anonymous ACL submission

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 predominantly 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** (**Tem**poral **Hal**lucination Detector), a novel detection framework that models the cross-layer evolutionary trajectory of internal representations. 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 universal representational patterns that differentiate factually consistent outputs from hallucinated ones. Extensive experimental results demonstrate that the proposed method achieves significant performance improvements across multiple mainstream model architectures and diverse benchmark datasets, while also exhibiting superior robustness and generalization in cross-model and cross-dataset evaluations. 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 generate 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., 2024) or naively aggregated combinations across all layers (Zhang et al., 2025b; Kadavath et al., 2022a). Conversely, training-free techniques compute heuristic metrics directly from raw activations within single layers (Ren et al., 2022a; Sriramanan et al., 2024a).

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** (**Tem**poral **Hal**lucination 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 concatenating 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 14 out of 15 evaluation settings. Furthermore, ablation studies validate our architectural design choices and elucidate the underlying mechanisms, revealing that temporal modeling contributes the largest performance gain.

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 representations**. Early approaches utilized confidence

scores (Kadavath et al., 2022b; Lin et al., 2022) and uncertainty metrics (Ren et al., 2022b; Sriramanan et al., 2024b), 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 (Nelson 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) to cross-layer aggregations (Sky et al., 2024; Orgad et al., 2024; Zhang et al., 2025b) and training-free metrics (Chen et al., 2024; Sriramanan et al., 2024b). 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., 2024) demonstrate that probing exact answer tokens (defined as generated tokens semantically aligning with the ground-truth answer)

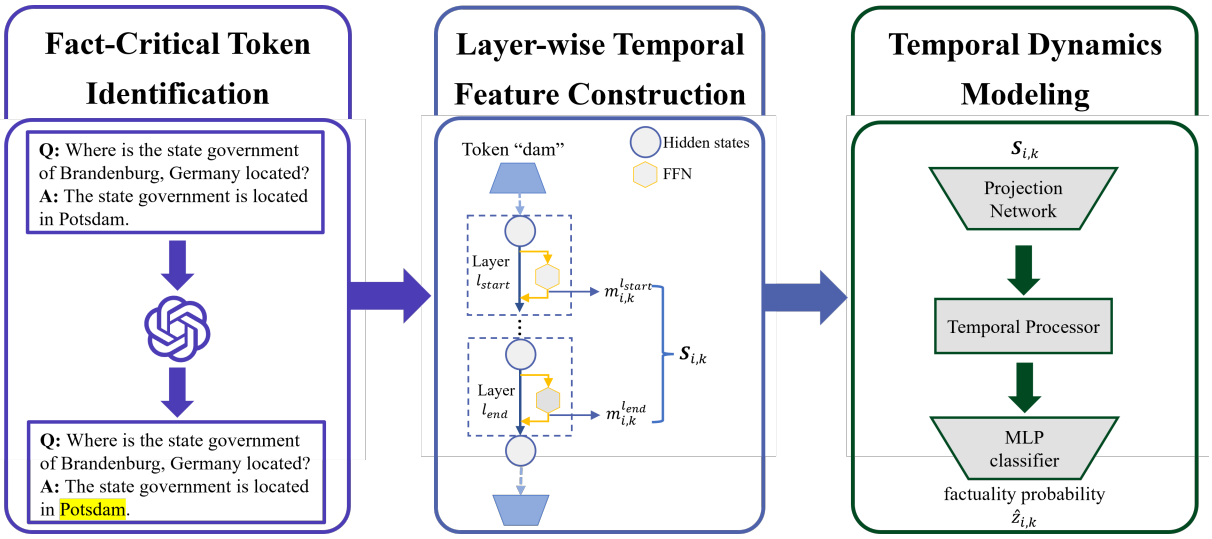

Figure 1: Overview of our framework. We extract exact answer tokens from the generated response and probe their MLP activations across selected transformer layers to capture the layer-wise evolution of factuality signals.

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 exact answer tokens through an LLM-guided extraction strategy. Given a question q_i and generated response $\hat{y}_i = F(q_i)$, we prompt the model to highlight spans in $\hat{y}_i$ that semantically align with the reference answer a_i. The resulting tokens are denoted as $\{t_{i,k}^{\text{exact}}\}_{k=1}^{n_i}$, where n_i is the count of exact answer tokens in sample i.

For multi-token factual spans, we focus on the *last token* of each phrase, as it typically aggregates semantic information from preceding tokens (Luo et al., 2026). Formally, let $\mathcal{K}_i \subseteq \{1, \ldots, n_i\}$ denote the indices of tokens serving as the final token of a multi-token factual phrase. Our analysis centers on the subset $\{t_{i,k}^{\text{exact}}\}_{k \in \mathcal{K}_i}$, probing their 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-cased (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 sequences of MLP outputs corresponding to the exact answer tokens $\{t_{i,k}^{\text{exact}}\}_{k \in \mathcal{K}_i}$ across a contiguous block of intermediate layers.

Specifically, let L be the total number of transformer layers in model F. For a chosen layer range $[l_{\text{start}}, l_{\text{end}}]$, we collect the MLP output vectors at position k from layers $l_{\text{start}}, \ldots, l_{\text{end}}$, yielding a sequence:

$$\mathbf{S}_{i,k} = \left[\mathbf{m}_{i,k}^{(l_{\text{start}})}, \ \ldots, \ \mathbf{m}_{i,k}^{(l_{\text{end}})}\right] \in \mathbb{R}^{w \times d},$$

where $w = l_{\text{end}} - l_{\text{start}} + 1$ is the 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 $\mathbf{S}_{i,k} \in \mathbb{R}^{w \times d}$ constructed in Section 4.2, we design a temporal discriminator to predict the factu-

ality of each exact answer token. As illustrated in Figure 1, the discriminator processes $\mathbf{S}_{i,k}$ through three stages: (1) dimensionality projection, (2) temporal encoding via a Temporal Processor, and (3) classification.

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

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

where $\mathbf{e}_{\text{temp}} \in \mathbb{R}^{d_{\text{hidden}}}$ is the aggregated temporal embedding. Finally, a multi-layer perceptron (MLP) classifier produces the token-level factuality probability:

$$\hat{z}_{i,k} = \text{MLP}(\mathbf{e}_{\text{temp}}), \tag{2}$$

where $\hat{z}_{i,k} \in [0, 1]$ estimates the probability that token $t_{i,k}^{\text{exact}}$ is factually consistent.

The discriminator is trained with binary cross-entropy loss using ground-truth labels $z_{i,k} \in \{0, 1\}$, where $z_{i,k} = 1$ indicates factual consistency and $z_{i,k} = 0$ denotes hallucination. The response-level factuality score $\hat{z}_i$ is obtained by averaging across all fact-critical tokens: $\hat{z}_i = \frac{1}{|\mathcal{K}_i|} \sum_{k \in \mathcal{K}_i} \hat{z}_{i,k}$.

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 best captured through temporal dynamics rather than static representations.

Bidirectional Temporal Encoding. The Bi-LSTM simultaneously captures forward propagation and backward contextual refinement, effectively distinguishing stable, convergent evolution patterns in factually correct responses from divergent trajectories in hallucinated outputs. Given the projected sequence $\bar{\mathbf{S}}_{i,k} = \{\bar{\mathbf{m}}_{i,k}^{(l)}\}_{l=l_{\text{start}}}^{l_{\text{end}}}$, the Bi-LSTM computes forward and backward LSTM hidden states $\{\overrightarrow{\mathbf{h}}_l\}$ and $\{\overleftarrow{\mathbf{h}}_l\}$, which are concatenated to form $\mathbf{h}_l = [\overrightarrow{\mathbf{h}}_l; \overleftarrow{\mathbf{h}}_l] \in \mathbb{R}^{2d_{\text{hidden}}}$. Our preliminary experiments confirm that this bidirectional setup outperforms unidirectional RNNs and Transformer variants under comparable parameter budgets, suggesting that gated recurrence better models the non-monotonic evolution of internal representations.

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 α_l are computed via a linear projection followed by softmax normalization over the sequence dimension:

$$\alpha_l = \frac{\exp(\mathbf{w}^T \mathbf{h}_l + b)}{\sum_{j=l_{\text{start}}}^{l_{\text{end}}} \exp(\mathbf{w}^T \mathbf{h}_j + b)}, \tag{3}$$

where $\mathbf{w} \in \mathbb{R}^{2d_{\text{hidden}}}$ and $b \in \mathbb{R}$ are learnable parameters. The final temporal embedding is obtained via weighted sum:

$$\mathbf{e}_{\text{temp}} = \sum_{l=l_{\text{start}}}^{l_{\text{end}}} \alpha_l \mathbf{h}_l. \tag{4}$$

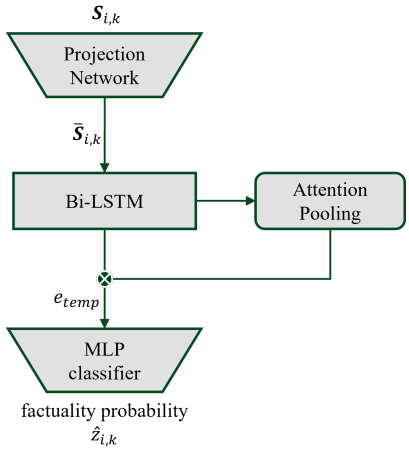

Figure 2: Architecture of the Factuality Discriminator. The model processes layer-wise MLP activation sequences through a projection network, Bi-LSTM encoder, attention pooling mechanism, and MLP classifier to predict factuality probability.

5 Experiment

5.1 Settings

Models. We evaluate our method on three representative 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** (Team, 2025) (hereafter *Qwen-3*). For detailed specifications (including organizations and Hugging Face repository paths), please refer to 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., 2024), we employ dataset-appropriate labeling strategies based on domain characteristics; comprehensive details, including labeling protocols, dataset statistics, prompt templates, and hallucination rates across models, are provided in Appendix A.2.

Baselines. To comprehensively evaluate the effectiveness of our proposed method, we compare it against representative hallucination detection baselines spanning three technical categories: (1) **classic generation probability-based methods**, including **Perplexity** (Ren et al., 2022b) and **P(True)** (Kadavath et al., 2022b), which estimate factual consistency directly from output logits or generation probabilities; (2) **training-free internal representation methods**, primarily **LLM-Check** (Sriramanan et al., 2024b), 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 the Table 6); and (3) **trainable internal representation methods**, including **Probe-Exact** (Orgad et al., 2024), which probes single-layer internal representations, and **ICR** (Zhang et al., 2025b), which jointly models representations across all transformer layers. Detailed implementation configurations for all baselines, including prompt templates and hyperparameter settings, are provided in Appendix A.3.

Evaluation Metric. Following established practice in hallucination detection research, we adopt the area under the receiver operating characteristic curve (AUROC) as our primary metric—a threshold-independent measure that ensures robust evaluation across varying decision thresholds. To ensure statistical reliability, we run each experiment with 5 random seeds (0, 5, 26, 42, 63).

5.2 Hallucination Detection Results

We conduct comprehensive evaluations of TemHal across five benchmark datasets and three mainstream open-source LLMs. Table 1 summarizes detection performance measured by AUROC across all evaluated model families. Empirical results demonstrate that TemHal achieves state-of-the-art performance in 14 out of 15 evaluation settings, consistently surpassing baseline methods by substantial margins. Notably, the most pronounced improvements are observed on complex reasoning tasks and context-dependent question answering, highlighting the method's particular efficacy in cognitively demanding scenarios.

Beyond detection accuracy, TemHal exhibits several distinctive advantages. First, it maintains robust and stable predictions across diverse model architectures and datasets, whereas certain baselines exhibit considerable variance or degrade to near-chance performance under distribution shift. Second, by leveraging multi-layer temporal aggregation, TemHal extracts more discriminative signals compared to approaches relying on single-layer probing or static attention patterns; this advantage is especially evident on reasoning-intensive benchmarks where modeling representational dynamics proves critical. Third, the method operates with a fixed intermediate one-third layer configuration determined via ablation studies (Section 5.4), thereby avoiding the computational overhead associated with exhaustive layer-wise search while preserving strong generalization capability across model families.

5.3 Generalization Analysis

We assess cross-dataset generalization by training on one dataset and evaluating on held-out benchmarks across three LLM families. Figure 3 illustrates the AUROC performance differences between TemHal and the baseline, where positive values indicate superior performance of our method. TemHal consistently outperforms the baseline on Llama-3.1 and Mistral-v0.3, achieving substantial gains across most transfer settings. Notably, generalization is most effective within task categories (e.g., Factual to Factual, Reasoning to Reasoning), suggesting that temporal dynamics of representations are highly transferable among tasks with similar cognitive demands.

For Qwen-3, certain cross-category transfers exhibit performance degradation. Further analysis reveals that under extreme 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 level when transferring between fundamentally different tasks. This observation highlights an inher-

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
	Probe-exact	0.8259	0.7441	0.7762	0.7996	0.8715
	TemHal (Ours)	0.8375	**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
	Probe-exact	0.8396	0.7674	0.7719	0.8537	0.7202
	TemHal (Ours)	**0.8506**	**0.7869**	**0.7901**	**0.8680**	**0.7255**
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	0.3598	0.5635	0.6201
	ICR	0.6765	0.5549	0.8038	0.6708	0.6450
	Probe-exact	0.8222	0.7315	0.8116	0.8252	0.7727
	TemHal (Ours)	**0.8812**	**0.7684**	**0.8462**	**0.8969**	**0.8234**

ent trade-off: while multi-layer temporal modeling yields stronger discriminative signals within task categories, it exhibits greater sensitivity to domain shift compared to static single-layer probes. Overall, by explicitly modeling representation dynamics across multiple knowledge-relevant layers, TemHal demonstrates robust cross-dataset generalization, particularly when source and target tasks share similar cognitive structures.

5.4 Ablation Study

We conduct comprehensive ablation studies on Qwen-3-8B-Instruct to analyze architectural contributions, 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. Empirical results demonstrate that intermediate layers consistently yield superior performance across most datasets, 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 (e.g., Winogrande), where the intermediate window substantially outperforms late layers, suggesting mid-level layers capture critical syntactic-to-semantic transitions. For knowledge-intensive tasks (e.g., Natural Questions, TriviaQA), intermediate layers remain optimal, though late layers exhibit competitive yet slightly inferior performance. Utilizing all layers introduces representational noise, consistently degrading performance compared to a focused intermediate window. We select a one-third layer window to optimize signal discriminability. Fine-grained tuning confirms performance stability within the 10–16 layer range around the middle section, validating that the exact boundary is not overly sensitive while confirming the intermediate region's importance. Although a 16-layer window shows marginal gains on HotpotQA-WC, the 12-layer configuration offers the best overall balance across diverse scenarios. Thus, we adopt the intermediate configuration as default, balancing reasoning dynamics and factual representations.

Contribution of Core Components. Table 3 quantifies component impact on a representative reasoning benchmark. Temporal modeling emerges as the most critical component; its removal leads to the most substantial performance decline, indicating hallucination signals reside in representation evolution rather than static states. The bidirectional architecture surpasses unidirectional and Transformer variants, highlighting the necessity of capturing contextual dependencies within the layer sequence. Finally, the attention mechanism yields marginal yet consistent improvements, underscor-

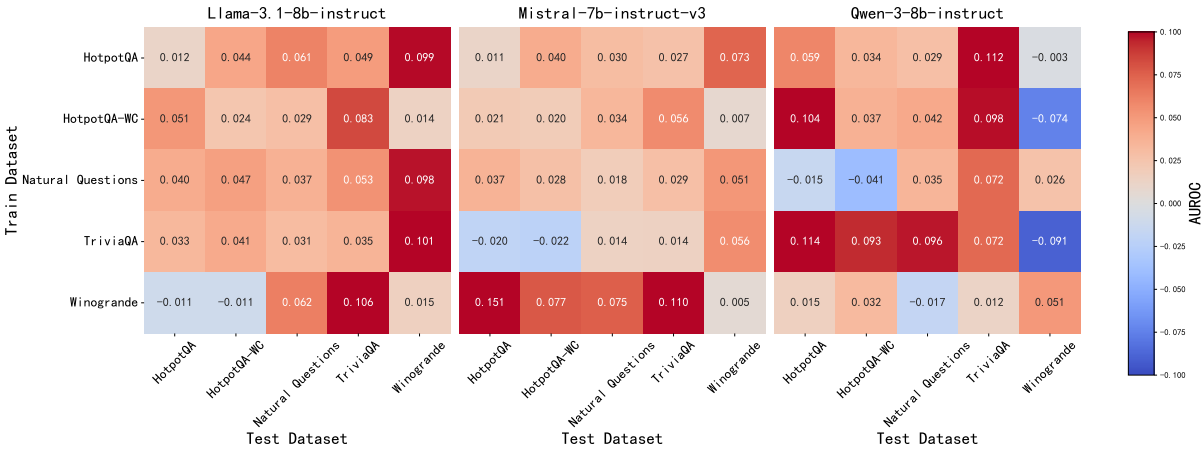

Figure 3: Cross-dataset generalization performance across three LLM families. Each cell shows the AUC difference (ours − Probe-Exact) for the corresponding train–test pair. Red indicates improvement, while blue indicates degradation relative to the baseline.

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
Position Selection (1/3 Split)					
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
Range Size Tuning (Around Mid Layers)					
8 layers [15-22]	0.8791	0.7684	0.8424	0.8957	0.8172
10 layers [14-23]	0.8796	0.7685	0.8452	0.8966	0.8162
14 layers [12-25]	0.8767	0.7704	0.8460	0.8939	0.8118
16 layers [11-26]	0.8753	**0.7705**	0.8454	0.8938	0.8129

ing its role in refining key factual tokens across the temporal dimension.

Table 3: Ablation study on core components (AUROC on **Winogrande**).

Variant	AUROC	Δ vs Full
Full TemHal	**0.8234**	–
Transformers	0.8183	-0.0051
RNN	0.8197	-0.0037
w/o Attention	0.8124	-0.0110
w/o Temporal (single-layer)	0.7727	-0.0507
w/o Bi (Uni-LSTM)	0.7901	-0.0333

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 achieves strong generalization across diverse tasks and models, offering a practical solution for enhancing LLM trustworthiness in high-stakes applications. 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 cognitive structures of target tasks.

Limitations

While TemHal demonstrates strong performance, several limitations warrant discussion. First, TemHal fundamentally relies on accessing layer-wise internal representations, which inherently restricts its deployment to models with exposed internals; developing black-box-compatible variants remains an important direction for future work. Second, the default layer window was empirically optimized; models with substantially different depths may require proportional adjustment, suggesting future work on automated layer selection. Third, although TemHal generalizes effectively within task categories, cross-category transfer under extreme domain shift can exhibit sensitivity, as temporal signals may adapt to task-specific dynamics; exploring task-invariant temporal representations or adaptive calibration mechanisms may provide potential solutions. Finally, while TemHal shows strong performance on reasoning benchmark, its application to more complex reasoning scenarios (such as multi-step mathematical proofs, chain-of-thought verification, or logical deduction with intermediate fact-checking) remains underexplored; adapting temporal dynamics modeling to capture reasoning-specific evolutionary patterns is a promising direction for future work.

Ethical Considerations

Our methodology does not rely on personal or sensitive data. All experiments use publicly available academic benchmarks (HotpotQA, Natural Questions, TriviaQA, Winogrande) curated to exclude personally identifiable information and offensive content. We follow original dataset licenses and perform no 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.

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, and 1 others. 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, and 1 others. 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.

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, and 1 others. 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, and 1 others. 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, 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 S Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *arXiv preprint arXiv:1705.03551*.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, and 1 others. 2022a. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, and 1 others. 2022b. 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, and 1 others. 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 Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, and 1 others. 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 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. 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, and 1 others. 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, and 1 others. 2026. Two pathways to truthfulness: On the intrinsic encoding of llm hallucinations. *arXiv preprint arXiv:2601.07422*.

Audrey Moores and Vânia G Zuin Zeidler. 2025. Don't let generative ai shape how we see chemistry. *Nature Reviews Chemistry*, 9(10):649–650.

Elhage Nelson, Nanda Neel, Olsson Catherine, Henighan Tom, Joseph Nicholas, Mann Ben, Askell Amanda, Bai Yuntao, Chen Anna, Conerly Tom, and 1 others. 2021. A mathematical framework for transformer circuits. *Transformer Circuits Thread*.

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

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. 2022a. Out-of-distribution detection and selective generation for conditional language models. *arXiv preprint arXiv:2209.15558*.

Jie Ren, Jiaming Luo, Yao Zhao, Kundan Krishna, Mohammad Saleh, Balaji Lakshminarayanan, and Peter J Liu. 2022b. 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 Skean, 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. 2024a. Llm-check: Investigating detection of hallucinations in large language models. *Advances in Neural Information Processing Systems*, 37:34188–34216.

Gaurang Sriramanan, Siddhant Bharti, Vinu Sankar Sadasivan, Shoumik Saha, Priyatham Kattakinda, and Soheil Feizi. 2024b. Llm-check: Investigating detection of hallucinations in large language models. *Advances in Neural Information Processing Systems*, 37:34188–34216.

Qwen Team. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

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.

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.

Zhenliang Zhang, Xinyu Hu, Huixuan Zhang, Junzhe Zhang, and Xiaojun Wan. 2025b. 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, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 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 4 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

Table 4: 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`
Qwen-3-8B-Instruct	Alibaba	`Qwen/Qwen-3-8B-Instruct`

generation length is set to 1024 tokens. We extract internal states from all transformer layers using the Hugging Face `transformers` library with `output_hidden_states=True`. MLP outputs are obtained by applying the model's feed-forward network to the final hidden state at each layer.

A.2 Dataset Details

Table 5 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 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 5, 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 (Orgad et al., 2024), we assign binary labels ($z_k \in \{0, 1\}$) to fact-critical tokens by comparing generated answer spans against reference answers:

- **Exact-match datasets** (HotpotQA, HotpotQA-WC, TriviaQA, Winogrande): A token is labeled as factual ($z_k = 0$) if its containing span exactly matches the reference answer; otherwise $z_k = 1$.
- **Free-form datasets** (Natural Questions): We employ Mistral-7B-v0.2 as an automatic ver-

Table 5: 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

ifier to assess factual consistency between generated answers and reference labels. The model evaluates each claim based on the input question, generated response, and ground-truth answer, assigning binary verification labels. Tokens in spans verified as factually consistent receive $z_k = 0$ (prompt details in repository).

A.3 Baseline Implementation Details

Perplexity (Ren et al., 2022b). 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., 2022b). 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., 2024b). The original work introduces three training-free detection methods operating on single-layer representations: *Window Entropy*, *Hidden Score (HS)*, and *Attention Score (AS)*. We systematically evaluate each variant across all layers of the assessed LLMs (32–36 layers) and report their optimal performance in Table 6, where the layer indices yielding peak results for HS and AS are indicated in parentheses. As shown in the table, the attention-based variant (AS) consistently achieves the strongest performance among the baselines. In contrast, both Window Entropy and HS exhibit comparatively limited efficacy, particularly on complex reasoning benchmarks such as HotpotQA-WC, where their discriminative capacity is notably constrained. All implementations follow the official repository [1] provided by the original authors.

Probe-Exact (Orgad et al., 2024). Following the original work, we train a linear classifier on the internal states at fact-critical token positions. While the authors recommend layer 16, we conduct a layer-wise search across the full layer range of each backbone and select the optimal layer based on validation AUROC. Optimization: AdamW, with early stopping on a held-out validation set (20% of training data). Our implementation follows the public codebase[2] released with the original paper.

As shown in Appendix Table 7, optimal layers exhibit significant variation across models and tasks: Llama-3.1 consistently selects mid-layers (12–13), Mistral-v0.3 favors mid-to-late layers (14–17), while Qwen-3 shows the highest variance (19–35), with knowledge-intensive tasks (HotpotQA, Natural Questions) preferring layers near the output. This model- and task-dependent distribution confirms that factual representations are not uniformly encoded across transformer depths, and that a fixed-layer protocol (e.g., layer 16) may suboptimally capture task-specific signals. These findings further motivate our multi-layer temporal modeling approach in **TemHal**, which aggregates evidence across layers rather than relying on a single potentially suboptimal representation.

[1] https://github.com/GaurangSriramanan/LLM_Check_Hallucination_Detection

[2] https://github.com/technion-cs-nlp/LLMsKnow

Table 6: Comparison of LLM-Check variants (Window Entropy, HS, AS) across LLMs and benchmarks (AUROC).

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**

* Values originally below 0.5 (negative correlation) are inverted via $1 - \text{AUC}$.
For HS/AS variants, numbers in parentheses denote the optimal layer index (1-based); WE variant reports best performance without layer annotation.
Bold: best result per column. All results computed on 10K test samples.

Table 7: 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

ICR (Zhang et al., 2025b). We follow the original implementation, which leverages multi-layer attention patterns and hidden state distribution divergence to detect hallucinations. Specifically, the method extracts attention head weights and computes Jensen-Shannon divergence of representations at fact-critical positions, then trains a linear classifier on these features. Optimization follows the original paper's protocol with AdamW optimizer and early stopping on validation data. We refer readers to the official code repository[3] for complete architectural details and implementation.

[3] https://github.com/XavierZhang2002/ICR_Probe