DUD: Decoupled Update Dynamics for Reliable Uncertainty Quantification in Large Language Models

Yixin Bu[1,2], **Runze Xia**[1,2], **Guanyun Zou**[1,2], **Yupeng Ji**[3], **Hongliang Dai**[1,2], **Haodong Liu**[3], **Piji Li**[1,2,†]
[1]College of Artificial Intelligence,
Nanjing University of Aeronautics and Astronautics, China
[2]The Key Laboratory of Brain-Machine Intelligence Technology,
Ministry of Education, Nanjing, China
[3]Researcher
{sx2416068,xiarunze,zouguanyun,hongldai,pjli}@nuaa.edu.cn,
{yupengjitju,liuhaodong2008}@gmail.com

Abstract

Accurate Uncertainty Quantification (UQ) is critical for reliable deployment of Large Language Models (LLMs), yet traditional probability-based metrics often fail to capture the model's true epistemic state. While recent mechanistic approaches leverage hidden state dynamics, they typically aggregate residual stream updates, conflating the distinct roles of parametric memory (Feed-Forward Networks) and contextual processing (Attention). We argue that this aggregation obscures fine-grained mechanistic conflicts, such as memory-context misalignment, that are fundamental indicators of uncertainty. To address this, we introduce **D**ecoupled **U**pdate **D**ynamics **(DUD)**, a framework that explicitly decouples FFN and Attention contributions via noise-induced causal interventions. By quantifying the independent restoration capabilities of each module, we construct a dual-stream dynamic profile that captures the model's internal fragility. Extensive experiments demonstrate that DUD significantly outperforms state-of-the-art baselines in both uncertainty estimation and calibration, while exhibiting superior cross-dataset generalization, validating decoupled dynamics as a robust proxy for model faithfulness.

1 Introduction

Reliable Uncertainty Quantification (UQ) is a prerequisite for deploying Large Language Models (LLMs) in high-stakes environments. Ideally, a model's predicted confidence should perfectly align with the correctness of its generation. However, achieving such calibration remains a formidable challenge. Traditional UQ approaches, ranging from black-box sampling consistency (Manakul et al., 2023) to white-box logit analysis (Kuhn et al., 2023), predominantly rely on the final output distribution. While effective for capturing data-level ambiguity (aleatoric uncertainty), these surface-level metrics often fail to reflect the model's true epistemic state. By treating the inference process as a black box and focusing solely on the end result, they overlook the internal computational dynamics where uncertainty actually originates. Consequently, a model may output a high probability score while its internal reasoning process is fraught with instability, leading to severe miscalibration.

To bridge this gap, Mechanistic Interpretability offers a fundamental theoretical grounding. Pioneering studies have mapped functional specialization within Transformer architectures, identifying Feed-Forward Networks (FFNs) as repositories for parametric knowledge (Meng et al., 2022) and Multi-Head Self-Attention (MHSA) as routers for contextual information (Elhage et al., 2021). These insights have been actively adapted in related fields, particularly in hallucination detection, where researchers utilize hidden state trajectories to pinpoint generation errors (Chen et al., 2024; Marks and Tegmark, 2023; Zhang et al., 2025b). However, we identify a critical methodological oversight in how these mechanistic signals are currently utilized for UQ: existing approaches typically employ aggregation strategies, merging the contributions of memory and context into unified representations. We argue that this aggregation obscures the essence of uncertainty. In complex reasoning, epistemic uncertainty often arises precisely from the mechanistic conflict between these components such as when parametric memory contradicts contextual cues. By aggregating these opposing signals, current methods neutralize the fine-grained evidence required for precise diagnosis.

In this paper, we propose Decoupled Update Dynamics (DUD), a framework that transforms UQ from passive estimation to active mechanistic diagnosis. DUD explicitly disentangles the

[†]Corresponding author.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 35972–35993

causal contributions of FFNs and MHSA modules by measuring their independent ability to restore prediction confidence from a noise-induced perturbed state. This decoupled tracing reveals a critical phenomenon overlooked by previous studies: uncertainty manifests as distinct spatiotemporal fragility patterns. We observe that generation failures are not uniform but stem from specific mechanistic breakdowns—manifesting as routing instability in early-layer MHSA during context encoding, and a precipitous collapse in late-layer FFNs during knowledge retrieval. By capturing these fine-grained conflicts, we construct a dual-stream dynamic profile that serves as a robust proxy for the model's epistemic state. Extensive experiments demonstrate that DUD outperforms state-of-the-art baselines in both uncertainty estimation and calibration. Notably, our method exhibits exceptional cross-dataset generalization, confirming that these mechanistic signatures are intrinsic to the model's reasoning process rather than artifacts of specific tasks.

Our contributions are summarized as follows:

- **Decoupled Causal Framework:** We introduce the first UQ framework that explicitly disentangles the causal contributions of FFNs and MHSA modules, demonstrating that internal module dynamics provide a significantly more faithful proxy for reliability than traditional output probabilities.
- **Mechanistic Fragility Discovery:** Through rigorous causal tracing, we uncover that prediction uncertainty physically manifests as systemic fragility and module conflict, particularly within middle-layer FFNs, offering a physical grounding for detecting model errors.
- **Empirical Performance:** Extensive experiments across diverse benchmarks demonstrate that our decoupled approach significantly outperforms both logits-based baselines and aggregated mechanistic probes, achieving superior calibration and error detection capabilities.

2 Related Work

2.1 Uncertainty Quantification in LLMs

Traditional uncertainty quantification predominantly relies on output probabilities, such as Semantic Entropy (Kuhn et al., 2023). However, these logit-based metrics remain superficial, assessing final confidence without inspecting the internal reasoning process, often failing to distinguish inherent data ambiguity from model capability failures. Recent approaches leverage intermediate representations to address this. Methods like SAPLMA (Chen et al., 2024), Lookback Lens (Chuang et al., 2024), and DoLa (Chuang et al., 2023) utilize hidden states, attention weights, or layer-wise contrasts. While richer than logits, these methods typically aggregate signals from different modules into unified representations. This conflation obscures fine-grained internal conflicts—specifically between parametric memory and contextual cues that are critical indicators of generation failure.

2.2 Mechanistic Interpretability of Transformer Components

Mechanistic interpretability research establishes that Multi-Head Self-Attention (MHSA) and Feed-Forward Networks (FFN) perform distinct computational roles: MHSA acts as a contextual signal router transferring information (Elhage et al., 2021), while FFNs serve as key-value memories for parametric knowledge (Geva et al., 2021; Meng et al., 2022). Stolfo et al. (2023) further demonstrated their dominance at different processing stages of inference. Building on these insights, our work diverges from previous aggregation-based UQ methods. We explicitly model the dynamic evolution of these decoupled streams, quantifying uncertainty as the mechanistic misalignment between parametric memory (FFN) and contextual processing (MHSA).

3 Methodology

We propose the Decoupled Update Dynamics (DUD) framework, a mechanistic approach designed to quantify model uncertainty by dissecting the internal tug-of-war between parametric memory and contextual processing. As illustrated in Figure 1, our method transforms uncertainty quantification from a passive observation task into an active causal interrogation process.

3.1 Preliminaries: Transformer Dynamics

Given an input sequence $x = [x_1, \ldots, x_T]$, the hidden state $\mathbf{h}_i^{(l)}$ at layer l and token position i is updated as the following illustrative logic:

$$\mathbf{h}_i^{(l)} = \mathbf{h}_i^{(l-1)} + \mathbf{a}_i^{(l)} + \mathbf{m}_i^{(l)} \quad (1)$$

where $\mathbf{a}_i^{(l)}$ is the output of the Multi-Head Self-Attention (MHSA) module, responsible for con-

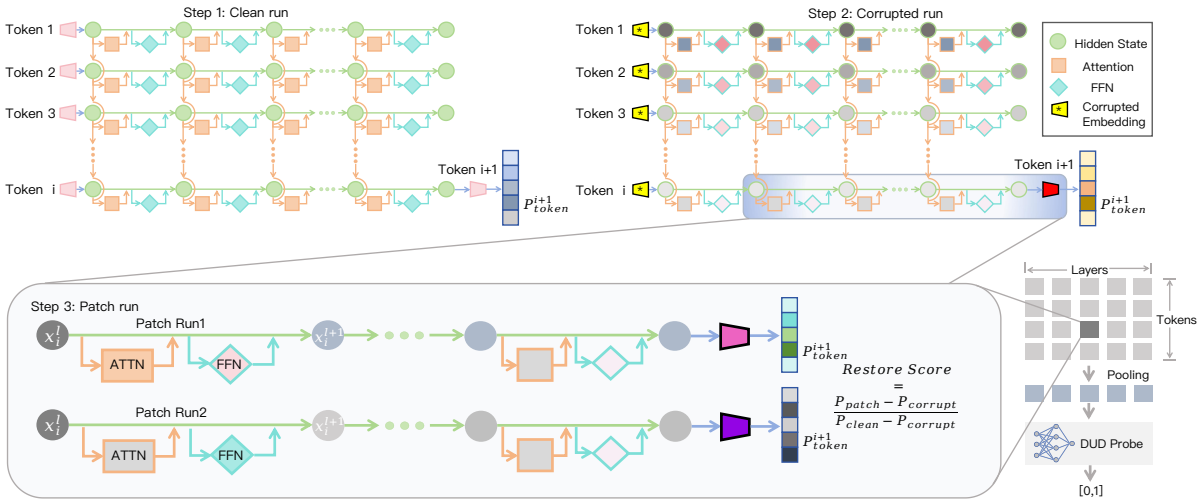

Figure 1: Overview of the **Decoupled Update Dynamics (DUD)** framework. The process involves three stages: (1) **Clean Run:** A standard forward pass to cache hidden states; (2) **Corrupted Run:** Noise is injected into embeddings to induce an information bottleneck; (3) **Patch Run:** We perform causal interventions by independently restoring the activations of Attention and FFN modules layer-by-layer in Step 2, followed by recovering Attention and FFN in Step 3. The resulting probability shifts are normalized into restoration scores, forming a $2L$-dimensional dynamic profile that trains the DUD-Probe to detect uncertainty.

textual routing; and $\mathbf{m}_i^{(l)}$ is the output of the Feed-Forward Network (FFN), which serves as a key-value memory for parametric knowledge (Geva et al., 2021). The final probability distribution for the next token is obtained by projecting the final layer state $\mathbf{h}_T^{(L)}$ through the unembedding matrix $\mathbf{E}$ and applying a softmax function:

$$P(y|x) = \text{Softmax}(\mathbf{E}\mathbf{h}_T^{(L)}) \tag{2}$$

3.2 Decoupled Causal Tracing

In this subsection, we introduce a decoupled causal tracing framework to isolate and quantify the causal contributions of memory- and context-related components under uncertainty. Traditional uncertainty metrics break down when memory and context provide conflicting evidence. We explicitly model this mismatch. Our method applies decoupled causal tracing under three execution settings: Clean, Corrupted, and Patch runs. Comparing these runs reveals how memory and context individually influence the final prediction (Figure 1).

Establishing Baselines via Clean and Corrupted Runs. We first define the upper and lower bounds of the model's performance to quantify the information bottleneck.

- **Clean Run:** We perform a standard forward pass with the original input x. We cache the clean activations (hidden states $\mathbf{h}$, MHSA outputs $\mathbf{a}$, and FFN outputs $\mathbf{m}$) for all layers and record the prediction probability $P_{\text{clean}}(y_t)$.
- **Corrupted Run:** We inject Gaussian noise $\epsilon \sim \mathcal{N}(0, \nu)$ into the input embeddings to simulate a loss of context ($\mathbf{e}^*(x) = \mathbf{e}(x) + \epsilon$). This degrades the prediction probability to $P_{\text{corr}}(y_t)$.

The probabilistic gap between these two runs represents the Total Effect (TE) of the noise, calculated as the average drop across the sequence:

$$\text{TE} = \frac{1}{T}\sum_{t=1}^{T}[P_{\text{clean}}(y_t) - P_{\text{corr}}(y_t)] \tag{3}$$

Decoupled Patch Run. To disentangle the roles of the two pathways, we perform Component-Specific Interventions. Unlike standard causal tracing which restores the entire hidden state, we execute separate restored runs for the FFN and Attention modules. For a specific layer l, we intervene in the corrupted forward pass by replacing the activation of a target module $\phi \in \{\text{MHSA}, \text{FFN}\}$ with its value cached from the Clean Run, while keeping all other components in the Corrupted state. Mathematically, restoring module ϕ at layer l yields a restored probability $P_{\text{restored}}^{(l,\phi)}(y_t)$. This operation isolates the causal contribution of that specific module in recovering the correct prediction from the noisy baseline.

Sequence-Level Dynamic Profiling. We quantify the contribution of each module by calculating the Restoration Score. The score $\mathcal{S}_\phi^{(l)}$ represents the normalized probability shift induced by

the Restored Run:

$$\mathcal{S}_{\phi}^{(l)} = \frac{1}{T}\sum_{t=1}^{T}\frac{P_{\text{restored}}^{(l,\phi)}(y_t) - P_{\text{corr}}(y_t)}{\text{TE}} \quad (4)$$

Empirically, we observe that $\mathcal{S}_{\phi}^{(l)}$ is predominantly negative, as restoring a single module within a corrupted stream introduces state incoherence. We therefore interpret the *magnitude* of this suppression as a proxy for uncertainty. Specifically, a small magnitude ($\mathcal{S} \approx 0$) indicates mechanistic robustness, where the model tolerates the internal mismatch, suggesting the generation is grounded in stable features. In contrast, a large negative magnitude ($|\mathcal{S}| \gg 0$) signals mechanistic fragility, where single-module restoration causes a catastrophic drop in probability, revealing that the generation relies on a precarious internal equilibrium.

3.3 DUD-Probe Construction

To translate these fine-grained internal dynamics into a comprehensive uncertainty estimate, we construct a lightweight, non-linear probe.

Feature Engineering. We concatenate the layer-wise restoration scores into a unified feature vector $\mathbf{v} \in \mathbb{R}^{2L}$, applying Z-score normalization to ensure numerical stability:

$$\mathbf{v} = \text{Norm}\left(\text{Concat}\left[\mathcal{S}_{\text{attn}}^{(1..L)}, \mathcal{S}_{\text{ffn}}^{(1..L)}\right]\right) \quad (5)$$

This vector preserves the structural relationship between memory and context streams.

Probe Design and Training. We formulate uncertainty quantification as a binary classification task. Ground truth labels are derived from the ROUGE-L score ($\mathcal{S}_{\text{rouge}}$) of the generated response:

$$y_i = \begin{cases} 0 \text{ (Correct)} & \text{if } \mathcal{S}_{\text{rouge}} \geq \tau \\ 1 \text{ (Uncertain)} & \text{otherwise} \end{cases} \quad (6)$$

Unlike linear classifiers, we employ a 4-layer MLP ($2L \to 128 \to 64 \to 32 \to 1$) to capture non-linear interactions. We incorporate LeakyReLU, Batch Normalization to handle negative restoration scores and prevent overfitting. The probe is trained using Binary Cross-Entropy (BCE) loss to predict generation errors.

4 Empirical Preliminary Analysis

In our Decoupled Update Dynamics (DUD) framework, the core objective of the decoupled restoration score is to quantify model uncertainty, specifically by assessing its ability to distinguish between correct and incorrect predictions, and to serve as a robust signal for guiding probe training. To this end, we conduct a series of experiments across four distinct cognitive task datasets: HaluEval and SQuAD (context-rich), and TriviaQA and HotpotQA (context-free). These experiments validate the effectiveness of the proposed decoupled restoration scores in distinguishing correct and incorrect predictions, and demonstrate their potential in providing valuable signals for effective probe training.

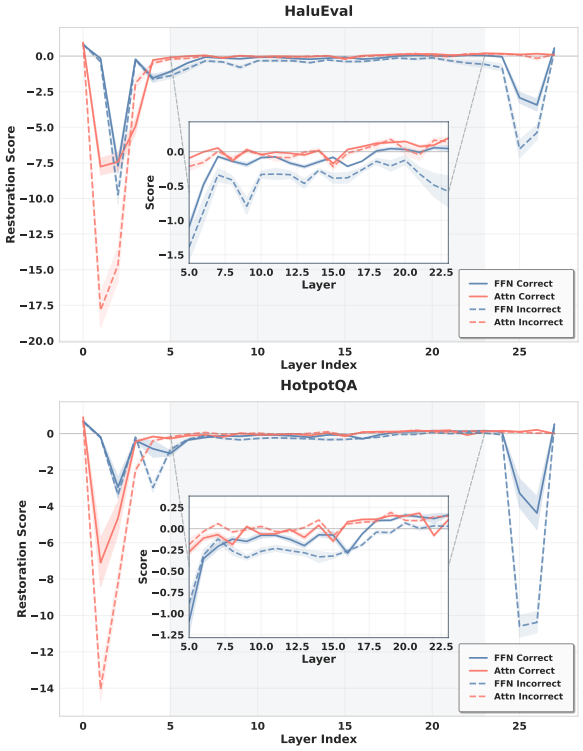

Figure 2: Layer-wise restoration dynamics on HaluEval (Top) and HotpotQA (Bottom). The curves show the restoration scores for correct and incorrect predictions across layers.

4.1 Restoration Dynamics: Fragility and Task Dependence

Figure 2 visualizes layer-wise restoration dynamics, revealing a distinct mechanistic signature of uncertainty. Across all datasets, we observe a consistent pattern in the restoration scores. In the *early layers (0-5)*, both Attention and FFN modules exhibit significant fluctuations, with restoration scores showing larger differences between correct and incorrect predictions. This indicates a high degree of model uncertainty during the initial encoding phase, where errors are often driven by

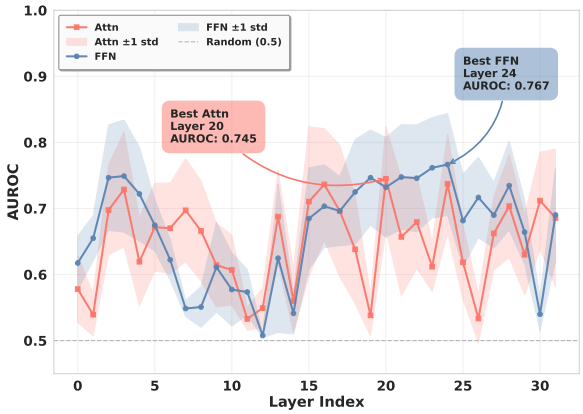

Figure 3: Layer-wise AUROC analysis for LLaMA-3.1-8B, averaged across four datasets. The FFN (blue) and Attention (red) curves demonstrate that restoration scores from both modules effectively distinguish uncertainty across the network depth.

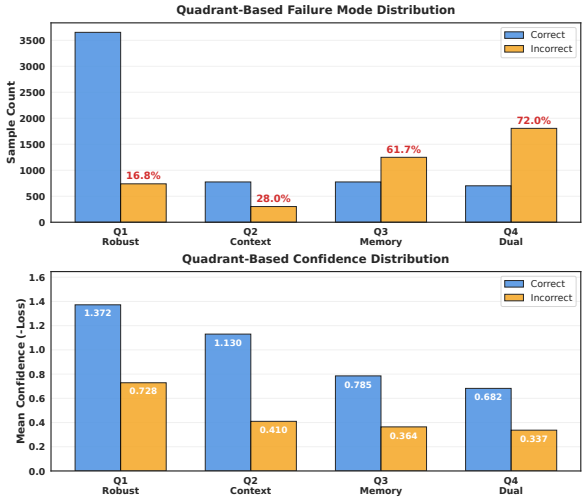

Figure 4: **Quadrant Analysis on SQuAD (Layer 24). Top:** Error rates across four mechanistic quadrants defined by FFN and Attention deviation. **Bottom:** Confidence score distributions for Correct vs. Incorrect predictions within each quadrant, highlighting the misalignment between surface confidence and internal mechanistic stability.

Attention failures, particularly in contextual routing. In contrast, the *middle layers* demonstrate relative stability, with restoration scores exhibiting smaller variations, which suggests a more stable integration of information as the model processes the input. However, a critical divergence emerges in the *late layers (20-28)*, where we observe a specific collapse in FFN scores while Attention remains stable. The sharp decline in FFN restoration scores highlights a fundamental failure in parametric memory during the later stages of generation, marking this as a key indicator of model uncertainty. These observed differences in the restoration scores across layers serve as crucial signals for guiding probe training, allowing for the effective identification of model uncertainty.

The magnitude of these patterns varies with task demands. In context-rich tasks like HaluEval, early layers show greater instability due to the challenge of processing long contexts. In contrast, HotpotQA, a knowledge-intensive task, exhibits particularly large changes in the late-layer FFN scores. This could be due to the model's reliance on internal synthesis for answering complex questions, as it lacks external context to ground its reasoning. As a result, late-layer FFN stability becomes more crucial, with large score shifts indicating the model's difficulty in maintaining internal coherence without sufficient external information.

4.2 Discriminative Power: AUROC Analysis

To assess whether decoupled restoration scores can reliably distinguish correct from incorrect generations and serve as effective signals for probe learning, Figure 3 shows that these scores exhibit strong and structured discriminative power across network depth. For LLaMA-3.1-8B, elevated AUROC values consistently emerge in the early layers (e.g., layers 2-3) as well as in the middle-to-late layers (approximately layers 16-24), closely mirroring the layer-wise mechanistic deviations identified in the restoration dynamics analysis. This alignment indicates that the proposed scores capture predictable and functionally meaningful uncertainty signals tied to distinct stages of representation processing, rather than arbitrary fluctuations. As a result, the restoration scores form informative and structured features for uncertainty probing, enabling reliable separation between correct and incorrect generations. Similar AUROC profiles are observed across Qwen-2.5-7B and Gemma-2-9B (Appendix B), suggesting that the discriminative capacity of the decoupled restoration scores generalizes across model architectures.

4.3 The Confidence Paradox: Mechanistic Deviations vs. Surface Certainty

To further examine why surface-level confidence often fails to reflect generation reliability, we conduct a quadrant analysis on SQuAD at Layer 24. By establishing deviation thresholds at the 75th percentile of correct predictions, we partition samples into four mechanistic regimes based on the stability of Attention and FFN modules. These

regimes span from Q1 (Robust), a stable state where both modules remain within nominal limits, through localized instability in Q2 (Context Deviant) where only the Attention score exceeds the threshold and Q3 (Memory Deviant) where only the FFN score exceeds it, to Q4 (Dual Deviant), which signals systemic fragility with both modules violating the stability bounds.

The results (Figure 4) reveal a clear misalignment between internal stability and logits-based confidence. Error rates surge with accumulated deviation, rising from 16.8% in the robust Q1 quadrant to 72.0% in the dual-deviant Q4, confirming that internal instability is a decisive predictor of failure. Crucially, surface confidence fails to track this risk: incorrect predictions in stable Q1 often retain deceptively high confidence (Stubborn Errors), while correct predictions in deviant Q4 suffer from low confidence (Humble Truths). This "Confidence Inversion" demonstrates that probability thresholds alone are insufficient. By grounding prediction risk in decoupled mechanistic states, DUD-Probe effectively identifies high-risk generations that surface confidence overlooks.

5 Experiments

5.1 Experimental Setup

Models, Datasets, and Baselines We evaluate three open-source LLMs Llama-3-8B-Instruct (Grattafiori et al., 2024), Qwen2.5-7B-Instruct (Hui et al., 2024), and Gemma-2-9B-it (Team et al., 2024) across four knowledge-intensive benchmarks: HaluEval (Li et al., 2023) (Question Answering) , SQuAD (Rajpurkar et al., 2016) (Reading Comprehension), TriviaQA (Joshi et al., 2017) (Retrieval), and HotpotQA (Yang et al., 2018) (Reasoning). We benchmark against diverse baselines, including training-free methods (**PPL** (Ren et al., 2022), **LN-Entropy** (Malinin and Gales, 2020), **SE** (Kuhn et al., 2023), **SAR** (Duan et al., 2024), **LLM-Check**) and training-based probes (**SAPLMA** (Chen et al., 2024), **SEP** (Kossen et al., 2024), and the aggregated mechanistic **ICR Probe** (Zhang et al., 2025b)). Detailed configurations are provided in Appendix C.

Evaluation Protocol and Implementation We employ a rigorous two-stage protocol. First, ground truth correctness is determined via ROUGE-L (Lin, 2004), where generations meeting the semantic threshold are labeled as *Correct* ($y = 0$), otherwise *Incorrect* ($y = 1$). Second, uncertainty performance is primarily assessed using AUROC (Fawcett, 2006), with auxiliary metrics (ECE (Guo et al., 2017), PRR (Malinin and Gales, 2018)) and sensitivity analyses (BERTScore (Zhang et al., 2019), EM (Rajpurkar et al., 2016)) detailed in the Appendix G. For implementation, the DUD-Probe is trained on decoupled features ($\mathbf{v} \in \mathbb{R}^{2L}$) using 5-fold cross-validation, ensuring identical data splits across all baselines for fair comparison.

5.2 Main Results: Uncertainty Quantification Performance

Table 1 demonstrates that DUD-Probe consistently achieves state-of-the-art uncertainty quantification performance across diverse models and datasets. Compared to logits-based Semantic Entropy, DUD-Probe yields substantial improvements (e.g., +17.4% on LLaMA-3.1 HaluEval), indicating that internal causal dynamics provide substantially richer reliability signals than surface probabilities. Moreover, DUD-Probe outperforms the aggregated ICR Probe (e.g., +5.0% on Qwen2.5), validating the necessity of decoupling Attention and FFN contributions. By separating these two update pathways, our method exposes fine-grained mechanistic conflicts that are obscured by aggregated representations. This advantage is consistent across architectures (LLaMA, Qwen, Gemma), with near-perfect detection on LLaMA-3.1 HaluEval (AUROC 0.949), suggesting that stronger models exhibit more distinguishable internal signatures of correctness.

To assess cross-dataset transferability, we train DUD-Probe on source datasets and evaluate on unseen targets (Figure 5). DUD-Probe exhibits markedly stronger generalization than baselines, maintaining high AUROC scores where SAPLMA and SEP frequently degrade below 0.6. For example, transferring from TriviaQA to SQuAD achieves an AUROC of 0.868, substantially surpassing ICR (0.613). Notably, out-of-distribution performance often approaches in-domain results (e.g., TriviaQA $\rightarrow$ HaluEval: 0.930), indicating that DUD-Probe captures task-agnostic mechanistic signatures such as FFN fragility rather than dataset artifacts. Overall, DUD-Probe exhibits minimal degradation ($< 5\%$), compared to $15\% - 20\%$ drops in baselines.

We further examine robustness to the inherent ambiguity in defining correctness for open-ended generation by varying the ROUGE-L threshold

Methods	Gemma-2-9B				Qwen2.5-7B				Llama-3.1-8B			
	Halu	**SQ**	**Hot**	**Triv**	**Halu**	**SQ**	**Hot**	**Triv**	**Halu**	**SQ**	**Hot**	**Triv**
PPL	0.5431	0.5415	0.7038	0.7535	0.5553	0.5201	0.6032	0.6649	0.5720	0.6231	0.6577	0.7238
LN-Entropy	0.7531	0.6491	0.7227	0.7388	0.7177	0.6342	0.6957	0.7431	0.6431	0.6203	0.6643	0.5994
Semantic Entropy	0.7745	0.7483	0.7361	0.7482	0.7293	0.7001	0.7142	0.7248	0.7745	0.7041	0.7143	0.7362
SAR	0.7856	0.7643	0.7217	0.7391	0.7647	0.7113	0.6845	0.7432	0.7793	0.7531	0.7335	0.7647
LLM-Check	0.5745	0.5703	0.5411	0.5593	0.5143	0.5555	0.5431	0.5729	0.5348	0.5331	0.5501	0.5275
SAPLMA	0.8100	0.7015	0.8201	0.7755	0.7901	0.6832	0.7507	0.8149	0.7328	0.7007	0.7638	0.7650
SEP	0.6331	0.6645	0.6218	0.7563	0.6647	0.6432	0.6507	0.7667	0.7547	0.7309	0.6537	0.7241
ICR Probe	0.8358	0.8204	0.8333	0.7917	0.8144	0.7203	0.7771	0.7544	0.7703	0.7548	0.8139	0.7557
DUD-Probe (Ours)	**0.8747**	**0.8218**	**0.8597**	**0.8199**	**0.8642**	**0.8204**	**0.8435**	**0.8252**	**0.9487**	**0.8959**	**0.8579**	**0.8031**

Table 1: Comparison of uncertainty quantification performance (AUROC) across three LLMs and four datasets. By displaying models horizontally, we highlight that **DUD-Probe** consistently achieves the best performance (bold) across different architectures. Halu: HaluEval, SQ: SQuAD, Hot: HotpotQA, Triv: TriviaQA.

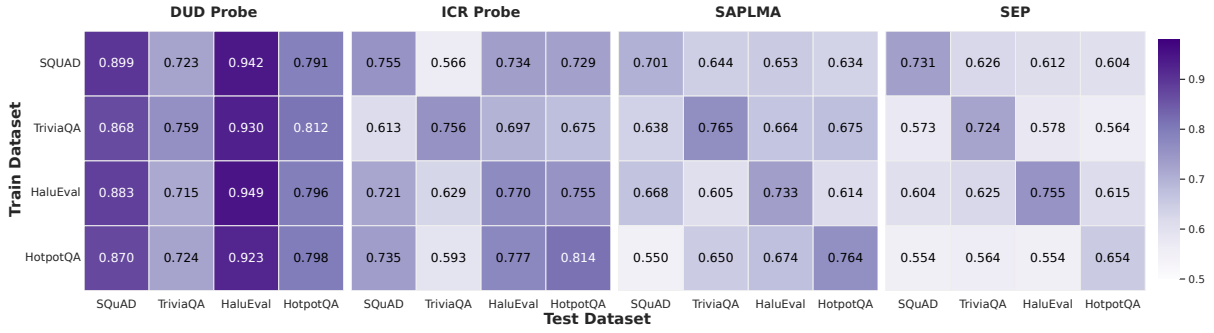

Figure 5: Cross-dataset generalization heatmaps for ICR Probe, SAPLMA, and SEP. Each subplot displays the AUROC when probe is trained on the row dataset and tested on column dataset, with values annotated in each cell.

used for labeling. As shown in Table 2, DUD-Probe consistently achieves the highest AUROC under Lenient (0.3), Standard (0.5), and Strict (0.7) criteria across all datasets. In contrast, baselines such as SEP and SAPLMA exhibit substantial performance fluctuations. Importantly, under the strictest setting, where generations must closely match references, DUD-Probe maintains strong performance, suggesting that the learned features capture intrinsic correctness rather than superficial lexical overlap.

Finally, to provide a comprehensive assessment, we report additional evaluation results in the Appendix. As detailed in Appendix F, DUD-Probe consistently improves calibration error (ECE) and rejection efficiency (PRR). We also analyze robustness across model scales (Appendix H), confirming stable performance from smaller to larger parameter regimes. Sensitivity analysis with respect to alternative labeling metrics (EM and BERTScore) further verifies that the observed gains are not artifacts of a particular evaluation choice (Appendix G).

TH	Method	Halu	SQ	Hot	Triv
Lenient (> 0.3)	SEP	0.6147	0.6493	0.6645	0.7503
	SAPLMA	0.7119	0.7135	0.7047	0.7531
	ICR Probe	0.7438	0.7001	0.7399	0.7769
	DUD (Ours)	**0.8528**	**0.7614**	**0.8579**	**0.8476**
Standard (> 0.5)	SEP	0.6647	0.6432	0.6507	0.7667
	SAPLMA	0.7901	0.6832	0.7507	0.8149
	ICR Probe	0.8144	0.7203	0.7771	0.7544
	DUD (Ours)	**0.8642**	**0.8204**	**0.8435**	**0.8252**
Strict (> 0.7)	SEP	0.6047	0.6214	0.6535	0.7443
	SAPLMA	0.7105	0.7031	0.7149	0.7348
	ICR Probe	0.7549	0.6943	0.7344	0.7599
	DUD (Ours)	**0.8135**	**0.8347**	**0.8636**	**0.8345**

Table 2: Robustness analysis on Qwen2.5-7B under varying ROUGE-L thresholds . DUD-Probe consistently outperforms baselines across all settings, demonstrating robustness to labeling criteria.

5.3 Mechanistic Ablation Analysis

To dissect the sources of DUD-Probe's performance, we conduct ablation studies on module contributions and layer hierarchy.

Impact of Decoupled Components. A core premise of DUD-Probe is that FFN (paramet-

ric memory) and Attention (contextual routing) provide distinct, complementary uncertainty signals. As shown in the upper section of Table 3, while single-stream probes (**Attn Only** and **FFN Only**) significantly outperform the random baseline, the Full dual-stream configuration consistently achieves the highest AUROC across all datasets. This confirms that FFN and Attention capture different aspects of generation failure for instance, a factual error might trigger an FFN collapse while leaving Attention relatively stable. By integrating both streams, the DUD-Probe effectively captures the full spectrum of uncertainty, validating the necessity of our decoupling strategy.

Impact of Layer Groups. We investigate the distribution of uncertainty signals by removing Early (EL, 1-10), Middle (ML, 11-21), and Deep (DL, 22-32) layer groups (Table 3, Part II). The **w/o EL** setting causes the most significant drop on HaluEval (0.95 $\rightarrow$ 0.85), confirming that early-layer instability reflects encoding failures in context-heavy tasks. Conversely, removing Middle and Deep layers universally degrades performance, particularly on reasoning tasks like HotpotQA. These results indicate that there is no single "golden layer" for uncertainty detection; instead, the probe relies on the holistic dynamic trajectory from initial encoding to final projection to make reliable judgments.

Configuration	Halu	SQ	Hot	Triv
I. Module Contribution				
Random	0.5000	0.5000	0.5000	0.5000
Attn Only	0.9144	0.8349	0.7933	0.7385
FFN Only	0.8968	0.8511	0.7872	0.7447
Full	**0.9487**	**0.8959**	**0.8579**	**0.8031**
II. Layer Contribution (removing from Full)				
w/o EL	0.8511	0.8351	0.7845	0.7164
w/o ML	0.8893	0.8241	0.7576	0.7281
w/o DL	0.9001	0.8024	0.8005	0.7542

Table 3: Mechanistic Ablation Study on Llama-3.1-8B. We analyze the impact of decoupled modules (I) and layer groups (II). Halu: HaluEval, SQ: SQuAD, Hot: HotpotQA, Triv: TriviaQA.

5.4 Case Study: Decoupling Faithfulness from Confidence

To intuitively understand how DUD-Probe operates beyond surface statistics, we present a token-level visualization of two representative cases in Table 4. We define the probe's output as a *Faithfulness Score*, displayed as subscripts.

The "Overconfidence" Trap *(Confident but Wrong)*

Q: What corresponds to the sport of stick gymnastics?

A: It$_{0.12}$ is$_{0.09}$ **Japan**$_{0.04}$.$_{0.08}$

Logits Conf: 0.96 (High) **DUD Score: 0.08** (Low) $\rightarrow$
DUD correctly flags mechanistic collapse.

The "Underconfidence" Save *(Hesitant but Correct)*

Q: What is the province traditionally known as?

A: Land$_{0.65}$ of$_{0.72}$ Fish$_{0.81}$ and$_{0.68}$ Rice$_{0.75}$

Logits Conf: 0.35 (Low) **DUD Score: 0.72** (High) $\rightarrow$
DUD confirms internal robustness.

Table 4: Token-level visualization of DUD faithfulness scores (subscripts). DUD effectively penalizes overconfident hallucinations (Case 1) while validating underconfident truths (Case 2).

Case 1: The Stubborn Error. The model answers the question about "stick gymnastics" with "Japan". Crucially, the model's output probability (Logits) is extremely high (≈ 0.96), indicating a state of deceptive certainty. However, our DUD-Probe assigns near-zero scores to the generated tokens (e.g., Japan$_{0.04}$). This confirms that despite the smooth output distribution, the internal mechanistic support for this answer is collapsed. The probe successfully punctures the veil of overconfidence to reveal the hidden epistemic failure.

Case 2: The Humble Truth. Conversely, for the "Land of Fish and Rice" query, the model's confidence is low (≈ 0.35), likely due to the rarity of the entity in the training distribution. A logits-based detector would flag this as uncertain. In contrast, DUD-Probe maintains high faithfulness scores (Avg 0.72) across the tokens. This indicates that the internal retrieval and routing mechanisms are actually stable, correctly identifying the generation as a valid, albeit "hesitant" truth.

These cases demonstrate that our decoupled dynamics capture the *intrinsic reliability* of the generation process, which is often effectively orthogonal to the model's calibrated confidence.

6 Conclusion

We introduced the **Decoupled Update Dynamics (DUD)** framework, shifting uncertainty quantification to active mechanistic diagnosis. By disentangling FFN and Attention contributions, we revealed that uncertainty manifests as systemic fragility and internal conflict. Our dual-stream probe effectively captures these signals, achieving

state-of-the-art performance and robust generalization. This demonstrates that monitoring the dynamic interplay of internal components offers a superior reliability proxy compared to surface-level probabilities or aggregated signals.

Limitations

While effective, our approach has limitations. First, as a white-box method requiring access to internal activations and causal interventions, DUD is inapplicable to closed-source API-based models (e.g., GPT-4). Second, although we optimized the probe efficiency, the causal tracing process involves multiple forward passes (clean, corrupted, and patched runs), which increases inference latency compared to simple logits-based methods. Future work will focus on distilling these mechanistic signals into lighter-weight detectors to enable real-time monitoring without heavy computational overhead.

Acknowledgments

This research is supported by the National Natural Science Foundation of China (No.62476127), the Natural Science Foundation of Jiangsu Province (No.BK20242039), the Basic Research Program of the Bureau of Science and Technology (ILF24001), the Scientific Research Starting Foundation of Nanjing University of Aeronautics and Astronautics (No.YQR21022), and the High Performance Computing Platform of Nanjing University of Aeronautics and Astronautics.

References

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*.

Yung-Sung Chuang, Linlu Qiu, Cheng-Yu Hsieh, Ranjay Krishna, Yoon Kim, and James Glass. 2024. Lookback lens: Detecting and mitigating contextual hallucinations in large language models using only attention maps. *arXiv preprint arXiv:2407.07071*.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James Glass, and Pengcheng He. 2023. Dola: Decoding by contrasting layers improves factuality in large language models. *arXiv preprint arXiv:2309.03883*.

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

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, and 1 others. 2021. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 1(1):12.

Tom Fawcett. 2006. An introduction to roc analysis. *Pattern recognition letters*, 27(8):861–874.

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.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR.

Binyuan Hui, Jian Yang, Zeyu Cui, Jiaxi Yang, Dayiheng Liu, Lei Zhang, Tianyu Liu, Jiajun Zhang, Bowen Yu, Keming Lu, and 1 others. 2024. Qwen2. 5-coder technical report. *arXiv preprint arXiv:2409.12186*.

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*.

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

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

Junyi Li, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023. Halueval: A large-scale hallucination evaluation benchmark for large language models. *arXiv preprint arXiv:2305.11747*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Andrey Malinin and Mark Gales. 2018. Predictive uncertainty estimation via prior networks. *Advances in neural information processing systems*, 31.

Andrey Malinin and Mark Gales. 2020. Uncertainty estimation in autoregressive structured prediction. *arXiv preprint arXiv:2002.07650*.

Potsawee Manakul, Adian Liusie, and Mark Gales. 2023. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pages 9004–9017.

Samuel Marks and Max Tegmark. 2023. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. *arXiv preprint arXiv:2310.06824*.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in gpt. *Advances in neural information processing systems*, 35:17359–17372.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

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*.

Alessandro Stolfo, Yonatan Belinkov, and Mrinmaya Sachan. 2023. A mechanistic interpretation of arithmetic reasoning in language models using causal mediation analysis. *arXiv preprint arXiv:2305.15054*.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, and 1 others. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

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.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, and 1 others. 2025a. sirens song in the ai ocean: A survey on hallucination in large language models. *Computational Linguistics*, pages 1–46.

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.

A Complete Causal Tracing Results

This section presents the complete layer-wise restoration score analysis across all four evaluation datasets, complementing the main text's focused discussion on representative datasets (HaluEval and HotpotQA).

Figure 6 provides a comprehensive view of the restoration dynamics across SQuAD, TriviaQA, HaluEval, and HotpotQA. Several consistent patterns emerge across all datasets:

Universal Layer-wise Patterns: Across all four datasets, we observe the characteristic "double-valley" structure discussed in Section 3, with Attention modules (red curves) exhibiting deeper negative scores in early layers (0-5), while FFN modules (blue curves) dominate in late layers (20-28). This universal pattern validates the temporal specialization of different modules in the prediction pipeline.

Middle-Layer Dynamics: The inset plots provide zoomed views of the middle layers (6-22), revealing critical differences between correct and incorrect predictions. In all datasets, correct predictions (solid lines) show restoration scores approaching zero or slightly positive, indicating successful semantic integration. In contrast, incorrect predictions (dashed lines) maintain persistent negative scores (approximately -0.3 to -0.5), suggesting locked-in erroneous trajectories. Notably, FFN scores consistently remain below Attention scores in this region, confirming FFN's dominant role in knowledge retrieval as discussed in Section 4.1.

Task-Specific Variations: Comparing context-rich tasks (SQuAD, HaluEval) with context-free tasks (TriviaQA, HotpotQA) reveals the computational burden shift identified in Section 4.2. Context-rich tasks exhibit deeper early-layer FFN valleys (reflecting input encoding costs), while context-free tasks show more pronounced late-layer FFN drops (reflecting parametric memory retrieval demands). These task-dependent dynamics underscore the adaptability of our method across different information sources.

Robustness vs. Fragility: The magnitude gap between correct (solid) and incorrect (dashed) predictions is consistent across all datasets, with incorrect predictions showing 2-3Œ deeper negative scores. This systematic difference evident in early valleys, middle-layer persistence, and late-layer drops provides the signal our probe leverages for hallucination detection. The consistency of this pattern across diverse tasks (reading comprehension in SQuAD, factual QA in TriviaQA, multi-hop reasoning in HotpotQA, and hallucination-focused evaluation in HaluEval) demonstrates the generalizability of restoration-based uncertainty quantification.

Cross-Dataset Consistency: Despite differences in task formulation and difficulty, all four datasets exhibit qualitatively similar restoration dynamics. Standard deviation analysis (not shown) confirms that the temporal ordering of module contributions (early Attention middle FFN late FFN) remains stable across datasets, analogous to the consistency observed in ICR scores (see ICR Probe Figure 5). This stability suggests that restoration scores capture intrinsic model dynamics rather than dataset-specific artifacts, supporting their use as a robust uncertainty signal.

B Layer-wise AUROC Analysis Across Models and Datasets

This section provides comprehensive layer-wise AUROC analysis across all three evaluated models (Qwen-2.5-7B, LLaMA-3.1-8B, Gemma-2-9B) and all four datasets (SQuAD, TriviaQA, HaluEval, HotpotQA). Each figure shows the discriminative power of individual layer restoration scores, demonstrating the consistency of temporal patterns across different model architectures and task types.

B.1 Qwen-2.5-7B-Instruct

Figure 7 presents the layer-wise AUROC analysis for Qwen-2.5-7B-Instruct. Key observations include:

Dataset-Specific Peak Patterns:

- **SQuAD (context-rich):** FFN peaks at Layer 15 (AUROC: 0.661), while Attention shows strong performance in late layers, peaking at Layer 26 (AUROC: 0.688).
- **TriviaQA (context-free):** Attention shows a pronounced peak at Layer 22 (AUROC: 0.735), whereas FFN peaks earlier at Layer 15 (AUROC: 0.679).
- **HaluEval (context-rich):** FFN exhibits strong middle-layer performance peaking at Layer 15 (AUROC: 0.679), while Attention outperforms in the final stages, peaking at Layer 26 (AUROC: 0.715).
- **HotpotQA (multi-hop):** FFN peaks at Layer 15 (AUROC: 0.698), while Attention reaches the highest discriminative power at Layer 26 (AUROC: 0.741).

Model-Specific Characteristics: Qwen-2.5-7B shows a consistent temporal pattern where FFN contributions peak in the middle layers (Layer 15), while Attention contributions tend to dominate in the late layers (Layers 22-26). The peak AUROC values generally range from 0.66 to 0.74.

B.2 LLaMA-3.1-8B-Instruct

Figure 8 presents the layer-wise AUROC analysis for LLaMA-3.1-8B-Instruct. Key observations include:

Dataset-Specific Peak Patterns:

- **SQuAD:** Attention peaks at Layer 15 (AUROC: 0.812), with FFN showing consistent performance across layers 10-25, peaking at Layer 24 (AUROC: 0.817).
- **TriviaQA:** FFN exhibits the strongest performance at Layer 24 (AUROC: 0.674), while Attention peaks at Layer 24 (AUROC: 0.649).
- **HaluEval:** FFN dominates with a peak at Layer 24 (AUROC: 0.864), while Attention peaks at Layer 24 (AUROC: 0.833), demonstrating clear temporal separation of module contributions.
- **HotpotQA:** Both modules show strong late-layer performance, with FFN peaking at Layer 23 (AUROC: 0.717) and Attention at Layer 24 (AUROC: 0.702).

Model-Specific Characteristics: LLaMA-3.1-8B achieves the highest peak AUROC values across all models, particularly in FFN modules.

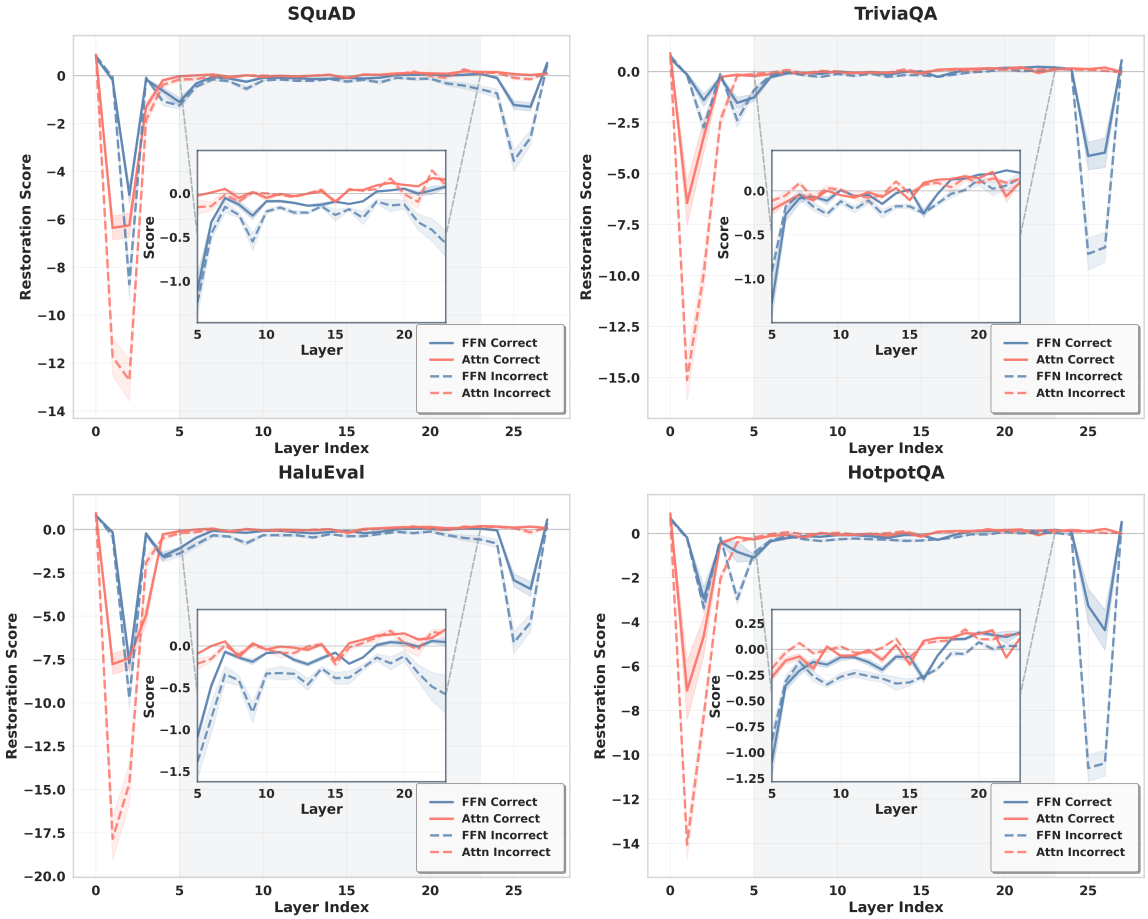

Figure 6: Layer-wise restoration scores across all four datasets: SQuAD (top-left), TriviaQA (top-right), HaluEval (bottom-left), and HotpotQA (bottom-right). Solid lines represent correct predictions; dashed lines represent incorrect predictions. Blue curves: FFN modules; Red curves: Attention modules. Inset plots show zoomed views of middle layers (6-22).

The model exhibits pronounced late-layer discriminative power, suggesting that its architecture emphasizes final-stage knowledge integration and projection for uncertainty-related signals.

B.3 Gemma-2-9B-IT

Figure 9 presents the layer-wise AUROC analysis for Gemma-2-9B-IT. Key observations include:

Dataset-Specific Peak Patterns:

- **SQuAD:** Attention shows a strong late-layer peak at Layer 38 (AUROC: 0.711), while FFN peaks earlier at Layer 20 (AUROC: 0.702).
- **TriviaQA:** Both modules show moderate performance, with Attention peaking at Layer 39 (AUROC: 0.604) and FFN at Layer 21 (AUROC: 0.602).
- **HaluEval:** Attention exhibits a peak at Layer 28 (AUROC: 0.654), while FFN shows a late peak at Layer 38 (AUROC: 0.648).
- **HotpotQA:** Attention achieves the highest performance at Layer 28 (AUROC: 0.651), with FFN peaking at Layer 21 (AUROC: 0.613).

Model-Specific Characteristics: Gemma-2-9B exhibits distinct behavior compared to Qwen and LLaMA, with Attention often outperforming FFN in peak AUROC values across datasets. The model generally shows lower overall AUROC scores (mostly in the 0.60-0.71 range) compared to LLaMA-3.1-8B, suggesting that uncertainty signals in Gemma might be more distributed or harder to isolate using single-layer restoration scores.

B.4 Cross-Model Summary

Table 5 summarizes the peak AUROC values for each model-dataset combination, highlighting the

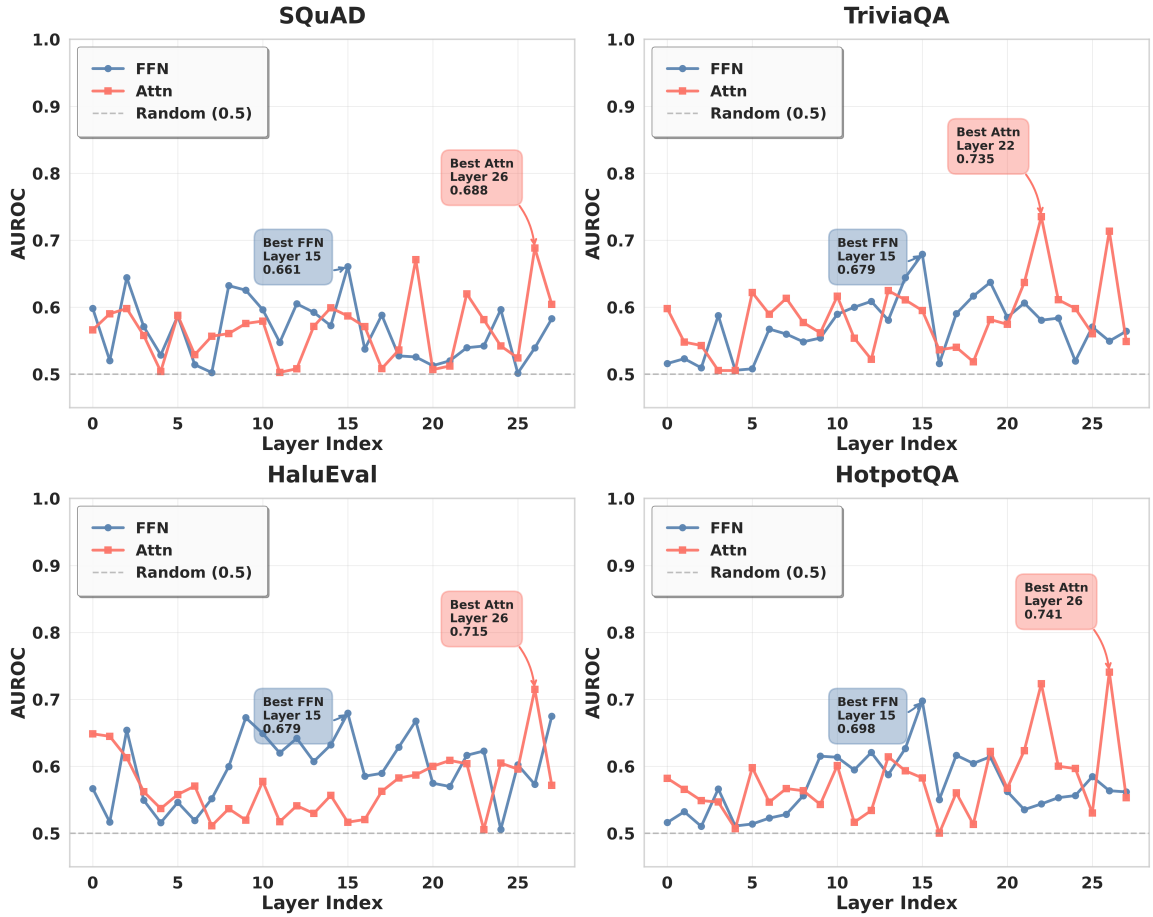

Figure 7: Layer-wise AUROC analysis for Qwen-2.5-7B-Instruct across four datasets: SQuAD (top-left), TriviaQA (top-right), HaluEval (bottom-left), and HotpotQA (bottom-right). Blue curves represent FFN modules; red curves represent Attention modules. Shaded regions indicate ± 1 standard deviation across 5-fold cross-validation. The horizontal dashed line at 0.5 represents random performance.

best-performing layer and module.

Universal Findings:

1. **Model-Dependent Module Dominance:** Contrary to a single universal rule, the dominant uncertainty signal varies by architecture. Qwen-2.5 and Gemma-2 consistently achieve peak performance using **Attention** restoration scores (dominating 8/8 cases combined), whereas LLaMA-3.1 relies exclusively on **FFN** signals (4/4 cases). This suggests that different architectures may localize failure modes in different components (contextual routing vs. parametric memory).

2. **Late-Layer Information Concentration:** Across all models, the most discriminative signals are concentrated in the deep layers (Layers 22-26 for Qwen/LLaMA, Layers 28-39 for Gemma). This validates that uncertainty is most effectively detected during the final stages of semantic integration and token projection.

3. **Performance Hierarchy:** LLaMA-3.1-8B exhibits the strongest internal signatures of correctness, achieving the highest peak AUROC (0.864), followed by Qwen-2.5 (0.741) and Gemma-2 (0.711). This correlates with the general capability of the base models, suggesting that stronger models possess more distinct mechanistic boundaries between correct and incorrect generations.

These results highlight that while the "late-layer" principle is universal, the specific module (Attention vs. FFN) serving as the best uncertainty proxy is architecture-dependent, necessitating the dual-stream approach of our DUD-Probe.

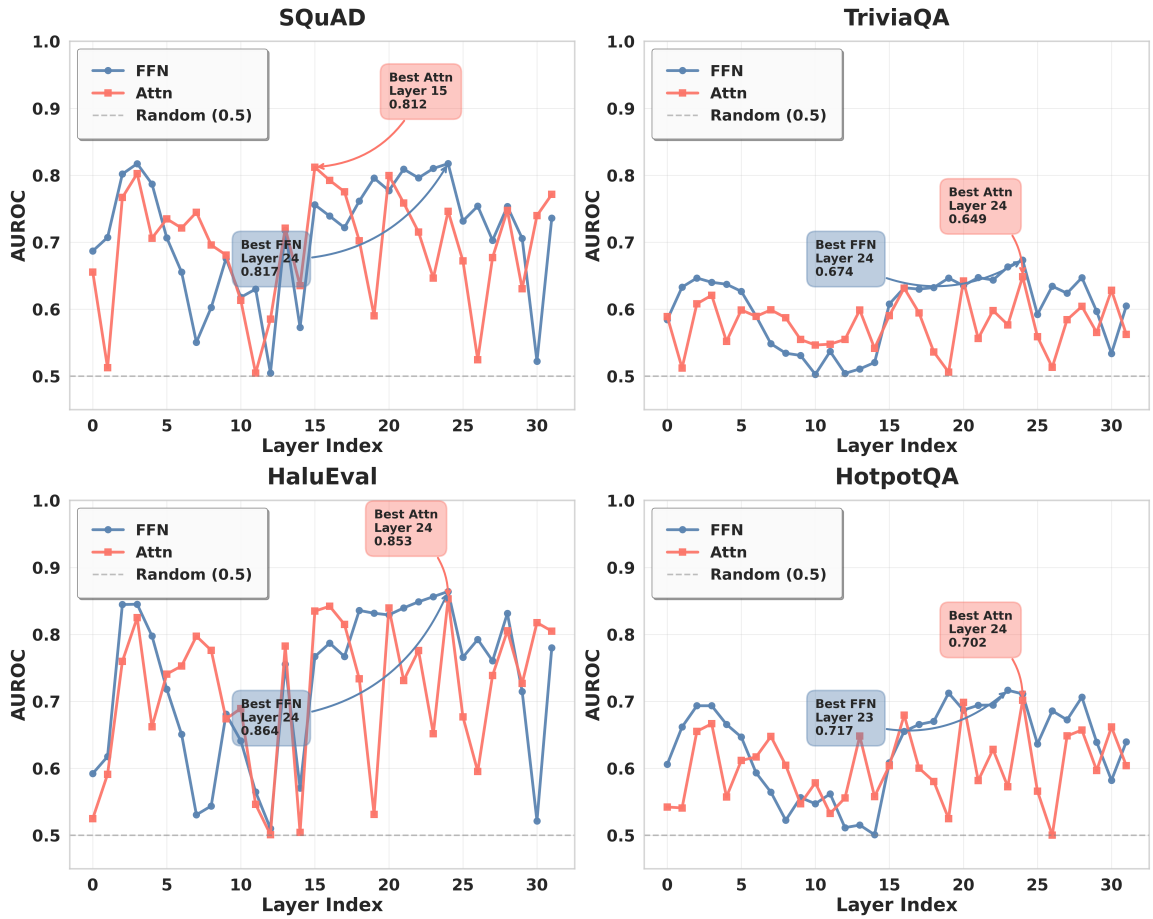

Figure 8: Layer-wise AUROC analysis for LLaMA-3.1-8B-Instruct across four datasets: SQuAD (top-left), TriviaQA (top-right), HaluEval (bottom-left), and HotpotQA (bottom-right). Blue curves represent FFN modules; red curves represent Attention modules. Shaded regions indicate ± 1 standard deviation across 5-fold cross-validation.

C Experimental Details

C.1 Computing Infrastructure

Our experiments were conducted on a server equipped with 4 NVIDIA A100 GPUs (80 GB memory each), running CUDA 12.4 and CentOS Linux 7 (Core). The system specifications are as follows:

```
NAME="CentOS Linux"
VERSION="7 (Core)"
ID="centos"
VERSION_ID="7"
PRETTY_NAME="CentOS Linux 7 (Core)"
```

Across multiple rounds of experiments, the total computational budget amounted to approximately 800-1000 GPU hours.

C.2 Prompt Templates

In our causal tracing experiments, we employ task-specific prompt templates with few-shot examples to ensure consistent evaluation across different datasets. This section details the exact prompts used for each dataset type.

C.2.1 Context-Rich Tasks (SQuAD & HaluEval)

For tasks where answers can be extracted from provided context passages, we use the following template structure:

System Prompt (Qwen/LLaMA models)

```
You are a helpful assistant. Read the context carefully and answer the question truthfully and concisely.
[Few-shot Examples]
Now answer the following question based on the context provided.
```

Few-shot Example Format:

```
Context: [Context Passage]
Question: [Question Text]
Answer: [Ground Truth Answer]
```

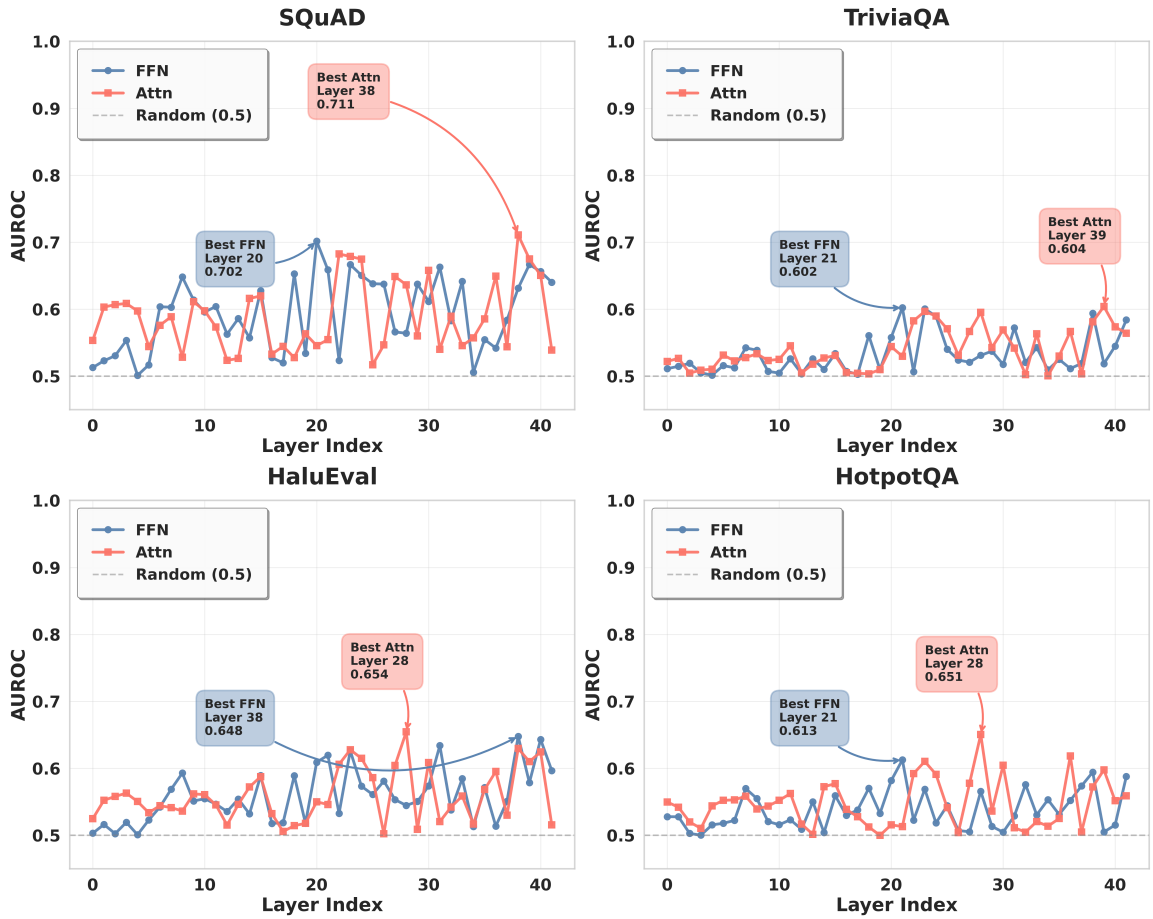

Figure 9: Layer-wise AUROC analysis for Gemma-2-9B-IT across four datasets: SQuAD (top-left), TriviaQA (top-right), HaluEval (bottom-left), and HotpotQA (bottom-right). Blue curves represent FFN modules; red curves represent Attention modules. Shaded regions indicate ±1 standard deviation across 5-fold cross-validation.

Test Query Format:

```
Context: [Test Context Passage]
Question: [Test Question]
Answer:
```

Note on Model Variants: For Gemma-2-9B, we consolidate the system prompt, few-shot examples, and transition text into a single user message, as this model architecture does not support separate system messages.

C.2.2 Context-Free Tasks (TriviaQA & HotpotQA)

For parametric knowledge tasks without external context, we use a simplified template:

System Prompt (Qwen/LLaMA models)

```
You are a helpful assistant. Answer the question
truthfully and concisely.
[Few-shot Examples]
Now answer the following question.
```

Few-shot Example Format:

```
Question: [Question Text]
Answer: [Ground Truth Answer]
```

Test Query Format:

```
Question: [Test Question]
Answer:
```

C.2.3 Implementation Details

Few-shot Configuration: We use the first 3 examples from each dataset as few-shot demonstrations, following standard practice in prompt-based evaluation. The remaining samples form the test pool.

Model-Specific Adaptations:

- **Qwen-2.5 & LLaMA-3.1:** Use separate `system` and `user` roles via the model's chat template API.
- **Gemma-2-9B:** Concatenate all prompt components into a single `user` message due to architecture constraints.

Model	Dataset	Best Module	Best Layer	Peak AUROC
Qwen-2.5-7B	SQuAD	Attention	26	0.69
	TriviaQA	Attention	22	0.74
	HaluEval	Attention	26	0.72
	HotpotQA	Attention	26	0.74
LLaMA-3.1-8B	SQuAD	FFN	24	0.82
	TriviaQA	FFN	24	0.67
	HaluEval	FFN	24	0.86
	HotpotQA	FFN	23	0.72
Gemma-2-9B	SQuAD	Attention	38	0.71
	TriviaQA	Attention	39	0.60
	HaluEval	Attention	28	0.65
	HotpotQA	Attention	28	0.65

Table 5: Summary of peak layer-wise AUROC values across all models and datasets. LLaMA-3.1-8B achieves the highest overall peak (0.86) on HaluEval using FFN restoration scores.

Generation Settings: All experiments use the following hyperparameters:

- Maximum new tokens: 32
- Temperature: Default (greedy decoding)
- Padding token: Model-specific EOS token
- Noise level: 0.15 for causal tracing

C.2.4 Complete Example

To illustrate the full prompt structure, we provide a concrete example from the SQuAD dataset:

Complete Prompt Example (SQuAD)

System:
```
You are a helpful assistant. Read the context carefully and answer the question truthfully and concisely.
```
[Few-shot Example 1]
```
Context: The Normans (Norman: Normaunds; French: Normands) were the people who in the 10th and 11th centuries gave their name to Normandy...
Question: In what country is Normandy located?
Answer: France
```
[Examples 2-3 omitted for brevity]
```
Now answer the following question based on the context provided.
```
User:
```
Context: [Test Context]
Question: [Test Question]
Answer:
```

Design Rationale: This prompt structure balances clarity, consistency, and practical constraints:

1. **Explicit Instructions:** Clear task definition reduces output ambiguity
2. **Standardized Format:** Uniform structure across samples minimizes confounding variables
3. **Moderate Few-shot:** Three examples provide sufficient task demonstration without excessive context length

C.3 Datasets

For our experiments, we use 10,000 instances from each dataset. Specifically, we use the QA subset of the HaluEval dataset and the standard splits for other datasets. The datasets used in this study are publicly available and adhere to their respective licenses, which allow for academic research and non-commercial use.

Dataset Details:

- **HaluEval** (Li et al., 2023): A large-scale hallucination evaluation benchmark focusing on QA tasks
- **SQuAD** (Rajpurkar et al., 2016): Stanford Question Answering Dataset for reading comprehension
- **TriviaQA** (Joshi et al., 2017): Large-scale reading comprehension dataset with evidence documents
- **HotpotQA** (Yang et al., 2018): Multi-hop question answering dataset requiring reasoning over multiple documents

Each dataset is split into 80%-20% for training and testing. We train and test on each dataset separately, reporting the corresponding results. This experimental setup for baseline methods matches ours exactly.

C.4 Baseline Methods

This section provides a brief overview of several baseline methods for hallucination detection.

PPL (Perplexity): PPL assesses the likelihood of hallucinations in LLM-generated responses by calculating perplexity (Ren et al., 2022). A higher perplexity value reflects increased uncertainty in the model's output, as hallucinations are often associated with the model's lack of confidence or inaccurate knowledge.

Length-Normalized Entropy (LN-Entropy): LN-Entropy is designed to quantify sequence-level uncertainty across multiple generations (Malinin and Gales, 2020). This metric normalizes entropy with respect to sequence length, facilitating a more precise assessment of the stability and reliability of generated content.

Shifting Attention to Relevance (SAR) : Proposed by (Duan et al., 2024), SAR is a single-pass method that computes uncertainty by focusing only on the "relevant" tokens in a generation. It first calculates token-level uncertainty (e.g., entropy) and then re-weights these scores based on each token's relevance to the core answer. This relevance is determined by analyzing the model's internal states (e.g., attention or gradients), allowing the method to disregard uncertainty arising from peripheral or non-essential tokens.

Semantic Entropy (SE) : Proposed by (Kuhn et al., 2023), this method first samples M candidate answers, groups them into clusters based on semantic equivalence, and then calculates the entropy over the probability distribution of these semantic clusters.

LLM-Check: LLM-check uses attention mechanism kernel similarity analysis to conduct hallucination detection (Zhang et al., 2025a). By calculating the log-determinant of attention kernel matrices and aggregating across heads, it captures characteristics related to hallucinations.

SAPLMA: SAPLMA trains a classifier to detect hallucinations using the hidden states of specific layers in LLMs (Chen et al., 2024). It captures internal signals from these layers to identify when the model is likely to generate hallucinated content.

SEP (Semantic Entropy Probe): SEP leverages linear probes trained on the hidden states of large language models (Kossen et al., 2024). It detects hallucinations by analyzing the semantic entropy of tokens before generation.

ICR (Internal Consistency Ratio): ICR tracks the dynamic evolution of hidden state updates across layers by measuring the contribution ratio between Attention and FFN modules to the residual stream (Zhang et al., 2025b). The ICR Probe trains a classifier on the layer-wise ICR scores to detect hallucinations, capturing the global update patterns throughout the model's forward pass rather than focusing on isolated layer features.

C.5 Causal Tracing Probe Training

Input and Output: The Causal Tracing Probe is a multi-layer perceptron (MLP) classifier trained to predict hallucinations using the pooled restoration scores. Given a model generation, the restoration score matrix $\mathbf{R}^{2L\times N}$ (where L is the number of layers and N is the number of generated tokens) is token-wise pooled to obtain a $1 \times 2L$ vector, which serves as the input to the probe. Specifically, dimensions 1 to L contain the pooled Attention restoration scores, and dimensions $L+1$ to $2L$ contain the pooled FFN restoration scores. The label for each instance is derived from annotated hallucination datasets.

Model Architecture: The probe consists of four fully connected layers, with batch normalization applied after each layer to stabilize training. The architecture follows the configuration $(2L, 128, 64, 32, 1)$, where $2L$ is the input dimension (twice the number of layers due to separate Attention and FFN scores). For models with approximately $L = 28$ layers, the total number of parameters remains below 16K, maintaining computational efficiency. ReLU activation functions are applied between layers, and the final layer uses a sigmoid activation to output a hallucination probability score.

Details of Training:

- The model is trained using the binary cross-entropy loss with the Adam optimizer (learning rate = 5×10^{-4}).
- We employ learning rate scheduling with a cosine annealing strategy.
- Training is performed for 50 epochs with early stopping based on validation AUROC.

- Batch size is set to 256, and we apply dropout (rate = 0.3) for regularization.
- The dataset is split into 80% training and 20% testing sets.
- We perform 5-fold cross-validation and report the average performance.

D Implementation Details of DUD-Probe

D.1 Probe Architecture and Training

Input Representation. Unlike previous methods that aggregate residual stream information into a single scalar per layer, the DUD-Probe operates on a high-dimensional feature space that preserves the independent contributions of different modules. For a model with L layers, the input is a concatenated vector $\mathbf{v} \in \mathbb{R}^{2L}$, constructed as:

$$\mathbf{v} = \text{Concat}\left([\mathcal{S}_{\text{ffn}}^{(1)}, \ldots, \mathcal{S}_{\text{ffn}}^{(L)}], [\mathcal{S}_{\text{attn}}^{(1)}, \ldots, \mathcal{S}_{\text{attn}}^{(L)}]\right) \tag{7}$$

where $\mathcal{S}_{\text{ffn}}^{(l)}$ and $\mathcal{S}_{\text{attn}}^{(l)}$ are the decoupled causal restoration scores for the FFN and Attention modules at layer l, respectively.

Model Configuration. We employ a Multi-Layer Perceptron (MLP) to capture the non-linear interactions between memory (FFN) and context (Attention) signals. Based on empirical tuning, we adopt a 4-layer architecture with the dimension progression: $2L \to 128 \to 64 \to 32 \to 1$. To ensure training stability and prevent overfitting on the probe dataset, we apply the following regularization techniques:

- **Batch Normalization:** Applied after each linear transformation to standardize layer inputs.
- **Activation Function:** We use **LeakyReLU** (negative slope $\alpha = 0.01$) for all hidden layers to handle potential negative values in the causal scores, while the final output uses a **Sigmoid** function to produce a probability $p \in [0, 1]$.
- **Dropout:** A dropout rate of $p = 0.3$ is applied after the activation function of each hidden layer.
- **Initialization:** Linear layers are initialized using Kaiming Uniform initialization.

Label Generation. To train the probe, we generate binary labels ($y \in \{0, 1\}$) for the model's responses. We use the ROUGE-L F-measure to compare the generated response with the ground truth answer. Following standard practices in hallucination detection benchmarks (e.g., HaluEval), a response is labeled as *Faithful* ($y = 0$) if the ROUGE-L score ≥ 0.5, and *Hallucinated* ($y = 1$) otherwise.

Training Hyperparameters. The probe is trained using the Binary Cross-Entropy (BCE) loss. We use the Adam optimizer with an initial learning rate of 5×10^{-4}. We employ a learning rate scheduler ('ReduceLROnPlateau') that reduces the learning rate by a factor of 0.5 if the validation loss does not improve for 5 consecutive epochs. The training is conducted with a batch size of 32 for up to 50 epochs, with early stopping triggered if validation performance stagnates.

E Algorithms

We summarize the procedures for extracting the decoupled dynamic features (Algorithm 1) and training the DUD-Probe (Algorithm 2).

F Additional Experimental Results

To further validate the reliability and practical utility of our uncertainty estimation, we present additional evaluation metrics on the Qwen2.5-7B model: **Expected Calibration Error (ECE)** and **Prediction Rejection Ratio (PRR)**.

F.1 Calibration Performance (ECE)

ECE measures the alignment between the model's predicted confidence and its actual accuracy. Lower ECE values indicate better calibration. As shown in Table 6, DUD-Probe achieves significantly lower ECE scores across all datasets compared to baselines. Notably, on HotpotQA, our method reaches an ECE of **0.0039**, which is orders of magnitude better than SEP (0.2879) and SAPLMA (0.1675). This indicates that the uncertainty scores produced by our probe can be interpreted as reliable probabilities of correctness, which is crucial for downstream decision-making.

F.2 Rejection Efficiency (PRR)

PRR evaluates how effectively an uncertainty metric can filter out incorrect predictions. Higher PRR values indicate that rejecting samples with high uncertainty leads to a greater improvement

Algorithm 1 Decoupled Feature Extraction via Causal Tracing

Require: LLM $\mathcal{M}$, Input prompt x, Answer tokens y, Number of layers L, Noise level ν
Ensure: Feature vector $\mathbf{v} \in \mathbb{R}^{2L}$

Step 1: Clean Run
Run $\mathcal{M}(x)$ to get clean probs P_{clean} and cache activations $\{\mathbf{m}^{(l)}, \mathbf{a}^{(l)}\}_{l=1}^{L}$
Step 2: Corrupted Run
$x^* \leftarrow x + \epsilon$, where $\epsilon \sim \mathcal{N}(0, \nu)$ ▷ Inject noise to embeddings
Run $\mathcal{M}(x^*)$ to get corrupted probs P_{corr}
TE $\leftarrow P_{\text{clean}} - P_{\text{corr}}$ ▷ Calculate Total Effect
Step 3: Decoupled Restoration (Patching)
Initialize score lists $S_{\text{ffn}} \leftarrow []$, $S_{\text{attn}} \leftarrow []$
for $l = 1$ to L **do**
    // Patch FFN Module
    Restore $\mathbf{m}^{(l)}$ into $\mathcal{M}(x^*)$
    $P^{\text{ffn}}_{\text{restored}} \leftarrow$ Probs from patched run
    $s_{\text{ffn}} \leftarrow (P^{\text{ffn}}_{\text{restored}} - P_{\text{corr}})/\text{TE}$
    Append s_{ffn} to S_{ffn}
    // Patch Attention Module
    Restore $\mathbf{a}^{(l)}$ into $\mathcal{M}(x^*)$
    $P^{\text{attn}}_{\text{restored}} \leftarrow$ Probs from patched run
    $s_{\text{attn}} \leftarrow (P^{\text{attn}}_{\text{restored}} - P_{\text{corr}})/\text{TE}$
    Append s_{attn} to S_{attn}
end for
Return $\mathbf{v} \leftarrow \text{Concat}(S_{\text{ffn}}, S_{\text{attn}})$

Method	HaluEval	SQuAD	HotpotQA	TriviaQA
SEP	0.1431	0.1537	0.0979	0.1019
SAPLMA	0.0997	0.1578	0.0675	0.0435
ICR Probe	0.0798	0.0314	0.0138	0.0662
DUD-Probe	**0.0425**	**0.0173**	**0.0039**	**0.0289**

Table 6: ECE performance on Qwen2.5-7B (ROUGE-L > 0.5). Lower values indicate better calibration.

in the overall accuracy of the remaining set. Table 7 demonstrates that DUD-Probe consistently achieves the highest PRR scores. For example, on HaluEval, our method achieves a PRR of **0.6346**, significantly outperforming ICR Probe (0.5003). This confirms that our decoupled features provide a highly discriminative signal for separating correct from incorrect generations.

G Sensitivity to Labeling Criteria

The definition of "correctness" in generation tasks can be subjective. To ensure that our performance gains are not artifacts of a specific labeling threshold (i.e., ROUGE-L > 0.5 used in the main text), we evaluate the robustness of the DUD-Probe on Qwen2.5-7B under varying strictness levels and alternative metrics.

G.1 Alternative Metrics: Exact Match and BERTScore

We further evaluate performance using **Exact Match (EM)**, which requires the generation to be identical to the ground truth, and **BERTScore-F1 (> 0.95)**, which measures semantic similarity.

Results in Table 8 demonstrate the superior adaptability of DUD-Probe.

- **Exact Match:** DUD-Probe achieves remarkable performance on SQuAD (**0.8801**) and HotpotQA (**0.8717**), significantly surpassing the ICR Probe. This suggests that decoupled traces are highly sensitive to the precise retrieval of correct entities required for exact matching.
- **BERTScore:** In semantic evaluation, our

Algorithm 2 DUD-Probe Training and Inference

Require: Training dataset $\mathcal{D} = \{(\mathbf{v}_i, y_i)\}_{i=1}^{N}$, Learning rate η, Batch size B
Ensure: Trained probe parameters θ

procedure TRAIN($\mathcal{D}$)
  Initialize MLP weights θ (Kaiming Uniform)
  for epoch $= 1$ to E **do**
    for batch $(\mathbf{V}, \mathbf{Y})$ in $\mathcal{D}$ **do**
      $\hat{\mathbf{Y}} \leftarrow \sigma(\text{MLP}_\theta(\mathbf{V}))$ ▷ Forward pass
      $\mathcal{L} \leftarrow \text{BCELoss}(\hat{\mathbf{Y}}, \mathbf{Y})$
      $\theta \leftarrow \theta - \eta \nabla_\theta \mathcal{L}$ ▷ Adam update
    end for
    Update η via Scheduler based on val loss
  end for
  return θ
end procedure

procedure INFERENCE(Prompt x, Answer y, $\mathcal{M}$, θ)
  $\mathbf{v} \leftarrow \text{FEATUREEXTRACTION}(x, y, \mathcal{M})$ ▷ See Alg. 1
  uncertainty $\leftarrow \sigma(\text{MLP}_\theta(\mathbf{v}))$
  return uncertainty
end procedure

Method	HaluEval	SQuAD	HotpotQA	TriviaQA
SEP	0.3927	0.3578	0.3243	0.3696
SAPLMA	0.4414	0.4675	0.4373	0.4459
ICR Probe	0.4989	0.5204	0.4361	0.5435
DUD-Probe	**0.6346**	**0.5669**	**0.5187**	**0.5945**

Table 7: PRR performance on Qwen2.5-7B (ROUGE-L > 0.5). Higher values indicate better rejection efficiency.

method maintains its lead (e.g., **0.8969** on HotpotQA), confirming that the probe captures semantic correctness rather than just lexical overlap.

H Model Size Comparison

To evaluate the robustness of our method across different model scales, we conduct experiments on two variants of the Qwen2.5 family: Qwen2.5-3B and Qwen2.5-14B. Table 9 presents the AUROC performance of DUD-Probe and baseline methods across both model sizes.

Key Observations:

1. **Consistent Superiority:** DUD-Probe achieves the highest AUROC across both model sizes and all datasets, confirming that the effectiveness of decoupled causal tracing is independent of model capacity.

2. **Scale Robustness:** While larger models (14B) generally improve baseline performance slightly, the relative advantage of DUD-Probe over ICR Probe remains substantial (e.g., +7.5% on HaluEval for 14B, +12.0% for 3B), indicating that our method scales well.

3. **Small Model Gains:** On the smaller Qwen2.5-3B model, DUD-Probe demonstrates particularly strong improvements over aggregated methods (ICR: 0.8017 DUD: 0.9212 on HaluEval), suggesting that explicit decoupling is especially valuable when model capacity is limited and internal conflicts are more pronounced.

I Cross-Model Generalization Analysis

To demonstrate the universality of our proposed framework, we extend the cross-dataset general-

Metric	Method	HaluEval	SQuAD	HotpotQA	TriviaQA
Exact Match	SEP	0.6544	0.6038	0.6359	0.7754
	SAPLMA	0.7341	0.7267	0.7148	0.7749
	ICR Probe	0.7554	0.7543	0.7497	0.7903
	DUD-Probe	**0.8204**	**0.8801**	**0.8717**	**0.8461**
BERTScore > 0.95	SEP	0.6643	0.6111	0.6799	0.7034
	SAPLMA	0.7267	0.7313	0.7524	0.7348
	ICR Probe	0.7347	0.7432	0.7781	0.7438
	DUD-Probe	**0.8111**	**0.8541**	**0.8969**	**0.8249**

Table 8: AUROC performance using Exact Match and BERTScore labeling on Qwen2.5-7B.

Model	Method	HaluEval	SQuAD	HotpotQA	TriviaQA
Qwen2.5-3B	SAPLMA	0.7638	0.7015	0.7547	0.7424
	SEP	0.6589	0.6606	0.6929	0.6745
	ICR Probe	0.8017	0.7484	0.7995	0.7731
	DUD-Probe	**0.9212**	**0.7894**	**0.8689**	**0.8475**
Qwen2.5-14B	SAPLMA	0.7527	0.7088	0.7541	0.7559
	SEP	0.6961	0.6955	0.6974	0.7003
	ICR Probe	0.8121	0.7432	0.7951	0.7546
	DUD-Probe	**0.8867**	**0.7911**	**0.8261**	**0.9209**

Table 9: AUROC comparison across different model sizes (ROUGE-L > 0.5). DUD-Probe consistently outperforms baselines regardless of model scale.

ization analysis to two additional state-of-the-art open-source model families: **Qwen2.5-7B** and **Gemma-2-9B**. This ensures that the effectiveness of DUD-Probe is not limited to a specific architecture but is applicable across different Transformer implementations.

Results on Qwen2.5. Table 10 presents the transferability results for Qwen2.5-7B. The model exhibits exceptional stability, with in-domain AUROC scores consistently exceeding 0.82. Notably, the transfer performance is remarkably high; for instance, a probe trained on *HotpotQA* achieves an AUROC of 0.8294 when tested on *TriviaQA*, a drop of less than 2% compared to the in-domain baseline. This suggests that Qwen2.5's internal uncertainty signatures are highly consistent across different knowledge-intensive tasks.

Results on Gemma-2. Table 11 details the performance on Gemma-2-9B. Despite architectural differences (e.g., sliding window attention, logit capping), DUD-Probe maintains robust generalization capabilities. The method achieves an average in-domain AUROC of approximately 0.84. While the transfer from *SQuAD* to other datasets shows a slight variance, the overall performance remains significantly superior to random guessing and competitive with in-domain training, further validating the robustness of decoupled dynamics as a generalized uncertainty proxy.

These findings, combined with the Llama-3 results in the main text, confirm that the mechanistic conflict between parametric memory and contextual processing is a fundamental indicator of uncertainty invariant across diverse LLM architectures.

Table 10: Cross-dataset generalization evaluation for **Qwen2.5-7B** (DUD-Probe). Each cell shows the AUROC when the probe is trained on the row dataset and evaluated on the column dataset. Diagonal entries (in-domain performance) are bolded.

Qwen2.5		**Test**			
		HaluEval	**SQuAD**	**HotpotQA**	**TriviaQA**
Train	HaluEval	**0.8642**	0.8059	0.8326	0.7685
	SQuAD	0.8270	**0.8204**	0.8275	0.7652
	HotpotQA	0.7860	0.7533	**0.8435**	0.7870
	TriviaQA	0.7543	0.7183	0.8294	**0.8252**

Table 11: Cross-dataset generalization evaluation for **Gemma-2-9B** (DUD-Probe). Each cell shows the AUROC when the probe is trained on the row dataset and evaluated on the column dataset. Diagonal entries (in-domain performance) are bolded.

Gemma-2		**Test**			
		HaluEval	**SQuAD**	**HotpotQA**	**TriviaQA**
Train	HaluEval	**0.8747**	0.7941	0.7668	0.7658
	SQuAD	0.7567	**0.8218**	0.7436	0.7555
	HotpotQA	0.7759	0.7749	**0.8597**	0.7632
	TriviaQA	0.7992	0.7625	0.7741	**0.8199**