

When Attention Escapes: Diagnosing and Mitigating Hallucination in Retrieval-Augmented Generation

Wencheng Su¹, Qinglong Wang⁴, Zhongjie Ba^{1,2*}, Zhixuan Chu^{1,2},
Zhenyu Zhong³, Liu Yan³, Kui Ren^{1,2}

¹The State Key Laboratory of Blockchain and Data Security, Zhejiang University

²Hangzhou High-Tech Zone (Binjiang) Institute of Blockchain and Data Security

³Ant Group ⁴Shanghai Xinyuan Innovation Center

{suwencheng, zhongjieba}@zju.edu.cn

Abstract

Retrieval-Augmented Generation (RAG) mitigates hallucinations by grounding generation in retrieved evidence, yet failures persist when models underutilize the retrieved context. While prior work attributes this to conflicts between external context and internal parametric knowledge encoded in Feed-Forward Networks (FFNs), our empirical analyses reveal that FFN-derived signals are unstable and insufficient for reliable diagnosis across diverse contexts. We therefore propose the **Attention Escape Score (AES)**, a training-free diagnostic framework that provides a mechanistic lens on RAG hallucinations via dynamic attention routing failures. We model the failures as weakened *Semantic Anchoring*, increased *Contextual Distraction*, and heightened *Generation Inertia*, moving beyond static component-level analyses. Extensive experiments show that AES significantly outperforms state-of-the-art baselines with robust cross-model generalization. Furthermore, we introduce **Pruning Escape Routes (PER)**, a lightweight inference-time intervention that suppresses hallucinations by attenuating identified escape pathways, achieving improved faithfulness with negligible degradation to complex reasoning and general capabilities. The code is available at <https://github.com/Eason8/AES-RAG>.

1 Introduction

Large Language Models (LLMs) (Llama Team, 2024; OpenAI, 2023; Yang et al., 2025) demonstrate remarkable performance in knowledge-intensive applications, while their real-world deployment remains fundamentally limited by hallucinations (Huang et al., 2025). Retrieval-Augmented Generation (RAG) has emerged as a widely adopted strategy to mitigate hallucinations by grounding generation in retrieved document (Ayala and Béchar, 2024; Gao et al., 2023).

*Corresponding author.

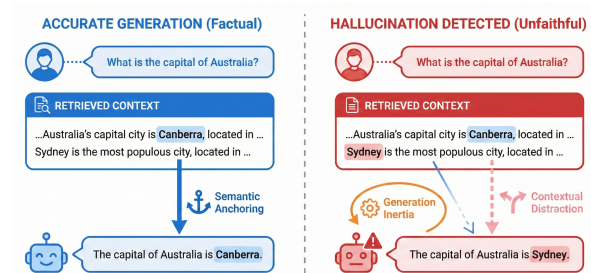


Figure 1: **Illustrative Examples of Attention Escape in RAG.** **Left:** accurate generation with attention anchored to the evidence token (*Canberra*). **Right:** hallucinated generation, where attention is influenced by contextual distraction toward an irrelevant token (*Sydney*) and shows signs of generation inertia.

Nevertheless, even with retrieved context, models may still underutilize or mis-handle contextual evidence, leading to unsupported or contradictory outputs (Niu et al., 2024).

Hallucination detection has shifted from black-box uncertainty heuristics to white-box diagnostics leveraging internal states. While prior work trained supervised probes on hidden representations (Zhang et al., 2025), RAG-oriented methods increasingly focus on *context-knowledge interaction* signals (Xu et al., 2024), often quantifying utilization via contrasting the external context with parametric contributions from Feed-Forward Networks (FFNs) (Sun et al., 2025; Yeh et al., 2025).

However, in RAG settings, FFNs inevitably process both attention-routed contextual information and internal parametric knowledge. This dual role entangles contextual transformation with memorized knowledge activation, making FFN-derived signals difficult to interpret reliably. Our empirical analysis confirms this limitation: FFN-based indicators exhibit substantial instability across layers and samples, failing to consistently distinguish whether the model relies on retrieved evidence or internal priors. Motivated by this ambiguity, we

instead focus on the **attention mechanism** (Bastings and Filippova, 2020), which directly governs information routing during generation. This perspective is especially important under the *lost-in-the-middle* phenomenon (Liu et al., 2024), where hallucinations arise when attention fails to reach relevant evidence in long contexts. Therefore, understanding attention routing dynamics is crucial for diagnosing hallucination in RAG systems.

Building on this insight, we propose the **Attention Escape Score (AES)**, a training-free framework that diagnoses hallucinations as dynamic routing failures along three dimensions: *Semantic Anchoring*, *Contextual Distraction*, and *Generation Inertia* (Figure 1). Furthermore, we introduce **Pruning Escape Routes (PER)**, a lightweight inference-time intervention that attenuates identified escape pathways. Crucially, PER suppresses hallucinations with negligible degradation to complex reasoning and general capabilities. Extensive experiments across diverse LLMs demonstrate that AES consistently outperforms strong baselines, while PER yields clear mitigation gains.

In summary, our core contributions are:

- An empirical study revealing that FFN dynamics in RAG systems lack stable, discriminative patterns for hallucination diagnosis, motivating a strategic shift toward attention routing.
- The **Attention Escape Score (AES)** for training-free detection via routing failures, and **Pruning Escape Routes (PER)** for lightweight, inference-time mitigation that preserves general model capabilities.
- Extensive evaluations demonstrating that AES robustly outperforms state-of-the-art baselines across diverse models and datasets, while PER significantly improves generation faithfulness.

2 Related Work

Hallucination Detection in LLMs. Early hallucination detectors mostly adopt *black-box* heuristics based on output uncertainty (Ren et al., 2023; Malinin and Gales, 2021; Manakul et al., 2023; Zhang et al., 2023). More recent *white-box* methods leverage internal states, often via supervised probes over hidden representations (Zhang et al., 2025; Orgad et al., 2025; CH-Wang et al., 2024; Li et al., 2023), but typically require labeled data and can be brittle under domain or model shifts.

In RAG systems, hallucination detection methods increasingly focus on *context-knowledge inter-*

action signals that reflect how models trade off retrieved evidence against parametric knowledge (Xu et al., 2024). ReDeEP (Sun et al., 2025) provides a mechanistic analysis of the roles of attention heads and FFNs, while LUMINA (Yeh et al., 2025) proposes utilization-based metrics; both link hallucinations to reduced context use and a stronger reliance on internal knowledge.

Attention Heads and Information Routing in RAG. Mechanistic interpretability studies (Olsson et al., 2022; Zheng et al., 2024; Ma et al., 2025; Jin et al., 2024) show that attention heads in Transformer-based LLMs exhibit diverse, functionally specialized behaviors and contribute unevenly to information routing. This is particularly salient in RAG systems, where long and heterogeneous contexts make evidence access non-trivial. Prior work (Ma et al., 2025; Cheng et al., 2025) finds that grounded generations often depend on a subset of attention heads maintaining focus on retrieved evidence, whereas hallucinations coincide with degraded routing and misdirected attention.

Building on these insights, we directly analyze dynamic attention routing through head-level patterns. Our method tracks how attention shifts among retrieved evidence, irrelevant context, and generated history during decoding, and leverages these routing dynamics to diagnose hallucination.

3 Preliminary Analysis: The Role of FFNs in RAG Hallucination

Recent studies hypothesize that RAG hallucinations arise from conflicts between retrieved context and FFN-hosted parametric knowledge (Geva et al., 2021; Sun et al., 2025; Yeh et al., 2025). Motivated by this view, we examine whether FFN dynamics provide reliable hallucination signals. Specifically, we analyze whether FFN updates exhibit discriminative patterns between hallucinated and factual generations using Llama2-7B (Touvron et al., 2023) on the RAGTruth benchmark (Niu et al., 2024).

3.1 Global Indistinguishability

We begin by examining whether FFN updates differ between hallucinated and factual samples at the level of *global distributional statistics*. Using Logit Lens (nostalgebraist, 2020), we decode intermediate residual states into vocabulary distributions to quantify FFN-induced semantic shifts.

Let $\mathbf{h}_{l,t}$ be the residual stream at position t before the FFN of layer l , then we have the post-FFN

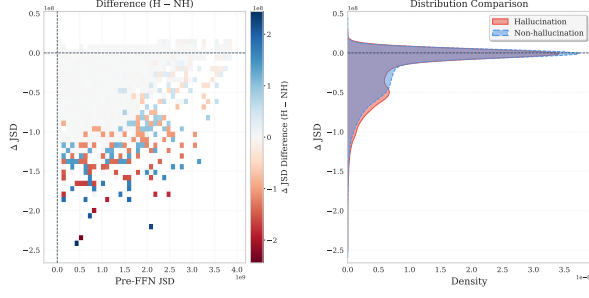


Figure 2: **Global Indistinguishability of FFN Injection.** (Left) Differential heatmap ($H - NH$) in the space of Pre-FFN semantic distance vs. ΔJSD . (Right) Marginal KDE plots of ΔJSD . The density curves for hallucinated (Red) and non-hallucinated (Blue).

state $\mathbf{h}'_{l,t} = \mathbf{h}_{l,t} + \text{FFN}(\mathbf{h}_{l,t})$. We project both representations using the unembedding matrix $\mathbf{E}$:

$$\begin{aligned} P_{\text{pre}}^{(l,t)} &= \text{softmax}(\mathbf{E}^\top \mathbf{h}_{l,t}), \\ P_{\text{post}}^{(l,t)} &= \text{softmax}(\mathbf{E}^\top \mathbf{h}'_{l,t}). \end{aligned} \quad (1)$$

We measure Jensen–Shannon Divergence (JSD) (Lin, 1991) to quantify how FFN updates shift intermediate predictions toward the final output distribution P_{final} . We aggregate token-level shifts by averaging over positions $t \in \{1, \dots, T\}$ and define the layer- l FFN-induced shift as:

$$\Delta\text{JSD}_l = \text{JSD}(P_{\text{post}}^{(l)} \| P_{\text{final}}) - \text{JSD}(P_{\text{pre}}^{(l)} \| P_{\text{final}}), \quad (2)$$

where $\Delta\text{JSD}_l < 0$ indicates a shift toward P_{final} .

Observation: Indistinguishable Distributional Patterns. Figure 2 shows that hallucinated and factual generations exhibit highly overlapping FFN-induced distributional shifts. Both the diffuse differential heatmap and overlapping KDE curves suggest that global FFN statistics alone provide limited discriminative power for hallucination diagnosis, motivating a finer-grained layer-wise analysis.

3.2 Layer-wise Instability

We further investigate whether specific layers exhibit distinctive pathological behaviors. We dissect FFN dynamics across network depth using two complementary metrics: one measuring the *magnitude* of the update, and the other measuring its *directional alignment* with hallucination.

Residual Contribution (η_l). We quantify the FFN update magnitude relative to the residual stream by computing the ratio of their L_2 norms.

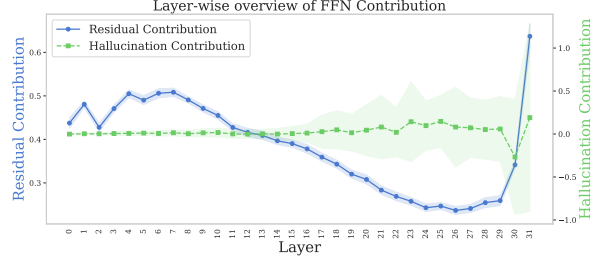


Figure 3: **Layer-wise FFN Contribution and Hallucination Alignment.** Left axis: Mean residual contribution magnitude η_l of FFN updates. Right axis: Directional hallucination contribution γ_l , with shaded regions indicating $\pm$ a standard deviation across tokens.

For a sequence of length T , the mean contribution at layer l reflects the amplitude of the FFN’s modification and is defined as:

$$\eta_l = \frac{1}{T} \sum_{t=1}^T \frac{\|\text{FFN}(\mathbf{h}_{l,t})\|_2}{\|\mathbf{h}_{l,t}\|_2}, \quad (3)$$

where $\mathbf{h}_{l,t}$ denotes the residual state at position t .

Directional Hallucination Contribution (γ_l). To access whether FFNs amplify hallucinated tokens, we compute their direct logit contributions. Let $\mathcal{H}$ and $\mathcal{R}$ denote hallucinated and factual token sets, respectively. For target token y_t , we examine the FFN update at the previous position:

$$\mathbf{d}_{l,t-1} = \text{FFN}(\mathbf{h}_{l,t-1}). \quad (4)$$

Let $\mathbf{W}_U \in \mathbb{R}^{V \times d}$ be the unembedding matrix and $\mathbf{w}_{y_t}$ the row corresponding to y_t . The average contribution over token set $\mathcal{S}$ is:

$$\text{score}(\mathcal{S}) = \frac{1}{|\mathcal{S}|} \sum_{t \in \mathcal{S}} \mathbf{d}_{l,t-1}^\top \mathbf{w}_{y_t}. \quad (5)$$

The directional hallucination contribution at layer l is then given by: $\gamma_l = \text{score}(\mathcal{H}) - \text{score}(\mathcal{R})$, where $\gamma_l > 0$ indicates FFNs amplify hallucinated tokens relative to factual ones.

Observation: Layer-wise Decoupling of Magnitude and Direction. Figure 3 reveals a clear decoupling between FFN update magnitude and hallucination-aligned direction. While η_l decreases smoothly with depth, γ_l fluctuates across layers and exhibits pronounced peaks mainly in deeper layers. Notably, layers with large FFN updates often contribute little to hallucination, whereas layers with stronger hallucination alignment tend to have relatively small residual contributions.

Overall, FFN updates do not provide a stable or separable signal for hallucination diagnosis. Similar trends on Llama2-13B and Mistral-7B suggest that this phenomenon generalizes across model scales and architectures (Appendix C).

Importantly, we do not argue that FFNs are irrelevant to hallucination. Rather, our findings suggest that aggregate FFN statistics can be inherently confounded in RAG, as FFNs simultaneously process retrieved contextual information and internal parametric knowledge. This motivates us to shift attention toward the attention mechanism, which directly controls information routing across the prompt, retrieved context, and generation history.

4 Methodology

We frame hallucination as an attention *routing failure*, where attention progressively *escapes* semantically grounded prompt regions toward irrelevant context or self-reinforcing generation. To characterize this, we propose the **Attention Escape Score (AES)**, a training-free framework modeling hallucination via three complementary dynamics: **Semantic Anchoring**, **Contextual Distraction**, and **Generation Inertia**. Formally, given a prompt $\mathcal{P}$ and response $\mathcal{R}$, let $C_{<t} = [\mathcal{P}; \mathcal{R}_{<t}]$ be the context at decoding step t , and $A^{(l,h)} \in \mathbb{R}^{T \times T}$ denote the attention matrix for layer l and head h .

4.1 Span-Level Semantic Alignment

Token-level attention is often noisy in long-context RAG systems. To establish a stable reference frame, we introduce a span-level semantic alignment. We partition $\mathcal{P}$ and $\mathcal{R}$ into semantically coherent spans $\{p_1, \dots, p_m\}$ and $\{r_1, \dots, r_n\}$ (construction strategies are compared in Appendix G). For each generated span r_i , its relevance to prompt span p_j is estimated via a sentence encoder $E(\cdot)$:

$$w_{i,j} = \text{ReLU}(\cos(E(r_i), E(p_j)) - \tau), \quad (6)$$

where τ filters weak associations.

Importantly, $w_{i,j}$ does not represent factual relevance; it only defines a soft response-conditioned semantic neighborhood in the prompt. Let $\mathcal{P}_{\text{rel}}^{(i)}$ denote the corresponding relevant prompt token indices. This semantic alignment serves solely as a coordinate system for measuring attention allocation, while all hallucination signals are derived entirely from internal attention statistics.

4.2 Identifying Functional Attention Heads

To disentangle generation dynamics, we analyze attention heads individually rather than relying on aggregate attention maps. Specifically, we characterize head functionality based on how attention mass is distributed relative to the response-conditioned semantic reference frame.

Semantic Union Coverage (SUC). For layer l and head h , let $A^{(l,h)} \in \mathbb{R}^{T \times T}$ denote the normalized attention matrix at decoding step t . Given a generated span r_i and context $C_{<t}$, we define the *Semantic Union Coverage* (SUC) as:

$$\text{SUC}_{l,h} = \frac{\sum_{j \in \mathcal{P}_{\text{rel}}} A^{(l,h)}[r_i, j]}{\sum_{k \in C_{<t}} A^{(l,h)}[r_i, k]}, \quad (7)$$

where $\mathcal{P}_{\text{rel}}$ denotes prompt token indices semantically aligned with r_i .

To handle attention sparsity, we define an activation score $\alpha_{l,h}$ as the mean top- k attention value from $\mathcal{R}_{<t}$ to $\mathcal{P}$. Heads with $\alpha_{l,h} < \delta$ are deemed *irrelevant*; those with $\alpha_{l,h} \geq \delta$ are candidates for functional classification based on SUC.

Head Functional Classification. Among heads with $\alpha_{l,h} \geq \delta$, we distinguish two functional roles based on SUC:

- **Useful Heads (Anchors):** high activation and high SUC (prompt-aligned routing);
- **Interfering Heads (Distractors):** high activation but low SUC (routing to weakly aligned prompt regions).

To reduce sensitivity to attention noise, we leave a margin between the two criteria and treat intermediate-SUC heads as neutral, excluding them from downstream analysis and intervention. We denote the resulting head sets as $\mathcal{H}_{\text{useful}}$ and $\mathcal{H}_{\text{int}}$, and define $\mathcal{H}_{\text{act}} = \mathcal{H}_{\text{useful}} \cup \mathcal{H}_{\text{int}}$. We report the layer-wise functional heads distributions in Appendix H.

4.3 Decomposed Attention Routing Dynamics

Motivated by layer-wise functional specialization (Skean et al., 2025), we focus on middle layers for routing analysis. Unlike early layers (surface features) or late layers (next-token prediction), middle layers support semantic integration and exhibit fewer extreme attention sinks (Xiao et al., 2024; Barbero et al., 2025), providing a stable window into information routing.¹

¹We use a contiguous middle range excluding the earliest and latest layers; details are in Appendix D.

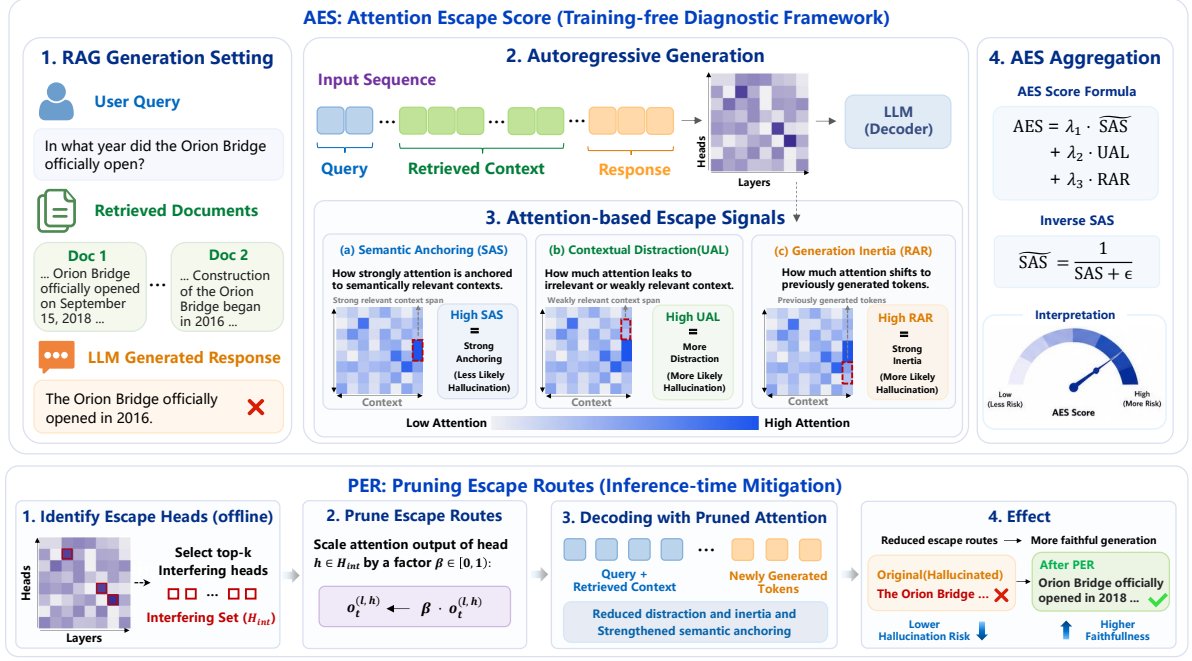


Figure 4: **Overview of AES and PER.** AES decomposes routing dynamics into three complementary signals: Semantic Anchoring (SAS), Contextual Distraction (UAL), and Generation Inertia (RAR), which are aggregated into the Attention Escape Score (AES). PER further mitigates hallucination by suppressing dominant interfering attention heads at inference time.

Accordingly, we characterize hallucination as progressive *attention escape* during decoding. As shown in Figure 4, we decompose routing dynamics into three complementary dimensions: *Semantic Anchoring* (alignment with relevant evidence), *Contextual Distraction* (leakage toward weakly related context), and *Generation Inertia* (reliance on generated history). Together, these signals capture how attention shifts away from grounded evidence, enabling fine-grained diagnosis of RAG routing failures.

4.3.1 Semantic Anchoring

The Semantic Alignment Score (SAS) quantifies anchoring strength by measuring attention mass routed to semantically aligned prompt regions. Aggregating over *useful* heads in middle layers $\mathcal{L}_{\text{mid}}$:

$$\text{SAS} = \sum_{l \in \mathcal{L}_{\text{mid}}} \sum_{h \in \mathcal{H}_{\text{useful}}} \alpha_{l,h} \cdot \frac{M(A^{(l,h)}, \mathcal{P}_{\text{rel}})}{M_{\text{total}}}, \quad (8)$$

where $\alpha_{l,h}$ is the activation score (Section 4.2), $M(A^{(l,h)}, \mathcal{P}_{\text{rel}})$ is the attention mass on relevant prompt tokens, and M_{total} is the total middle-layer attention mass for the span. Higher SAS indicates stronger evidence grounding.

4.3.2 Contextual Distraction

Beyond anchoring, attention may *leak* to prompt regions that are only weakly aligned with the generated content. We quantify such contextual distraction with Unrelated Attention Leakage (UAL):

$$\text{UAL} = \frac{\sum_{l \in \mathcal{L}_{\text{mid}}} \sum_{h \in \mathcal{H}_{\text{act}}} \Delta M(A^{(l,h)})}{M_{\text{total}}}, \quad (9)$$

where $\Delta M(A^{(l,h)}) = M(A^{(l,h)}, \mathcal{P}) - M(A^{(l,h)}, \mathcal{P}_{\text{rel}})$ and $\mathcal{P}$ is the full prompt token set. Higher UAL reflects greater contextual distraction disrupting anchored routing.

4.3.3 Generation Inertia

Response Attention Ratio (RAR) measures the extent to which attention shifts from the prompt to the generated history. Specifically, it is the fraction of middle-layer attention mass (aggregated over active heads) that is allocated to generated tokens:

$$\text{RAR} = 1 - \frac{1}{M_{\text{total}}} \sum_{l \in \mathcal{L}_{\text{mid}}} \sum_{h \in \mathcal{H}_{\text{act}}} M(A^{(l,h)}, \mathcal{P}). \quad (10)$$

Higher RAR indicates stronger self-reinforcement and generation inertia.

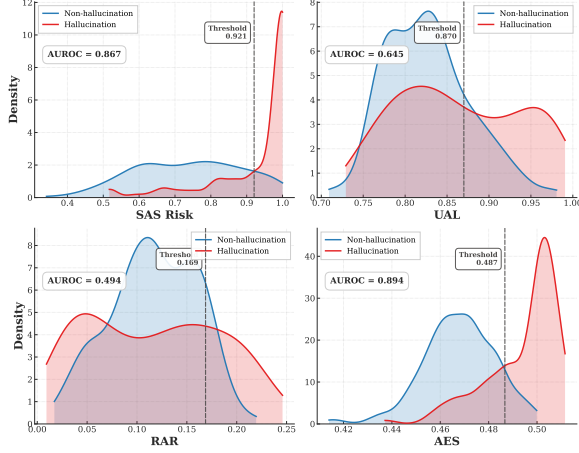


Figure 5: **Distributions of routing signals.** KDEs of the signals for hallucinated vs. non-hallucinated samples; dashed lines mark training-selected thresholds.

4.4 Detection: Attention Escape Score

These signals capture complementary hallucination risks: higher SAS lowers risk, while higher UAL and RAR increase it. We apply a monotonic transform $\widetilde{\text{SAS}} = 1/(\text{SAS} + \epsilon)$ for numerical stability and define the *Attention Escape Score (AES)* as:

$$\text{AES} = \lambda_1 \cdot \widetilde{\text{SAS}} + \lambda_2 \cdot \text{UAL} + \lambda_3 \cdot \text{RAR}, \quad (11)$$

where $\lambda_1, \lambda_2, \lambda_3 > 0$ are fixed coefficients. Component contributions are analyzed in Section 5.4.

Figure 5 provides a distributional view of separability. **SAS** achieves the strongest separation, underscoring semantic anchoring as the key signal. **UAL** and **RAR** are less discriminative in the bulk but capture complementary tail cases (diffuse attention and atypical routing). By integrating these cues, **AES** enhances separability and stabilizes risk scores across long contexts. Appendix I further demonstrates AES’s robustness to moderate coefficient perturbations, indicating precise weight tuning is unnecessary.

4.5 Mitigation: Pruning Escape Routes

Building on the above analysis, we further propose **PER (Pruning Escape Routes)**, a lightweight inference-time intervention method designed to suppress dominant *attention escape routes*. PER adopts a static selection strategy by pre-identifying a compact, model-specific set of heads that consistently exhibit *interfering* behavior on hallucinated samples. Specifically, we apply our head-role classifier on the hallucination split, aggregate per-span head labels, and select the top- k heads (e.g., $k=10$)

with the highest interfering frequency, forming a fixed intervention set $\mathcal{H}_{\text{int}}$.

At inference, for each head $(l, h) \in \mathcal{H}_{\text{int}}$, we attenuate its output vector $\mathbf{o}_t^{(l,h)}$ by a constant scalar $\beta \in [0, 1]$: $\tilde{\mathbf{o}}_t^{(l,h)} = \beta \cdot \mathbf{o}_t^{(l,h)}$, where $\tilde{\mathbf{o}}_t^{(l,h)}$ denotes the modified head output before the final linear projection. This static attenuation weakens dominant escape pathways by pruning distraction-prone channels, effectively reducing hallucinatory drift with negligible runtime overhead.

5 Experiments

5.1 Experimental Settings

Datasets. We evaluated two public RAG hallucination benchmarks: **RAGTruth** (Niu et al., 2024), which provides span-level hallucination annotations across multiple domains and generation tasks; and **Dolly (AC)** (Hu et al., 2024), an instruction-following benchmark under the Accurate Context setting with manually curated hallucination labels from Sun et al. (2025). Detailed statistics of the datasets are provided in Appendix A.

Models. We evaluated AES in various LLM families, including **Llama2** (7B, 13B), **Llama3-8B**, and **Mistral-7B**, to assess whether attention escape dynamics is generalized across architectures and scales. Importantly, unless otherwise specified, all experiments use the **same configuration and aggregation coefficients** across models and datasets without per-model or per-dataset retuning.

Baselines. We compare against eight representative detectors spanning four paradigms: *uncertainty-based* methods (Perplexity (Ren et al., 2023), LN-Entropy (Malinin and Gales, 2021)); *cross-sample consistency* methods (Self-CheckGPT (Manakul et al., 2023), EigenScore (Chen et al., 2024)); *self-evaluation* methods (P(True) (Kadavath et al., 2022)); and *internal-signal* methods, including supervised probing over hidden states (ICR Probe (Zhang et al., 2025)) and utilization-based diagnostics of contextual grounding and parametric contribution (ReDeEP (Sun et al., 2025), LUMINA (Yeh et al., 2025)). Details are provided in Appendix B.

Evaluation Metrics. We evaluate detection performance using four metrics: **AUROC** for ranking separability, **Accuracy (ACC)** for classification correctness, **Pearson Correlation Coefficient**

LLM	Approach	RAGTruth				Dolly (AC)			
		AUROC $\uparrow$	ACC $\uparrow$	PCC $\uparrow$	F1 $\uparrow$	AUROC $\uparrow$	ACC $\uparrow$	PCC $\uparrow$	F1 $\uparrow$
Llama2-7B	Perplexity	0.5091	0.5444	-0.0156	0.6677	0.3085	0.5859	-0.2028	0.7389
	LN-Entropy	0.6965	0.6933	0.3318	0.7000	0.6682	0.6667	0.2857	0.7481
	SelfCheckGPT	0.5230	0.5289	0.0549	0.6667	0.7283	0.6970	0.4004	0.7541
	EigenScore	0.5114	0.5333	0.0250	0.6635	0.7523	0.7071	0.4363	0.7680
	P(True)	0.6394	0.6022	0.1815	0.6703	0.5225	0.5859	0.0562	0.7389
	ICR Probe	0.7274	<u>0.7333</u>	0.3947	0.7273	0.4479	0.5000	-0.0969	0.5000
	ReDeEP	0.7474	0.6889	0.4207	0.7166	0.7788	<u>0.7374</u>	0.4864	<u>0.7816</u>
	LUMINA	<u>0.7646</u>	-	<u>0.4546</u>	<u>0.7404</u>	-	-	-	-
	AES	0.8829	0.8467	0.6370	0.7857	<u>0.7668</u>	0.7802	<u>0.4855</u>	0.8000
Llama2-13B	Perplexity	0.4121	0.5111	-0.1407	0.6463	0.3177	0.5455	-0.2495	0.7059
	LN-Entropy	0.7676	0.7244	0.4447	0.7293	0.6216	0.6263	0.2086	0.7176
	SelfCheckGPT	0.5475	0.5600	0.1740	0.6574	0.7510	0.7374	0.4648	0.7400
	EigenScore	0.7853	0.7533	0.4136	0.7259	<u>0.7770</u>	<u>0.7475</u>	<u>0.4699</u>	<u>0.7788</u>
	P(True)	0.5332	0.5156	0.0851	0.6305	0.6354	0.6364	0.2070	0.7000
	ICR Probe	<u>0.8900</u>	<u>0.8111</u>	0.6967	<u>0.7848</u>	0.6768	0.5500	0.3881	0.5263
	ReDeEP	0.7977	0.7356	0.4728	0.7302	0.7613	0.7172	0.4335	0.7667
	LUMINA	0.8569	-	0.6041	0.7797	-	-	-	-
	AES	0.9250	0.8763	<u>0.6407</u>	0.8205	0.8471	0.8222	0.6326	0.8400
Llama3-8B	Perplexity	0.7127	<u>0.6889</u>	0.3566	0.7083	0.6152	0.6364	0.2376	0.5909
	LN-Entropy	0.7072	0.6844	0.3500	0.7160	0.6451	0.6566	0.3154	0.5854
	SelfCheckGPT	0.5269	0.5711	0.0510	0.6932	0.5622	0.5556	0.0941	0.5686
	EigenScore	0.4561	0.5400	-0.0702	0.7013	<u>0.7204</u>	0.6768	<u>0.3405</u>	<u>0.6923</u>
	P(True)	0.6963	0.6822	0.2863	0.6635	0.6577	0.6465	0.2634	0.5977
	ICR Probe	<u>0.7695</u>	0.6778	<u>0.4674</u>	<u>0.7339</u>	0.4896	0.5500	-0.0648	0.4706
	ReDeEP	0.7139	0.5911	0.3653	0.7107	0.6846	<u>0.7071</u>	0.3073	0.6813
	LUMINA	0.7446	-	0.4236	0.7211	-	-	-	-
	AES	0.8807	0.7830	0.6388	0.7566	0.7995	0.7931	0.5483	0.7353

Table 1: **Main Results of Hallucination Detection.** We compare **AES** against representative baselines on **RAGTruth** and **Dolly (AC)**. In all result tables, the best score is **bolded**, the second-best is underlined, and gray shading indicates our method.

(**PCC**) for score–label correlation, and **F1-score** for the precision–recall trade-off.

5.2 Main Results

Detection Results. Table 1 summarizes the main hallucination detection results. AES consistently achieves the strongest or near-strongest performance across models and datasets, with particularly large gains in AUROC and PCC.

Importantly, AES is evaluated using a *single globally fixed configuration* across all experiments, without dataset-specific tuning or model-specific recalibration. Despite this, AES transfers effectively across both architecture families and parameter scales. In particular, when applied directly to Mistral-7B under the same configuration (Table 2), AES still substantially outperforms the previous baselines, demonstrating that attention escape dynamics remain stable across heterogeneous LLMs.

Furthermore, additional experiments on Llama-2-70B and the Qwen-3 family (Detailed results are in Appendix J) show that the observed routing behaviors persist in **larger and more modern**

Approach	AUROC $\uparrow$	ACC $\uparrow$	PCC $\uparrow$	F1 $\uparrow$
Perplexity	0.6416	0.6511	0.1842	0.7260
LN-Entropy	0.7312	0.7067	0.3808	0.7700
SelfCheckGPT	0.5955	0.6067	0.2107	0.7330
EigenScore	0.6892	0.6644	0.3390	0.7392
P(True)	0.2845	0.5578	-0.2968	0.7161
ICR Probe	<u>0.8040</u>	<u>0.7444</u>	<u>0.5371</u>	<u>0.7965</u>
LUMINA	0.7685	-	0.4623	0.7728
AES	0.8931	0.8673	0.6689	0.8379

Table 2: **Hallucination detection results on Mistral-7B.** We evaluate AES and baselines on Mistral-7B-Instruct-v0.1 under the same settings as Table 1.

models, suggesting that attention escape reflects a general property of autoregressive RAG generation rather than a small-model artifact.

Comparison with Internal-Signal Baselines. AES demonstrates stronger robustness than prior internal-signal methods. Unlike supervised approaches such as ICR Probe (Zhang et al., 2025), whose performance degrades substantially under distribution shift (e.g., on Dolly), the training-

Setting	Win	Tie	Lose	Net Improv.
PER vs original	67.3%	2.7%	30.0%	+37.3%
PER vs AARF	65.3%	5.6%	29.1%	+36.2%

Table 3: **Hallucination mitigation performance.** **Win/Tie/Lose** denote the percentage of cases where the *first* method is judged better/equal/worse than the *second*. Net Improvement is the win-loss margin.

free AES maintains consistent effectiveness across datasets and architectures. AES also matches or outperforms context-knowledge utilization methods such as ReDeEP (Sun et al., 2025) and LUMINA (Yeh et al., 2025), despite using a single globally fixed aggregation scheme across all experiments. In contrast, ReDeEP requires dataset-specific component selection and tuning.

These results suggest that attention routing dynamics provide a more stable and transferable diagnostic signal than proxy measurements of parametric-versus-contextual reliance.

Mitigation Results. We evaluate PER by attenuating escape-related attention pathways during inference and comparing the resulting generations with the original outputs. To reduce potential judge-specific bias, we conduct pairwise evaluations using three independent LLM judges, **DeepSeek-V3.2**, **GPT-5.1**, and **Qwen3-32B**. As shown in Table 3, all three judges consistently prefer PER over the original model outputs, yielding positive net improvements and supporting the robustness of the observed gains. PER also consistently outperforms AARF (Sun et al., 2025) with a substantial win-loss margin.

Notably, the intervention heads identified on RAGTruth are directly transferred to Dolly without re-selection, while keeping the same head set and hyperparameter configuration. PER remains effective on Dolly (details are provided in Appendix F), providing further evidence that the identified attention escape pathways exhibit cross-dataset stability.

Reasoning Preservation. A key concern for head-level intervention is whether suppressing specific heads may inadvertently impair reasoning or general capabilities, especially when attention heads exhibit multifunctional behaviors.

To evaluate this, we assess PER on **HoVer** (Jiang et al., 2020), a multi-hop verification benchmark requiring complex reasoning, using 1,000 uniformly sampled dev instances under greedy decoding setting. We also evaluate general knowledge retention

Benchmark	Baseline	+PER	Δ
HoVer (Multi-hop)	98.90	99.70	+0.80
MMLU (General)	40.00	39.43	-0.57

Table 4: **Reasoning and capability preservation.** Accuracy (%) of Llama-2-7B-Chat with and without PER.

on **MMLU** (Hendrycks et al., 2021).

As shown in Table 4, PER slightly improves HoVer accuracy (+0.80%) while causing only a marginal drop on MMLU (-0.57%) for Llama-2-7B-Chat. These results demonstrate that PER selectively attenuates escape-prone routing pathways while preserving the multifunctional attention heads essential for complex reasoning and general knowledge retention.

5.3 Analysis of Alignment and Attention Baselines

Beyond Semantic Alignment and Coarse Attention. Since AES uses semantic alignment to establish a reference frame for attention analysis, we further examine whether its performance can be explained by semantic similarity alone. We compare AES with two alignment-only baselines: Align-MaxSim, which uses the maximum similarity between a generated span and context spans, and Align-MeanTop3, which averages similarities over the three most aligned spans. We further include three simple attention baselines that measure attention allocated to retrieved evidence, top-aligned spans, or previously generated tokens.

As shown in Table 5, both alignment-only and attention-only baselines substantially underperform AES. This indicates that neither semantic matching nor coarse attention allocation alone is sufficient for reliable hallucination detection. Instead, AES benefits from jointly modeling complementary routing behaviors through Semantic Anchoring, Contextual Distraction, and Generation Inertia.

5.4 Ablation Study

Component Ablation. Table 6 shows that individual routing signals are insufficient in isolation. Among them, SAS is the strongest standalone component, while UAL and RAR primarily capture complementary tail behaviors.

Although certain partial combinations can achieve competitive **Best AUROC** under carefully tuned weights, their **Top-10 Mean AUROC** exhibits substantially larger variance, indicating poor

Method	AUROC	F1
<i>Alignment-only</i>		
Align-MaxSim	0.6676	0.3006
Align-MeanTop3	0.6488	0.2964
<i>Attention-only</i>		
Attn-Evidence-Ratio	0.6410	0.2906
Attn-TopAligned-Ratio	0.6245	0.2782
Attn-Self-Ratio	0.5559	0.2504
AES (Ours)	0.8829	0.7857

Table 5: **Comparison with alignment-only and simple attention baselines.** AES substantially outperforms both types of controls, indicating that its performance cannot be explained by semantic matching or coarse attention allocation alone.

Method Variant	Best AUROC	Top-10 Mean AUROC
AES	0.8941	0.8812 $\pm$ 0.004
SAS (Only / w/o)	0.8670 / 0.8869	0.8571 / 0.7729 $\pm$ 0.072
UAL (Only / w/o)	0.6453 / 0.8680	0.6453 / 0.8525 $\pm$ 0.019
RAR (Only / w/o)	0.4942 / 0.8571	0.4942 / 0.8185 $\pm$ 0.024

Table 6: **Component Ablation of AES with Robustness.** For each component X , we report *Only X / w/o X* (using the other two components). **Top-10 AUROC** reports mean $\pm$ std over the top-10 weight configurations ranked by AUROC (best AUROC for *Only* variants).

robustness to coefficient selection. In contrast, full AES maintains consistently strong performance with low variance, suggesting that the three routing signals provide complementary and stable diagnostics under a single fixed aggregation scheme.

6 Conclusion

In this work, we study RAG hallucinations from the perspective of attention routing and show that failures to utilize relevant retrieved evidence can be characterized through attention dynamics. We propose AES, a training-free framework that captures three complementary aspects of attention escape: weakened *Semantic Anchoring*, increased *Contextual Distraction*, and heightened *Generation Inertia*. Experiments across multiple datasets and LLM families show robust detection under a fixed hyperparameter configuration, with cross-dataset results supporting the stability of the identified routing patterns. We further introduce PER, a lightweight inference-time intervention that attenuates attention heads associated with systematic escape routes. PER effectively reduces hallucination while largely preserving general knowledge and multi-hop reasoning capabilities, suggesting that targeted attention intervention can provide a practical direction

for improving RAG reliability.

Limitations

Despite the promising results of our framework for diagnosing and mitigating RAG hallucinations, several limitations remain. **First**, our method requires access to internal activations and attention states, limiting its applicability to closed-source or API-only models. **Second**, AES relies on response-conditioned span-level alignment to define a semantic reference frame for measuring attention routing. Although this approach is robust across encoders and span granularities (Appendix G), the alignment may not match gold supporting evidence, especially in domains with technical or implicit evidence. **Finally**, our evaluation focuses on standard RAG settings with sufficiently relevant and reliable retrieved evidence. Extending AES/PER to noisy retrieval and developing causal measurements to disentangle contextual and parametric effects remain important directions for future work.

Ethical Considerations

This work does not involve the collection, processing, or inference of sensitive personal information. All experiments are conducted on publicly available benchmarks in compliance with standard research ethics and data usage policies. Our goal is to improve the reliability and transparency of LLMs in RAG systems by identifying and mitigating hallucinated outputs. Although such techniques could be misused to optimize deceptive systems, our contributions are intended to support safer deployment and evaluation of AI systems. AI-assisted tools were used exclusively for language refinement and presentation clarity, they did not contribute to the design of the proposed methods.

Acknowledgments

We thank the anonymous reviewers for their constructive feedback. This work was supported in part by the Zhejiang Provincial Natural Science Foundation of China (Grant No. LD24F020010), Ant Group, the ZJU-Ant Group Joint Lab of Data Security and Privacy Protection, the "Pionner" and "Leading Goose" R&D Program of Zhejiang Province (No.2026LDC01026(XC)) and the Key Research and Development Program of Hangzhou City (Grant No. 2024SZD1A27).

References

- Orlando Ayala and Patrice B  chard. 2024. [Reducing hallucination in structured outputs via retrieval-augmented generation](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Track, NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 228–238. Association for Computational Linguistics.
- Federico Barbero, Alvaro Arroyo, Xiangming Gu, Christos Perivolaropoulos, Michael M. Bronstein, Petar Velickovic, and Razvan Pascanu. 2025. [Why do llms attend to the first token?](#) *CoRR*, abs/2504.02732.
- Jasmijn Bastings and Katja Filippova. 2020. [The elephant in the interpretability room: Why use attention as explanation when we have saliency methods?](#) In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP, BlackboxNLP@EMNLP 2020, Online, November 2020*, pages 149–155. Association for Computational Linguistics.
- Sky CH-Wang, Benjamin Van Durme, Jason Eisner, and Chris Kedzie. 2024. [Do androids know they’re only dreaming of electric sheep?](#) In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 4401–4420. Association for Computational Linguistics.
- 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](#). In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.
- Ziling Cheng, Meng Cao, Marc-Antoine Rondeau, and Jackie CK Cheung. 2025. [Stochastic chameleons: Irrelevant context hallucinations reveal class-based \(mis\)generalization in llms](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 30187–30214. Association for Computational Linguistics.
- Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Qianyu Guo, Meng Wang, and Haofen Wang. 2023. [Retrieval-augmented generation for large language models: A survey](#). *CoRR*, abs/2312.10997.
- 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, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 5484–5495. Association for Computational Linguistics.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. [Measuring massive multitask language understanding](#). In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.
- Xiangkun Hu, Dongyu Ru, Lin Qiu, Qipeng Guo, Tianhang Zhang, Yang Xu, Yun Luo, Pengfei Liu, Yue Zhang, and Zheng Zhang. 2024. [Refchecker: Reference-based fine-grained hallucination checker and benchmark for large language models](#). *CoRR*, abs/2405.14486.
- Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2025. [A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions](#). *ACM Trans. Inf. Syst.*, 43(2):42:1–42:55.
- Yichen Jiang, Shikha Bordia, Zheng Zhong, Charles Dognin, Maneesh Kumar Singh, and Mohit Bansal. 2020. [Hover: A dataset for many-hop fact extraction and claim verification](#). In *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020*, Findings of ACL, pages 3441–3460. Association for Computational Linguistics.
- Zhuoran Jin, Pengfei Cao, Hongbang Yuan, Yubo Chen, Jixin Xu, Huaijun Li, Xiaojian Jiang, Kang Liu, and Jun Zhao. 2024. [Cutting off the head ends the conflict: A mechanism for interpreting and mitigating knowledge conflicts in language models](#). In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 1193–1215. Association for Computational Linguistics.
- Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, and 17 others. 2022. [Language models \(mostly\) know what they know](#). *CoRR*, abs/2207.05221.
- Kenneth Li, Oam Patel, Fernanda B. Vi  gas, Hanspeter Pfister, and Martin Wattenberg. 2023. [Inference-time intervention: Eliciting truthful answers from a language model](#). In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.
- Jianhua Lin. 1991. [Divergence measures based on the shannon entropy](#). *IEEE Trans. Inf. Theory*, 37(1):145–151.
- Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paran-jape, Michele Bevilacqua, Fabio Petroni, and Percy

- Liang. 2024. [Lost in the middle: How language models use long contexts](#). *Trans. Assoc. Comput. Linguistics*, 12:157–173.
- Llama Team. 2024. [The llama 3 herd of models](#). *CoRR*, abs/2407.21783.
- Xueqi Ma, Jun Wang, Yanbei Jiang, Sarah Monazam Erfani, Tongliang Liu, and James Bailey. 2025. Cognitive mirrors: Exploring the diverse functional roles of attention heads in llm reasoning. *arXiv preprint arXiv:2512.10978*.
- Andrey Malinin and Mark J. F. Gales. 2021. [Uncertainty estimation in autoregressive structured prediction](#). In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.
- Potsawee Manakul, Adian Liusie, and Mark J. F. 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, EMNLP 2023, Singapore, December 6-10, 2023*, pages 9004–9017. Association for Computational Linguistics.
- Cheng Niu, Yuanhao Wu, Juno Zhu, Siliang Xu, Kashun Shum, Randy Zhong, Juntong Song, and Tong Zhang. 2024. [Ragtruth: A hallucination corpus for developing trustworthy retrieval-augmented language models](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 10862–10878. Association for Computational Linguistics.
- nostalgebraist. 2020. [Interpreting GPT: the logit lens](#). AI Alignment Forum.
- Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, and 7 others. 2022. In-context learning and induction heads. *Transformer Circuits Thread*. <https://transformer-circuits.pub/2022/in-context-learning-and-induction-heads/index.html>.
- OpenAI. 2023. [GPT-4 technical report](#). *CoRR*, abs/2303.08774.
- Hadas Orgad, Michael Toker, Zorik Gekhman, Roi Reichart, Idan Szepktor, Hadas Kotek, and Yonatan Belinkov. 2025. [LLMs know more than they show: On the intrinsic representation of LLM hallucinations](#). In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.
- Jie Ren, Jiaming Luo, Yao Zhao, Kundan Krishna, Mohammad Saleh, Balaji Lakshminarayanan, and Peter J. Liu. 2023. [Out-of-distribution detection and selective generation for conditional language models](#). In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.
- 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](#). In *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*. OpenReview.net.
- Zhongxiang Sun, Xiaoxue Zang, Kai Zheng, Jun Xu, Xiao Zhang, Weijie Yu, Yang Song, and Han Li. 2025. [Redeep: Detecting hallucination in retrieval-augmented generation via mechanistic interpretability](#). In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and 49 others. 2023. [Llama 2: Open foundation and fine-tuned chat models](#). *CoRR*, abs/2307.09288.
- Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. 2024. [Efficient streaming language models with attention sinks](#). In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.
- Rongwu Xu, Zehan Qi, Zhijiang Guo, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. 2024. [Knowledge conflicts for llms: A survey](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 8541–8565. Association for Computational Linguistics.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 40 others. 2025. [Qwen3 technical report](#). *CoRR*, abs/2505.09388.
- Min-Hsuan Yeh, Yixuan Li, and Tanwi Mallick. 2025. [LUMINA: detecting hallucinations in RAG system with context-knowledge signals](#). *CoRR*, abs/2509.21875.
- Tianhang Zhang, Lin Qiu, Qipeng Guo, Cheng Deng, Yue Zhang, Zheng Zhang, Chenghu Zhou, Xinbing Wang, and Luoyi Fu. 2023. [Enhancing uncertainty-based hallucination detection with stronger focus](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 915–932. Association for Computational Linguistics.

Zhenliang Zhang, Xinyu Hu, Huixuan Zhang, Junzhe Zhang, and Xiaojun Wan. 2025. [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)*, ACL 2025, Vienna, Austria, July 27 - August 1, 2025, pages 17986–18002. Association for Computational Linguistics.

Zifan Zheng, Yezhaohui Wang, Yuxin Huang, Shichao Song, Mingchuan Yang, Bo Tang, Feiyu Xiong, and Zhiyu Li. 2024. [Attention heads of large language models: A survey](#). *CoRR*, abs/2409.03752.

A Details of Datasets

We conduct experiments on two widely used hallucination benchmarks covering both retrieval-augmented and general generation scenarios.

RAGTruth. RAGTruth is a benchmark specifically designed for hallucination detection in retrieval-augmented generation (RAG) settings, containing 15,090 training data and 2,700 testing data. Each instance consists of a user query, a set of retrieved documents, and an LLM-generated response annotated with span-level hallucination labels. The dataset covers multiple generation tasks, including question answering, summarization and data-to-text generation, enabling comprehensive evaluation across diverse reasoning scenarios. RAGTruth provides responses generated by six different LLMs: Llama2-7B, Llama2-13B, Llama2-70B, Mistral-7B, GPT-3.5 and GPT-4. In our experiments, we adopt the extended test set created by [Sun et al. \(2025\)](#), which includes annotations for Llama3-8b.

Dolly (AC). The Dolly (AC) benchmark is derived from the *databricks-dolly-15k*², an instruction-following dataset covering a wide range of tasks such as summarization, closed-domain question answering and information extraction. Following [Hu et al. \(2024\)](#), we adopt the *Accurate Context (AC)* setting, where the provided reference context is assumed to be largely noise-free. Since hallucination annotations are not natively available for generated responses in Dolly, we adopt the manually curated hallucination labels provided by [Sun et al. \(2025\)](#), which contains 300 data. These annotations are available for responses generated by Llama2-7B, Llama2-13B and Llama3-8B, which are the models evaluated on the Dolly (AC) benchmark in our experiments.

B Details of Baselines

We compare our method against representative hallucination detection baselines spanning four paradigms.

Uncertainty-based Methods. These approaches estimate hallucination likelihood from output uncertainty. **Perplexity** measures sequence-level generation confidence based on token probabilities.

²<https://huggingface.co/datasets/databricks/databricks-dolly-15k>

LN-Entropy computes length-normalized entropy across generated tokens to reduce length bias.

Cross-sample Consistency Methods. These methods detect hallucinations by measuring inconsistency across multiple stochastic generations. **SelfCheckGPT** evaluates logical consistency among sampled responses. **EigenScore** quantifies semantic divergence in the embedding space, where higher dispersion indicates unstable or hallucinated reasoning.

Self-evaluation Methods These methods rely on an external LLM to verify factual correctness. **P(True)** adopts a self-verification strategy, querying the language model itself to assess the truthfulness of its generated content. The hallucination score is derived from the model’s confidence in labeling its own output as true.

Internal-signal Methods. These approaches leverage internal activations or utilization signals to identify hallucinations. **ICR Probe** tracks hidden-state trajectory shifts using a supervised probe to identify unreliable generation patterns. **ReDeEP** detects conflicts between parametric knowledge and contextual reasoning by analyzing FFN-layer dynamics. **LUMINA** estimates hallucination risk by measuring the relative utilization of retrieved context versus parametric knowledge, using internal representation similarity signals as a proxy for context reliance.

Unless otherwise specified, all baselines are evaluated using either official implementations or faithfully re-implemented versions following the original papers.

C Cross-Model Analysis of FFN-Induced Updates

FFN Dynamics on Larger and Alternative Models. To examine whether the FFN behaviors observed in the main analysis are specific to a single model, we replicate the same layer-wise FFN contribution analysis on two additional models: Llama2-13B-Chat and Mistral-7B-Instruct-v0.1. Figure 6 report the corresponding residual contribution magnitude η_l and directional hallucination contribution γ_l across layers.

Across both models, we observe patterns highly consistent with those reported for Llama2-7B-Chat. In particular, the magnitude of FFN updates exhibits a smooth, monotonic decay with depth, while

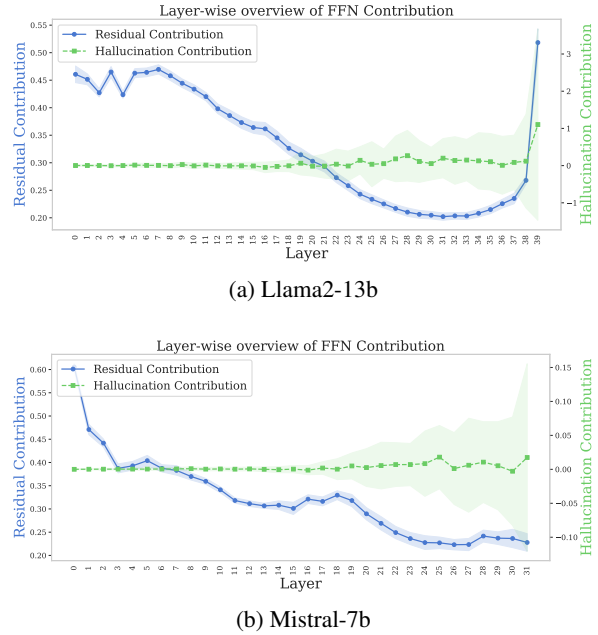


Figure 6: **Layer-wise FFN Contribution and Hallucination Alignment across models.** Plots showing the mean and standard deviation for (a) Llama2-13b and (b) Mistral-7b.

the directional contribution to hallucination remains non-monotonic and exhibits substantial variance across layers. As in the main model, layers with stronger FFN updates do not systematically align with stronger hallucination bias, and directional effects tend to appear in deeper layers where update magnitudes are relatively smaller.

These results suggest that the decoupling between FFN update magnitude and hallucination alignment is not an artifact of a particular model size or architecture. Rather, similar FFN dynamics emerge across models with different parameter scales and training regimes, supporting the robustness of our observations.

D Implementation Details

All experiments are conducted using open-weight Transformer-based LLMs. Inference is performed on NVIDIA H200 GPUs. Attention statistics are extracted at the final decoding step using a streaming implementation to avoid recomputation overhead.

For the AES in Eq. 11, we fix the coefficients as $\lambda_1 = 0.25$, $\lambda_2 = 0.35$ and $\lambda_3 = 0.4$ across all experiments. We select these values on the **training** split (via validation on held-out training data) and report all results on the **test** split, keeping the coefficients unchanged across models and datasets. More parameter analysis can be found in Appendix I.

For the threshold τ in Eq. 6, we set $\tau = 0.65$ to filter out weak or semantically unrelated span pairs. This value is chosen empirically to suppress incidental semantic overlap while preserving meaningful prompt–response alignment. We observe that AES performance is not highly sensitive to moderate variations of τ , and additional analysis is provided in Appendix I.

We define $\mathcal{L}_{\text{mid}} = \{\lfloor \rho_s L \rfloor, \dots, \lfloor \rho_e L \rfloor\}$ in Section 4.3.1 with $\rho_s = 0.25$ and $\rho_e = 0.875$ for all models (e.g., for 32 layers, including layer 8-28). We set $\epsilon = 1$ in Section 4.4 across all experiments.

Span segmentation. In our implementation, spans are constructed using sentence-level segmentation based on regular expressions. We also experiment with alternative segmentation strategies, including syntactic parsing and finer-grained chunking. While finer-grained spans theoretically increase the number of alignment computations, we find that the overall detection overhead remains low in practice, even with substantially increased span counts. Detailed comparisons of different span granularities are reported in Appendix G.

Attention head activation and role thresholds. Following Section 4.2, we identify functionally active attention heads based on their response-to-prompt attention strength. Specifically, we set the activation threshold to 0.025, and use a top- k selection strategy with $\text{top_k_ratio} = 0.05$ to account for the sparsity of attention distributions. A head is classified as *useful* if its Semantic Union Coverage (SUC) exceeds 0.2, and as *interfering* if its SUC is below 0.05. These thresholds are fixed across all experiments and models. The gap between the useful and interfering thresholds is intentionally introduced to improve robustness to attention variability and to avoid forcing ambiguous heads into hard categories. Heads with intermediate SUC values are treated as neutral and excluded from intervention, as they often serve auxiliary roles (e.g., syntactic organization) and are neither consistently beneficial nor harmful to semantic grounding.

Intervention details. We instantiate the proposed *escape-route pruning* with a fixed intervention set and a constant attenuation factor. To construct the intervention set $\mathcal{H}_{\text{int}}$, we run our head-role classifier on the **hallucination split** and aggregate head labels across response spans. For each head (l, h) , we compute its *interfering frequency*, i.e., the fraction of spans for which the head is clas-

sified as *interfering*. We then select the **top- k** heads with the highest interfering frequency and fix them as $\mathcal{H}_{\text{int}}$; unless otherwise stated, we use $k=20$.

At inference time, we apply a static attenuation to each intervened head:

$$\tilde{o}_t^{(l,h)} = \beta \cdot o_t^{(l,h)}, \quad (l, h) \in \mathcal{H}_{\text{int}}, \quad (12)$$

We set the attenuation factor to $\beta=0.2$ for all experiments, which provides a strong yet non-destructive suppression of dominant escape routes while incurring negligible runtime overhead.

E Prompt for Intervention Evaluation

In our hallucination mitigation experiments, we evaluate the truthfulness of model outputs *before* and *after* intervention using a fixed LLM-judge prompt template. The exact prompt template used for this pairwise evaluation is shown below.

You are a professional text hallucination detection expert. Please compare and evaluate the responses generated by two models to determine which response is more truthful and has fewer hallucinations.

Evaluation Criteria: 1. Consistency with the given context (whether it can be inferred from the context) 2. Factual Accuracy (whether it contains erroneous information) 3. Logical Reasonableness (whether it is self-contradictory)

Original Prompt: {prompt}

Context: {context_section}

Response A (Baseline): {baseline_response}

Response B (Intervened): {intervened_response}

Respond with ONLY a valid JSON object in the following format:

```
{
  "winner": "baseline" | "intervened" | "tie",
  "baseline_hallucination_score": Float (0.0-1.0),
  "intervened_hallucination_score": Float (0.0-1.0),
  "improvement_magnitude": "significant" | "moderate" | "slight" | "none" | "worse",
  "comparison_analysis": "Detailed analysis..."
}
```

F Mitigation Results on DOLLY (AC)

To assess cross-dataset transferability, we apply PER on **DOLLY (AC)** using the *same fixed intervention set* $\mathcal{H}_{\text{int}}$ (i.e., the top- k interfering heads identified from the hallucination split in our main setup) and the same attenuation factor β . We keep all intervention and evaluation settings unchanged, and perform pairwise judging with **DeepSeek-V3.2**. Table 7 reports **Win/Tie/Lose** rates and the **Net Improvement** (win–loss margin) on both the hallucination-only subset and the full test set.

Setting	Win	Tie	Lose	Net Improv.
PER vs original (Hallucination Set)	27.6%	62.1%	10.3%	+17.3%
PER vs original (Full Set)	17.2%	75.8%	7.0%	+10.2%
PER vs AARF (Hallucination Set)	50.0%	28.1%	21.9%	+28.1%
PER vs AARF (Full Set)	39.1%	47.9%	13.0%	+26.1%

Table 7: **Hallucination mitigation on DOLLY (AC).** **Win/Tie/Lose** denote the percentage of cases where the *first* method is judged better/equal/worse than the *second*. Net Improvement is the win–loss margin.

G Robustness to Semantic Alignment Design Choices

Our span-level semantic alignment relies on design choices that define a response-conditioned semantic reference frame.

In this section, we evaluate the robustness of AES with respect to two key factors: (i) the choice of sentence encoder and (ii) the granularity of span segmentation. Unless otherwise specified, all robustness results in this section are reported on the **Llama2-7B-chat** model evaluated on the **RAGTruth** benchmark, with all other components kept fixed.

Robustness to sentence encoder choice. Our span-level semantic alignment uses a sentence encoder $E(\cdot)$ to define a semantic reference frame for measuring attention allocation (Section 4.1). Importantly, the encoder is not used to assess factual correctness or provide supervision; all detection signals (SAS/UAL/RAR) are computed exclusively from internal attention statistics.

We evaluate robustness to the encoder choice by repeating the full pipeline with three encoders: all-MiniLM-L6-v2, embeddinggemma-300m, and qwen3-embedding-0.6b. For all settings, we keep the same span segmentation, cosine similarity and denoising threshold τ , so the only change is the embedding model used to form $\mathcal{P}_{\text{rel}}$.

Table 8 reports the corresponding detection results. Across datasets and models, we observe consistent performance with similar relative trends, indicating that AES is not sensitive to the specific choice of sentence encoder.

Robustness to span granularity. Beyond the encoder choice, the granularity of span segmentation may also influence the construction of the semantic reference frame. To assess this factor, we compare AES performance under different span construction strategies, while keeping the encoder, denoising threshold τ and attention statistics fixed.

Dataset	Model	Encoder	AUROC	PCC
RAGTruth	Llama2-7B-chat	all-MiniLM-L6-v2	0.8717	0.5800
		embeddinggemma-300m	0.8845	0.5944
		qwen3-embedding-0.6b	0.8538	0.5491
Dolly	Llama2-7B-chat	all-MiniLM-L6-v2	0.7821	0.4855
		embeddinggemma-300m	0.7589	0.4425
		qwen3-embedding-0.6b	0.7638	0.4429

Table 8: **Robustness to sentence encoder choice.** AES performance computed using three different encoders for span-level alignment.

Specifically, we consider three representative span segmentation methods: (i) syntactic segmentation using spaCy with the en_core_web_sm pipeline, (ii) coarse-grained chunking via RecursiveCharacterTextSplitter in LangChain³, and (iii) fine-grained sentence-level segmentation based on regular expressions. Table 9 reports the corresponding results.

Overall, we observe that finer-grained spans yield more stable and discriminative detection signals, suggesting that reduced semantic mixing within spans leads to a sharper reference frame for attention interpretation.

Span Segmentation Method	Avg. Span Length	AUROC	PCC
RecursiveCharacterTextSplitter	364.77	0.6709	0.2953
spaCy	113.74	0.8407	0.5416
Regex-based fine-grained	90.06	0.8717	0.5800

Table 9: **Effect of span granularity on hallucination detection.** AES performance under different span construction methods.

H Additional Head Distribution Results

Cross-model distributions of useful and interfering heads. Figures 7 report the layer-wise distributions of *useful* and *interfering* heads under hallucinated versus non-hallucinated samples for each model. All results are obtained under the same experimental setting. Non-hallucinated responses exhibit higher activity in both head types, indicating robust anchoring alongside active context scanning. In contrast, hallucinated responses show a systemic degradation of this evidence-seeking behavior, shifting toward self-reinforcement. This suggests that hallucination is better characterized as a **breakdown in prompt-aligned routing dynamics** rather than the emergence of a single pathological head type. Overall, the qualitative trends are consistent across model families: non-hallucinated generations allocate more attention mass to *useful*

³<https://www.langchain.com/>

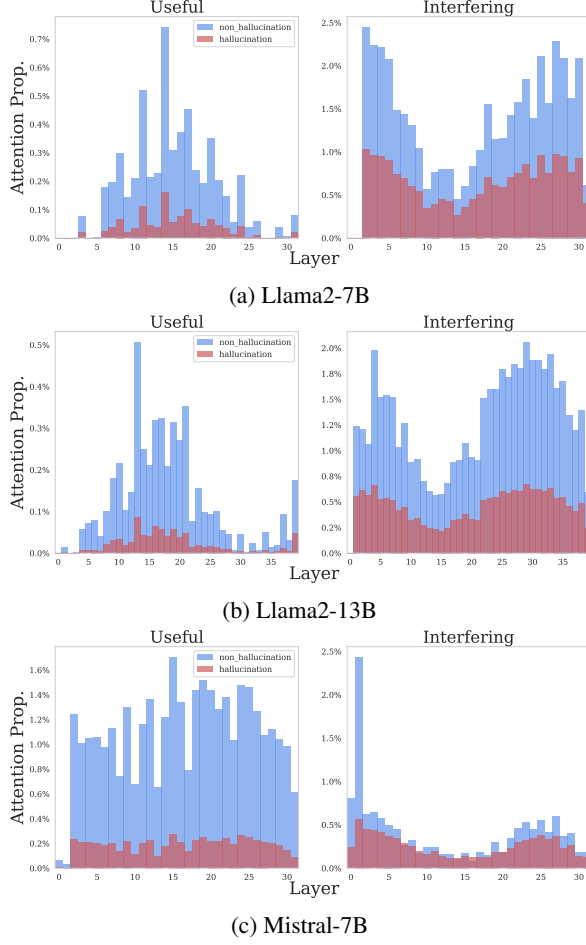


Figure 7: **Layer-wise distribution of head roles across models.** For each model, we plot histograms of *useful* and *interfering* heads over layers for hallucinated vs. non-hallucinated samples. The y-axis shows the *normalized proportion of heads* in each layer, where normalization is performed within each condition (hallucination / non-hallucination).

heads in middle layers, while hallucinated generations exhibit weakened prompt-aligned routing and a greater tendency toward non-evidence-seeking attention allocation.

I Parameter Sensitivity Analysis

We analyze the robustness of AES with respect to two types of parameters: (i) aggregation coefficients that balance different attention signals and (ii) the semantic alignment threshold that defines span-level relevance. Unless otherwise stated, all experiments in this section are conducted using the **Llama2-7B-chat** model on the **RAGTruth** dataset, with all other components kept fixed.

Sensitivity to aggregation coefficients. The **Attention Escape Score (AES)** aggregates three com-

plementary attention-based signals via a linear combination:

$$\text{AES} = \lambda_1 \cdot \frac{1}{\text{SAS} + 1} + \lambda_2 \cdot \text{UAL} + \lambda_3 \cdot \text{RAR}. \quad (13)$$

To assess whether the effectiveness of AES depends on precise coefficient values or reflects intrinsic properties of the proposed signals, we conduct a parameter sensitivity analysis. Rather than tuning for optimal performance, this analysis aims to evaluate the *robustness* of AES under moderate coefficient perturbations.

Specifically, we vary each coefficient independently over a deliberately wider range than the values observed in our tuned settings, while keeping the other coefficients fixed. This serves as a stress test, verifying that AES remains stable even under substantial weight perturbations.

- Semantic Anchoring Weight: $\lambda_1 \in [0.5, 0.8]$
- Contextual Distraction Weight: $\lambda_2 \in [0.05, 0.2]$
- Generation Inertia Weight: $\lambda_3 \in [0.15, 0.4]$

This results in a total of 196 coefficient combinations. For each configuration, we evaluate hallucination detection performance on the RAGTruth dataset using the same experimental protocol as in the main results.

Robustness Statistics. As summarized in Table 10, AES exhibits strong robustness to coefficient variation. Across all 196 configurations, the standard deviation of AUROC is only **0.0028**, and the total performance range (maximum minus minimum) is approximately **1.1%**. A similarly low variance is observed for Pearson correlation.

Notably, even under the worst-performing coefficient setting within the explored range, AES achieves an AUROC of **0.8639**, which remains substantially higher than the strongest baseline reported in our experiments. These results indicate that AES performance is not driven by fine-grained coefficient tuning, but instead arises from the stable and complementary nature of the underlying attention routing signals.

Sensitivity to Semantic Alignment Threshold.

We further analyze the sensitivity of AES to the semantic alignment threshold τ in Eq. 6. As shown in Table 11, AES exhibits strong robustness to moderate variations in the similarity threshold τ , with performance metrics remaining stable across the tested

Metric	Mean	Std. Dev (σ)	Min	Max	Range (Δ)
AUROC	0.8700	0.0028	0.8639	0.8748	0.0109
PCC	0.5784	0.0060	0.5661	0.5906	0.0245

Table 10: **Sensitivity Analysis Statistics.** Aggregate performance statistics across 196 coefficient configurations. Low variance in both AUROC and PCC indicates that AES is insensitive to moderate coefficient perturbations.

range. Notably, we observe that higher thresholds can yield a more selective semantic reference frame, further enhancing separability. This trend is mechanistically supported by two key observations. As illustrated in Figure 8, the probability of finding no aligned prompt spans ($\Pr(\mathcal{P}_{\text{rel}} = \emptyset)$) remains consistently higher for hallucinated spans as τ increases, indicating a fundamental lack of highly matched evidence. Complementing this, Figure 9 reveals a distinct distributional shift in maximum cosine similarity, where non-hallucinated spans cluster at significantly higher similarity values. Collectively, these results suggest that increasing τ effectively filters out weak, incidental associations, thereby sharpening the contrast between anchored and unanchored attention routing captured by the AES framework.

Default Choice of τ . Despite the potential gains at very strict thresholds (e.g., $\tau \geq 0.9$ on our benchmarks), overly large τ may under-cover semantically grounded but non-literal evidence (e.g., paraphrase/implicit grounding), distributed evidence in multi-hop settings, or domain shifts that change the encoder similarity scale. We therefore use $\tau = 0.65$ as a conservative default that balances selectivity and coverage, and keep it fixed across experiments for consistent evaluation.

τ	AUROC	PCC
0.5	0.8250	0.5429
0.55	0.8424	0.5608
0.6	0.8514	0.5585
0.65	0.8717	0.5800
0.7	0.8901	0.5916
0.75	0.9027	0.5993

Table 11: **Sensitivity to semantic alignment threshold τ .** AES performance under different τ values, with all other components fixed.

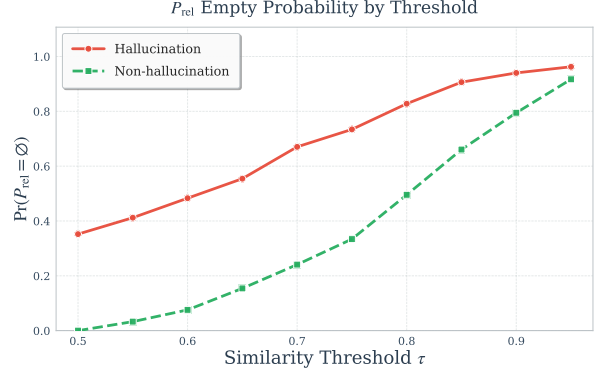


Figure 8: **Empty-alignment probability under varying τ .** Class-conditional probability that the aligned prompt set is empty, $\Pr(\mathcal{P}_{\text{rel}} = \emptyset)$, for hallucinated vs. non-hallucinated spans.

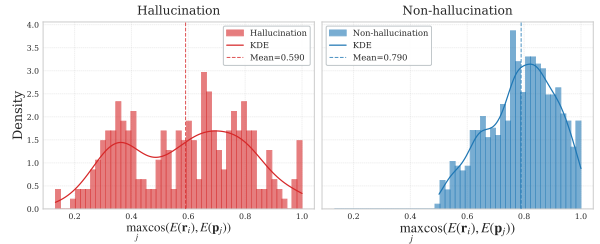


Figure 9: **Distribution of max semantic similarity.** Histogram/KDE of $s_i = \max_j \cos(E(r_i), E(p_j))$ for hallucinated vs. non-hallucinated spans.

J Scalability and Cross-Architecture Generalization

A natural concern is whether the ‘‘Attention Escape’’ phenomenon is merely a transient artifact of smaller models that diminishes with scaling, or whether our framework remains effective on models with the latest architectural features (e.g., Qwen-3). To address these questions, we conduct targeted supplementary experiments on massive-scale and state-of-the-art architectures.

J.1 Persistence of Attention Escape at Scale

To directly answer whether attention escape persists in large-scale parameter regimes, we benchmark our AES approach against standard baselines on the massive **Llama-2-70B** model using the RAGTruth dataset.

As detailed in Table 12, our method significantly outperforms all baseline detection strategies. The fact that AES achieves an exceptional AUROC of **0.9332** on Llama-2-70B proves that the conflict between retrieved context and internal parametric knowledge remains a fundamental architectural challenge even at the 70B scale. Attention

escape does not spontaneously vanish as parameter counts increase; rather, capturing the specific routing mechanisms of attention remains the most precise way to detect hallucinations in large-scale models.

Approach	AUROC $\uparrow$	ACC $\uparrow$	F1 $\uparrow$	Precision $\uparrow$	PCC $\uparrow$
Perplexity	0.5669	0.5222	0.5873	0.4371	0.0412
LN-Entropy	0.6801	0.6311	0.6210	0.5094	0.2531
SelfCheckGPT	0.5084	0.4089	0.5596	0.3903	0.0252
EigenScore	0.6673	0.6156	0.6263	0.4966	0.2566
P(True)	0.6394	0.6533	0.5517	0.5424	0.1795
AES	0.9332	0.8838	0.8114	0.7717	0.5942

Table 12: **Hallucination detection results on Llama-2-70B.** AES significantly outperforms baselines, confirming that attention escape persists as a primary driver of hallucination even in large-scale parameter regimes.

J.2 Generalization to Latest Architectures

Our initial focus on Llama and Mistral was strictly due to the lack of annotated data for the Qwen series in standard benchmarks (RAGTruth and Dolly-AC). However, mechanistically, newer models like Qwen-3 still rely on standard attention mechanisms to process context, meaning our principles of capturing and intervening in attention routing remain fundamentally applicable.

To empirically validate this, we generated responses using RAGTruth contexts and prompts from **Qwen3-8B** and **Qwen3-32B**, both with (+PER) and without (Baseline) our intervention. We employed **DeepSeek-V3.2** as an impartial judge to conduct a pairwise comparison on context faithfulness.

As shown in Table 13, the intervened models demonstrate a massive advantage. The intervened Qwen3-8B outperforms the baseline by a factor of four, while Qwen3-32B achieves a dominant 61.6% win rate. These compelling results strongly confirm the universal applicability of our framework: the attention escape phenomenon persists in the latest architectures, and PER effectively blocks these routes to enhance context utilization regardless of the specific model family.

Evaluation Model	Win (+PER)	Tie	Lose (Baseline)
Qwen3-8B	40.7%	49.3%	10.0%
Qwen3-32B	61.6%	7.3%	31.1%

Table 13: **Hallucination mitigation performance on Qwen-3 family.** Win/Tie/Lose denote the percentage of cases where PER is judged better/equal/worse than the baseline. PER shows overwhelming advantages on the latest architectures.

Task Type	ACC $\uparrow$	Precision $\uparrow$	Recall $\uparrow$	F1 $\uparrow$
Data-to-text	90.32%	93.10%	96.43%	0.9474
QA	80.67%	74.47%	67.31%	0.7071
Summary	83.87%	71.43%	74.07%	0.7273

Table 14: **Task-wise response-level performance.** Metrics are computed under the same evaluation protocol; we omit sample counts to emphasize relative difficulty across task types.

K Error Analysis

We analyze representative failure cases of AES at the response level. Our analysis proceeds in two steps: (i) stratifying performance by task type to identify systematic weaknesses and (ii) examining false positives/negatives to summarize recurring patterns.

Task-type sensitivity. Performance varies across task types, with Data2txt being the most stable and QA being the most challenging. Table 14 summarizes this trend. A plausible explanation is that QA often involves implicit reasoning or sparse evidence usage, which can dilute span-level anchoring signals at the response level, while Data2txt tends to align more directly with localized evidence.

Typical errors. False positives often arise when grounded responses rely on *implicit* or *diffuse* evidence, leading to weaker concentration on explicitly aligned spans and mildly elevated HRS. False negatives typically occur when unsupported content is brief, weakly expressed, or tightly interwoven with factual statements, making attention escape less pronounced.

Overall, these results suggest that AES is broadly applicable across task types: it remains effective beyond Data2txt. For the more challenging QA task, it still provides a strong and useful signal. Importantly, the remaining errors are dominated by borderline cases near the decision threshold rather than task-specific failure modes, indicating that the same attention-routing diagnostic transfers across tasks without requiring task-specific redesign.

L Compute Resources

Computational Cost of Detection. AES is training-free and derives diagnostics from a single inference pass. Evaluated on **Llama-2-7B-Chat** using the RAGTruth benchmark under the same setting with the main experiments, the detection pipeline achieves an average latency of **1.25s** per

Setting	Latency (s)	Throughput (tok/s)	Avg. Tokens	Overhead
Baseline	4.806	39.02	187.3	–
Intervention	4.867	37.91	184.3	+1.27%

Table 15: **Inference overhead of PER.** Comparison of average latency per response and generation throughput on Llama-2-7B-Chat ($N = 450$). The intervention incurs negligible computational cost.

sample with an average GPU memory footprint of **16.9 GB**. The runtime scales linearly with the length of the sequence, ranging from ~ 0.9 s for short responses (200–250 tokens) to ~ 3.8 s for long sequences (500–1000 tokens).

Inference Latency of Mitigation. PER applies a static, element-wise scaling to identified interfering heads, avoiding iterative optimization. As shown in Table 15, we quantify the overhead by comparing generation statistics with and without intervention. The intervention maintains a robust throughput of **37.91** tokens/s compared to the baseline’s **39.02** tokens/s. This corresponds to a marginal latency increase of approximately **1.27%**, confirming that PER-based suppression is highly efficient and viable for real-time deployment.