TriLens:
Per-Layer Logit-Lens Entropy for White-Box Hallucination Detection

Anonymous ACL submission

Abstract

When a language model hallucinates, the final answer is wrong, but the mistake is not necessarily invisible inside the model. Different internal pathways may remain uncertain, disagree in how quickly they sharpen, or commit to competing continuations before the output is produced. We introduce *TriLens*, a white-box detector that turns this intuition into a compact representation: at every layer, it reads the multi-head self-attention output, the feed-forward output, and the residual stream through the model's own logit lens, then records only the entropy of each readout. The resulting $3L$-dimensional trajectory describes how certainty forms across depth and across modules, without storing high-dimensional hidden states or sampling multiple generations. This simple signal yields a strong detector across instruction-tuned LLMs and QA benchmarks, and our analyses show that the three module-wise entropy trajectories provide complementary evidence. TriLens suggests that hallucination detection can benefit from tracking how internal computation settles, not only what the final layer predicts. The code is available at https://anonymous.4open.science/w/TriLens/.

1 Introduction

Large language models (LLMs) have reshaped a wide range of NLP tasks, but their tendency to generate hallucinations—outputs that are fluent yet factually wrong or inconsistent with provided context—remains a fundamental obstacle to deployment in knowledge-sensitive settings (Ji et al., 2023; Li et al., 2023). Reliable *detection* enables selective abstention, confidence calibration, and targeted retrieval.

Among existing detection approaches, *white-box* methods that analyze an LLM's internal state during a single forward pass are particularly attractive: they avoid the need for multiple samples or external references while enabling efficient inference-time scoring. Recent work has shown that hallucination-relevant information can be extracted from hidden states, gradients, semantic-entropy probes, spectral summaries, attention kernels, and update-consistency features (Azaria and Mitchell, 2023; Kossen et al., 2024; Hu et al., 2024; Chen et al., 2024; Sriramanan et al., 2024; Zhang et al., 2025). However, these approaches still largely treat internal computation as a high-dimensional representation to be classified, rather than asking which low-dimensional aspect of the computation changes when a model moves toward an unsupported answer.

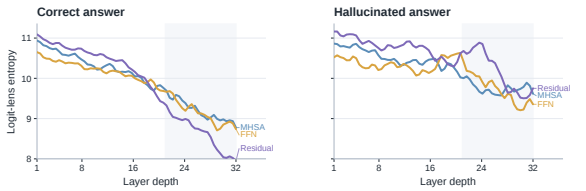

Figure 1: **TriLens: Per-layer logit-lens entropy tracks internal certainty during generation.** Supported answers show coordinated entropy sharpening across MHSA, FFN, and residual-stream readouts, whereas hallucinated answers tend to retain higher, less stable, and less synchronized entropy across depth.

This leads to two empirical questions. First, can hallucination be detected from how internal certainty forms across depth, rather than from final-layer confidence or high-dimensional hidden states? Second, do the main transformer pathways provide redundant views of that uncertainty, or do attention, feed-forward, and residual-stream readouts carry complementary evidence? A useful detector should answer these questions under a single forward pass, expose where uncertainty enters the computation, and avoid turning every hidden state into a large learned representation. We therefore test a simpler alternative: instead of storing activations themselves, we ask how uncertain different parts of the computation look when read through

the model's own vocabulary lens.

TriLens records one scalar from each of three internal locations at every layer: the multi-head self-attention output, the feed-forward output, and the residual stream. Each location is projected through the logit lens, converted to a Shannon entropy, and concatenated across depth. The result is a compact $3L$-dimensional trajectory that tracks how certainty forms across pathways. It is small enough for a simple probe, but structured enough to distinguish whether contextual routing, parametric recall, and the composed residual state sharpen together or remain diffuse.

Why should a first-order statistic suffice? Our working hypothesis is that hallucination-relevant errors are often accompanied by elevated uncertainty across internal pathways of an LLM, and that per-layer Shannon entropy of logit-lens distributions is a compact diagnostic correlate of this effect. In a decoder transformer, MHSA routes contextual information while the FFN pathway retrieves parametric content (Geva et al., 2021, 2023; Elhage et al., 2021). When these pathways align, their logit-lens distributions typically sharpen with depth; when they diverge, refinement can become slower and higher-entropy. TriLens tracks this behavior with three scalars per layer, yielding a compact alternative to high-dimensional hidden-state features.

Contributions. Our main contribution is to test these questions through a compact pathway-wise entropy feature, supported by ablations and diagnostic comparisons.

- We introduce **TriLens**, a per-layer logit-lens entropy feature over MHSA, FFN, and residual-stream states.
- Across three instruction-tuned LLMs and four benchmark settings, this $3L$-dimensional feature improves over the strong prior white-box baseline ICR Probe on *all 12* cells, averaging $+12.1$ AUROC, and attaining the highest AUROC in 11 of the 12 cells.
- Ablations show that MHSA, FFN, and residual-stream entropies contribute complementary signal, while an intra-layer module-disagreement term adds little.
- A head-to-head comparison with DoLa-style cross-layer contrast features shows that per-layer entropy captures much of the useful signal in that family.
- Layer-wise analysis shows that the peak-discriminative depth varies systematically across architectures and benchmarks, consistent with differences in how models integrate contextual and parametric information.

2 Related Work

White-box hallucination detection. The dominant white-box line trains detectors on internal activations from a single forward pass. SAPLMA (Azaria and Mitchell, 2023), layer-combination probes (CH-Wang et al., 2024), SEP (Kossen et al., 2024), gradient-aware detectors (Hu et al., 2024), INSIDE (Chen et al., 2024), and LLM-Check (Sriramanan et al., 2024) all fit this template, while ICR Probe (Zhang et al., 2025) reframes detection around residual-stream updates. TriLens differs by using a compact $3L$-scalar feature built from pathway-specific logit-lens entropy rather than high-dimensional hidden states or update-consistency scores.

Uncertainty-based detection. A parallel black-box line scores hallucination risk from a model's own uncertainty, including calibration-style signals (Kadavath et al., 2022; Malinin and Gales, 2020), semantic uncertainty and semantic entropy (Kuhn et al., 2023; Farquhar et al., 2024), and recent refinements based on Bayesian estimation, pairwise semantic similarity, fact-level self-consistency, or log-probability time series (Ciosek et al., 2025; Nguyen et al., 2025; Sawczyn et al., 2026; Shapiro et al., 2026). TriLens instead uses a single white-box pass over internal activations, without multiple sampled generations or output-layer uncertainty alone.

Logit-lens and mechanistic motivation. The logit lens and Tuned Lens project intermediate states into vocabulary space (nostalgebraist, 2020; Belrose et al., 2023); DoLa (Chuang et al., 2024) contrasts these distributions across layers. Our design is further motivated by mechanistic views of FFNs as key-value memory, MHSA as residual-stream routing, and depth-varying module influence (Geva et al., 2021, 2023; Elhage et al., 2021; Stolfo et al., 2023). TriLens combines these ingredients into a supervised detection feature based on pathway-specific entropy trajectories; §4.6 compares it directly with DoLa-style cross-layer contrast.

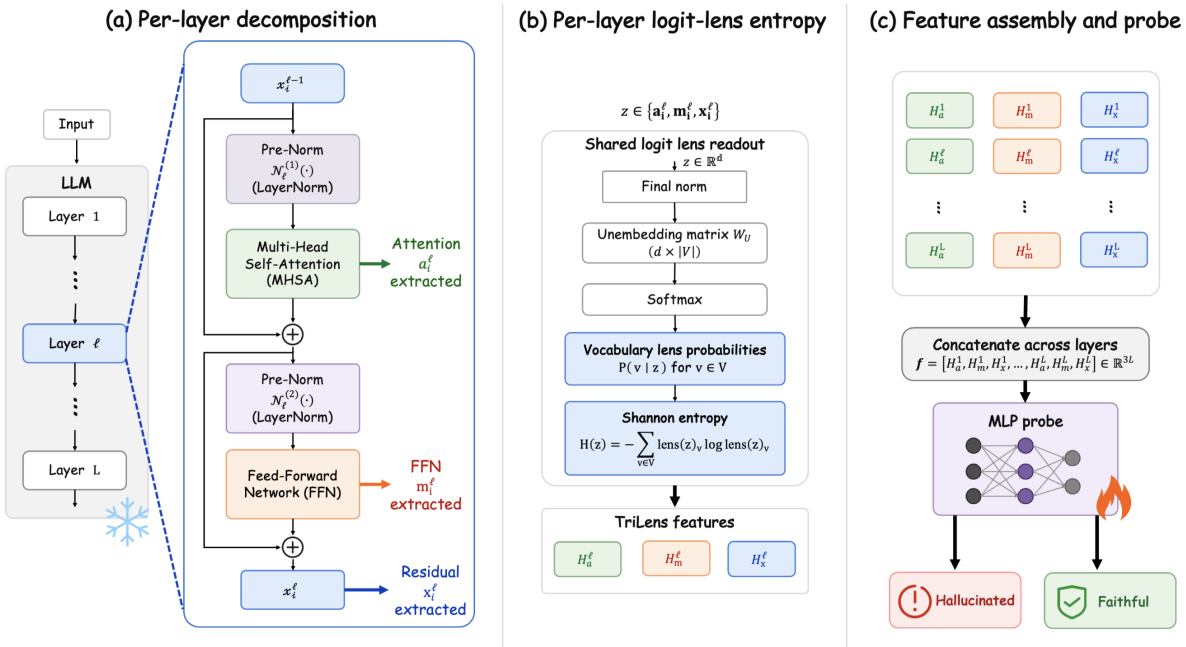

Figure 2: TriLens overview. A single forward pass over the evaluation sequence extracts the MHSA output, FFN output, and residual stream at every layer; each is projected through the model's logit lens, converted to a Shannon entropy, concatenated into a $3L$-dimensional feature, and fed to a probe.

3 Method

3.1 Per-Module Logit-Lens Entropy

We consider an LLM with L decoder layers and vocabulary V. At layer ℓ and token position i, the residual stream updates as

$$x_i^\ell = x_i^{\ell-1} + \underbrace{\mathrm{MHSA}^\ell(\mathcal{N}_\ell(x_i^{\ell-1}))}_{a_i^\ell} + \underbrace{\mathrm{FFN}^\ell(\mathcal{N}_\ell(x_i^{\ell-1} + a_i^\ell))}_{m_i^\ell}. \quad (1)$$

For a hidden activation $z \in \{x_i^\ell, a_i^\ell, m_i^\ell\}$, the *logit lens* (nostalgebraist, 2020) projects z into vocabulary-space probabilities

$$\mathrm{lens}(z) = \mathrm{softmax}\big(W_U\,\mathrm{Norm}_{\mathrm{final}}(z)\big), \qquad H(z) = -\sum_{v\in V} \mathrm{lens}(z)_v \log \mathrm{lens}(z)_v. \quad (2)$$

using the model's own unembedding matrix W_U and final normalization layer. Here $\mathcal{N}_\ell$ denotes the layer's pre-submodule normalization (e.g., RMSNorm in the model families studied here), while the lens itself always uses the model's *final* normalization layer before unembedding. This mirrors the standard logit-lens construction on the residual stream and keeps a^ℓ, m^ℓ, and x^ℓ in the same readout coordinate. In implementation we apply the final normalization directly to each of a^ℓ, m^ℓ, and x^ℓ, with no additional centering or rescaling. The *per-layer logit-lens entropy* is the Shannon entropy of this distribution; we compute it at all three positions, yielding $H_a^\ell, H_m^\ell, H_x^\ell$ for every layer (Eq. 2). TriLens applies the same readout not only to the residual stream but also to the isolated MHSA and FFN writes, using the resulting three-pathway trajectory as a detection feature.

3.2 Mechanistic Motivation

The choice to measure entropy *per module* rather than only on x^ℓ, and to use *Shannon entropy* rather than richer summary statistics, is not arbitrary. It follows from three established views of transformer dynamics.

Dual-pathway decomposition. Under the residual-stream framework of Elhage et al. (2021), a^ℓ and m^ℓ are functionally distinct pathways: the OV circuit of MHSA copies content from earlier positions, while FFN functions as a read-out from a parametric key-value memory (Geva et al., 2021, 2023). In factual completion, both pathways often converge on the same target token—MHSA because the context uniquely determines the answer, FFN because the parametric memory has stored the relevant key–value pair. In hallucinated completion under our benchmark setting, the pathways can

diverge: a^ℓ may still route contextually correct information, while m^ℓ can partially recall the same correct alternative even though the observed token is wrong. Projecting a^ℓ and m^ℓ independently through the logit lens exposes this divergence, which a single measurement on the composed x^ℓ can mask whenever the two pathways cancel in direction but not in magnitude. Although a^ℓ and m^ℓ are not themselves full residual states, they are additive writes in the same residual space and are therefore legitimate objects for the model's own final readout. The resulting distributions should be interpreted as *readout preferences of an isolated write* rather than as standalone next-token predictions; our claim is empirical and diagnostic, not that a^ℓ or m^ℓ alone constitute a full decoder state. Accordingly, the role of H_a^ℓ and H_m^ℓ in TriLens is not to replace H_x^ℓ but to supply complementary pathway-specific readouts under a shared lens.

Superposition and commitment. The residual stream at any layer is a linear superposition of all prior module writes. When the writes are coherent on a single vocabulary direction, the lens distribution is sharply peaked; when they are incoherent, the distribution spreads over multiple candidates. Shannon entropy is a compact statistic that tracks this spreading: it is invariant to which candidate tokens are involved, makes no assumption about whether the vocabulary is flat or structured, and is a calibrated function of the distribution's concentration. Richer alternatives such as KL divergence to the final layer (DoLa's signal) or a spectral summary lose information about within-layer commitment, as we show empirically in §4.6.

Iterative refinement trajectories. The logit-lens view treats each layer as producing a refinement of a running next-token prediction (nostalgebraist, 2020; Belrose et al., 2023). Under this view, $H(\mathrm{lens}(z^\ell))$ is the uncertainty of the model's belief at depth ℓ along that module's pathway. For a correct continuation, uncertainty often decreases with depth as module contributions reinforce each other. For a hallucinated continuation, the refinement trajectory is more often non-monotone: uncertainty can re-enter the distribution at the depth where the parametric pathway recalls the true token while the forced context fixes the output on the wrong one. The per-layer vector $(H_a^\ell, H_m^\ell, H_x^\ell)_\ell$ can thus be viewed as a trajectory of this three-pathway refinement process, and a linear classifier on this trajectory can separate the two regimes in our benchmark setting.

These three views explain why the feature is minimal (three location-specific scalars per layer), why it is non-redundant (each scalar measures a different pathway's commitment), and why it is architecturally natural (it uses only the model's own unembedding and pre-trained norm). Section 4.4 examines the resulting empirical layer-wise trajectories.

3.3 Feature Construction

Given an evaluation sequence, we run the LLM once and extract $(H_a^\ell, H_m^\ell, H_x^\ell)$ at the token positions used for scoring. Following the standard probe-input convention, we obtain a fixed-length vector per sample by applying a fixed readout rule consistently throughout the paper. The resulting feature is

$$\mathbf{f} = (H_a^1, H_m^1, H_x^1, \ldots, H_a^L, H_m^L, H_x^L) \in \mathbb{R}^{3L}. \tag{3}$$

For the models we consider, $3L$ ranges from 84 (Qwen2.5-7B, L=28) to 126 (Gemma-2-9B, L=42)—roughly 30× smaller than the raw hidden-state features used by SAPLMA and its descendants.

3.4 Probes

We evaluate two classifiers on $\mathbf{f}$: (i) a **linear probe** (L2-regularized logistic regression), which highlights that our feature is already separable without a learned transformation; and (ii) an **MLP probe** ($3L \to 128 \to 64 \to 32 \to 1$; LeakyReLU(0.01), BatchNorm, Dropout 0.3), using the same hidden-layer topology as the architecture used by the strong prior baseline (Zhang et al., 2025). Because TriLens uses a $3L$-dimensional input whereas ICR Probe uses an L-dimensional input, this comparison controls probe family and training recipe but does not fully equalize input dimensionality. We therefore treat the MLP comparison as a matched probe-family comparison rather than as a dimension-controlled capacity study, and we also report linear-probe results plus feature-subset ablations to separate feature quality from probe size. As an additional dimension-matched control, Appendix B compares TriLens against a repeated-feature baseline $H_x^{\times 3}$ that simply copies the residual-stream entropy three times to match the same $3L$ input size without adding new information, and further reports stricter MLP controls

that reduce the three-pathway feature back to L dimensions via simple summaries or train-split PCA. Training details (BCE loss, Adam, 50 epochs, LR scheduler) are identical to those in Zhang et al. (2025) and are listed in Appendix H.

4 Experiments

4.1 Experimental Setup

Models. We evaluate on three recent instruction-tuned LLMs spanning the ∼7–9B parameter range and three architecture families: Qwen2.5-7B-Instruct (Yang et al., 2024), Meta-Llama-3-8B-Instruct (Grattafiori et al., 2024), and Gemma-2-9B-it (Team et al., 2024). The first two use the standard RMSNorm+GQA decoder stack; Gemma-2 alternates full and sliding-window attention.

Datasets. We evaluate on four benchmarks commonly used in prior white-box detection work: HaluEval-QA (Li et al., 2023), SQuAD2.0 (Rajpurkar et al., 2018), HotpotQA (Yang et al., 2018), and TriviaQA (Joshi et al., 2017). For datasets that natively provide positive/negative supervision (HaluEval, SQuAD2), we use the provided supervision directly. For HotpotQA and TriviaQA, we follow a fixed benchmark-specific preprocessing pipeline to instantiate the evaluation examples used in our experiments. The same preprocessed instances are shared by TriLens and all baselines.

Metrics and protocol. Each benchmark is split 80/20 into train and test sets, and we report test-set AUROC averaged over five random seeds. Unless otherwise stated, we randomly sample 10,000 instances from each dataset, keeping the final label distribution balanced within every dataset. All model states are extracted from a single forward pass over the evaluation sequence, and the same sampled instances, splits, model suite, and model–dataset grid are shared by TriLens and all baselines. Further implementation details are listed in Appendix H.

Baselines. We compare against six baselines spanning both training-free and trainable white-box detectors. The training-free baselines are perplexity (**PPL**) (Ren et al., 2022), length-normalized entropy (**LN-Entropy**) (Malinin and Gales, 2020), and **LLM-Check** (Sriramanan et al., 2024); the trainable baselines are **SAPLMA** (Azaria and Mitchell, 2023), **SEP** (Kossen et al., 2024), and the previous state-of-the-art **ICR Probe** (Zhang et al., 2025). Appendix A.4 summarizes these baselines and the evaluation protocol used in our runs. All baseline results are produced under the same evaluation protocol, group-preserving 80/20 split, model suite, and model–dataset grid as TriLens.

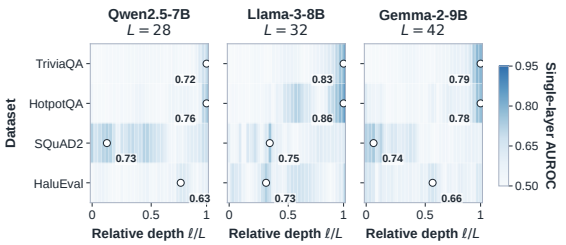

Figure 3: Single-layer AUROC heatmap from H_x^ℓ. Rows are benchmarks, columns are relative depth ℓ/L, and the peak cell in each row is circled. Peak-discriminative depth varies systematically across models.

4.2 Main Results

Table 1 reports test AUROC on all (model, dataset) cells against the six baselines. Our primary method, **TriLens (MLP)** (last row per model), uses the feature $\mathbf{f} = (H_a^\ell, H_m^\ell, H_x^\ell)_\ell$ with the same MLP probe topology used by ICR Probe. Unless otherwise noted, *TriLens* refers to this MLP variant used in Table 1.

Main finding. Hallucination risk is visible in the way internal certainty forms across layers: a compact trajectory of logit-lens entropies is sufficient to separate supported from hallucinated answers in a single forward pass. TriLens improves over ICR Probe on all 12 cells and is the highest-AUROC entry in 11 of the 12 cells, with gains ranging from $+4.2$ AUROC (Qwen2.5-7B on TriviaQA) to $+16.0$ AUROC (Qwen2.5-7B on SQuAD2). The improvement is consistent across model families ($+10.2$ averaged over Gemma-2, $+11.0$ over Qwen2.5, and $+15.1$ over Llama-3) and across benchmark regimes ($+12.8$ over the three context-grounded tasks and $+10.2$ on TriviaQA).

Probe and capacity controls. The pattern is not explained solely by the MLP probe. A linear probe on the same features also outperforms ICR Probe on all 12 cells, with the MLP adding another ~ 3.2 AUROC points (Appendix B). Dimension-matched controls further show that simply repeating H_x three times is nearly identical to H_x alone, whereas TriLens remains better on all 12 linear-probe cells. A stricter MLP sweep with L-dimensional reduc-

LLM	Methods	HaluEval	SQuAD2	HotpotQA	TriviaQA
Gemma-2-9B	PPL	0.5538	0.5348	0.7239	0.7536
	LN-Entropy	0.7357	0.6818	0.7349	0.7023
	LLM-Check	0.5780	0.5517	0.5102	0.5911
	SAPLMA	0.8123	0.7409	0.8329	0.7766
	SEP	0.6439	0.6627	0.6149	0.7726
	ICR Probe	0.8436	0.8142	0.8409	0.8001
	TriLens (Ours)	**0.9277**	**0.9270**	**0.9433**	**0.9106**
Qwen2.5-7B	PPL	0.5512	0.5278	0.6069	0.7041
	LN-Entropy	0.7286	0.6564	0.6913	0.6938
	LLM-Check	0.5367	0.5639	0.5518	0.5604
	SAPLMA	0.7725	0.7016	0.7689	**0.8297**
	SEP	0.6634	0.6418	0.6536	0.7449
	ICR Probe	0.8076	0.7382	0.7865	0.7751
	TriLens (Ours)	**0.9053**	**0.9061**	**0.9242**	0.8103
Llama-3-8B	PPL	0.5867	0.6472	0.6658	0.7085
	LN-Entropy	0.6574	0.6249	0.6661	0.5928
	LLM-Check	0.5263	0.5356	0.5559	0.5482
	SAPLMA	0.7315	0.7043	0.7768	0.7586
	SEP	0.7309	0.7274	0.6607	0.7048
	ICR Probe	0.7671	0.7568	0.7909	0.7396
	TriLens (Ours)	**0.9136**	**0.9169**	**0.9425**	**0.8861**

Table 1: **Main results**: test AUROC on four benchmark datasets under a matched evaluation protocol. **Bold** marks the best method in each model–dataset cell; underlining marks the second-best. TriLens improves over ICR Probe on all 12 cells and is best in 11, with average gain $+12.1$ AUROC. Per-cell standard deviations are listed in Appendix D.

tions of the three-pathway feature gives the same conclusion, indicating that the gain is not explained by input dimensionality alone.

4.3 Feature Ablation

The ablation answers whether the three readouts provide redundant or complementary evidence. We train the MLP probe on progressively larger subsets of $\mathbf{f}$ and additionally include an intra-layer module-disagreement term $\mathrm{JSD}^{\ell}_{am} = \mathrm{JSD}(\mathrm{lens}(a^{\ell}_i), \mathrm{lens}(m^{\ell}_i))$ as a fifth feature, motivated by an intra-layer analog of ICR Probe's consistency construction. Detailed ablation numbers are reported in Appendix E. The useful signal is not confined to the residual stream: adding either H_a or H_m to H_x improves performance on all 12 cells, and the full 3-entropy feature improves further in every cell. This indicates complementary predictive evidence across the three locations. Adding the intra-layer disagreement term changes AUROC by at most $+0.006$ per cell and $+0.001$ on average, suggesting little value beyond the entropy feature itself. The relative importance of H_a and H_m is architecture-dependent: the $H_a{+}H_x$ branch is stronger on Gemma-2, whereas $H_m{+}H_x$ is stronger on Llama-3 and part of Qwen2.5.

Figure 4 visualizes this complementarity from two angles. The left panel shows that the three entropies open hallucination–correct separation at different depths and with architecture-specific pathway asymmetry; the right panel shows that both two-feature branches improve over H_x alone, while the full 3-entropy feature attains the highest overall mean.

4.4 Per-Layer Analysis

Detailed trajectory plots and peak-layer densities are deferred to Appendix F. In the main text, we focus on the summary heatmap in Figure 3. It shows that the peak-discriminative depth varies substantially across model–dataset pairs: some open-book cells peak in early or middle layers, while TriviaQA and several HotpotQA cells peak only near the final layer. This benchmark-conditional variation is consistent with prior mechanistic analyses (Geva et al., 2023; Zhang et al., 2025), which found that different models integrate knowledge at different depths.

4.5 Cross-Dataset Diagnostics

The main results in Table 1 use a separate probe per benchmark. We next use two benchmark-shift diagnostics; detailed results appear in Appendix C. First, we train one probe per model on the union of all four benchmark training splits and evaluate it on each benchmark's held-out test split. This

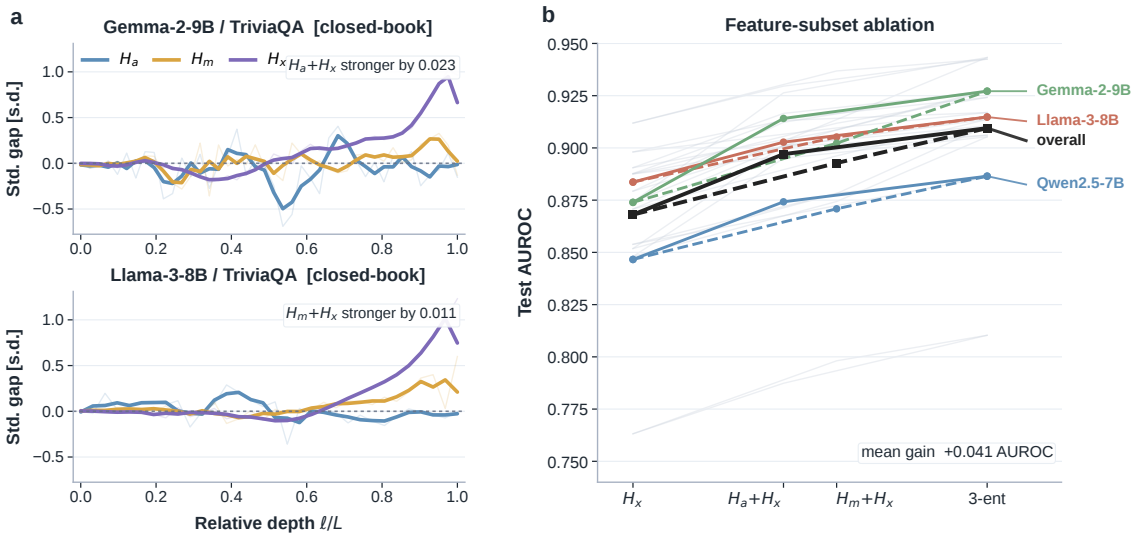

Figure 4: **Complementarity of the three entropy features. Left:** standardized per-layer separation between hallucinated and correct samples for H_a^ℓ, H_m^ℓ, and H_x^ℓ on two representative TriviaQA cells. The stronger two-feature branch is architecture-dependent. **Right:** feature-subset ablation across all 12 (model, dataset) cells. Colored lines show per-model means; grey lines show individual cells. Both two-feature branches improve over H_x alone, and the full three-entropy feature attains the best mean.

shared-probe setting still exceeds per-dataset ICR Probe on all 12 cells, with average gain +10.7 AUROC, while trailing our own per-dataset probes by only $\sim$ 1.4 AUROC on average. Second, we train on one benchmark and test on another to form benchmark-level transfer heatmaps. In this cross-dataset transfer diagnostic, TriLens is the strongest method on all 12 off-diagonal cells and on all 16 cells overall; its off-diagonal mean AUROC is 0.848, compared with 0.738 for ICR Probe and 0.678 for SAPLMA. We therefore view the transfer result as further evidence that the gain is not confined to in-domain fitting, while still treating it as a distribution-shift diagnostic rather than as evidence of universal cross-benchmark generalization.

4.6 Comparison with Cross-Layer Contrast Signals

TriLens and DoLa (Chuang et al., 2024) both derive signals from logit-lens distributions but differ in what they measure: DoLa contrasts logit-lens distributions *across* layers and uses the result to modify decoding, while we measure entropy *within* each layer and use it for detection. To assess whether our per-layer entropy feature could be explained as a reformulation of the cross-layer contrast signal, we repurpose DoLa's quantity as a detection feature and compare it head to head against ours.

Concretely, define the DoLa-style detection feature as

$$f^\ell_{\mathrm{DoLa}} = \mathrm{JSD}\big(\mathrm{lens}(x^\ell), \mathrm{lens}(x^L)\big), \quad \ell = 0, \ldots, L-1. \tag{4}$$

and assemble the L-dimensional vector $\mathbf{f}_{\mathrm{DoLa}}$. We train the same MLP probe on (a) $\mathbf{f}_{\mathrm{DoLa}}$, (b) the TriLens feature $\mathbf{f}_{\text{3-ent}}$, and (c) their concatenation $\mathbf{f}_{\text{3-ent}} \oplus \mathbf{f}_{\mathrm{DoLa}}$. Experiments are run under the same group-preserving 80/20 evaluation protocol used in our main experiments, using the same 10k-scale feature files as the main paper. Table 2 reports the main-setting comparison, while Appendix G gives the full configuration grid.

Dataset	DoLa-JSD	TriLens	TriLens+DoLa
HaluEval	0.7623	0.9136	**0.9155**
SQuAD2	0.8538	0.9158	**0.9172**
HotpotQA	0.8928	**0.9425**	0.9422
TriviaQA	0.8679	**0.8861**	0.8860

Table 2: Head-to-head comparison against DoLa-style cross-layer contrast features (MLP probe, test AUROC). Full results appear in Appendix G.

Findings. Table 2 presents three observations. First, **our per-layer entropy consistently outperforms the DoLa-style cross-layer contrast feature on every dataset**, with gaps ranging from +0.02 to +0.15 AUROC and an average gain of roughly +0.07 in the main setting. Second, **their**

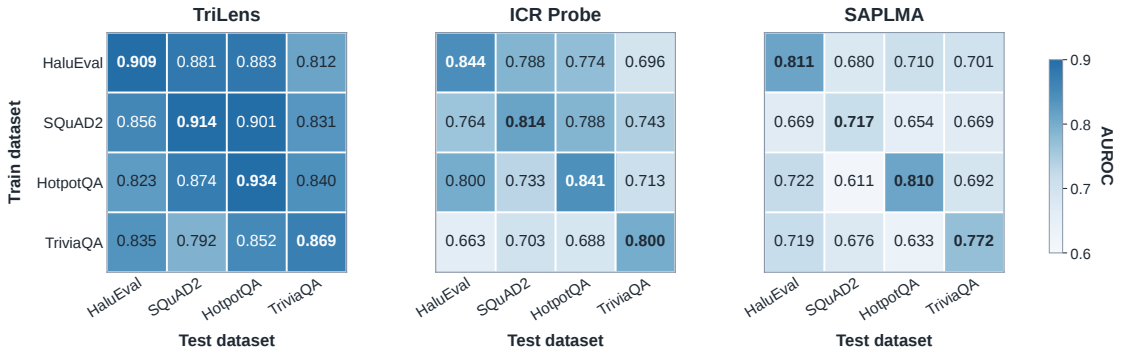

Figure 5: Cross-dataset generalization heatmaps for TriLens, ICR Probe, and SAPLMA. Each cell reports AUROC for the train-on-row, test-on-column setting. TriLens is strongest on all 12 off-diagonal cells and on all 16 cells overall, indicating more stable transfer under benchmark shift.

union changes performance only slightly and without a consistent upward pattern relative to our feature alone: averaged across the four datasets, the union changes AUROC by less than $+0.001$. The cross-layer contrast signal therefore contributes little beyond what is already captured by per-layer entropy. Third, the DoLa-style feature degrades most severely on HaluEval, where in our default setup, intermediate layers have already converged onto the final layer's prediction (so the JSD collapses toward zero regardless of whether the response is hallucinated or correct); per-layer entropy is not subject to this degeneracy because it measures the shape of each layer's distribution independently.

5 Conclusion

We propose *TriLens*, a $3L$-dimensional feature that computes Shannon entropies of logit-lens distributions at the MHSA output, FFN output, and residual stream at every layer. Across 3 LLMs $\times$ 4 benchmarks, under the same MLP probe family used by a strong prior baseline, this feature improves over ICR Probe on all 12 cells with an average gain of $+12.1$ AUROC and attains the highest AUROC in 11 of the 12 cells. Ablation identifies all three module entropies as independently informative and rules out intra-layer module-disagreement as redundant. A head-to-head comparison suggests that our feature captures much of the useful signal in DoLa-style cross-layer contrast. Layer-wise analysis further indicates that the most informative depth varies across architectures and benchmarks rather than concentrating at a universal layer. Taken together, these findings suggest that simple, layer-wise features merit consideration alongside more elaborate probe architectures in white-box hallucination detection.

Limitations

TriLens provides an effective white-box detection signal, but it has several limitations. First, it requires access to internal activations and is therefore restricted to open-source or otherwise transparent models. Second, our study focuses on detection rather than mitigation, so it does not directly address how such signals should be used to reduce hallucinations during generation. Finally, although our experiments span multiple benchmarks and model families, extending the evaluation to broader model scales and more complex generation settings would provide a stronger test of robustness and generality.

Ethics Statement

Our study is conducted entirely on publicly available LLMs and established public benchmarks, introduces no new human-subject data collection, and requires no additional annotation. We do not release any new text dataset containing personally identifying information or offensive content; all reported results are aggregate metrics over benchmark splits. We use the public models and benchmarks only under their publicly stated access conditions and licenses or terms of use, and our released code contains no redistributed model weights or benchmark text. Our use of these artifacts is limited to research evaluation of hallucination detection and is consistent with their benchmark or model-evaluation purpose and access conditions. We acknowledge the standard dual-use consideration that more accurate hallucination detection could be used to identify, and thereby refine, adversarial gen-

erations. However, as a detection-only capability without any generation or mitigation component, the net effect is defensive.

References

Amos Azaria and Tom Mitchell. 2023. The internal state of an LLM knows when it's lying. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 967–976, Singapore. Association for Computational Linguistics.

Nora Belrose, Igor Ostrovsky, Lev McKinney, Zach Furman, Logan Smith, Danny Halawi, Stella Biderman, and Jacob Steinhardt. 2023. Eliciting latent predictions from transformers with the tuned lens. *arXiv preprint arXiv:2303.08112*.

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*, pages 4401–4420, Bangkok, Thailand. 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*.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James R. Glass, and Pengcheng He. 2024. Dola: Decoding by contrasting layers improves factuality in large language models. In *The Twelfth International Conference on Learning Representations*.

Kamil Ciosek, Nicolò Felicioni, and Sina Ghiassian. 2025. Hallucination detection on a budget: Efficient bayesian estimation of semantic entropy. *arXiv preprint arXiv:2504.03579*.

Nelson Elhage, Neel Nanda, Catherine Olsson, and 1 others. 2021. A mathematical framework for transformer circuits. *Anthropic Technical Report*.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. 2024. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630:625 – 630.

Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. 2023. Dissecting recall of factual associations in auto-regressive language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12216–12235, Singapore. Association for Computational Linguistics.

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, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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*.

Xiaomeng Hu, Yiming Zhang, Ru Peng, Haozhe Zhang, Chenwei Wu, Gang Chen, and Junbo Zhao. 2024. Embedding and gradient say wrong: A white-box method for hallucination detection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1950–1959, Miami, Florida, USA. Association for Computational Linguistics.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM computing surveys*, 55(12):1–38.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada. Association for Computational Linguistics.

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

Jannik Kossen, Jiatong Han, Muhammed Razzak, Lisa Schut, Shreshth A. Malik, and Yarin Gal. 2024. Semantic entropy probes: Robust and cheap hallucination detection in llms. *ArXiv*, abs/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, Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023. HaluEval: A large-scale hallucination evaluation benchmark for large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6449–6464, Singapore. Association for Computational Linguistics.

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

Dang Nguyen, Ali Payani, and Baharan Mirzasoleiman. 2025. Beyond semantic entropy: Boosting llm uncertainty quantification with pairwise semantic similarity. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 4530–4540.

nostalgebraist. 2020. Interpreting GPT: The logit lens. https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.

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*.

Albert Sawczyn, Jakub Binkowski, Denis Janiak, Bogdan Gabrys, and Tomasz Jan Kajdanowicz. 2026. FactSelfCheck: Fact-level black-box hallucination detection for LLMs. In *Findings of the Association for Computational Linguistics: EACL 2026*, pages 5603–5621, Rabat, Morocco. Association for Computational Linguistics.

Ahmad Shapiro, Karan Taneja, and Ashok Goel. 2026. Halt: Hallucination assessment via log-probs as time series. *arXiv preprint arXiv:2602.02888*.

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

Alessandro Stolfo, Yonatan Belinkov, and Mrinmaya Sachan. 2023. A mechanistic interpretation of arithmetic reasoning in language models using causal mediation analysis. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7035–7052, Singapore. Association for Computational Linguistics.

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*.

Qwen An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxin Yang, Jingren Zhou, Junyang Lin, and 25 others. 2024. Qwen2.5 technical report. *ArXiv*, abs/2412.15115.

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, Brussels, Belgium. 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 (ACL)*.

A Experimental Details

A.1 Computing Infrastructure

Our experiments were conducted on a server equipped with 6 NVIDIA GeForce RTX 4090 GPUs (24 GB memory each), running PyTorch with CUDA acceleration. Across multiple rounds of feature extraction, ablation, and evaluation experiments, the total computational budget amounted to 100–200 GPU hours. Feature extraction requires one teacher-forced forward pass per scored sequence; probe training is lightweight relative to LLM-state extraction and is repeated across seeds and evaluation configurations reported in the main text and appendix.

A.2 Details about Datasets

For our experiments, we randomly sampled 10,000 instances from each dataset. Specifically, we used the QA subset of HaluEval and the 'rc.nocontext' subset of TriviaQA. These benchmarks are English-language QA-style evaluation sets covering factoid question answering, unanswerable questions, multi-hop reasoning, and hallucinated QA responses; we do not use demographic annotations or perform demographic subgroup analysis. Each dataset is split into 80%-20% for training and testing. We train and test on each dataset, reporting the corresponding results. This experimental setup for baseline methods matches ours exactly.

A.3 Sampling, Splits, and Supervision

All trainable methods use group-preserving 80/20 train/test splits, and the same split assignment is shared by TriLens and all baselines. All reported results are averaged over five random seeds.

A.4 Details about Baselines

We summarize the six baselines in Table 1. The first three are training-free sample-level scorers, whereas the latter three are trainable white-box detectors. All are evaluated in the same pipeline as

TriLens, on the same model–dataset grid and group-preserving 80/20 split. Scalar baselines are used directly as sample-level scores; trainable baselines retain their original feature definition and are fit on the corresponding training split. For the 10k-scale evaluation sets used in the main paper, HaluEval and SQuAD2 follow the dataset sizes reported in the main text.

PPL. Perplexity of the scored sequence under the same forward-pass setup. Higher perplexity indicates lower model confidence and is used directly as a hallucination score.

LN-Entropy. Length-normalized predictive entropy of the scored sequence. This baseline measures output uncertainty while partially controlling for raw response length.

LLM-Check. A training-free attention-kernel baseline that scores responses from structural statistics of self-attention kernels rather than hidden states.

SAPLMA. A supervised hidden-state baseline that trains an MLP classifier on internal activations from selected layers.

SEP. A probe-based hidden-state baseline motivated by semantic entropy, using learned detectors over internal representations rather than black-box semantic clustering over sampled outputs.

ICR Probe. The strongest prior white-box baseline. It derives one scalar feature per layer from residual-stream update statistics and feeds the resulting layer-wise vector to the same 4-layer MLP architecture used throughout our comparisons; the shared probe hyperparameters are listed in Appendix H.

A.5 Probe Comparability Note

TriLens and ICR Probe are compared under the same training protocol and the same hidden-layer MLP topology, but not under equal input dimensionality: TriLens uses $3L$ input features whereas ICR Probe uses L. The main text therefore interprets the comparison as a feature-plus-probe comparison within a matched probe family, rather than as a fully dimension-controlled capacity study.

B Probe Controls and Fairness Checks

Table 3 reports the same TriLens feature used in the main paper, replacing the MLP probe with L2-regularized logistic regression under the same feature configuration as the main results. The qualitative conclusion is unchanged: the linear probe already improves over ICR Probe on all 12 cells, with the MLP adding a further $\sim$ 3.2 AUROC points on average. To isolate the effect of input dimensionality, Table 4 adds a dimension-matched control $H_x^{\times 3}$ that repeats the residual-stream entropy feature three times, yielding the same $3L$ input size as TriLens without introducing any new signal. Table 5 extends this check under the same MLP probe family with stricter dimensionality controls: three handcrafted L-dimensional reductions of the three-pathway feature and a train-split PCA projection of the full $3L$-dimensional TriLens vector back to L. We further report alternative readout controls in Appendix B.1, comparing the pathway-separated TriLens readout against intermediate-state and additive-write readouts built from the same layer-wise logit-lens pipeline.

B.1 Alternative Readout Controls

The main TriLens feature applies the final logit-lens readout to the isolated MHSA write a^ℓ, the isolated FFN write m^ℓ, and the composed residual state x^ℓ. To test whether the gains depend on this particular choice of readout location, we compare it with two more conservative alternatives built from the same layer-wise pipeline: (i) H_{pre}, the entropy of the intermediate state $x^{\ell-1} + a^\ell$ after the attention write but before the FFN write, and (ii) H_p, the entropy of the additive-write state $a^\ell + m^\ell$. We then pair each alternative with the standard residual-stream readout H_x and evaluate the resulting features under the same protocol as the main paper.

Three patterns are consistent across probe families. First, performance improves monotonically as the readout moves from residual-only H_x to $H_{\mathrm{pre}} + H_x$, then to $H_p + H_x$, and finally to the full pathway-separated TriLens feature. Under the linear probe, the corresponding average AUROC increases from 0.814 to 0.833, 0.860, and 0.878, with TriLens strongest on all 12 cells. Second, the same ordering largely persists under the MLP probe, where TriLens is strongest on 11 of 12 cells and improves over H_x by $+0.052$ AUROC on average. Third, the mixed-state alternatives do help, but they remain consistently below the explicit three-way decomposition. We therefore interpret this sweep as evidence that TriLens benefits from separating MHSA, FFN, and residual-stream readouts, rather than from using an arbitrary auxiliary lens

LLM	Methods	HaluEval	SQuAD2	HotpotQA	TriviaQA
Gemma-2-9B	ICR Probe	0.8436	0.8142	0.8409	0.8001
	TriLens (Linear)	**0.8963**	**0.9004**	**0.9203**	**0.8864**
Qwen2.5-7B	ICR Probe	0.8003	0.7456	0.7917	0.7684
	TriLens (Linear)	**0.8597**	**0.8696**	**0.8746**	**0.7780**
Llama-3-8B	ICR Probe	0.7603	0.7634	0.7982	0.7325
	TriLens (Linear)	**0.8731**	**0.8806**	**0.9253**	**0.8700**

Table 3: Linear-probe results on the same 12 cells as Table 1. All TriLens entries are 5-seed means; seed standard deviation is ≤ 0.007 in every cell.

LLM	Dataset	H_x	$H_x^{\times 3}$	TriLens
Gemma-2-9B	HaluEval	0.7970	0.7970	**0.8963**
	SQuAD2	0.8497	0.8497	**0.9004**
	HotpotQA	0.7812	0.7813	**0.9203**
	TriviaQA	0.8238	0.8238	**0.8864**
Qwen2.5-7B	HaluEval	0.7746	0.7746	**0.8597**
	SQuAD2	0.8389	0.8389	**0.8696**
	HotpotQA	0.7992	0.7992	**0.8746**
	TriviaQA	0.7384	0.7384	**0.7780**
Llama-3-8B	HaluEval	0.8247	0.8248	**0.8731**
	SQuAD2	0.8254	0.8255	**0.8806**
	HotpotQA	0.8717	0.8717	**0.9253**
	TriviaQA	0.8385	0.8385	**0.8700**

Table 4: Dimension-matched fairness control under the linear probe. $H_x^{\times 3}$ repeats the L-dimensional residual-stream entropy feature three times to match TriLens's $3L$ input size. The repeated baseline is nearly identical to H_x, while TriLens remains better on all 12 cells.

location.

C Cross-Dataset Generalization Heatmaps

The main paper briefly reports the shared-probe variant of TriLens. Table 7 gives the corresponding per-benchmark numbers. We additionally report benchmark-level cross-dataset generalization heatmaps for TriLens, ICR Probe, and SAPLMA. In each 4×4 matrix, a probe is trained on the row benchmark and evaluated on the column benchmark under the same train/test transfer protocol.

Aggregate statistics. Across the 12 off-diagonal cells, TriLens reaches a mean AUROC of 0.848, compared with 0.738 for ICR Probe and 0.678 for SAPLMA. Its diagonal mean is also higher (0.907 vs. 0.825 and 0.778), so the advantage is not limited to either in-domain or out-of-domain evaluation alone.

Transfer pattern. The transfer gap is not perfectly uniform across source benchmarks. For TriLens, probes trained on SQuAD2 or HaluEval give the strongest off-diagonal averages (0.863 and 0.859), while TriviaQA is the weakest source row (0.826 off-diagonal mean). Even so, all 12 off-diagonal TriLens cells remain above 0.79, whereas the baselines show broader degradation. We therefore interpret Figure 5 as evidence of comparatively stable transfer under benchmark shift, while still viewing cross-dataset evaluation as a diagnostic rather than as a substitute for the in-domain results in Table 1.

D Main-Table Standard Deviations

Table 8 lists the standard deviations for the TriLens entries in Table 1, across the same five seeds used in the main-paper evaluation.

E Feature Ablation

Table 9 reports the full feature-set ablation from the main paper discussion.

Table 10 lists the standard deviations across the same five seeds used in Table 9.

F Per-Layer Diagnostics and Mechanistic Reading

This appendix collects the layer-wise plots and the fuller interpretation omitted from the compressed main text.

Peak-layer analysis is included only as a descriptive diagnostic rather than as a tuned component of TriLens. The main results always use the full per-layer feature vector, without selecting or reweighting layers based on validation performance. We report the peak-discriminative layer only to summarize where the single-feature H_x^ℓ detector reaches its highest AUROC in each model–dataset cell.

Context-grounded vs. closed-book regimes. On context-grounded benchmarks such as SQuAD2 and HotpotQA, the correct–hallucinated gap often

LLM	Dataset	H_x	$H_x^{\times 3}$	TriMeanL	TriMaxL	TriMinL	TriLens+PCA$\to L$	TriLens
Gemma-2-9B	HaluEval	0.8325	0.8420	0.8364	0.8140	0.8188	0.8794	**0.8967**
	SQuAD2	0.8532	0.8638	0.8461	0.8319	0.8349	0.8773	**0.8951**
	HotpotQA	0.8051	0.8272	0.8326	0.7991	0.8050	0.8927	**0.9145**
	TriviaQA	0.8237	0.8313	0.8057	0.7681	0.8053	0.8714	**0.8838**
Qwen2.5-7B	HaluEval	0.8000	0.8131	0.8004	0.7534	0.7877	0.8379	**0.8576**
	SQuAD2	0.8523	0.8606	0.8358	0.8134	0.8202	0.8669	**0.8743**
	HotpotQA	0.8376	0.8529	0.8155	0.7813	0.7585	0.8655	**0.8789**
	TriviaQA	0.7330	0.7387	0.6933	0.6218	0.6934	0.7667	**0.7700**
Llama-3-8B	HaluEval	0.7635	0.7783	0.7874	0.7480	0.7840	0.8209	**0.8379**
	SQuAD2	0.8100	0.8383	0.7902	0.7469	0.8077	0.8093	**0.8448**
	HotpotQA	0.8612	0.8710	0.8737	0.8057	0.8539	0.9034	**0.9156**
	TriviaQA	0.8305	0.8326	0.8130	0.6753	0.8110	0.8471	**0.8578**

Table 5: Stricter dimensionality controls under the MLP probe. Besides $H_x^{\times 3}$, we compare three handcrafted L-dimensional reductions of the three-pathway feature and a train-split PCA projection of the full $3L$-dimensional TriLens vector back to L. Full TriLens remains strongest on all 12 cells.

Readout	Linear		MLP	
	Avg	Wins	Avg	Wins
H_x	0.8137	0/12	0.8169	0/12
$H_{\text{pre}} + H_x$	0.8334	0/12	0.8268	0/12
$H_p + H_x$	0.8604	0/12	0.8496	1/12
TriLens	**0.8779**	**12/12**	**0.8690**	**11/12**

Table 6: Readout-location robustness summary across the 12 model–dataset cells. "Avg" denotes mean test AUROC; "Wins" counts how often a readout is the strongest cell-wise configuration. All entries are 5-seed means under the same protocol as the main paper.

Model	Test	ICR†	Multi (ours)	Δ
Qwen2.5-7B	HaluEval	0.8003	0.8856	+0.085
	SQuAD2	0.7456	0.8915	+0.146
	HotpotQA	0.7917	0.9102	+0.118
	TriviaQA	0.7684	0.7998	+0.031
Llama-3-8B	HaluEval	0.7603	0.8965	+0.136
	SQuAD2	0.7634	0.9045	+0.141
	HotpotQA	0.7982	0.9329	+0.135
	TriviaQA	0.7325	0.8739	+0.141
Gemma-2-9B	HaluEval	0.8436	0.9132	+0.070
	SQuAD2	0.8142	0.9129	+0.099
	HotpotQA	0.8409	0.9225	+0.082
	TriviaQA	0.8001	0.8971	+0.097
avg		0.7883	**0.8950**	**+0.107**

Table 7: **Multi-dataset probe**: one probe per model, trained on the union of all four benchmarks and evaluated on each held-out test split. Each cell is a 5-seed mean (std ≤ 0.008). †ICR entries are our per-dataset reruns from Table 1.

opens earlier and remains visible across much of the upper stack. On TriviaQA, by contrast, the trajectories tend to remain close until the final layers, consistent with a closed-book setting in which the relevant parametric recall signal is resolved late.

Architecture-dependent peak depth. Figure 3 shows that the peak-discriminative layer is not universal across architectures. Some Qwen2.5 cells peak in earlier or mid layers, whereas Llama-3 and Gemma-2 frequently peak closer to the final layers. This pattern is consistent with prior observations that different model families accumulate factual recall at different depths.

Pathway asymmetry. The reversal between Gemma-2 and Llama-3 in Figure 4 suggests that the relative usefulness of H_a and H_m depends on how readable each pathway's contribution is under the model's own unembedding. Gemma-2's hybrid attention stack may shift that balance toward MHSA relative to the other two models.

Smallest-gain cell. The smallest gain in the main table occurs on Qwen2.5-7B / TriviaQA. Its late-layer trajectories show weaker separation than the corresponding cells for Llama-3 and Gemma-2, suggesting a more diffuse parametric recall signal in this closed-book setting.

Table 11 reports the peak-discriminative layer for the single-feature detector based on H_x^ℓ, using the same setup as Figures 6–3. Layer indices are zero-based to match the figure annotations.

G DoLa Defense: Full Configuration Grid

Table 12 provides the full configuration grid behind Table 2: both probe families and all four datasets, using the same feature configuration as the main

Model	Dataset	Std
Gemma-2-9B	HaluEval	0.0022
Gemma-2-9B	SQuAD2	0.0025
Gemma-2-9B	HotpotQA	0.0032
Gemma-2-9B	TriviaQA	0.0027
Qwen2.5-7B	HaluEval	0.0053
Qwen2.5-7B	SQuAD2	0.0025
Qwen2.5-7B	HotpotQA	0.0015
Qwen2.5-7B	TriviaQA	0.0051
Llama-3-8B	HaluEval	0.0014
Llama-3-8B	SQuAD2	0.0037
Llama-3-8B	HotpotQA	0.0033
Llama-3-8B	TriviaQA	0.0037

Table 8: Seed standard deviations for the TriLens entries in Table 1.

Model	Data	H_x	$H_a{+}H_x$	$H_m{+}H_x$	**3-ent**	**3-ent**+JSD^{ℓ}_{am}
Gemma-2	HaluEval	.8793	.9123	.9087	.9277	**.9286**
	SQuAD2	.8905	.9165	.9085	.9270	**.9285**
	HotpotQA	.8741	.9264	.9135	.9433	**.9451**
	TriviaQA	.8518	.9013	.8782	.9106	**.9115**
Qwen2.5	HaluEval	.8480	.8950	.8743	.9053	**.9067**
	SQuAD2	.8876	.9006	.8992	.9061	**.9065**
	HotpotQA	.8877	.9138	.9118	.9242	**.9256**
	TriviaQA	.7632	.7874	.7981	**.8103**	.8102
Llama-3	HaluEval	.8708	.9008	.8962	.9136	**.9194**
	SQuAD2	.8981	.9132	.9095	.9169	**.9194**
	HotpotQA	.9119	.9296	.9368	.9425	**.9444**
	TriviaQA	.8538	.8675	.8783	**.8861**	.8846

Table 9: Feature-set ablation (MLP probe, test AUROC). **3-ent** $= (H_a, H_m, H_x)$; the last column adds JSD^{ℓ}_{am}.

paper.

H Efficiency and Complexity Analysis

H.1 Training Hyperparameters

H.2 Time Complexity

For a decoder with L layers and hidden size d, TriLens adds a logit-lens readout at three locations per layer and reduces each readout to a scalar entropy. Relative to methods that store or classify high-dimensional hidden states, the resulting sample representation is compact: TriLens produces a $3L$-dimensional vector, whereas hidden-state baselines typically operate on features that scale with d or selected layer subsets of size comparable to d. Probe-time complexity is therefore negligible compared with the upstream LLM forward pass; in practice, extraction dominates runtime and the downstream classifier is cheap to retrain across seeds.

Model	Data	H_x	$H_a{+}H_x$	$H_m{+}H_x$	**3-ent**	**3-ent**+JSD^{ℓ}_{am}
Gemma-2	HaluEval	0.004	0.002	0.003	0.002	0.002
	SQuAD2	0.006	0.004	0.006	0.004	0.004
	HotpotQA	0.003	0.003	0.003	0.004	0.004
	TriviaQA	0.004	0.004	0.004	0.004	0.005
Qwen2.5	HaluEval	0.005	0.002	0.005	0.003	0.003
	SQuAD2	0.006	0.005	0.004	0.006	0.005
	HotpotQA	0.005	0.005	0.003	0.004	0.005
	TriviaQA	0.006	0.004	0.008	0.006	0.006
Llama-3	HaluEval	0.003	0.003	0.001	0.002	0.002
	SQuAD2	0.003	0.004	0.004	0.004	0.004
	HotpotQA	0.004	0.003	0.004	0.002	0.003
	TriviaQA	0.003	0.003	0.005	0.004	0.004

Table 10: Standard deviations corresponding to Table 9 (MLP probe, same aggregation as the main results).

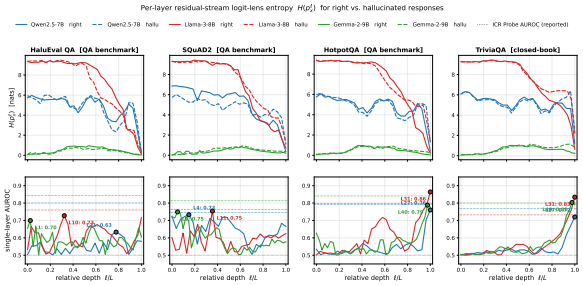

Figure 6: **Per-layer residual-stream logit-lens entropy H_x^ℓ (top) and single-layer AUROC (bottom)** across three models and four benchmarks. Solid lines denote correct responses; dashed lines denote hallucinated responses.

H.3 Memory Footprint of Logit-Lens Extraction

For a model with vocabulary size $|V|$, a naive implementation could materialize, for each scored token, $3L$ intermediate logit-lens distributions of the form $\mathrm{softmax}(W_U \cdot \mathrm{Norm}(z)) \in \mathbb{R}^{|V|}$. In float32, this corresponds to a peak memory cost of $3L|V| \times 4$ bytes per token if all such distributions are retained simultaneously. For example, for Llama-3-8B (L=32, $|V|\approx$128K), the raw peak is approximately $3 \times 32 \times 128000 \times 4 \approx 49$ MB per token. In our implementation, however, each distribution is used immediately to compute its Shannon entropy and is then discarded, so the persistent storage scales as $O(3L)$ rather than $O(3L|V|)$. For Llama-3-8B, this reduces the retained per-token state to only 96 scalar entropy values.

Model	Dataset	ℓ^*	AUROC
Qwen2.5-7B	HaluEval	L21	0.6332
Qwen2.5-7B	SQuAD2	L4	0.7329
Qwen2.5-7B	HotpotQA	L27	0.7603
Qwen2.5-7B	TriviaQA	L27	0.7199
Llama-3-8B	HaluEval	L10	0.7274
Llama-3-8B	SQuAD2	L11	0.7541
Llama-3-8B	HotpotQA	L31	0.8638
Llama-3-8B	TriviaQA	L31	0.8340
Gemma-2-9B	HaluEval	L1	0.6994
Gemma-2-9B	SQuAD2	L2	0.7491
Gemma-2-9B	HotpotQA	L40	0.7881
Gemma-2-9B	TriviaQA	L40	0.8033

Table 11: Peak-discriminative layer ℓ^* and the corresponding single-layer AUROC for H_x^ℓ.

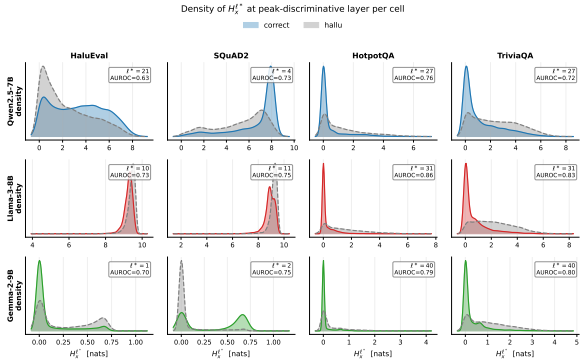

Figure 7: Probability density of $H_x^{\ell^*}$ at the peak-discriminative layer for correct vs. hallucinated responses.

Probe	Dataset	DoLa-JSD	TriLens	TriLens+DoLa
MLP	HaluEval	0.7623	0.9136	**0.9155**
	SQuAD2	0.8538	0.9158	**0.9172**
	HotpotQA	0.8928	**0.9425**	0.9422
	TriviaQA	0.8679	**0.8861**	0.8860
Linear	HaluEval	0.6790	0.8731	**0.8738**
	SQuAD2	0.7189	0.8778	**0.8798**
	HotpotQA	0.8111	0.9253	**0.9256**
	TriviaQA	0.8383	0.8700	**0.8710**

Table 12: Full DoLa comparison grid under both probe families.

Model weights dtype	bfloat16
Logit-lens projection dtype	float32
Attention implementation	SDPA
Temperature τ	1.0
Max response tokens	32
Train/test split	80/20 group-preserving split
Seeds	5
Main-paper dataset sizes	reported in Section 4.1
Row labels per dataset	exactly balanced correct/hallucinated labels
Linear probe	L2 logreg, $C = 1$
MLP probe architecture	$3L \to 128 \to 64 \to 32 \to 1$
MLP activation	LeakyReLU(0.01)
MLP regularization	BatchNorm + Dropout 0.3
MLP optimizer	Adam, lr $= 5 \times 10^{-4}$
MLP training	50 epochs, batch 32
MLP LR schedule	ReduceLROnPlateau (p=5, f=0.5)

Table 13: Training hyperparameters used in the main experiments.