BEYOND THE ENDPOINT: HALLUCINATION DETECTION VIA PREDICTIVE-PATH GEOMETRY

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models (LLMs) are increasingly deployed where factual reliability matters, making hallucination detection a pressing need. Existing detectors draw on output confidence, response consistency, and internal representations, yet the geometry of ordered intermediate predictions remains underexplored. We introduce *Hierarchical Path Excess* (HiPEx), a post-hoc white-box detector that characterizes predictive-path non-directness. HiPEx projects intermediate residual states into a shared vocabulary probability space and measures *path excess*—the two-step path length minus the direct endpoint distance. The same measure is applied across neighboring decoder blocks and within their attention–feed-forward transitions, yielding complementary layer-wise profiles. A lightweight probe scores these profiles, extracted in one teacher-forced pass through the frozen model without auxiliary generations. Across two instruction-tuned models and four question-answering benchmarks under BLEURT-20 labels, HiPEx outperforms all eight baselines in AUROC in every setting, with an average gain of 2.32 percentage points over the strongest baseline in each setting. Controlled ablations demonstrate the benefits of subtracting endpoint distance and combining both scales. Cross-benchmark transfer and alternative-label evaluations further support its robustness. These results highlight the value of predictive-path geometry for hallucination detection: the route to a prediction matters alongside its endpoint.

1 INTRODUCTION

Large language models (LLMs) can generate fluent yet incorrect or unsupported answers, highlighting the need for effective hallucination detection (Huang et al., 2025). Existing detectors draw on output confidence, response consistency, and internal representations (Kadavath et al., 2022; Farquhar et al., 2024; Azaria & Mitchell, 2023). Recent white-box studies also examine internal dynamics across layers and modules (Zhang et al., 2025; Yang et al., 2026). As token preferences evolve across depth, similar net predictive changes can arise from different intermediate paths. We argue that *the non-directness of these paths provides a useful signal for hallucination detection beyond endpoint distance alone.* We capture this signal through its layer-wise patterns across decoder blocks and within Attention–FFN transitions.

Logit-lens and tuned-lens analyses reveal how token preferences evolve across Transformer depth (nostalgebraist, 2020; Belrose et al., 2023). At each answer-token position, projecting successive residual states into a shared vocabulary probability space yields an ordered predictive trajectory. We focus on how directly local segments of this trajectory connect their endpoints. The endpoint distance summarizes net predictive change, whereas the path length accounts for movement through intermediate states. Relating these two quantities provides a geometric description of *predictive-path non-directness*, capturing intermediate deviations that endpoint distance alone does not reveal.

We introduce *Hierarchical Path Excess* (HiPEx), a post-hoc white-box detector. For three states ordered by computation, *path excess* measures the extra distance introduced by passing through the intermediate state: the two-step path length minus the direct endpoint distance.

The same distinction matters within a decoder block: its input and output expose the net change, while retaining the post-attention state reveals how attention and feed-forward updates compose. HiPEx therefore applies the same path-excess functional at two nested scales. The *inter-block* scale

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

operates across neighboring block outputs; the *intra-block* scale operates along cumulative residual states before attention, after attention, and after the feed-forward update. The resulting profiles are averaged over answer tokens while preserving their depth coordinates, then scored by a lightweight probe. All features are extracted from a completed answer in one teacher-forced pass through the frozen model, with no auxiliary generations at detection time.

Across Llama-3 and Gemma-2 on four question-answering benchmarks, HiPEx achieves the highest AUROC among nine methods in all eight model–benchmark settings under BLEURT-20 labels, with a macro-average of 87.79% and a mean gain of 2.32 percentage points over the strongest baseline in each setting. Controlled ablations support subtracting the endpoint distance and combining both scales, including comparisons that match feature dimensionality or probe capacity. Under source-only fitting, HiPEx also achieves the highest mean cross-benchmark transfer AUROC averaged across both backbones among the evaluated learned white-box detectors. It retains its detection lead under ROUGE-L relabeling on two Llama-3 benchmarks.

Our contributions are threefold:

- We introduce predictive-path excess as a geometric descriptor of intermediate predictive movement relative to net endpoint displacement for hallucination detection.
- We develop HiPEx, which applies one path-excess functional across decoder blocks and within their Attention–FFN transitions to construct a compact, depth-preserving representation for single-response detection.
- We show consistent detection gains and strong source-only cross-benchmark transfer, supported by controlled geometric and hierarchical ablations and alternative-label evaluations.

2 Related Work

2.1 Output-Space and Sampling-Based Uncertainty

Output-space approaches estimate answer reliability through token likelihood, perplexity, predictive entropy, and verbalized confidence (Malinin & Gales, 2021; Ren et al., 2023; Kadavath et al., 2022). Sampling-based methods compare multiple generations. SelfCheckGPT checks consistency between a candidate response and additional samples from the same model (Manakul et al., 2023), while semantic entropy groups sampled answers by meaning and measures uncertainty over semantic equivalence classes (Farquhar et al., 2024). These approaches summarize output confidence or agreement among completed responses. HiPEx complements them by examining the ordered predictive states induced when a single completed answer is replayed through the model.

2.2 White-Box Hallucination Detection

White-box methods complement output confidence with internal signals. Hidden-state probes and representation-geometry studies identify truth-related information in LLM activations (Azaria & Mitchell, 2023; CH-Wang et al., 2024; Marks & Tegmark, 2024; Bürger et al., 2024). INSIDE uses hidden-state geometry for eigenvalue-based self-consistency (Chen et al., 2024), while MIND derives unsupervised token-level signals (Su et al., 2024). HaloScope identifies a latent subspace from unlabeled generations, estimates truthful or hallucinated membership through projection scores, and trains a classifier on these assignments (Du et al., 2024). Semantic Entropy Probes distill sampling-derived semantic uncertainty into single-response hidden-state predictors (Kossen et al., 2024).

LLM-Check also studies statistics of hidden states, attention maps, and output probabilities (Sriramanan et al., 2024). Other approaches exploit embedding gradients (Hu et al., 2024), attention allocation or spectral structure (Chuang et al., 2024a; Binkowski et al., 2025), and the use of external evidence versus parametric knowledge in retrieval-augmented generation (Sun et al., 2025).

2.3 Intermediate Predictions and Internal Dynamics

The logit lens reveals token preferences across depth by applying the output head to residual states (nostalgebraist, 2020). Tuned Lens improves intermediate-to-final prediction correspondence with

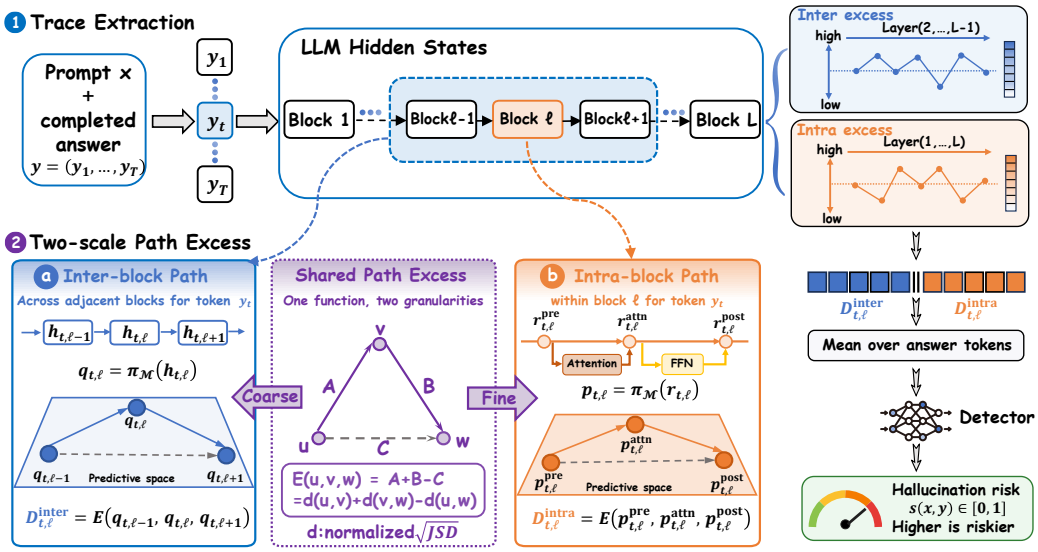

Figure 1: Overview of Hierarchical Path Excess. A single teacher-forced forward pass extracts residual-stream states, which are projected into a shared vocabulary probability simplex. Path excess measures predictive-path non-directness at inter-block and intra-block scales. The resulting layer-wise profiles are concatenated, mean-pooled over answer tokens, and scored by an MLP probe.

learned affine translators (Belrose et al., 2023), while DoLa contrasts layer-wise predictions to improve factual generation (Chuang et al., 2024b). These analyses align with the residual-stream view of successive attention and feed-forward updates (Elhage et al., 2021; Geva et al., 2021; 2023).

Internal dynamics also inform hallucination detection through hidden-state transitions (Zhu et al., 2024) and cross-layer dispersion (Srey et al., 2026). ICR Probe measures agreement between attention and residual-update projections over context positions (Zhang et al., 2025). TriLens records per-layer logit-lens entropies of attention and FFN outputs and residual states, and also examines pairwise branch-disagreement and intermediate-to-final contrast features (Yang et al., 2026). HiPEx focuses on three-point geometry along cumulative residual states: distances computed from full vocabulary distributions measure the two-step path length, from which the direct endpoint distance is subtracted. The same functional operates across neighboring block outputs and within each block's executed Attention–FFN sequence. Sections 4.3 and 4.4 evaluate the contributions of endpoint subtraction and combining these two scales.

3 Hierarchical Path Excess

Given a prompt x and a completed answer $y = (y_1, \ldots, y_T)$, Hierarchical Path Excess (HiPEx) uses one teacher-forced pass through the frozen language model to extract answer-token residual states. It measures predictive-path excess at two computational scales and aggregates the resulting profiles for answer-level detection (Figure 1).

3.1 Predictive Path Excess

Logit-Lens Projection. Consider an autoregressive model $\mathcal{M}$ with vocabulary $\mathcal{V}$, residual dimension d, and L decoder blocks. Let $h_{t,\ell} \in \mathbb{R}^d$ denote the residual-stream state at answer position t after block ℓ. We project a residual state r into the vocabulary probability simplex:

$$\pi_{\mathcal{M}}(r) = \mathrm{softmax}\left(\frac{g_{\mathcal{M}}(W_U \,\mathrm{Norm}_f(r))}{\tau}\right) \in \Delta^{|\mathcal{V}|-1}. \tag{1}$$

Here, Norm_f and W_U are the native final normalization and output head, $g_{\mathcal{M}}$ follows the model's logit convention, and $\tau > 0$ is fixed. This shared coordinate system makes predictive effects comparable across layers and module locations; model-specific definitions appear in Appendix J.

Distance between Predictive States. For vocabulary distributions p, q, we use the normalized square-root Jensen–Shannon distance:

$$d(p, q) = \sqrt{\frac{\mathrm{JSD}(p, q)}{\log 2}} \in [0, 1], \tag{2}$$

where $\mathrm{JSD}(p, q) = \frac{1}{2}\,\mathrm{KL}(p\|\bar{p}) + \frac{1}{2}\,\mathrm{KL}(q\|\bar{p})$ and $\bar{p} = (p + q)/2$. This symmetric, bounded metric satisfies the triangle inequality, which ensures non-negative path excess. It compares predictive distributions without specifying pairwise distances between vocabulary items.

Path Excess. For three predictive distributions $u \to v \to w$ ordered by computation, define

$$E(u, v, w) = d(u, v) + d(v, w) - d(u, w). \tag{3}$$

The two-step length measures total predictive movement but alone does not distinguish a long, nearly direct path from a redirected one. Conversely, the endpoint distance retains only the net change and discards the intermediate state. Their difference subtracts the endpoint distance from the realized path length, emphasizing the additional distance introduced by passing through the intermediate state. Thus, path excess characterizes non-directness rather than movement magnitude alone. This is the slack in the triangle inequality, so $E \geq 0$. For fixed step lengths, E increases with directional disagreement in the metric's implicit Hilbert-space embedding. Aligned displacements yield zero excess, whereas opposing displacements yield twice the shorter step length (Appendix E.1).

3.2 HiPEx at Two Computational Scales

HiPEx applies the same functional in Equation 3 across decoder blocks and within each block.

Inter-Block Construction. Let $q_{t,\ell} = \pi_{\mathcal{M}}(h_{t,\ell})$ for $\ell = 1, \ldots, L$. For every interior block, define

$$D^{\mathrm{inter}}_{t,\ell} = E(q_{t,\ell-1}, q_{t,\ell}, q_{t,\ell+1}), \qquad \ell = 2, \ldots, L-1. \tag{4}$$

This measures the additional path length through the output of block ℓ, relative to the direct distance between its neighboring block outputs. The profile $\mathbf{D}^{\mathrm{inter}}_t = [D^{\mathrm{inter}}_{t,2}, \ldots, D^{\mathrm{inter}}_{t,L-1}] \in \mathbb{R}^{L-2}$ retains the layer index of path excess.

Intra-Block Construction. In each block, self-attention aggregates information from causally accessible positions, and the FFN then applies a learned position-wise nonlinear transformation to the resulting attention-updated state (Elhage et al., 2021; Geva et al., 2021; 2023). Because the FFN acts on the attention-updated residual rather than on the block's original input, the two branches form a coupled, sequential computation rather than two independent contributions to the residual stream. Keeping the intermediate post-attention state therefore lets us see how their predictive effects compose, beyond what is visible from the block's net input–output change alone. We write $a_{t,\ell}$ and $m_{t,\ell}$ for the effective increments that the attention and FFN branches add to the residual stream, respectively, including any native branch-output normalization; the cumulative residual states are then

$$\begin{aligned} r^{\mathrm{pre}}_{t,\ell} &= h_{t,\ell-1}, \\ r^{\mathrm{attn}}_{t,\ell} &= r^{\mathrm{pre}}_{t,\ell} + a_{t,\ell}, \\ r^{\mathrm{post}}_{t,\ell} &= r^{\mathrm{attn}}_{t,\ell} + m_{t,\ell} = h_{t,\ell}. \end{aligned} \tag{5}$$

Here $h_{t,0}$ is the first block input, and $m_{t,\ell} = \mathcal{F}_\ell(r^{\mathrm{attn}}_{t,\ell})$, where $\mathcal{F}_\ell$ includes the FFN branch's native normalization operations. We directly capture these cumulative states at model-specific execution points (Appendix J). Projecting the states as $p^c_{t,\ell} = \pi_{\mathcal{M}}(r^c_{t,\ell})$ for $c \in \{\mathrm{pre}, \mathrm{attn}, \mathrm{post}\}$ gives

$$D^{\mathrm{intra}}_{t,\ell} = E\Big(p^{\mathrm{pre}}_{t,\ell}, p^{\mathrm{attn}}_{t,\ell}, p^{\mathrm{post}}_{t,\ell}\Big), \qquad \ell = 1, \ldots, L. \tag{6}$$

This measures the extra predictive distance along the executed Attention–FFN path relative to its input–output distance. The projected points are cumulative residual states rather than isolated branch outputs. The profile $\mathbf{D}^{\mathrm{intra}}_t = [D^{\mathrm{intra}}_{t,1}, \ldots, D^{\mathrm{intra}}_{t,L}] \in \mathbb{R}^L$ retains all L block coordinates.

Hierarchical Representation. We concatenate the two layer-wise profiles:

$$\mathbf{F}_t = [\mathbf{D}^{\mathrm{inter}}_t; \mathbf{D}^{\mathrm{intra}}_t] \in \mathbb{R}^{2L-2}. \tag{7}$$

The representation preserves both block-to-block evolution and within-block Attention–FFN composition, with 62 dimensions for 32-block models and 82 for 42-block models.

3.3 HIPEX PROBE

We mean-pool over all T valid answer tokens, preserving the layer coordinates of both scales:

$$\mathbf{F}(x, y) = \frac{1}{T} \sum_{t=1}^{T} \mathbf{F}_t \in \mathbb{R}^{2L-2}. \tag{8}$$

Path excess is computed separately at each answer-token position before pooling; we average the resulting features, not the predictive distributions. We standardize each feature using training-split statistics and apply an MLP probe:

$$s_\phi(x, y) = f_\phi(\mathrm{Scale}(\mathbf{F}(x, y))) \in \mathbb{R}. \tag{9}$$

The probe is trained with unweighted binary cross-entropy, using $z_i = 1$ for answers labeled as hallucinated according to the protocol in Appendix B. Larger logits indicate higher hallucination risk; reported scores use the sigmoid of the logit. Only the probe is trained, while the language model remains frozen. Feature-extraction details appear in Appendix J, and probe architecture and training settings in Appendix K.

4 EXPERIMENTS

We evaluate detection performance, geometric and hierarchical design choices, cross-benchmark transfer, and robustness to alternative labels.

4.1 EXPERIMENTAL SETUP

Models and Generations. We evaluate Meta-Llama-3-8B-Instruct (Grattafiori et al., 2024) (32 decoder blocks) and Gemma-2-9B-it (Gemma Team et al., 2024) (42 blocks), keeping both backbones frozen. Each model produces one candidate answer per question with its native chat template and deterministic five-beam decoding, capped at 64 new tokens. This frozen answer bank is shared across all detectors. Exact prompts and generation settings are provided in Appendix A.

Benchmarks and Labels. Four question-answering benchmarks span complementary regimes: closed-book TriviaQA (Joshi et al., 2017) and NQ-Open (Kwiatkowski et al., 2019), and passage-grounded SQuAD 1.1 (Rajpurkar et al., 2016) and English TyDiQA-GoldP (Clark et al., 2020), totaling 67,312 generated answers across both models. A response is labeled correct when its maximum BLEURT-20 score over references exceeds 0.5, and hallucinated otherwise (Sellam et al., 2020). Dataset construction, reference-based labeling, and split protocols are detailed in Appendix B; ROUGE-L label robustness is evaluated in Section 4.6.

Baselines. Eight baselines cover three categories: single-response scores, including Perplexity (Ren et al., 2023) and LLM-Check (Sriramanan et al., 2024); sampling-based uncertainty estimators, including LN-Entropy (Malinin & Gales, 2021) and Semantic Entropy (Kuhn et al., 2023; Farquhar et al., 2024); and learned white-box detectors, including SEP (Kossen et al., 2024), HaloScope (Du et al., 2024), ICR Probe (Zhang et al., 2025), and TriLens (Yang et al., 2026). All methods share the candidate answers and data splits; sampling-based methods also share five auxiliary generations per question. Baseline-specific supervision, model selection, and adaptations are summarized in Appendix C, alongside each method's feature construction and scoring procedure.

Evaluation. AUROC is our primary metric. We also report hallucination-positive and correct-answer-positive AUPRC (Manakul et al., 2023) for all methods to characterize performance under class imbalance; the latter treats correct answers as positive and reverses the hallucination-risk score. We use five label-stratified 72/8/20 train/validation/test splits. For SQuAD and TyDiQA, each source paragraph is confined to a single partition. Preprocessing and fitting use the training split, model selection uses validation, and the test split is reserved for final evaluation. The main table reports means for all three metrics over five test splits. HiPEx sample standard deviations are provided in Appendix D for all three metrics across the eight model–benchmark settings.

4.2 MAIN RESULTS

HiPEx ranks first on all three metrics in all eight model–benchmark settings (Table 1), reaching a macro-average AUROC of 87.79%. Compared with the strongest baseline listed in Table 1 for each

Table 1: Hallucination detection under BLEURT-20 labels. Entries are five-split means (%); higher is better. H-PR and C-PR denote hallucination-positive and correct-answer-positive AUPRC. † indicates auxiliary generations at test time. Best values are bold and blue-shaded; lighter shading indicates second-best values.

Methods	TriviaQA			NQ-Open			SQuAD 1.1			TyDiQA		
	AUROC	H-PR	C-PR	AUROC	H-PR	C-PR	AUROC	H-PR	C-PR	AUROC	H-PR	C-PR
Meta-Llama-3-8B-Instruct												
Perplexity	72.95	61.51	81.04	70.84	87.82	49.76	68.86	63.35	70.77	59.25	90.94	18.80
LLM-Check	69.46	61.94	74.31	82.21	95.49	43.05	56.17	54.68	56.39	53.27	89.22	13.68
LN-Entropy†	81.92	73.59	86.23	77.57	92.10	50.81	76.61	75.43	75.52	67.17	93.41	20.76
Semantic Entropy†	80.63	72.01	81.44	82.38	93.87	45.03	70.26	68.29	63.10	64.16	91.80	17.45
SEP	77.06	70.27	80.79	80.30	94.30	44.24	75.01	73.03	73.33	65.28	91.94	22.55
HaloScope	80.57	77.45	83.04	90.29	97.67	59.36	88.36	86.61	86.55	85.13	97.28	51.28
ICR Probe	84.53	82.93	86.81	89.73	97.61	59.05	88.20	87.35	87.37	82.41	96.37	55.08
TriLens	80.91	75.25	84.84	82.62	95.23	51.95	78.90	79.53	77.17	81.93	96.34	50.85
HiPEx (Ours)	**86.44**	**84.31**	**88.59**	**91.25**	**97.90**	**64.54**	**89.78**	**89.29**	**88.49**	**88.01**	**97.88**	**63.53**
Gemma-2-9B-it												
Perplexity	78.44	61.86	88.06	69.62	82.78	50.66	71.42	54.48	82.20	67.14	67.13	65.33
LLM-Check	62.00	46.21	75.40	71.85	86.93	44.47	57.25	42.93	68.17	66.09	69.01	60.58
LN-Entropy†	80.91	68.17	88.70	75.21	87.41	53.53	72.38	59.49	82.05	69.52	71.44	66.79
Semantic Entropy†	78.56	64.60	84.50	75.46	85.72	48.52	63.95	52.51	69.45	68.29	70.72	56.45
SEP	78.58	64.57	86.57	74.78	87.53	50.52	73.03	62.59	78.46	76.66	79.15	71.30
HaloScope	77.43	65.01	86.16	61.44	78.70	38.07	84.92	78.41	87.48	79.81	82.15	75.23
ICR Probe	77.10	68.19	85.99	80.83	91.93	57.27	85.61	78.46	89.36	88.10	89.70	85.38
TriLens	74.13	59.33	84.66	76.48	88.76	53.35	75.09	63.67	82.09	82.66	85.41	78.50
HiPEx (Ours)	**82.91**	**74.06**	**90.52**	**84.81**	**93.70**	**64.90**	**88.76**	**83.23**	**91.57**	**90.39**	**91.79**	**87.32**

setting and metric, its mean gains are 2.32 AUROC points (0.96–3.98), 2.33 hallucination-positive AUPRC points, and 3.77 correct-answer-positive AUPRC points. The strongest AUROC competitor varies among ICR Probe, HaloScope, and LN-Entropy, while HiPEx's lead spans both backbones and both closed-book and passage-grounded QA. The gains in both AUPRC directions support using the same risk score to prioritize hallucinated responses and select correct answers.

Paired Comparison Evidence. A separate paired cluster-bootstrap analysis supports HiPEx's AUROC gains over ICR Probe and HaloScope in all 16 comparisons (Holm-adjusted $p < 0.05$), conditional on the fitted models (Appendix L).

Supplementary Representation Controls. To compare representations under a matched downstream budget, we compress HiPEx to ICR Probe's input dimensionality and match probe parameter counts. HiPEx-PCA achieves higher mean AUROC than ICR Probe in seven of eight settings, with a macro-average AUROC gain of 1.35 points (Appendix M.1). A supplementary pooling control applies answer-token mean pooling to TriLens, matching HiPEx's aggregation. HiPEx remains ahead in all eight settings, averaging a 2.44-point AUROC gain (Appendix M.2).

4.3 Why Path Excess?

We isolate the effect of subtracting the endpoint distance by replacing $E = A + B - C$ at every inter- and intra-block location with either the realized two-step path length $A + B$ or the endpoint distance C, while leaving the remainder of HiPEx unchanged. The three variants therefore share the same feature dimensionality, data splits, standardization, and probe architecture. Figure 2 compares these three geometric summaries: total path length, endpoint distance, and path excess.

Subtracting Endpoint Distance Yields Consistent Gains. Across the eight settings, path excess improves mean AUROC by 1.81 percentage points over $A + B$ and 1.89 over C, averaging a 1.63-point gain over the stronger control in each setting, as shown in Figure 2. $A + B$ leads the controls on all four Llama-3 benchmarks, while C leads on three Gemma-2 benchmarks. These gains favor path excess over total predictive movement or net endpoint displacement alone.

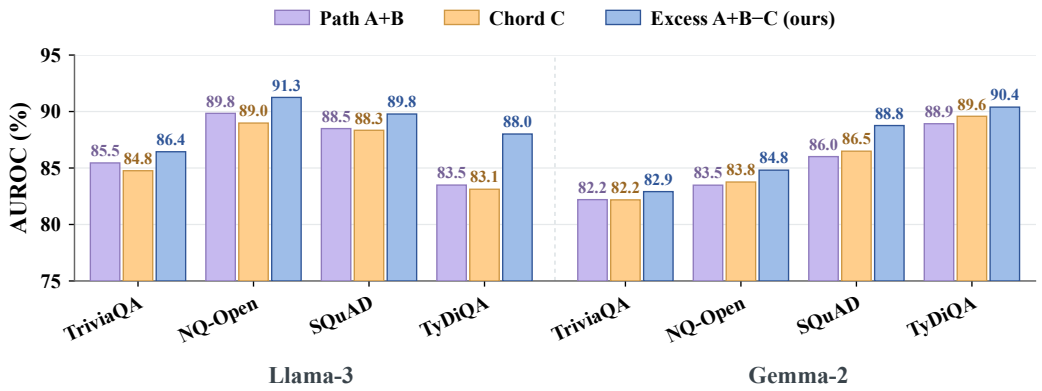

Figure 2: Geometric ablation across both backbones. Bars show five-split mean AUROC (%) for $A + B$, C, and $A + B - C$ with other settings fixed. Bars show means only.

Explicit Geometry versus Learned Combination. We compare HiPEx with joint features $[A + B, C]$ and Joint-PCA, which compresses them to HiPEx's input dimensionality. Under answer-token mean pooling, HiPEx is the coordinate-wise difference of the two component profiles. Using the uncompressed joint input, we compare explicit endpoint subtraction with learned combination under the same training protocol. HiPEx achieves higher mean AUROC in seven of eight settings against joint features and all eight against Joint-PCA, with macro-average gains of 1.29 and 1.84 percentage points, respectively. These results support endpoint subtraction as an effective geometric inductive bias, including at matched dimensionality (Appendix E.2).

4.4 Why Two Computational Scales?

We examine layer-wise class differences and test whether the two computational scales provide complementary detection signals.

4.4.1 Layer-Wise Path Profiles

Figure 3 reveals layer- and scale-specific class differences in path excess on Llama-3 TriviaQA. We first average each feature over tokens within each answer, then compute separate class means for hallucinated and correct answers. Panel (b) shows a shared overall decline and late rebound at each scale, while panel (a) reveals the class differences within this common depth pattern.

Using the signed gap $\Delta_\ell = D_\ell^H - D_\ell^N$, panel (a) highlights two aspects of these differences. Across depth, hallucinated answers have greater inter-block path excess through much of the early and middle stack, but lower excess near the output, with the most negative gap at block 31. Across scales, the intra-block gap follows a distinct pattern, changing sign repeatedly and remaining predominantly negative in late blocks. At blocks 11–13, positive inter-block and negative intra-block gaps show that cross-block evolution and within-block updates exhibit opposite class contrasts even in the same depth region.

These observations motivate HiPEx's depth-preserving, two-scale representation. Retaining layer coordinates preserves opposing class contrasts that depth averaging could attenuate, while retaining both scales makes their distinct patterns available to the probe. HiPEx thus captures both where predictive-path non-directness differs between classes and how those differences depend on computational scale. Gemma-2 also exhibits layer- and scale-dependent contrasts (Appendix F); Section 4.4.2 quantifies the detection benefit of combining both scales.

4.4.2 Complementarity Across Computational Scales

Table 2 compares Inter-only, Intra-only, and Full HiPEx with the standard $128/64/32$ probe. We additionally widen only the first hidden layer of each single-scale probe to match Full's trainable parameter count, keeping Full unchanged (Appendix K).

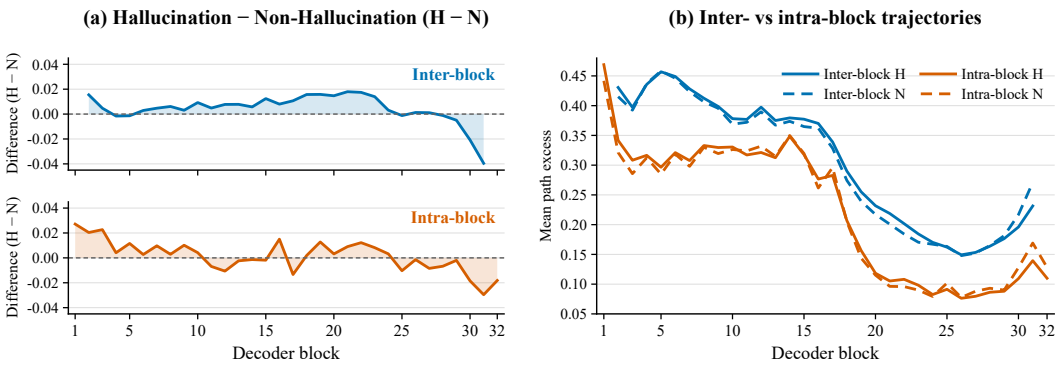

Figure 3: HiPEx profiles on Llama-3 TriviaQA after within-answer token averaging. Panel (a) shows the hallucinated-minus-correct gap at each scale; positive values indicate larger hallucinated-class means. Panel (b) shows the corresponding class-conditional profiles.

Full leads in all eight settings under both standard and matched probe capacity, gaining 2.19 and 2.10 AUROC points on average over the stronger single-scale variant, respectively. Gains over the stronger standard variant range from 0.70 to 5.79 points. Intra outperforms Inter on all four Llama-3 benchmarks, while Inter leads on all four Gemma-2 benchmarks. Combining the scales therefore provides a consistently effective representation without backbone-specific scale selection. Full also exceeds dimension-matched Repeated Inter and sample-shuffled fusion in all eight settings, by 3.32 and 4.89 points on average (Appendix G), supporting the value of answer-aligned information beyond redundant feature expansion.

Table 2: HiPEx scale ablation with standard and parameter-matched probes. Entries are five-split mean AUROC (%); row-wise best values are bold.

Model	Benchmark	Standard		Matched params.		Full
		Inter	Intra	Inter	Intra	
Llama-3	TriviaQA	81.66	85.20	81.62	85.40	**86.44**
	NQ-Open	87.24	90.55	87.36	90.54	**91.25**
	SQuAD	86.33	87.29	86.53	87.36	**89.78**
	TyDiQA	81.22	82.22	81.23	82.32	**88.01**
Gemma-2	TriviaQA	80.92	79.51	81.18	79.66	**82.91**
	NQ-Open	83.47	83.20	83.21	83.18	**84.81**
	SQuAD	86.40	85.56	86.61	85.74	**88.76**
	TyDiQA	88.75	87.92	88.92	88.01	**90.39**

4.5 Cross-Benchmark Transfer

We compare HiPEx with ICR Probe and HaloScope under source-only transfer. For each backbone and source benchmark, preprocessing and fitting use source training data, while model selection uses source validation data. The frozen pipeline is evaluated on all target test sets, with target labels used only for final evaluation.

On Gemma-2 (Figure 4), HiPEx leads all 12 cross-benchmark directions, with mean off-diagonal AUROC of 80.44% versus 70.82% for ICR Probe and 71.52% for HaloScope. The lead covers transfers within and between closed-book and passage-grounded QA. Across both backbones, HiPEx leads in 16 of 24 transfer directions and attains the highest mean off-diagonal AUROC at 78.80%, compared with 74.90% for HaloScope and 71.16% for ICR Probe. Performance varies by backbone: HaloScope leads HiPEx by 1.12 points in mean off-diagonal AUROC on Llama-3 (Appendix I). Overall, HiPEx leads the evaluated learned white-box detectors in aggregate transfer.

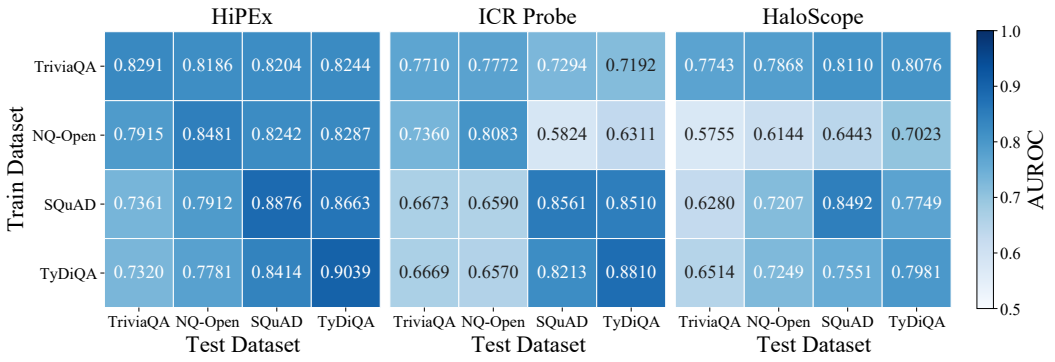

Figure 4: Source-only transfer on Gemma-2. Rows indicate training benchmarks and columns test benchmarks. Entries are AUROC; off-diagonal cells measure transfer without target adaptation.

4.6 Robustness to Alternative Labels

We assess label robustness by replacing BLEURT-20 with ROUGE-L F_1 labels on Llama-3 TriviaQA and TyDiQA, covering closed-book and passage-grounded QA. An answer is labeled correct when its maximum ROUGE-L F_1 score over references exceeds 0.5. We retain the frozen answers and seed identifiers, with partitions stratified under the alternative labels. Probes trained on correctness labels are retrained, while training-free scores are evaluated against the new labels.

Table 3: Hallucination detection under ROUGE-L labels on Llama-3. Values are mean ± sample standard deviation AUROC (%) over five splits.

Method	**TriviaQA**	**TyDiQA**	**Average**
PPL	73.72 ± 1.04	62.73 ± 2.56	68.23
LN-Entropy	82.55 ± 1.09	69.45 ± 1.85	76.00
LLM-Check	69.61 ± 0.40	49.33 ± 1.91	59.47
Semantic Entropy	81.50 ± 1.04	65.97 ± 1.62	73.73
SEP	77.61 ± 0.94	64.09 ± 5.62	70.85
HaloScope	81.30 ± 1.62	81.96 ± 3.60	81.63
ICR Probe	84.72 ± 0.64	82.44 ± 1.13	83.58
TriLens	80.67 ± 1.15	83.13 ± 1.51	81.90
HiPEx	$\mathbf{86.04 \pm 0.90}$	$\mathbf{85.96 \pm 1.60}$	**86.00**

HiPEx ranks first among all nine methods on both benchmarks under ROUGE-L labels (Table 3), with a macro-average AUROC of 86.00%. Under human answer–reference agreement labels on TriviaQA and TyDiQA across both backbones, HiPEx leads the four evaluated methods in three of four settings and in macro-average AUROC (75.59%); HaloScope leads on Llama-3 TyDiQA (Appendix H.2). HiPEx also exceeds a length-only control in all eight settings, with a macro-average AUROC gain of 19.56 percentage points (Appendix H.3).

5 Conclusion

We introduced Hierarchical Path Excess, a white-box detector that measures predictive-path non-directness in a shared vocabulary probability space at inter-block and intra-block resolutions. Across two language models and four question-answering benchmarks under BLEURT-20 labels, HiPEx led the compared methods in AUROC in all eight settings. Controlled ablations support subtracting the endpoint distance and combining both scales, with gains retained under matched probe capacity. HiPEx also led the evaluated learned white-box detectors in cross-benchmark transfer AUROC averaged across both backbones. These results highlight predictive-path geometry as a useful signal for hallucination detection: the route to a prediction matters alongside its endpoint.

AI USE STATEMENT

The authors employed generative AI tools exclusively for light language polishing of text originally written by the authors. This includes grammar checks, sentence rephrasing, improving clarity and consistency of academic prose. No generative AI was used to conceive research ideas, design experiments, derive technical claims, analyze experimental results, or draft any core technical content of this work. All modifications proposed by the AI tool were carefully examined and verified by the authors before adoption. The authors hold full responsibility for all content, claims, and conclusions presented in this manuscript.

REPRODUCIBILITY STATEMENT

We include an anonymized implementation of the main HiPEx pipeline in the supplementary materials. The package covers data preparation, answer generation and BLEURT-20 labeling, model-specific residual-state extraction, inter- and intra-block path-excess computation, answer-token pooling, feature preprocessing, and probe training and evaluation.

Configuration files and execution commands are provided for the main experiments with Llama-3-8B-Instruct and Gemma-2-9B-it on TriviaQA, NQ-Open, SQuAD 1.1, and TyDiQA-GoldP English, using five fixed random seeds per setting. The documentation specifies data sources, partitioning rules, generation settings, optimization hyperparameters, checkpoint selection, evaluation metrics, and software dependencies, complementing Appendices A–B and J–K. Reference hashes, unit tests, and a CPU-only synthetic example support artifact verification and basic workflow checks. The README also documents external resource requirements and clarifies the scope of reproducibility, noting that the original frozen answers, extracted features, and trained checkpoints are not included.

Upon publication, we plan to publicly release the implementation in a maintained repository, together with configurations for controlled ablations, fixed split indices, and scripts for reproducing the reported tables and figures.

REFERENCES

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

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

Jakub Binkowski, Denis Janiak, Albert Sawczyn, Bogdan Gabrys, and Tomasz Kajdanowicz. Hallucination detection in llms using spectral features of attention maps, 2025. Accepted to EMNLP 2025.

Lennart Bürger, Fred A. Hamprecht, and Boaz Nadler. Truth is universal: Robust detection of lies in llms. In *Advances in Neural Information Processing Systems*, 2024.

Sky CH-Wang, Benjamin Van Durme, Jason Eisner, and Chris Kedzie. Do androids know they're only dreaming of electric sheep? In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 4401–4420. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-acl.260. URL https://aclanthology.org/2024.findings-acl.260/.

Chao Chen, Kai Liu, Ze Chen, Yi Gu, Yue Wu, Mingyuan Tao, Zhihang Fu, and Jieping Ye. Inside: Llms' internal states retain the power of hallucination detection. In *International Conference on Learning Representations*, 2024.

Yung-Sung Chuang, Linlu Qiu, Cheng-Yu Hsieh, Ranjay Krishna, Yoon Kim, and James R. Glass. Lookback lens: Detecting and mitigating contextual hallucinations in large language models using only attention maps. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 1419–1436. Association for Computational Linguistics, 2024a. doi: 10.18653/v1/2024.emnlp-main.84. URL https://aclanthology.org/2024.emnlp-main.84/.

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

Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. TyDi QA: A benchmark for information-seeking question answering in typologically diverse languages. *Transactions of the Association for Computational Linguistics*, 8:454–470, 2020. doi: 10.1162/tacl_a_00317. URL https://aclanthology.org/2020.tacl-1.30/.

Xuefeng Du, Chaowei Xiao, and Yixuan Li. Haloscope: Harnessing unlabeled llm generations for hallucination detection. In *Advances in Neural Information Processing Systems*, 2024. Spotlight.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. A mathematical framework for transformer circuits. Transformer Circuits Thread, 2021.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 2024. doi: 10.1038/s41586-024-07421-0.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, et al. Gemma 2: Improving open language models at a practical size, 2024.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 5484–5495, 2021. doi: 10.18653/v1/2021.emnlp-main.446.

Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. Dissecting recall of factual associations in auto-regressive language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12216–12235, 2023. doi: 10.18653/v1/2023.emnlp-main.751.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, et al. The Llama 3 herd of models, 2024.

Xiaomeng Hu, Yiming Zhang, Ru Peng, Haozhe Zhang, Chenwei Wu, Gang Chen, and Junbo Zhao. 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*, 2024. doi: 10.18653/v1/2024.emnlp-main.116.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2), 2025. doi: 10.1145/3703155.

Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. 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)*, pp. 1601–1611. Association for Computational Linguistics, 2017. doi: 10.18653/v1/P17-1147. URL https://aclanthology.org/P17-1147/.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. Language models (mostly) know what they know, 2022.

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

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=VD-AYtP0dve.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466, 2019. doi: 10.1162/tacl_a_00276. URL https://aclanthology.org/Q19-1026/.

Andrey Malinin and Mark J. F. Gales. Uncertainty estimation in autoregressive structured prediction. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=jN5y-zb5Q7m.

Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. 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*, 2023.

Samuel Marks and Max Tegmark. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets, 2024. Conference on Language Modeling.

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

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 2383–2392. Association for Computational Linguistics, 2016. doi: 10.18653/v1/D16-1264. URL https://aclanthology.org/D16-1264/.

Jie Ren, Jiaming Luo, Yao Zhao, Kundan Krishna, Mohammad Saleh, Balaji Lakshminarayanan, and Peter J. Liu. Out-of-distribution detection and selective generation for conditional language models. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=kJUS5nD0vPB.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7881–7892. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.704. URL https://aclanthology.org/2020.acl-main.704/.

Ponhvoan Srey, Xiaobao Wu, Cong-Duy T. Nguyen, and Anh Tuan Luu. Learning uncertainty from sequential internal dispersion in large language models. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 40088–40106. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.1862. URL https://aclanthology.org/2026.acl-long.1862/.

Gaurang Sriramanan, Siddhant Bharti, Vinu Sankar Sadasivan, Shoumik Saha, Priyatham Kattakinda, and Soheil Feizi. LLM-check: Investigating detection of hallucinations in large language models. In *Advances in Neural Information Processing Systems*, 2024. doi: 10.52202/079017-1077.

Weihang Su, Changyue Wang, Qingyao Ai, Yiran Hu, Zhijing Wu, Yujia Zhou, and Yiqun Liu. Unsupervised real-time hallucination detection based on the internal states of large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, 2024. doi: 10.18653/v1/2024.findings-acl.854.

Zhongxiang Sun, Xiaoxue Zang, Kai Zheng, Jun Xu, Xiao Zhang, Weijie Yu, Yang Song, and Han Li. ReDeEP: Detecting hallucination in retrieval-augmented generation via mechanistic interpretability. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/7daf60e805e596c3bd1e843e72ea5560-Abstract-Conference.html.

Bohan Yang, Yijun Gong, Zhi Zhang, Ge Zhang, Wenpeng Xing, and Meng Han. Trilens: Per-layer logit-lens entropy for white-box hallucination detection, 2026. URL https://arxiv.org/abs/2606.01033.

Zhenliang Zhang, Xinyu Hu, Huixuan Zhang, Junzhe Zhang, and Xiaojun Wan. 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*, 2025. doi: 10.48550/arXiv.2507.16488.

Derui Zhu, Dingfan Chen, Qing Li, Zongxiong Chen, Lei Ma, Jens Grossklags, and Mario Fritz. PoLLMgraph: Unraveling hallucinations in large language models via state transition dynamics. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 4737–4751. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-naacl.294. URL https://aclanthology.org/2024.findings-naacl.294/.

APPENDIX

CONTENTS

A Prompt Templates and Answer Generation **15**

B Datasets, Labels, and Data Splits **15**

C Baseline Methods **15**

D HiPEx Stability Across Data Splits **16**

E Additional Path-Geometry Details **16**

E.1 Geometric Derivation 16

E.2 Joint Path-Geometry Controls 17

F Additional Layer-Wise Path Profiles **17**

G Additional Hierarchical-Design Controls **18**

G.1 Equal-Dimensional PCA Control 18

G.2 Redundant-Dimension and Alignment Controls 18

H Robustness Analyses **19**

H.1 Alternative ROUGE-L Labels 19

H.2 Human Answer–Reference Agreement Labels 19

H.3 Answer-Length Control 20

I Cross-Benchmark Transfer on Llama-3 **20**

J HiPEx Feature-Extraction Details **20**

K Probe Architecture and Training Details **21**

L Paired AUROC Comparisons **22**

M Additional Representation Controls **23**

M.1 Feature Dimensionality and Matched-Budget Comparison 23

M.2 Answer-Token Pooling Control 24

A Prompt Templates and Answer Generation

Prompt Templates. Following HaloScope (Du et al., 2024), we use the following template for closed-book QA (TriviaQA and NQ-Open):

```
Answer the question concisely. Q: [question] A:
```

For passage-grounded QA (SQuAD 1.1 and TyDiQA-GoldP), we use:

```
Answer these questions concisely based on the context:
Context: [context] Q: [question] A:
```

Each prompt is wrapped in the model's native chat template, without demonstrations or reference answers.

Answer Generation. We generate one candidate answer per question using deterministic five-beam search with a maximum of 64 new tokens. The exact prompt and answer token IDs are retained, and the frozen candidate-answer bank is shared across all detectors. LN-Entropy, Semantic Entropy, and SEP additionally share a separate bank of five stochastic generations per question, sampled at temperature 1.0, top-$p = 0.9$, and top-$k = 50$.

B Datasets, Labels, and Data Splits

Dataset Construction. For TriviaQA, we use 9,960 examples from the deduplicated `rc.nocontext` validation split. NQ-Open uses a fixed, deduplicated subset of 10,000 examples from the official training split. SQuAD 1.1 uses a fixed sample of 10,000 training examples from 7,956 source paragraphs, and TyDiQA-GoldP uses all 3,696 English training examples from 3,429 source paragraphs. Both backbones share the question sets but have model-specific generations and labels.

Answer Labeling. For an answer y_i with reference set $\mathcal{A}_i$, the primary label is

$$z_i = \mathbb{I}\left[\max_{a \in \mathcal{A}_i} \mathrm{BLEURT20}(y_i, a) \leq 0.5\right], \tag{10}$$

where $z_i = 1$ denotes hallucinated and $z_i = 0$ denotes correct. These labels reflect answer–reference mismatch rather than exhaustive human verification of factuality.

Splits. Each model–benchmark setting uses five frozen, label-stratified 72/8/20 train/validation/test splits. TriviaQA and NQ-Open are split by question; SQuAD 1.1 and TyDiQA-GoldP are split by source paragraph, with each paragraph confined to a single partition. The two backbones are evaluated as separate settings rather than as paired observations.

C Baseline Methods

All baselines share HiPEx's frozen candidate answers and test partitions. PPL, LN-Entropy, LLM-Check, and Semantic Entropy are training-free and use no correctness labels.

Perplexity (PPL). We compute answer-only teacher-forced perplexity as the exponential of the mean negative log-likelihood over valid answer tokens. Higher perplexity indicates greater hallucination risk.

Length-Normalized Entropy (LN-Entropy). We average the answer-only, length-normalized negative log-likelihood over the five shared auxiliary generations.

LLM-Check. For each attention head, we average the log diagonal attention weights over valid tokens, then sum over heads and negate the result to obtain a risk score (Sriramanan et al., 2024). We use the full prompt–answer sequence at fixed layer 23 for Llama-3 and layer 30 for Gemma-2.

Semantic Entropy. We cluster the five auxiliary answers using strict bidirectional NLI over question–answer pairs and compute the entropy of the resulting discrete semantic clusters (Kuhn et al., 2023; Farquhar et al., 2024). The NLI input excludes the passage context.

SEP. We concatenate hidden states at the final valid answer token from the last five layers and train a logistic classifier to predict high versus low Semantic Entropy (Kossen et al., 2024). The entropy threshold is fitted by sum-of-squared-error minimization on the training split; neither threshold fitting nor classifier training uses correctness labels.

HaloScope. We use the code-based HaloScope configuration (Du et al., 2024), rather than the paper's Equation 7 sensitivity variant. Our adaptation applies training-only PCA membership estimation to layer-wise hidden states from the final answer token and trains a probe on the resulting pseudo-labels. Validation correctness labels are used only to select the membership direction, layer, rank, threshold, and classifier layer.

ICR Probe. Following the released code, we compare the attention top-p distribution with the distribution induced by projecting the residual update onto attended hidden states (Zhang et al., 2025). Averaging scores over answer tokens yields one feature per layer (32 for Llama-3 and 42 for Gemma-2), on which a supervised probe is trained for each training split.

TriLens. We retain the first-answer-token readout used in the authors' released implementation of TriLens (Yang et al., 2026). At this position, we concatenate the attention, feed-forward, and residual-stream logit-lens entropies (H_a, H_m, H_x) across all layers and use its native MLP probe. An answer-token-mean variant is evaluated separately as a supplementary pooling control in Appendix M.2.

D HiPEx Stability Across Data Splits

Table 4 complements the HiPEx results in Table 1 with variability across data splits.

Table 4: HiPEx performance under BLEURT-20 labels across five predefined train/validation/test splits. Entries are mean (%) $\pm$ sample standard deviation (percentage points).

Model	Benchmark	AUROC	Hall. AUPRC	Correct AUPRC
Llama-3	TriviaQA	86.44 ± 0.84	84.31 ± 1.04	88.59 ± 0.95
	NQ-Open	91.25 ± 0.55	97.90 ± 0.18	64.54 ± 1.13
	SQuAD	89.78 ± 0.64	89.29 ± 0.67	88.49 ± 0.70
	TyDiQA	88.01 ± 1.64	97.88 ± 0.48	63.53 ± 1.60
Gemma-2	TriviaQA	82.91 ± 0.43	74.06 ± 0.59	90.52 ± 0.29
	NQ-Open	84.81 ± 0.86	93.70 ± 0.48	64.90 ± 1.90
	SQuAD	88.76 ± 0.52	83.23 ± 1.05	91.57 ± 0.70
	TyDiQA	90.39 ± 0.36	91.79 ± 0.32	87.32 ± 1.08

Sample standard deviations remain below two percentage points for all three metrics in all eight model–benchmark settings, indicating limited variability across the evaluated splits.

E Additional Path-Geometry Details

E.1 Geometric Derivation

The normalized square-root Jensen–Shannon metric d admits an isometric embedding Φ in a Hilbert space. For predictive distributions $u \to v \to w$, define the successive displacement vectors

$$\mathbf{a} = \Phi(v) - \Phi(u), \qquad \mathbf{b} = \Phi(w) - \Phi(v). \tag{11}$$

Their lengths are $A = \|\mathbf{a}\| = d(u, v)$ and $B = \|\mathbf{b}\| = d(v, w)$. For $A, B > 0$, let θ denote the angle between them. The endpoint distance is

$$C = d(u, w) = \|\mathbf{a} + \mathbf{b}\| = \sqrt{A^2 + B^2 + 2AB\cos\theta}. \tag{12}$$

The path excess therefore satisfies

$$E = A + B - C \tag{13}$$

$$= \frac{(A+B)^2 - C^2}{A+B+C} \tag{14}$$

$$= \frac{2AB(1-\cos\theta)}{A+B+\sqrt{A^2+B^2+2AB\cos\theta}}. \tag{15}$$

If either step has zero length, $E = 0$. In all cases, $0 \le E \le 2\min(A, B) \le 2$. For fixed positive step lengths, E increases with θ, from zero for aligned displacements to $2\min(A, B)$ for opposing displacements. Here θ is defined in the implicit metric embedding, not in residual space, and does not represent a causal attribution.

E.2 Joint Path-Geometry Controls

Joint-standard concatenates the answer-token-averaged path-length and endpoint-distance profiles, $[A + B, C]$, at both computational scales, allowing the probe to learn their combination. The joint representation has 124 dimensions for Llama-3 and 164 for Gemma-2. Joint-PCA compresses these features to 62 and 82 dimensions, respectively, matching the original HiPEx representation rather than the 30/40-dimensional budget used in the hierarchical PCA control in Appendix G.1.

Both controls reuse the five predefined splits and the $128/64/32$ probe architecture and training configuration in Appendix K. For each split, Joint-PCA applies $\texttt{StandardScaler}_1 \to$ PCA $\to$ $\texttt{StandardScaler}_2 \to$ the probe, using full SVD with `whiten=False`. All preprocessing transforms are fitted only on the training partition and applied unchanged to validation and test data. Joint-PCA and HiPEx therefore have identical downstream probe parameter counts: 18,881 for Llama-3 and 21,441 for Gemma-2, including biases and trainable Batch Normalization parameters.

Table 5: Joint path-geometry controls. Joint-PCA matches HiPEx's input dimensionality. Entries are mean AUROC (%) ± sample standard deviation (percentage points) across five predefined splits. The last row reports equally weighted macro-averages over the eight settings; row-wise best means are bold.

Model	Benchmark	Joint-standard	Joint-PCA	HiPEx
Llama-3	TriviaQA	86.09 ± 0.25	85.08 ± 0.42	$\mathbf{86.44 \pm 0.84}$
	NQ-Open	90.10 ± 0.23	90.01 ± 0.16	$\mathbf{91.25 \pm 0.55}$
	SQuAD	88.82 ± 0.36	88.59 ± 0.70	$\mathbf{89.78 \pm 0.64}$
	TyDiQA	83.63 ± 2.90	83.60 ± 3.88	$\mathbf{88.01 \pm 1.64}$
Gemma-2	TriviaQA	$\mathbf{83.02 \pm 0.40}$	82.23 ± 0.24	82.91 ± 0.43
	NQ-Open	84.14 ± 0.44	83.42 ± 0.48	$\mathbf{84.81 \pm 0.86}$
	SQuAD	86.71 ± 0.44	85.90 ± 0.31	$\mathbf{88.76 \pm 0.52}$
	TyDiQA	89.50 ± 0.33	88.77 ± 0.39	$\mathbf{90.39 \pm 0.36}$
Macro-average		86.50	85.95	**87.79**

Table 5 shows that HiPEx achieves higher mean AUROC than Joint-standard in seven of eight settings and than Joint-PCA in all eight, with macro-average gains of 1.29 and 1.84 percentage points, respectively. Under the shared answer-token mean aggregation, HiPEx is the coordinate-wise difference of the two component profiles. These results support explicit endpoint subtraction as an effective geometric inductive bias under the evaluated training conditions.

F Additional Layer-Wise Path Profiles

Figure 5 extends the Llama-3 analysis in Figure 3 to Gemma-2 on TriviaQA.

Gemma-2 exhibits pronounced negative inter-block gaps in early and near-final blocks, with smaller sign-varying differences in between. The intra-block profile follows a distinct layer-wise pattern. Together with Figure 3, these depth- and scale-dependent class differences motivate retaining layer coordinates at both scales rather than collapsing path excess into a single global magnitude.

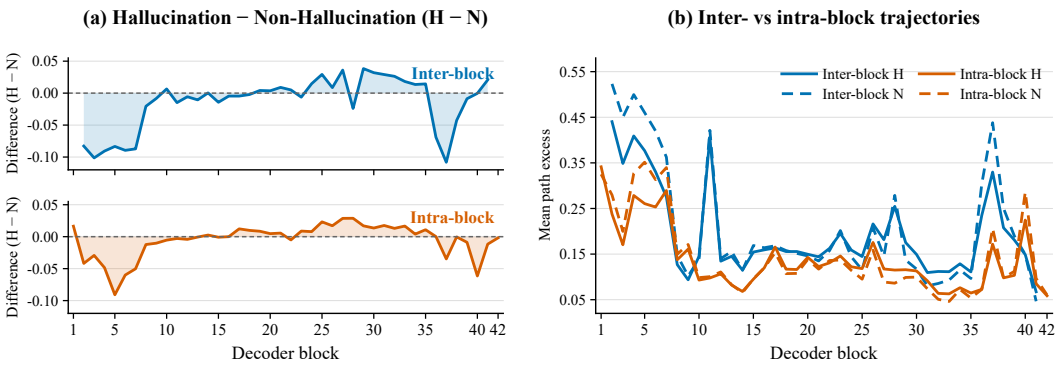

Figure 5: Layer-wise path profiles for Gemma-2-9B-it on TriviaQA after within-answer token averaging. Panel (a) shows the signed class gap $\Delta_\ell = D_\ell^{\mathrm{H}} - D_\ell^{\mathrm{N}}$ at each computational scale; shading marks the area between each curve and zero, not confidence intervals. Panel (b) shows the corresponding class-conditional mean profiles. H/N denote hallucinated/non-hallucinated answers.

G Additional Hierarchical-Design Controls

We complement the parameter-matched comparisons in Section 4.4.2 with controls for input dimensionality, redundant feature expansion, and answer-level alignment.

G.1 Equal-Dimensional PCA Control

For each backbone, we fit PCA separately to the Inter, Intra, and Full HiPEx features using only the training split. All three representations are projected to 30 dimensions for Llama-3 and 40 for Gemma-2 and evaluated with the same $128/64/32$ probe.

Table 6: Equal-dimensional PCA control. Entries are mean AUROC (%) $\pm$ sample standard deviation (percentage points) across five predefined splits.

Model	Benchmark	PCA Inter	PCA Intra	PCA Full
Llama-3	TriviaQA	81.37 ± 0.56	85.07 ± 0.69	85.43 ± 0.99
	NQ-Open	87.27 ± 0.43	90.44 ± 0.68	90.29 ± 0.70
	SQuAD	86.55 ± 0.60	86.98 ± 0.33	88.17 ± 0.26
	TyDiQA	79.53 ± 2.49	80.49 ± 4.05	83.68 ± 2.40
Gemma-2	TriviaQA	80.73 ± 1.44	78.65 ± 0.95	81.50 ± 1.06
	NQ-Open	82.54 ± 1.52	82.46 ± 1.30	83.79 ± 1.37
	SQuAD	86.28 ± 0.90	84.33 ± 1.64	86.13 ± 1.36
	TyDiQA	87.34 ± 0.89	86.86 ± 0.78	88.42 ± 0.25

At equal input dimensionality, PCA Full improves macro-average AUROC over PCA Inter and PCA Intra by 1.98 and 1.52 percentage points, respectively (Table 6). It also exceeds the stronger single-scale variant in six of eight settings, with a 0.15-point shortfall in each of the remaining two.

G.2 Redundant-Dimension and Alignment Controls

Both controls match Full HiPEx's input dimensionality and probe architecture. Repeated Inter tiles and truncates the Inter vector to this dimensionality without introducing new information. Shuffled Intra randomly pairs Inter vectors with complete Intra vectors from different answers, preserving the Intra vectors and their marginal distribution while breaking answer-level alignment.

Full HiPEx achieves higher mean AUROC than both controls in all eight settings (Table 7). Repeated Inter remains close to Inter-only (Table 2), whereas Shuffled Intra incurs a larger average loss relative to Full. These results support the value of answer-aligned information from both scales beyond redundant feature expansion.

Table 7: Redundant-dimension and answer-alignment controls. Entries are mean AUROC (%) across five predefined splits. Repeated and Shuffled denote Repeated Inter and Shuffled Intra. The Δ columns report Full minus the corresponding control in percentage points.

Model	Benchmark	Full	Repeated	Shuffled	ΔRep.	ΔShuf.
Llama-3	TriviaQA	86.44	81.57	79.94	4.87	6.50
	NQ-Open	91.25	87.16	85.64	4.09	5.61
	SQuAD	89.78	86.57	85.38	3.21	4.40
	TyDiQA	88.01	80.89	77.65	7.12	10.36
Gemma-2	TriviaQA	82.91	81.31	79.72	1.60	3.19
	NQ-Open	84.81	83.28	82.08	1.53	2.73
	SQuAD	88.76	86.30	85.56	2.46	3.20
	TyDiQA	90.39	88.69	87.30	1.70	3.09
Mean advantage of Full HiPEx					3.32	4.89

H Robustness Analyses

H.1 Alternative ROUGE-L Labels

For the alternative-label evaluation on Llama-3 TriviaQA and TyDiQA, we replace BLEURT-20 labels with

$$z_i^{\text{ROUGE}} = \mathbb{I}\left[\max_{a \in \mathcal{A}_i} \text{ROUGE-L}_{F_1}(y_i, a) \leq 0.5\right]. \tag{16}$$

The evaluation protocol and nine-method comparison are reported in Section 4.6 and Table 3.

H.2 Human Answer–Reference Agreement Labels

Sample and Annotation. We manually annotate fixed half-sized subsets of frozen answers from Llama-3 and Gemma-2: 4,980 TriviaQA and 1,848 TyDiQA answers per backbone (13,656 total). Each record is assessed by three annotators using only the question, generated answer, and references, with disagreements resolved by majority vote. Label 0 (correct) denotes agreement in core meaning, allowing reasonable precision differences and text-supported aliases. Label 1 (incorrect) covers mismatched or contradictory core answers, names without textual evidence of alias equivalence, refusals, and non-answers. Cases where corrupted references or missing information prevent comparison are marked indeterminate. These labels measure answer–reference agreement rather than independent factuality verification.

Table 8: Human annotation outcomes. Percentages use all sampled records in each setting as the denominator; retained counts exclude indeterminate records.

Model	Benchmark	Correct (0)	Incorrect (1)	Indeterminate	Retained
Llama-3	TriviaQA	3,680 (73.90%)	1,298 (26.06%)	2 (0.04%)	4,978
	TyDiQA	1,657 (89.66%)	179 (9.69%)	12 (0.65%)	1,836
Gemma-2	TriviaQA	3,751 (75.32%)	1,228 (24.66%)	1 (0.02%)	4,979
	TyDiQA	1,617 (87.50%)	211 (11.42%)	20 (1.08%)	1,828
Total		10,705 (78.39%)	2,916 (21.35%)	35 (0.26%)	13,621

Evaluation Protocol. For each setting and seed 0–4, we intersect the sampled answers with the original train/validation/test partitions, preserving question-disjoint TriviaQA and paragraph-disjoint TyDiQA assignments without re-stratifying on human labels. The same 35 indeterminate records (0.26%) are excluded across all methods and seeds without replacement, leaving 13,621 answers (Table 8). We compare refitted HaloScope, ICR Probe, TriLens (first-answer-token readout), and HiPEx on identical retained partitions and labels.

Table 9: Detection under human answer–reference agreement labels. Entries are mean AUROC (%) ± sample standard deviation (percentage points) over seeds 0–4. Macro-averages equally weight the four settings within each seed. Column-wise best means are bold.

Method	Llama-3		Gemma-2		Macro-average
	TriviaQA	TyDiQA	TriviaQA	TyDiQA	
HaloScope	73.10 ± 3.29	$\mathbf{78.12 \pm 2.86}$	64.69 ± 13.64	63.28 ± 11.50	69.80 ± 2.03
ICR Probe	68.91 ± 1.51	71.32 ± 2.33	70.21 ± 2.37	73.31 ± 5.15	70.94 ± 0.80
TriLens	74.08 ± 1.54	73.06 ± 5.45	67.21 ± 1.37	71.22 ± 5.29	71.39 ± 1.46
HiPEx	$\mathbf{77.10 \pm 1.24}$	72.43 ± 6.43	$\mathbf{76.16 \pm 1.37}$	$\mathbf{76.69 \pm 4.89}$	$\mathbf{75.59 \pm 2.25}$

Results. HiPEx achieves the highest mean AUROC in three of four settings and the highest macro-average (75.59%), exceeding TriLens, the strongest baseline by macro-average, by 4.20 percentage points (Table 9). HaloScope leads on Llama-3 TyDiQA.

H.3 Answer-Length Control

To assess the predictive value of answer length, we use the number of generated answer tokens as the sole input, keeping HiPEx's data splits and fitting protocol unchanged.

Table 10: Answer-length control under BLEURT-20 labels across five predefined splits. Length-only reports mean AUROC (%) ± sample standard deviation (percentage points); HiPEx reports the corresponding means from Table 1.

Model	Benchmark	Length only	HiPEx
Llama-3	TriviaQA	69.80 ± 0.18	86.44
	NQ-Open	77.80 ± 0.46	91.25
	SQuAD	75.80 ± 0.56	89.78
	TyDiQA	52.04 ± 4.05	88.01
Gemma-2	TriviaQA	62.11 ± 1.02	82.91
	NQ-Open	67.33 ± 1.28	84.81
	SQuAD	69.09 ± 1.80	88.76
	TyDiQA	71.86 ± 0.58	90.39

Length-only achieves a macro-average AUROC of 68.23%, indicating a length–label association in the frozen answers. HiPEx exceeds this control in all eight settings by 13.45–35.97 percentage points, with a macro-average gain of 19.56 points (Table 10).

I Cross-Benchmark Transfer on Llama-3

Figure 6 complements the Gemma-2 results in Figure 4 using the same source-only protocol described in Section 4.5.

Across the 12 cross-benchmark directions, HiPEx exceeds ICR Probe by 5.67 percentage points in mean off-diagonal AUROC and trails HaloScope by 1.12 points. The comparison with Figure 4 highlights a backbone-dependent ranking relative to HaloScope.

J HiPEx Feature-Extraction Details

Teacher-Forced Extraction. We concatenate the stored prompt and answer token IDs and run one teacher-forced forward pass through the frozen model, without regenerating or retokenizing the answer. Features are retained only at valid answer-token positions, excluding prompt and padding positions.

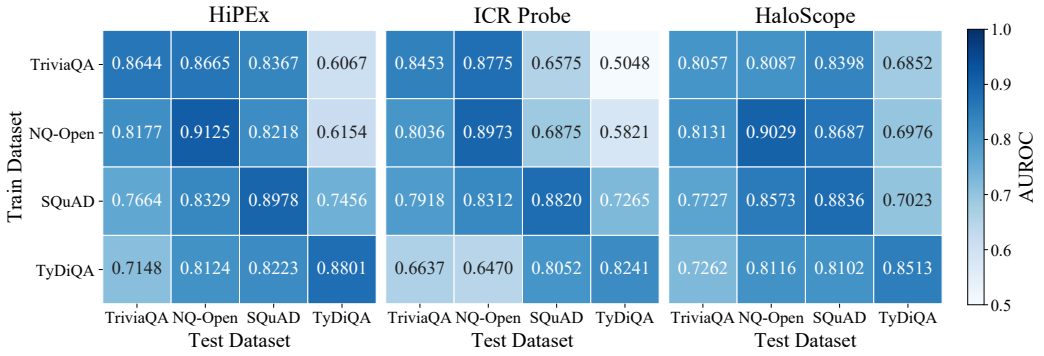

Figure 6: Source-only cross-benchmark transfer on Llama-3. Rows and columns denote source and target benchmarks, respectively. Entries are AUROC on the 0–1 scale; diagonal cells show within-benchmark performance, while off-diagonal cells measure transfer without target-benchmark adaptation.

Model-Specific Logit-Lens Projection. We apply each model's native final normalization and output head at temperature $\tau = 1$. Llama-3 leaves the resulting logits unchanged,

$$g_{\text{Llama-3}}(\mathbf{a}) = \mathbf{a}, \tag{17}$$

whereas Gemma-2 applies its native soft-capping,

$$g_{\text{Gemma-2}}(\mathbf{a}) = 30 \tanh(\mathbf{a}/30). \tag{18}$$

Softmax maps the transformed logits to vocabulary distributions.

Residual-State Capture. Inter-block features use decoder-block output residuals, excluding the embedding state from the decoder-layer sequence. For the intra-block midpoint, forward-pre hooks read `inputs[0]` at `post_attention_layernorm` in Llama-3 and `pre_feedforward_layernorm` in Gemma-2. Both locations capture the cumulative residual after the attention increment is added and before FFN input normalization. In Gemma-2, the attention and FFN increments include their respective native branch-output normalization. Together with block inputs and outputs, these midpoints give the executed pre-attention → post-attention → post-FFN path.

Dimensions, Aggregation, and Precision. The Inter/Intra/Full feature dimensions are 30/32/62 for Llama-3 and 40/42/82 for Gemma-2. Unless an ablation specifies otherwise, each coordinate is mean-pooled over all valid answer tokens. The decoder runs in BF16 with SDPA attention and `use_cache=false`, while final normalization, the output head, logit transformation, softmax, and JSD are computed in FP32. We apply $\mathrm{ReLU}(A+B-C)$ solely to remove negligible negative values caused by floating-point roundoff.

K Probe Architecture and Training Details

The HiPEx probe is an MLP with architecture

$$d_{\text{in}} \to h_1 \to h_2 \to h_3 \to d_{\text{out}},$$

where d_{in} is the input dimensionality, h_1, h_2, h_3 are the hidden-layer widths, and d_{out} is the output dimensionality. Each hidden linear layer is followed by Batch Normalization, LeakyReLU with negative slope α, and dropout with probability ρ.

For the standard probe, we set $(h_1, h_2, h_3) = (128, 64, 32)$, $d_{\text{out}} = 1$, $\alpha = 0.01$, and $\rho = 0.3$. The output is a single hallucination-risk logit.

Table 11 summarizes the training configuration. The `StandardScaler` and probe are fitted independently for each backbone, benchmark, and split seed. Validation data are used only for learning-rate scheduling and checkpoint selection.

Table 11: HiPEx probe-training configuration.

Component	Configuration
Feature preprocessing	Training-only StandardScaler
Objective	Unweighted BCEWithLogitsLoss
Positive class	Hallucinated answer
Optimizer	Adam
Initial learning rate	5×10^{-4}
Batch size	32
Maximum epochs	50
Learning-rate schedule	ReduceLROnPlateau
Scheduler patience/factor	5 / 0.5
Model selection	Minimum validation loss
Test evaluation	Once, after checkpoint selection
Reported score	Sigmoid of the restored probe logit

Parameter-Matched Probes. For the control in Table 2, we keep Full HiPEx unchanged and widen only the first hidden layer of each single-scale probe, retaining the original input features and all later layer widths. For the architecture above, the trainable parameter count, including linear biases and Batch Normalization scale and shift parameters, is

$$N_{\text{probe}} = d_{\text{in}}h_1 + h_1h_2 + h_2h_3 + h_3d_{\text{out}} + 3(h_1 + h_2 + h_3) + d_{\text{out}}. \tag{19}$$

For parameter matching, h_2, h_3, and d_{out} retain their standard values, while h_1 is increased to approximately match Full HiPEx's trainable parameter count. Table 12 lists the resulting configurations.

Table 12: Probe configurations and trainable parameter counts for the parameter-matched hierarchical control in Table 2. Small parameter-count differences arise from integer hidden-layer widths.

Model	Variant	Input d_{in}	First width h_1	Parameters
Llama-3	Inter	30	170	18,859
	Intra	32	167	18,902
	Full	62	128	18,881
Gemma-2	Inter	40	178	21,415
	Intra	42	175	21,444
	Full	82	128	21,441

L Paired AUROC Comparisons

Relation to the Main Results. Tables 13–14 compare HiPEx with ICR Probe and HaloScope across the eight model–benchmark settings. The displayed baseline means agree with Table 1, while HiPEx's paired-analysis means differ from Table 1 by at most 0.03 percentage points. Intervals correspond to the paired estimates reported here. Reported differences may differ by 0.01 percentage points from subtracting the displayed rounded means.

Paired Cluster-Bootstrap Protocol. We fix the fitted models and predictions for seeds 0–4 and align test IDs, labels, and evaluation masks across methods within each split. For each model–benchmark setting, we sample clusters with replacement from the union of the five test sets: question IDs for TriviaQA and NQ-Open, and paragraph IDs for SQuAD and TyDiQA. Each sampled cluster's multiplicity defines a shared weight across methods and every split containing that cluster. For each resample, we recompute the weighted AUROC within each split and average the five HiPEx-minus-baseline differences to obtain Δ_b^*. We collect 10,000 valid resamples and use their 2.5th and 97.5th percentiles as the pointwise 95% interval. With $\widehat{\Delta}$ denoting the observed mean

paired difference, the unadjusted two-sided, null-centered bootstrap p-value is

$$p = \frac{1 + \sum_{b=1}^{10000} \mathbf{1}\left\{\left|\Delta_b^* - \widehat{\Delta}\right| \geq \left|\widehat{\Delta}\right|\right\}}{10001}. \tag{20}$$

We apply Holm correction jointly across all 16 comparisons in Tables 13–14: two backbones, four benchmarks, and two baselines.

Inferential Scope. Shared cluster weights account for overlap between test splits without treating the five splits as independent experimental replicates. Inference quantifies test-sample uncertainty conditional on the five fitted models, rather than variability from retraining or newly drawn training splits. The intervals are pointwise; multiplicity adjustment applies to the p-values within the specified 16-comparison family.

Table 13: Paired AUROC comparison with ICR Probe. Mean AUROCs are percentages; Δ (HiPEx minus baseline) and CI limits are percentage points. CIs are pointwise 95% paired cluster-bootstrap intervals. p is unadjusted; p_{H} is Holm-adjusted jointly across all 16 comparisons in Tables 13–14. Pos. counts splits with a positive AUROC difference. L/G denote Llama-3/Gemma-2.

Setting	HiPEx	Baseline	Δ	95% CI	p	p_{H}	Pos.
L-TriviaQA	86.43	84.53	+1.90	[+0.89, +2.88]	0.0001	0.0016	5/5
L-NQ-Open	91.25	89.73	+1.53	[+0.85, +2.23]	0.0002	0.0016	5/5
L-SQuAD	89.76	88.20	+1.57	[+0.66, +2.48]	0.0006	0.0030	5/5
L-TyDiQA	88.04	82.41	+5.64	[+2.75, +8.60]	0.0001	0.0016	5/5
G-TriviaQA	82.89	77.10	+5.79	[+4.46, +7.07]	0.0001	0.0016	5/5
G-NQ-Open	84.82	80.83	+3.99	[+2.98, +5.00]	0.0001	0.0016	5/5
G-SQuAD	88.79	85.61	+3.18	[+1.87, +4.49]	0.0001	0.0016	5/5
G-TyDiQA	90.40	88.10	+2.30	[+0.89, +3.67]	0.0013	0.0052	5/5

Table 14: Paired AUROC comparison with HaloScope. Mean AUROCs are percentages; Δ and CI limits are percentage points. Statistical conventions and abbreviations follow Table 13.

Setting	HiPEx	Baseline	Δ	95% CI	p	p_{H}	Pos.
L-TriviaQA	86.43	80.57	+5.86	[+4.79, +6.91]	0.0001	0.0016	5/5
L-NQ-Open	91.25	90.29	+0.97	[+0.23, +1.68]	0.0091	0.0182	5/5
L-SQuAD	89.76	88.36	+1.40	[+0.53, +2.30]	0.0022	0.0066	5/5
L-TyDiQA	88.04	85.13	+2.91	[+0.50, +5.28]	0.0194	0.0194	5/5
G-TriviaQA	82.89	77.43	+5.47	[+4.35, +6.58]	0.0001	0.0016	5/5
G-NQ-Open	84.82	61.44	+23.38	[+22.46, +24.25]	0.0001	0.0016	5/5
G-SQuAD	88.79	84.92	+3.87	[+2.55, +5.18]	0.0001	0.0016	5/5
G-TyDiQA	90.40	79.81	+10.58	[+9.32, +11.83]	0.0001	0.0016	5/5

Observed Pattern. HiPEx achieves higher AUROC in all five splits for every comparison. All 16 pointwise 95% intervals lie above zero, and all Holm-adjusted p-values are below 0.05 (Tables 13–14).

M Additional Representation Controls

M.1 Feature Dimensionality and Matched-Budget Comparison

Table 15 summarizes representation dimensionality. For a matched-budget comparison with ICR Probe, we compress Full HiPEx to 32 dimensions for Llama-3 and 42 for Gemma-2 using PCA

fitted only on the training split. The comparison uses fixed split seeds 0–4 and matches ICR Probe's input dimensionality and downstream probe parameter count.

Table 15: Feature dimensionality of the evaluated representations.

Representation	Llama-3	Gemma-2
HiPEx	62	82
ICR Probe	32	42
HaloScope	4096	3584
TriLens	96	126
HiPEx-PCA	32	42

HiPEx-PCA achieves higher mean AUROC than ICR Probe in seven of eight settings, with a macro-average gain of 1.35 percentage points (Table 16). Relative to Full HiPEx, compression reduces input dimensionality by approximately 48% at a cost of 1.88 percentage points in macro-average AUROC.

Table 16: Matched-budget comparison with ICR Probe. Entries are mean AUROC (%) over seeds 0–4; Δ is HiPEx-PCA minus ICR Probe in percentage points. Bold marks the better of these two matched-budget configurations in each setting. Full HiPEx is the uncompressed reference from Table 1. The reported HiPEx-PCA macro-average is 85.91%; the rounded entries average to 85.915%.

Model	Benchmark	ICR Probe	HiPEx-PCA	Full HiPEx	Δ
Llama-3	TriviaQA	84.53	**85.39**	86.44	+0.86
	NQ-Open	89.73	**90.39**	91.25	+0.66
	SQuAD	**88.20**	88.02	89.78	−0.18
	TyDiQA	82.41	**83.45**	88.01	+1.04
Gemma-2	TriviaQA	77.10	**81.93**	82.91	+4.83
	NQ-Open	80.83	**83.83**	84.81	+3.00
	SQuAD	85.61	**86.10**	88.76	+0.49
	TyDiQA	88.10	**88.21**	90.39	+0.11
Macro-average		84.56	85.91	87.79	+1.35

M.2 Answer-Token Pooling Control

Table 1 uses the first-answer-token readout from the authors' released TriLens implementation. To test whether HiPEx's advantage persists under matched answer-token aggregation, we additionally evaluate a mean-pooled TriLens variant as a supplementary control.

Mean pooling improves TriLens in all eight settings, with a macro-average AUROC gain of 6.26 percentage points over its first-token configuration. HiPEx achieves higher mean AUROC in all eight settings, with a macro-average advantage of 2.44 percentage points (Table 17).

Table 17: Answer-token pooling control for TriLens. First-token is the primary baseline used in Table 1; mean-token is a supplementary variant. HiPEx results are reproduced from Table 1. Entries are mean AUROC (%); Δ is HiPEx minus mean-token TriLens in percentage points. Row-wise best means are bold.

		TriLens			
Model	Benchmark	First-token	Mean-token	HiPEx	Δ
Llama-3	TriviaQA	80.91	85.44	**86.44**	+1.00
	NQ-Open	82.62	89.91	**91.25**	+1.34
	SQuAD	78.90	87.73	**89.78**	+2.05
	TyDiQA	81.93	82.93	**88.01**	+5.08
Gemma-2	TriviaQA	74.13	80.24	**82.91**	+2.67
	NQ-Open	76.48	82.68	**84.81**	+2.13
	SQuAD	75.09	84.93	**88.76**	+3.83
	TyDiQA	82.66	88.94	**90.39**	+1.45
Macro-average		79.09	85.35	**87.79**	+2.44