HALLUTRACER: PRE-DECODING TRUTHFULNESS PREDICTION VIA DEPTH-AVERAGED PROBE-LOGIT

Anonymous authors
Paper under double-blind review

ABSTRACT

Internal-state probes enable truthfulness prediction before a large language model generates an answer. When detectors change both the layers they read and the rules used to combine them, the source of improved prediction becomes difficult to identify. We separate these choices and find that retaining more layers improves prediction even under fixed equal weighting. An exact Fisher-ratio decomposition explains why the additional benefit of linear reweighting is limited on these probe scores: information in layer-wise differences largely overlaps with that captured by the depth mean. Estimating additional weights can then offset this small benefit when the data used to fit them are limited. These findings motivate our proposed method HalluTracer, which averages layer-wise probe logits to predict truthfulness before decoding. Across six models and four benchmarks, HalluTracer achieves the highest AUROC in 23 of 24 model–benchmark pairs among the compared methods. The results support using evidence from across the network without requiring a correspondingly more flexible aggregation rule, clarifying the distinct roles of layer selection and weighting in pre-decoding detection.

1 INTRODUCTION

Large language models can produce fluent text that contains factual errors (Ji et al., 2023; Zhang et al., 2025; Huang et al., 2025). Such errors need not be caught only after the text is produced, because truth-relevant evidence is already present before generation: linear probes can recover truthfulness directions from hidden activations (Marks and Tegmark, 2023), and prompted models can report calibrated beliefs about their own answers (Kadavath et al., 2022). Pre-decoding detectors build on this observation and read the evidence before any token is emitted (Ji et al., 2024; Alnuhait et al., 2025). This evidence, however, is spread over many internal readouts, so a detector must decide which readouts to retain and how to combine them into one scalar score.

Existing white-box studies use selected layers, attention heads, or hidden-state representations (Burns et al., 2023; Li et al., 2023; Han et al., 2025), while other work combines information across depth through layer contrasts and learned modules (Srey et al., 2025; Chuang et al., 2024; Chen et al., 2024; Mir, 2025; Beigi et al., 2024; Bar-Shalom et al., 2025). Recent uncertainty and deception probes also average layer-wise scores across depth (Xiao et al., 2026; Taufeeque et al., 2026). These approaches motivate understanding how much learning aggregation weights can improve on a simple average. For a fixed set of layer-wise scores, Fisher's linear discriminant provides a reference by maximising the pooled signal-to-noise ratio (Fisher, 1936; Su and Liu, 1993; Fan et al., 2012). Retaining more layers increases the number of cross-layer covariances that Fisher weighting must estimate from the same labelled examples. With limited training data, estimation error in this covariance matrix can undermine the additional benefit of learned weighting (Bickel and Levina, 2004; Ledoit and Wolf, 2004). The remaining question is whether the additional gain over equal weighting outweighs this estimation cost in pre-decoding truthfulness detection.

We find that averaging scores across all layers improves truthfulness prediction over a single layer selected using training data, consistent with recent evidence on multi-layer probing (Nordby et al., 2026). However, learning aggregation weights provides limited additional benefit. Our results show that the depth mean retains at least 91.7 percent of the held-out pooled signal-to-noise ratio (SNR) of the fitted Fisher score, with a difference of at most 0.72 points in the area under the receiver operating characteristic curve (AUROC). To understand this limited gain, we derive an exact Fisher-ratio

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.

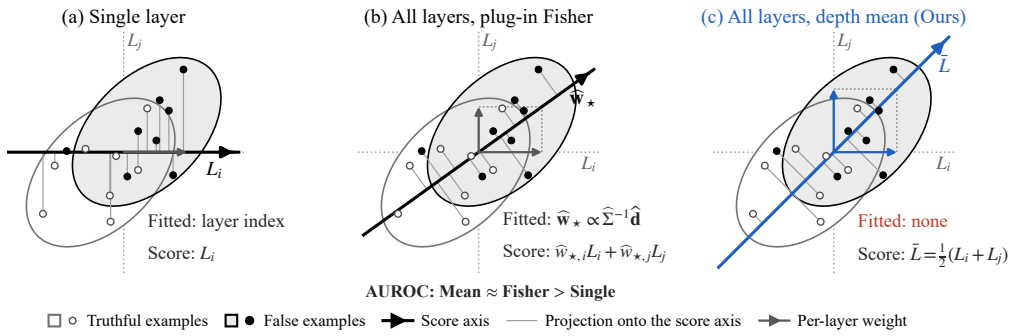


Figure 1: Three readouts of the same examples. A single layer retains one coordinate. The plug-in Fisher score estimates a direction from the mean gap and equal-class pooled covariance. The depth mean uses fixed equal weights. All three share the same training-fitted probes.

decomposition in terms of class separation and cross-layer covariance. On the evaluated probe scores, it indicates that truthfulness information in layer-wise differences largely overlaps with that captured by the mean. With less training data, the unshrunk Fisher score deteriorates, while covariance shrinkage keeps its AUROC close to the depth mean. These findings motivate using equal weighting, which combines evidence across layers without estimating their covariance.

Guided by this analysis, we introduce HalluTracer, as shown in Figure 1(c), which trains layer-wise linear probes and averages their logits to predict truthfulness before decoding. Across six models and four benchmarks, HalluTracer achieves the highest AUROC in 23 of 24 model–benchmark pairs and the highest AUPRC in 22 among the compared methods. Our contributions are as follows:

1. **Aggregation formulation.** We distinguish the benefit of retaining more layers from the gain available through linear weighting and the effects of estimating those weights.
2. **Uniform aggregation.** We derive an exact Fisher ratio decomposition relative to the depth mean and relate the resulting pooled-SNR gain to AUROC under stated score-difference conditions.
3. **Experimental validation.** Across 18 conditions, we evaluate the gains from combining layers and learning their weights, and examine how the fitted scores change with training size. The depth mean retains at least 91.7 percent of the held-out pooled SNR of the fitted Fisher score.

2 Related Work

Hallucination Detection and Internal Probing. Methods operating on generated responses assess factuality through sampled consistency (Manakul et al., 2023), semantic or token-level uncertainty (Farquhar et al., 2024; Fadeeva et al., 2024), and response representations (Chen et al., 2024; Zeng et al., 2026). Probing studies recover truthfulness signals from hidden activations and attention heads (Azaria and Mitchell, 2023; Burns et al., 2023; Li et al., 2023; Marks and Tegmark, 2023; Han et al., 2025), while IRIS (Srey et al., 2025) learns from self-verification representations using uncertainty-based pseudo labels. Prompt representations further support hallucination-risk prediction before answer generation (Ji et al., 2024; Alnuhait et al., 2025; Miftakhova and Zaytsev, 2026; Mrykhin and Malykh, 2026). We build on this pre-decoding setting to examine how truthfulness evidence from different layers can be combined.

Cross-Layer Evidence and Aggregation. Intermediate prediction methods expose information across Transformer depth (nostalgebraist, 2020; Belrose et al., 2025). DoLa (Chuang et al., 2024) and SLED (Zhang et al., 2024) use contrasts between layers to improve factuality during decoding. For detection, cross-layer information is represented through logit statistics, semantic trajectories, or information-flow measures (Sun, 2025; Mir, 2025; Kim et al., 2025). Aggregation ranges from fixed pooling to learned combinations. EAGLE (Xiao et al., 2026) averages vocabulary logits across layers during self-evaluation, while deception probes also average classifier probabilities over selected layers (Taufeeque et al., 2026). Learned approaches combine features from multiple layers (Beigi et al., 2024; Srey et al., 2026) or learn combinations of selected layer-wise probes for deception detection (Nordby et al., 2026). Building on these approaches, we hold the layer-wise probe scores fixed to distinguish the gain from retaining more layers from the additional benefit of learning aggregation weights.

Linear Aggregation and Weight Estimation. Fisher's criterion provides a classical reference for linear aggregation by maximising pooled SNR (Fisher, 1936). Research on diagnostic markers studies linear combinations that maximise AUROC and the evaluation of fitted combinations (Su and Liu, 1993; Huang et al., 2011). Classifier-combination theory relates the ideal gain of weighted over simple averaging to individual classifier performance and correlations between their outputs (Fumera and Roli, 2005). High-dimensional discriminant analysis characterises how covariance and estimation error affect fitted discriminants (Bickel and Levina, 2004; Fan et al., 2012; Dobriban and Wager, 2018), motivating covariance regularisation (Ledoit and Wolf, 2004). Equal weighting can remain competitive when estimation error offsets the potential gains from optimising weights (Dawes, 1979; Smith and Wallis, 2009; DeMiguel et al., 2009). We study this trade-off in pre-decoding truthfulness detection, using an exact Fisher-ratio decomposition to characterise the potential gain over the depth mean and held-out evaluations to measure how much remains after fitting weights on limited data.

3 METHOD

3.1 OVERVIEW

Layer-wise probing supplies a score vector. A detector turns that vector into one scalar through two decisions: which coordinates to retain and how to combine them. Our method HalluTracer takes the full support and deploys the depth mean (Figure 1). § 3.4 instantiates the coordinates as answer-onset probe logits, so the support mean is the deployed depth mean. Proofs appear in Appendix D.

3.2 SCORE VECTOR, SUPPORT, AND SUPPORT MEAN

Let a score-extraction procedure map an example to an indexed vector of scalar scores $\mathbf{Z} = [Z_0, \ldots, Z_{m-1}]^\top \in \mathbb{R}^m$. Here $m \geq 1$ is the number of score coordinates and $\mathcal{K}_m := \{0, \ldots, m-1\}$ is their index set. At evaluation time, the extractor, coordinate standardisation, score orientation, and support are fixed using the training data. We use the label convention $Y = 0$ for truthful examples and $Y = 1$ for false examples.

The analysis concerns linear scores $\mathbf{Z} \mapsto \mathbf{w}^\top \mathbf{Z}$ with $\mathbf{w} \in \mathbb{R}^m \setminus \{\mathbf{0}\}$. The extraction procedure fixes the meaning, ordering, and standardisation of the coordinates. Here the ordering is Transformer depth. We interpret uniform weights relative to this fixed coordinate system.

Fix a nonempty support $I \subseteq \mathcal{K}_m$ and let $\mathbf{1}_I$ be its indicator vector. The unit uniform direction and support mean are

$$\mathbf{u}_I := |I|^{-1/2}\mathbf{1}_I, \qquad \bar{Z}_I := |I|^{-1}\sum_{j\in I} Z_j = |I|^{-1/2}\mathbf{u}_I^\top \mathbf{Z}. \tag{1}$$

All scale-invariant comparisons below therefore give the same value for $\mathbf{u}_I$ and $\bar{Z}_I$.

The support and the aggregation rule are separate decisions. A singleton support retains one coordinate and discards the rest. A larger support contains more coordinates, after which uniform or fitted weights determine how that evidence enters the scalar score. The first choice changes the available coordinates, and the second changes only their linear combination. The theoretical and empirical comparisons keep these two choices separate.

We distinguish three effects on this score vector. The support gain is the held-out gain from enlarging the support under the uniform rule. The potential gain is the pooled-SNR gain of the Fisher-optimal direction over the full-support mean, quantified by the Fisher ratio before aggregation weights are fitted. Estimation cost is the performance loss incurred by estimating that direction from finite training data, assessed on held-out folds. All claims concern linear aggregation of this fixed score vector.

3.3 FISHER RATIO AND AUROC

We fix $I = \mathcal{K}_m$ to compare aggregation rules on the same full score vector. Write $\boldsymbol{\mu}_c := \mathbb{E}[\mathbf{Z} \mid Y = c]$ and $\Sigma_c := \mathrm{Cov}(\mathbf{Z} \mid Y = c)$. The mean gap and the *equal-class* pooled covariance are

$$\mathbf{d} := \boldsymbol{\mu}_1 - \boldsymbol{\mu}_0, \qquad \Sigma := \frac{\Sigma_0 + \Sigma_1}{2}. \tag{2}$$

The equal class weights are part of the definition: they give $\Sigma_0 + \Sigma_1 = 2\Sigma$, which is the covariance identity used by the pairwise AUROC result below.

For every nonzero $\mathbf{w} \in \mathbb{R}^m$, define the Fisher criterion

$$J(\mathbf{w}) := \frac{(\mathbf{w}^\top \mathbf{d})^2}{\mathbf{w}^\top \Sigma \mathbf{w}}. \tag{3}$$

Thus $J(\mathbf{w})$ is the squared pooled signal-to-noise ratio of the linear score, while $\sqrt{J(\mathbf{w})}$ is its pooled SNR. When $\Sigma \succ 0$, the Fisher-optimal direction is $\mathbf{w}_\star := \Sigma^{-1}\mathbf{d}$. The criterion J and the Gaussian relation between SNR and AUROC are classical. Their combination measures the gain available on the full support deployed below.

To compare $\mathbf{w}_\star$ with the uniform direction, extend $\mathbf{u}_I$ to an orthogonal matrix $U_I = [\mathbf{u}_I, Q_I]$, where $Q_I \in \mathbb{R}^{m\times(m-1)}$, $Q_I^\top Q_I = \mathbf{I}_{m-1}$, and $Q_I^\top \mathbf{u}_I = \mathbf{0}$. Define the coordinates along the uniform direction and its orthogonal complement

$$U_I^\top \mathbf{d} = \begin{pmatrix} s \\ \mathbf{z} \end{pmatrix}, \qquad U_I^\top \Sigma U_I = \begin{pmatrix} a & \mathbf{b}^\top \\ \mathbf{b} & C \end{pmatrix}. \tag{4}$$

Here $s = \mathbf{u}_I^\top \mathbf{d}$ is the mean gap along the uniform direction, $\mathbf{z} = Q_I^\top \mathbf{d}$ is its component in the orthogonal complement, $a = \mathbf{u}_I^\top \Sigma \mathbf{u}_I$ is the variance along the uniform direction, $\mathbf{b} = Q_I^\top \Sigma \mathbf{u}_I$ is the corresponding cross-covariance, and $C = Q_I^\top \Sigma Q_I$ is the covariance on the orthogonal complement. When $\Sigma \succ 0$, this change of coordinates gives $a > 0$, $C \succ 0$, and $a - \mathbf{b}^\top C^{-1}\mathbf{b} > 0$. Hence, for $s \neq 0$, define

$$\theta_I := \frac{\mathbf{b}^\top C^{-1}\mathbf{b}}{a}, \qquad \chi_I^2 := \frac{a\,\mathbf{z}^\top C^{-1}\mathbf{z}}{s^2}, \qquad t_I := \frac{\mathbf{b}^\top C^{-1}\mathbf{z}}{s}, \tag{5}$$

where $\chi_I \geq 0$. These scalars are invariant to the chosen orthonormal basis Q_I of the orthogonal complement. When $m = 1$, that complement is zero-dimensional and $\theta_I = \chi_I = t_I = 0$.

These quantities characterise the gain of Fisher weighting over the depth mean. χ_I measures class separation in the orthogonal complement after covariance scaling. θ_I measures the normalised cross-covariance between the uniform direction and that complement, while t_I records their interaction. The decomposition separates additional mean difference from covariance-dependent reweighting.

Theorem 3.1 (Fisher Ratio)**.** *Let the score vector $\mathbf{Z} \in \mathbb{R}^m$ be measurable and have finite class-conditional second moments. Fix the full support $I = \mathcal{K}_m$. Assume $\Sigma \succ 0$, $\mathbf{d} \neq \mathbf{0}$, and $s = \mathbf{u}_I^\top \mathbf{d} \neq 0$. Define*

$$\Delta_I := \sqrt{J(\mathbf{u}_I)}, \qquad R_I := \sqrt{\frac{J(\mathbf{w}_\star)}{J(\mathbf{u}_I)}}. \tag{6}$$

Here Δ_I is the uniform pooled SNR and R_I is the Fisher ratio: the Fisher-optimal pooled SNR divided by Δ_I. Then $\mathbf{w}_\star$ maximises J over all nonzero linear weights, and the squared SNR ratio admits the decomposition

$$R_I^2 = \chi_I^2 + \frac{(1 - t_I)^2}{1 - \theta_I}, \qquad 0 \leq \theta_I < 1, \qquad |t_I| \leq \sqrt{\theta_I}\,\chi_I. \tag{7}$$

In particular, $R_I^2 \geq 1$.

When $m = 1$, $R_I = 1$, since the score vector fixes every nonzero linear score up to scale. The ratio R_I compares the depth mean with all linear rules on the same score vector. Comparing the Fisher-optimal direction with a fixed reference direction through a pooled-SNR ratio bounded below by one is classical, with the independence rule as the reference direction (Fan et al., 2012). Equation 7 states that comparison for the uniform direction.

To connect the Fisher criterion to ranking performance, take independent draws $\mathbf{Z}_1 \sim (\mathbf{Z} \mid Y = 1)$ and $\mathbf{Z}_0 \sim (\mathbf{Z} \mid Y = 0)$, and write $D_\mathbf{w} := \mathbf{w}^\top(\mathbf{Z}_1 - \mathbf{Z}_0)$. The tie-aware pairwise AUROC is $\mathbb{P}(D_\mathbf{w} > 0) + \frac{1}{2}\mathbb{P}(D_\mathbf{w} = 0)$. For $\Sigma \succ 0$ and $\mathbf{w} \neq \mathbf{0}$, the signed pairwise SNR is

$$\Gamma(\mathbf{w}) := \frac{\mathbb{E}[D_\mathbf{w}]}{\sqrt{\mathrm{Var}(D_\mathbf{w})}} = \frac{\mathbf{w}^\top \mathbf{d}}{\sqrt{2\mathbf{w}^\top \Sigma \mathbf{w}}}. \tag{8}$$

When $\mathbf{w}^\top \mathbf{d} \geq 0$, this becomes $\Gamma(\mathbf{w}) = \sqrt{J(\mathbf{w})/2}$. The factor of two comes from the variance of the difference between the two independent class draws.

Corollary 3.2 (AUROC under Gaussian Score Differences)**.** *Under the assumptions of Theorem 3.1, assume $s > 0$. Suppose, for $\mathbf{w} \in \{\mathbf{u}_I, \mathbf{w}_\star\}$, that the one-dimensional difference $D_{\mathbf{w}}$ follows the positive-variance law $\mathcal{N}(\mathbf{w}^\top \mathbf{d}, 2\mathbf{w}^\top \Sigma \mathbf{w})$. Then, with Φ denoting the standard normal cumulative distribution function,*

$$\mathrm{AUROC}(\mathbf{w}) = \Phi(\Gamma(\mathbf{w})) = \Phi\left(\sqrt{\frac{J(\mathbf{w})}{2}}\right). \tag{9}$$

In particular, the AUROC difference is

$$\mathrm{AUROC}(\mathbf{w}_\star) - \mathrm{AUROC}(\mathbf{u}_I) = \Phi\left(\frac{R_I \Delta_I}{\sqrt{2}}\right) - \Phi\left(\frac{\Delta_I}{\sqrt{2}}\right), \tag{10}$$

and satisfies

$$0 \le \mathrm{AUROC}(\mathbf{w}_\star) - \mathrm{AUROC}(\mathbf{u}_I) \le \frac{(R_I - 1)\Delta_I}{2\sqrt{\pi}}. \tag{11}$$

The Gaussian condition is imposed on the two scalar differences. The joint score distribution may remain non-Gaussian, with unequal class covariances and dependent coordinates. The next result retains an explicit error term for an approximate Gaussian AUROC formula.

Corollary 3.3 (AUROC Bound with Gaussian Approximation Error)**.** *Under the assumptions of Theorem 3.1, let $\kappa_\star, \kappa_I \ge 0$ satisfy*

$$\left|\mathrm{AUROC}(\mathbf{w}_\star) - \Phi\left(\frac{R_I \Delta_I}{\sqrt{2}}\right)\right| \le \kappa_\star, \quad \left|\mathrm{AUROC}(\mathbf{u}_I) - \Phi\left(\frac{\Delta_I}{\sqrt{2}}\right)\right| \le \kappa_I. \tag{12}$$

Then

$$\mathrm{AUROC}(\mathbf{w}_\star) - \mathrm{AUROC}(\mathbf{u}_I) \le \frac{(R_I - 1)\Delta_I}{2\sqrt{\pi}} + \kappa_\star + \kappa_I. \tag{13}$$

The last inequality records deviations from the Gaussian AUROC formula through the two explicit error terms. The proofs appear in Appendix D.3 and the two following subsections. Appendix D.7 gives the train-fitted plug-in protocol and separates its held-out score from the theoretical optimum.

The Fisher ratio isolates the multiplicative pooled-SNR gain, while the uniform pooled SNR Δ_I sets the scale of the corresponding AUROC gap. Appendix D.3 gives an explicit weak bound and records the limiting cases of the decomposition.

3.4 HALLUTRACER PROBE-LOGIT TRAJECTORY

HalluTracer applies the preceding analysis to one score per layer. Let $l \in \mathcal{K}_m$ index Transformer depth and let $p^\star$ be the answer-onset position, namely the final prompt token immediately preceding answer generation. Let H be the number of attention heads and d_h their width. At layer l, we take the pre-output-projection attention vector at $p^\star$ and partition it into the indexed family of head-sized slices $(\mathbf{a}_l^{(h)})_{h=0}^{H-1}$, where $\mathbf{a}_l^{(h)} \in \mathbb{R}^{d_h}$. For each bootstrap resample of the training fold, we fit regularised logistic probes to all heads at each layer. We select the head with the highest training AUROC, denoted by $k_l^\star$, giving the layer-l readout $\mathbf{a}_l := \mathbf{a}_l^{(k_l^\star)} \in \mathbb{R}^{d_h}$. Because $p^\star$ is determined by the prompt, all coordinates are available before the first answer token is emitted.

The selected probe uses coordinate-wise standardisation $\mathcal{S}_l$ fitted on the same resample. Its decision logit

$$\ell_l := \widetilde{\boldsymbol{\omega}}_l^\top \mathcal{S}_l(\mathbf{a}_l) + \widetilde{b}_l \tag{14}$$

is the *probe logit*, oriented so that larger values point toward class 1. We average these logits across bootstrap resamples to obtain the layer score L_l. The raw-coordinate parametrisation, head-selection rules, and interpretation of the probe-logit scale are given in Appendix A.1.

The resulting score vector is the *probe-logit trajectory*

$$\boldsymbol{\tau} := [L_0, \ldots, L_{m-1}]^\top, \qquad \mathbf{Z} := \boldsymbol{\tau}. \tag{15}$$

Network depth supplies its ordering. For the deployed detector, HalluTracer takes the full support $I = \mathcal{K}_m$. Since positive rescaling changes neither J nor pairwise AUROC, its score is written as the arithmetic depth mean

$$\bar{L} := \frac{1}{m} \sum_{l=0}^{m-1} L_l. \tag{16}$$

One-dimensional logistic calibration is fitted and oriented within the training fold. Its decision logit is affine in $\bar{L}$, while the aggregation weights remain fixed. The retained layers determine the support, and the depth mean determines the aggregation rule. The experiments vary the support, compare aggregation rules on the full probe-logit trajectory, and evaluate the depth-mean detector.

Remark 3.4 (Scale of the Probe-Logit Coordinates). The coordinates used by HalluTracer are bootstrap averages of logits from training-fitted logistic probes, rather than raw activation magnitudes. They are dimensionless scores on a common fitted log-odds scale, so the depth mean is defined in a common score coordinate system across layers. Appendix A.1 records the corresponding product-rule interpretation, the Gaussian equal-variance case, and the effect of coordinate rescaling.

By averaging probe logits across depth, HalluTracer combines truthfulness evidence from all layers before generation without estimating cross-layer aggregation weights. This design retains evidence that selecting a single layer would discard, while avoiding the weight-estimation cost identified by our analysis. § 4 evaluates the resulting depth-mean detector on held-out data.

4 Experiments

4.1 Experimental Setup

Benchmarks and Backbones. The evaluation covers six LLMs and four benchmarks to assess the effectiveness of depth averaging across model families, scales, and prediction settings. The model suite comprises Llama-2-7B, Llama-3.1-8B, Qwen2.5-7B/14B/32B and Mistral-Small-24B. HaluEval2 (Li et al., 2024), TrueFalse (Azaria and Mitchell, 2023), and HELM (Su et al., 2024) evaluate statement-level truth classification: the statement appears in the prompt, and Y is its released truth label, following the probing settings of Inference-Time Intervention (Li et al., 2023) and hidden-state truth classification (Azaria and Mitchell, 2023). These benchmarks provide the 18 conditions for examining layer selection and aggregation as defined in § 3.2, using the released labels to measure pooled SNR and AUROC. TruthfulQA (Lin et al., 2022) extends the evaluation to predicting factual errors in the model's own answers. Following the pre-decoding setting in § 3.4, HalluTracer computes its score before the answer is generated. Each model then produces a response by greedy decoding, and GPT-5.4 evaluates it using the official TruthfulQA judging prompt, assigning $Y = 1$ if the response contains any factual error. All methods are evaluated against the same response labels for each model. The main comparison covers all 24 conditions across the four benchmarks.

Baselines. The comparison includes Fact-Probe (Han et al., 2025), ITI-Probe (Li et al., 2023), IRIS (Srey et al., 2025), SIVR-PD (Srey et al., 2026), and HalluGuard (Zeng et al., 2026). The first four construct classifiers from internal states. SIVR-PD adapts a state-of-the-art method for learned multi-layer aggregation to answer onset. HalluGuard scores the complete generated response after decoding. We also compare against the plug-in Fisher score.

Evaluation Protocol. We evaluate all methods using the same stratified five-fold split. Detector fitting and feature selection use only the training folds. The shrinkage analysis is a separate best-case comparison, reporting the highest held-out AUROC across the tested strengths.

4.2 Gain from Aggregating Across Layers

We examine whether retaining more layers improves truthfulness prediction, using the same averaging rule in Equation 1 for every layer subset. We compare subsets of two, four, and eight uniformly spaced layers with all layers and the first and second halves of the network. As shown in Figure 3, AUROC increases from two to eight layers in all nine conditions, and averaging over all layers matches or exceeds the eight-layer result. The stronger half varies across models, while the full-depth average outperforms both halves in every condition. These comparisons support collecting truthfulness evidence across the full network depth. To assess whether a single informative layer can match this

Table 1: Single-layer selection and full-support aggregation. Held-out pooled SNR and AUROC of the training-selected single layer (Single), the plug-in Fisher score (Fisher), and the depth mean (Mean). $j^\star$ is the modal selected layer and n_j the number of distinct layers selected across the five folds. Bold marks the largest AUROC in each row and underline the second largest.

		Pooled SNR			AUROC (%)			Selection	
Model	Dataset	Single	Fisher	Mean	Single	Fisher	Mean	$j^\star$	n_j
Llama-2-7B	HaluEval2	1.108	1.230	1.247	78.98	80.42	**80.85**	18	3
	TrueFalse	2.330	2.548	2.479	94.29	**95.23**	95.19	18	1
	HELM	1.389	1.687	1.664	83.55	**87.89**	87.70	12	3
Llama-3.1-8B	HaluEval2	1.335	1.431	1.415	82.38	**83.83**	83.60	28	2
	TrueFalse	2.841	3.147	3.038	96.27	97.19	**97.30**	21	2
	HELM	1.515	1.754	1.753	85.64	88.66	**88.80**	13	1
Qwen2.5-7B	HaluEval2	1.216	1.363	1.326	80.02	**82.59**	82.03	19	3
	TrueFalse	2.978	3.321	3.070	96.84	**97.47**	97.43	19	1
	HELM	1.400	1.669	1.667	83.83	87.69	**87.85**	4	5
Qwen2.5-14B	HaluEval2	1.297	1.406	1.420	81.49	83.30	**83.56**	35	3
	TrueFalse	3.328	3.758	3.512	97.79	98.25	**98.34**	44	3
	HELM	1.456	1.719	1.713	84.48	88.14	**88.34**	45	4
Qwen2.5-32B	HaluEval2	1.369	1.467	1.450	82.65	**84.23**	84.03	45	2
	TrueFalse	3.250	3.897	3.573	97.51	98.49	**98.51**	60	1
	HELM	1.375	1.727	1.740	83.36	88.32	**88.51**	42	4
Mistral-Small-24B	HaluEval2	1.358	1.448	1.497	82.47	83.99	**84.71**	39	2
	TrueFalse	3.192	3.724	3.590	97.54	98.13	**98.22**	36	4
	HELM	1.528	1.819	1.813	85.76	89.27	**89.50**	18	3
Average		1.904	2.173	2.109	87.49	89.62	**89.69**		

performance, we select the layer with the highest pooled SNR on each training fold. In Table 1, the depth mean achieves higher held-out AUROC than this selected layer in all 18 conditions, with gains of 0.55–5.15 points and pooled-SNR increases of 3.1%–26.5%. The largest gains occur on HELM, where all six models improve by 3.16–5.15 AUROC points. The gain persists both when the selected layer changes across folds and when it remains fixed.

4.3 Aggregation on the Probe-Logit Trajectory

Limited gains over the depth mean. We examine whether learning aggregation weights improves prediction when the full trajectory in Equation 15 is retained. We fit the Fisher direction from Theorem 3.1 and compare its held-out performance with the depth mean. As shown in Table 1, the depth mean retains at least 91.7% of the fitted Fisher score's pooled SNR across all 18 conditions. Their AUROC differs by at most 0.72 points, and the Fisher score averages 0.08 points below the mean. Using the Gaussian relation in Equation 9, the measured SNR values predict AUROC differences below one point, consistent with the observed differences. On HaluEval2 and HELM, the predicted and observed gaps agree within 0.3 points. The absolute approximation error $\widehat{\kappa}$ for individual scores ranges from 0.2 to 1.6 AUROC points and is largest on TrueFalse, where AUROC exceeds 95 (Table 8).

Redundancy across layers. To explain the limited gain from reweighting, we evaluate the decomposition in Equation 7. In Table 8, the estimated Fisher ratio $\widehat{R}_I$ ranges from 1.02 to 1.12, while $\widehat{\chi}_I$ ranges from 0.56 to 1.09. The nonzero values of $\widehat{\chi}_I$ show that differences between layer-wise scores carry class separation. This separation is strongly coupled to the depth mean: $\widehat{t}_I$ reaches at least 87% of its Cauchy-Schwarz bound $\sqrt{\widehat{\theta}_I\,\widehat{\chi}_I}$ in every condition. Let $\widehat{t}_I = 0$ in the decomposition while retaining the measured $\widehat{\theta}_I$ and $\widehat{\chi}_I$ gives ratios of 1.33–4.93 (Appendix D.7). The measured coupling therefore explains why the class separation in layer-wise differences adds little beyond the depth mean.

Sensitivity to training size. We assess the data requirements of learned aggregation by re-estimating the Fisher weights from smaller fractions of each training fold, keeping the probes, depth mean, and held-out folds fixed. The Fisher rule estimates $m(m+1)/2$ covariance entries, making covariance estimation sensitive to the available training data (Bickel and Levina, 2004). At full training size, the ratio of covariance parameters to minority-class training examples ranges from 0.16 to 1.74, and

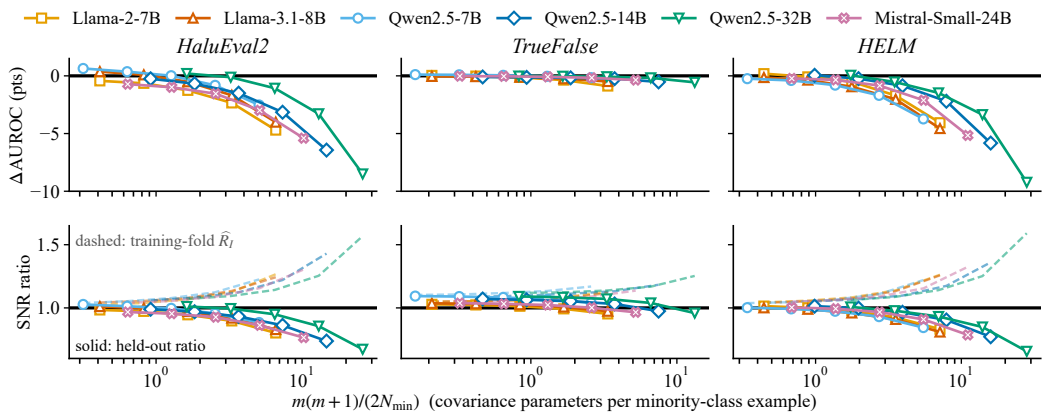


Figure 2: Estimation cost of the unshrunk plug-in Fisher score. The score is re-estimated on a fraction of the training fold, with probes, held-out folds, and the depth mean held fixed. Top: held-out AUROC difference to the depth mean. Bottom: held-out SNR ratio (solid) and training-fold $\widehat{R}_I$ (dashed). Horizontal axis: covariance parameters per minority-class training example, log scale.

Table 2: Truthfulness prediction results (AUROC / AUPRC, %) across four benchmarks and six LLMs. Bold: largest point estimate; underline: second largest. Where shown, $\pm$ denotes the standard deviation over five folds.

		Qwen2.5-7B		Qwen2.5-14B		Qwen2.5-32B		Mistral-Small-24B		Llama-2-7B		Llama-3.1-8B	
Dataset	**Method**	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC
TruthfulQA	HalluGuard	60.57	31.95	59.84	35.33	52.25	18.21	53.82	34.05	54.98	46.25	51.60	42.96
	Fact-Probe	67.26	48.40	60.30	22.00	75.53	40.14	69.23	47.79	69.20	62.98	66.07	59.33
	ITI-Probe	64.46	44.71	69.92	42.95	74.57	42.72	65.95	46.82	64.90	56.09	63.43	57.04
	IRIS	68.44	50.10	72.54	43.51	74.29	40.63	65.86	45.59	71.12	60.64	65.04	58.17
	SIVR-PD	73.69	59.73	78.95	**60.34**	**81.29**	**53.62**	69.74	51.60	72.77	72.23	69.72	72.96
	HalluTracer	75.38$_{\pm 1.3}$	63.95$_{\pm 3.2}$	79.67$_{\pm 3.9}$	58.80$_{\pm 8.2}$	80.97$_{\pm 2.6}$	51.18$_{\pm 3.1}$	**73.38**$_{\pm 3.0}$	53.36$_{\pm 6.5}$	**74.01**$_{\pm 2.6}$	**73.71**$_{\pm 3.7}$	71.24$_{\pm 4.4}$	**73.35**$_{\pm 4.8}$
TrueFalse	HalluGuard	52.17	52.13	50.38	50.23	58.55	56.40	53.52	52.85	50.84	50.77	51.11	49.78
	Fact-Probe	88.95	88.35	57.78	53.24	61.98	61.30	97.73	97.31	52.66	52.48	54.10	53.68
	ITI-Probe	96.98	96.76	98.08	97.87	98.05	97.98	98.01	97.83	94.59	94.19	96.97	96.69
	IRIS	95.36	94.41	94.86	90.38	97.78	97.53	97.28	96.54	93.94	93.49	96.73	96.42
	SIVR-PD	97.27	97.41	97.85	97.91	98.19	98.27	97.82	98.00	94.67	94.88	96.64	96.86
	HalluTracer	**97.59**$_{\pm 0.3}$	**97.49**$_{\pm 0.3}$	**98.40**$_{\pm 0.3}$	**98.23**$_{\pm 0.3}$	**98.54**$_{\pm 0.3}$	**98.45**$_{\pm 0.3}$	**98.26**$_{\pm 0.3}$	**98.11**$_{\pm 0.4}$	**95.34**$_{\pm 0.1}$	**95.00**$_{\pm 0.1}$	**97.44**$_{\pm 0.4}$	**97.26**$_{\pm 0.6}$
HaluEval2	HalluGuard	50.46	45.91	53.19	47.73	51.29	47.46	55.66	50.97	50.48	45.63	50.61	46.26
	Fact-Probe	57.13	53.73	58.20	54.49	62.37	57.73	77.29	70.18	58.63	55.61	59.71	56.26
	ITI-Probe	81.42	76.31	82.34	76.37	82.98	76.87	83.04	77.18	79.11	74.10	82.24	77.39
	IRIS	79.96	74.68	81.78	74.53	81.83	77.98	84.02	79.23	78.86	74.31	81.43	75.59
	SIVR-PD	80.85	76.32	82.98	77.72	82.54	76.77	84.12	78.85	79.27	74.15	81.55	75.98
	HalluTracer	**82.42**$_{\pm 1.4}$	**78.14**$_{\pm 0.3}$	**83.88**$_{\pm 0.3}$	**79.11**$_{\pm 0.3}$	**84.53**$_{\pm 1.4}$	**79.39**$_{\pm 1.3}$	**84.71**$_{\pm 1.3}$	**79.88**$_{\pm 1.5}$	**80.68**$_{\pm 1.2}$	**76.25**$_{\pm 1.1}$	**83.85**$_{\pm 1.0}$	**79.46**$_{\pm 0.1}$
HELM	HalluGuard	52.18	51.61	51.89	49.08	51.56	55.09	51.47	54.69	51.61	54.27	54.31	56.72
	Fact-Probe	59.14	59.89	57.78	59.55	67.76	67.00	86.78	86.75	57.53	59.40	58.14	58.94
	ITI-Probe	85.45	86.48	85.54	85.51	86.07	86.49	87.22	87.63	84.25	84.50	80.38	80.05
	IRIS	86.08	86.49	87.40	87.33	87.45	87.82	88.09	87.42	87.15	86.69	79.51	80.38
	SIVR-PD	82.89	83.85	83.51	83.17	83.10	83.26	83.51	83.14	82.32	82.66	84.24	83.75
	HalluTracer	**87.84**$_{\pm 1.2}$	**88.50**$_{\pm 1.1}$	**88.34**$_{\pm 0.4}$	**89.15**$_{\pm 0.2}$	**88.59**$_{\pm 1.3}$	**89.23**$_{\pm 1.0}$	**89.51**$_{\pm 1.9}$	**90.20**$_{\pm 2.1}$	**87.65**$_{\pm 0.1}$	**88.06**$_{\pm 0.4}$	**88.73**$_{\pm 1.0}$	**89.21**$_{\pm 0.4}$

the training estimate $\widehat{R}_I$ exceeds the held-out SNR ratio in all 18 conditions (Table 8). As shown in Figure 2, reducing the training data to one sixteenth increases the parameter ratio to 2.6–28.1 and the training-fold $\widehat{R}_I$ to 1.12–1.59. At the same time, the unshrunk Fisher score falls 0.2–9.2 AUROC points below the depth mean in every condition. The growing training ratio thus accompanies a loss in held-out ranking performance. Under the best-case shrinkage comparison, the fitted score remains within 0.58 points of the depth mean at one sixteenth of the training data (Table 9). The selected strengths range from 0.15 to 1.00 at full training size and from 0.55 to 1.00 at one sixteenth. These comparisons support the depth mean as a way to retain comparable AUROC without estimating additional aggregation weights. Appendix D.7 gives the estimation and shrinkage protocols.

4.4 Main Results

Overall detection performance. We evaluate whether the depth mean in Equation 16 yields competitive truthfulness prediction across model families and benchmarks. As shown in Table 2, HalluTracer achieves the highest AUROC in 23 of 24 conditions and the highest AUPRC in 22 among the compared methods. The largest margins over the strongest baseline are 4.49 AUROC points and 5.46 AUPRC points, both on HELM with Llama-3.1-8B. These results complement the controlled aggregation comparisons in § 4.2 and 4.3, showing that the resulting detector is competitive across the evaluated model families and benchmarks.

Prediction before generation. On TruthfulQA, HalluGuard (Zeng et al., 2026) scores the complete generated response, whereas HalluTracer predicts response truthfulness from activations available before decoding (§ 3.4). In Table 2, HalluTracer achieves higher AUROC and AUPRC than HalluGuard on all six models, with AUROC gains of 14.81–28.72 points. These results support the depth mean as an early indicator of response factuality.

4.5 Ablations and Controls

Trajectory Features and Aggregation. We test whether richer features or alternative aggregation rules improve prediction beyond the depth mean in Equation 16. The feature comparison includes classifiers using trajectory statistics, increment features, or an expanded input containing all layer-wise probe logits, their first differences, and six summary statistics (Appendix C). As shown in Table 6, the largest improvement over the mean is 0.6 AUROC points, obtained with the trajectory statistics, while the classifier using the expanded input can fall up to 1.0 point below it. Replacing the mean with the median or a 10%-trimmed mean changes AUROC by at most 0.3 points across nine conditions (Appendix D.8). At full training size, the best-case shrinkage comparison yields a gain of at most 0.79 points over the mean, averaging 0.24 points across the 18 conditions (Table 9). These comparisons complement the linear weighting analysis in § 4.3 and support the depth mean as an effective detector input, with limited gains from the tested feature expansions and aggregation alternatives.

Readout Controls. We examine how prediction depends on the head selection, internal representation, and probe-logit scale used in § 3.4. In the head-selection control, which averages probe probabilities, Random-1 remains within 0.7 AUROC points of Top-1 across nine conditions (Table 4). Using residual-stream activations with the same depth mean changes AUROC by at most 0.5 points on HaluEval2 and TrueFalse, and improves HELM by 1.8–2.3 points across five models (Table 5). The HELM gains show that changing the internal representation can improve prediction while retaining the same aggregation rule. The largest AUROC gain from standardising the probe logits or averaging their sigmoid-transformed probabilities is 0.18 points across 18 conditions (Table 10), supporting the use of the fitted logit scale described in Remark 3.4. A classifier using statement length alone reaches at most 55.5 AUROC on the three statement benchmarks, at least 25 points below the depth-mean detector (Appendix E.2).

5 Conclusion

Reliable use of language models depends on identifying answers that need further factual verification. We study how to provide this signal before generation by aggregating truthfulness evidence from internal states. Retaining more layers improves prediction even with fixed equal weights, while our analysis explains the limited reweighting gain through overlap between the depth mean and layer-wise differences. Across 18 conditions, the depth mean remains within 0.72 AUROC points of the fitted Fisher linear discriminant, without learning aggregation weights. This finding motivates HalluTracer, which averages layer-wise probe logits to predict truthfulness before decoding. Across six models and four benchmarks, it achieves the highest AUROC among the compared methods in 23 of 24 conditions. On TruthfulQA, it achieves higher AUROC and AUPRC than HalluGuard on all six models, while HalluGuard uses the complete generated response. By making a factuality score available at answer onset, this approach could inform whether additional verification is needed before the model generates an answer.

LLM USE DISCLOSURE

We used LLM-based assistants to support manuscript drafting and polishing, and coding. These activities were conducted under the authors' instruct, with suggestions reviewed and revised against the some proof. All literature reviews are completed under the guidance of the author. The ideas were generated without the use of LLMs. All experimental data were automatically generated by code, ensuring a strict separation between the LLMs and the experimental results.

REPRODUCIBILITY STATEMENT

This work evaluates pre-decoding truthfulness prediction on four publicly released benchmarks using six publicly released model checkpoints. Some state-of-the-art methods are using their best configures and full generation. The probe construction, head-selection procedure, feature definitions, Fisher diagnostics, covariance-shrinkage comparisons, and depth-support controls are documented in Appendices A.1 - D. The theoretical statements are made under the regularity conditions stated in the Method, with proofs in Appendices D.3–D.5. All the experimental data and formulas have been carefully verified and revised. Code and analysis artifacts will be released upon acceptance.

REFERENCES

Deema Alnuhait, Neeraja Kirtane, Muhammad Khalifa, and Hao Peng. FactCheckmate: Preemptively detecting and mitigating hallucinations in LMs. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, Suzhou, China, 2025. Association for Computational Linguistics.

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*. Association for Computational Linguistics, 2023.

Guy Bar-Shalom, Fabrizio Frasca, Yaniv Galron, Yftah Ziser, and Haggai Maron. Beyond token probes: Hallucination detection via activation tensors with ACT-ViT. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Mohammad Beigi, Ying Shen, Runing Yang, Zihao Lin, Qifan Wang, Ankith Mohan, Jianfeng He, Ming Jin, Chang-Tien Lu, and Lifu Huang. InternalInspector i^2: Robust confidence estimation in LLMs through internal states. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 12847–12865. Association for Computational Linguistics, 2024.

Nora Belrose, Igor Ostrovsky, Lev McKinney, Zach Furman, Logan Smith, Danny Halawi, Stella Biderman, and Jacob Steinhardt. Eliciting latent predictions from transformers with the tuned lens, 2025.

Peter J. Bickel and Elizaveta Levina. Some theory for Fisher's linear discriminant function, 'naive Bayes', and some alternatives when there are many more variables than observations. *Bernoulli*, 10(6):989–1010, 2004.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering latent knowledge in language models without supervision. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023.

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 *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James R. Glass, and Pengcheng He. Dola: Decoding by contrasting layers improves factuality in large language models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024.

Robyn M. Dawes. The robust beauty of improper linear models in decision making. *American Psychologist*, 34(7):571–582, 1979. doi: 10.1037/0003-066X.34.7.571.

Victor DeMiguel, Lorenzo Garlappi, and Raman Uppal. Optimal versus naive diversification: How inefficient is the 1/N portfolio strategy? *The Review of Financial Studies*, 22(5):1915–1953, 2009. doi: 10.1093/rfs/hhm075.

Edgar Dobriban and Stefan Wager. High-dimensional asymptotics of prediction: Ridge regression and classification. *The Annals of Statistics*, 46(1):247–279, 2018.

Ekaterina Fadeeva, Aleksandr Rubashevskii, Artem Shelmanov, Sergey Petrakov, Haonan Li, Hamdy Mubarak, Evgenii Tsymbalov, Gleb Kuzmin, Alexander Panchenko, Timothy Baldwin, Preslav Nakov, and Maxim Panov. Fact-checking the output of large language models via token-level uncertainty quantification. In *Findings of the Association for Computational Linguistics: ACL 2024*. Association for Computational Linguistics, 2024.

Jianqing Fan, Yang Feng, and Xin Tong. A road to classification in high dimensional space: The regularized optimal affine discriminant. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(4):745–771, 2012.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630:625–630, 2024. doi: 10.1038/s41586-024-07421-0.

Ronald A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7 (2):179–188, 1936.

Giorgio Fumera and Fabio Roli. A theoretical and experimental analysis of linear combiners for multiple classifier systems. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27 (6):942–956, 2005. doi: 10.1109/TPAMI.2005.109.

I. J. Good. Studies in the history of probability and statistics. XXXVII A. M. Turing's statistical work in World War II. *Biometrika*, 66(2):393–396, 1979.

Jiatong Han, Neil Band, Muhammed Razzak, Jannik Kossen, Tim G.J. Rudner, and Yarin Gal. Simple factuality probes detect hallucinations in long-form natural language generation. In *Findings of the Association for Computational Linguistics: EMNLP 2025*. Association for Computational Linguistics, 2025.

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 Trans. Inf. Syst.*, 43 (2):42:1–42:55, 2025. doi: 10.1145/3703155.

Xin Huang, Gengsheng Qin, and Yixin Fang. Optimal combinations of diagnostic tests based on AUC. *Biometrics*, 67(2):568–576, 2011.

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

Ziwei Ji, Delong Chen, Etsuko Ishii, Samuel Cahyawijaya, Yejin Bang, Bryan Wilie, and Pascale Fung. LLM internal states reveal hallucination risk faced with a query. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 88–104. Association for Computational Linguistics, 2024.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. Language models (mostly) know what they know. *CoRR*, abs/2207.05221, 2022. doi: 10.48550/ARXIV.2207.05221.

Hazel Kim, Tom A. Lamb, Adel Bibi, Philip Torr, and Yarin Gal. Detecting LLM hallucination through layer-wise information deficiency: Analysis of ambiguous prompts and unanswerable questions. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 32310–32322. Association for Computational Linguistics, 2025.

Josef Kittler, Mohamad Hatef, Robert P. W. Duin, and Jiri Matas. On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(3):226–239, 1998.

Olivier Ledoit and Michael Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88(2):365–411, 2004.

Junyi Li, Jie Chen, Ruiyang Ren, Xiaoxue Cheng, Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. The dawn after the dark: An empirical study on factuality hallucination in large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10879–10899, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.586.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pages 41451–41530, 2023.

Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252. Association for Computational Linguistics, 2022.

Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 9004–9017. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.EMNLP-MAIN.557.

Samuel Marks and Max Tegmark. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. *CoRR*, abs/2310.06824, 2023. doi: 10.48550/ARXIV.2310.06824.

Amina Miftakhova and Alexey Zaytsev. Pre-generation hallucination detection in large language models via soft-target attention probing. *arXiv preprint arXiv:2606.21917*, 2026.

Amir Hameed Mir. The geometry of truth: Layer-wise semantic dynamics for hallucination detection in large language models. *CoRR*, abs/2510.04933, 2025. doi: 10.48550/ARXIV.2510.04933.

Zakhar Mrykhin and Valentin Malykh. Prompt embedding probes (PEP): Hallucination detection in LLMs from hidden states. *arXiv preprint arXiv:2608.08024*, 2026.

Alexandru Niculescu-Mizil and Rich Caruana. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd International Conference on Machine Learning*, pages 625–632, 2005.

Erik Nordby, Tasha Pais, and Aviel Parrack. Linear probe accuracy scales with model size and benefits from multi-layer ensembling. *arXiv preprint arXiv:2604.13386*, 2026.

nostalgebraist. Interpreting GPT: the logit lens. LessWrong, 2020.

Jeremy Smith and Kenneth F. Wallis. A simple explanation of the forecast combination puzzle. *Oxford Bulletin of Economics and Statistics*, 71(3):331–355, 2009.

Ponhvoan Srey, Xiaobao Wu, and Anh Tuan Luu. Unsupervised hallucination detection by inspecting reasoning processes. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng, editors, *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 22117–22129, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.1124.

Ponhvoan Srey, Xiaobao Wu, Cong-Duy T Nguyen, and Anh Tuan Luu. Learning uncertainty from sequential internal dispersion in large language models. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2026.

John Q. Su and Jun S. Liu. Linear combinations of multiple diagnostic markers. *Journal of the American Statistical Association*, 88(424):1350–1355, 1993.

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 Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14379–14391, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.854.

Ao Sun. CHAIR: Classifier of hallucination as improver. *arXiv preprint arXiv:2501.02518*, 2025.

Mohammad Taufeeque, Stefan Heimersheim, Adam Gleave, and Chris Cundy. The obfuscation atlas: Mapping where honesty emerges in RLVR with deception probes. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026.

Zeguan Xiao, Diyang Dou, Boya Xiong, Yun Chen, and Guanhua Chen. Enhancing uncertainty estimation in LLMs with expectation of aggregated internal belief. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 34043–34051, 2026. doi: 10.1609/aaai.v40i40.40698.

Xinyue Zeng, Junhong Lin, Yujun Yan, Feng Guo, Liang Shi, Jun Wu, and Dawei Zhou. Halluguard: Demystifying data-driven and reasoning-driven hallucinations in llms. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust, editors, *International Conference on Learning Representations*, volume 2026, pages 87473–87505, 2026.

Jianyi Zhang, Da-Cheng Juan, Cyrus Rashtchian, Chun-Sung Ferng, Heinrich Jiang, and Yiran Chen. SLED: Self logits evolution decoding for improving factuality in large language models. In *Advances in Neural Information Processing Systems*, 2024.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Chen Xu, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, and Shuming Shi. Siren's song in the ai ocean: A survey on hallucination in large language models, 2025.

A Implementation Details

This section specifies how we standardise head activations, fit layer-wise probes, and select heads to construct the probe-logit trajectory in Equation 15. Its depth mean is the input to the scalar logistic classifier.

Training settings. Layer-wise probes use L_2-regularised logistic regression with `C=1` and the `liblinear` solver. We use three bootstrap resamples per training fold, each drawn uniformly with replacement at the original fold size. The logits are averaged before fitting a single scalar calibration model.

A.1 Probe Training and Head Selection

For an autoregressive Transformer with m layers and H attention heads, let $\mathbf{a}_l^{(h)} \in \mathbb{R}^{d_h}$ denote the raw pre-output-projection attention slice at the answer-onset position for $l \in \mathcal{K}_m$ and $h \in \{0, \dots, H-1\}$.

Supervised projection. For each bootstrap member and candidate pair (l, h), the resample supplies a coordinate-wise mean $\widehat{\boldsymbol{\mu}}_{l,h}$ and positive scale $\widehat{\boldsymbol{\sigma}}_{l,h}$. The fixed standardisation map is

$$\mathcal{S}_{l,h}(\mathbf{a}) := \operatorname{diag}(\widehat{\boldsymbol{\sigma}}_{l,h})^{-1}(\mathbf{a} - \widehat{\boldsymbol{\mu}}_{l,h}). \tag{17}$$

A regularised logistic regression fitted in this coordinate system has decision score, for $\widetilde{\mathbf{a}} \in \mathbb{R}^{d_h}$,

$$\widetilde{\phi}_{l,h}(\widetilde{\mathbf{a}}) = \widetilde{\boldsymbol{\omega}}_{l,h}^{\top}\widetilde{\mathbf{a}} + \widetilde{b}_{l,h}. \tag{18}$$

Composing the two maps gives the same score in raw activation coordinates:

$$\widehat{\phi}_{l,h}(\mathbf{a}) = \widehat{\boldsymbol{\omega}}_{l,h}^{\top}\mathbf{a} + \widehat{b}_{l,h}, \qquad \widehat{\boldsymbol{\omega}}_{l,h} = \operatorname{diag}(\widehat{\boldsymbol{\sigma}}_{l,h})^{-1}\widetilde{\boldsymbol{\omega}}_{l,h}, \tag{19}$$

$$\widehat{b}_{l,h} = \widetilde{b}_{l,h} - \widehat{\boldsymbol{\omega}}_{l,h}^{\top}\widehat{\boldsymbol{\mu}}_{l,h}. \tag{20}$$

After selecting $k_l^\star$, we abbreviate

$$\begin{aligned} \mathcal{S}_l &:= \mathcal{S}_{l,k_l^\star}, \quad \widetilde{\boldsymbol{\omega}}_l := \widetilde{\boldsymbol{\omega}}_{l,k_l^\star}, \quad \widetilde{b}_l := \widetilde{b}_{l,k_l^\star}, \\ \widehat{\boldsymbol{\omega}}_l &:= \widehat{\boldsymbol{\omega}}_{l,k_l^\star}, \quad \widehat{b}_l := \widehat{b}_{l,k_l^\star}, \end{aligned} \tag{21}$$

and use

$$\phi_l(\mathbf{a}) := \widetilde{\boldsymbol{\omega}}_l^{\top}\mathcal{S}_l(\mathbf{a}) + \widetilde{b}_l = \widehat{\phi}_{l,k_l^\star}(\mathbf{a}), \qquad \ell_l := \phi_l(\mathbf{a}_l^{(k_l^\star)}). \tag{22}$$

These unnormalised probe logits are averaged across bootstrap members to form L_l. Their orientation follows the binary probe convention $Y = 1$ for false examples. The corresponding unit raw-activation direction is

$$\mathbf{v}_{l,h} := \frac{\widehat{\boldsymbol{\omega}}_{l,h}}{\|\widehat{\boldsymbol{\omega}}_{l,h}\|_2}, \qquad \mathbf{v}_l := \mathbf{v}_{l,k_l^\star}. \tag{23}$$

The two scales serve different roles: L_l is the score coordinate used by the detector, and $\mathbf{v}_l$ is the unit direction used by the probe-alignment feature in Appendix C.

Head selection protocol. In the main experiments, we fit a logistic probe for every head at each layer and select the head with the highest training AUROC as $k_l^\star$. The head-selection ablation uses the same AUROC ranking for its Top-K and Bottom-1 controls. Head choices are fixed before the held-out fold is scored. An alternative fast implementation selects the head with the largest raw class-mean separation,

$$g_{l,h} := \left\| \mathbb{E}_{\mathrm{tr}}[\mathbf{a}_l^{(h)} \mid Y = 1] - \mathbb{E}_{\mathrm{tr}}[\mathbf{a}_l^{(h)} \mid Y = 0] \right\|_2. \tag{24}$$

Bootstrap probe average. The main experiments use bootstrap probe averaging. Each bootstrap resample of the training fold produces a trajectory with its own head choices and standardisers, and the logits are averaged at each layer. Because averaging is linear, the W/K identity holds for every ensemble member and for their coordinate-wise average.

Scale of the probe-logit coordinates. Each probe logit is the training-fitted logistic decision score at that layer, a dimensionless quantity independent of the activation scale of the selected head. Logistic regression places these scores on a log-odds scale (Niculescu-Mizil and Caruana, 2005). Their equal-weight sum is the product-rule combination of the layer-wise readouts (Kittler et al., 1998). Under a Gaussian model with equal class variances, a calibrated log-odds coordinate satisfies $d_l = \Sigma_{ll}$ (Good, 1979). If the coordinates are also uncorrelated, the uniform and Fisher-optimal directions coincide. More generally, the Fisher ratio measures the pooled-SNR gain available from the remaining mean and covariance structure. A coordinate-wise rescaling D changes the uniform reference direction to $D\mathbf{1}$.

B Head Selection Sensitivity

This section measures how detection changes with the selected attention head through exhaustive per-head probing and multi-head averaging.

Per-Head Discriminability. For each of the $m \times H$ layer-head coordinates, we train an independent logistic regression probe and record its training AUROC. Table 3 reports the results across three models. The per-head AUROC has CV below 0.08 across the three models.

Raw head-wise class-mean separation is scale-sensitive. Every probe is fitted after head-specific standardisation (Equation 17), so per-head AUROC reflects separation relative to the fitted coordinate scale. On the tested HaluEval2 representation, the small AUROC dispersion is consistent with separation distributed across several heads.

Table 3: Head-level probe statistics on pooled HaluEval2. AUC CV: within-layer AUROC coefficient of variation; $\bar{r}_{\mathrm{head}}$: mean pairwise Pearson correlation of head probabilities; Eff. Rank: exponential entropy of normalised correlation eigenvalues. CV, correlation, and effective rank are averaged over layers.

Model	$m \times H$	AUC CV	$\bar{r}_{\mathbf{head}}$	Eff. Rank	Rank/H
Qwen2.5-7B	28×28	0.047	0.75	3.5	12.4%
Llama-2-7B	32×32	0.073	0.65	5.7	17.8%
Llama-3.1-8B	32×32	0.061	0.74	3.7	11.5%

Multi-Head Ensemble and Random-Head Baseline. For this selection ablation, let $P_l^{(h)}$ denote the class-1 probability returned by the probe for head h at layer l. The evaluated score averages the selected probabilities within each layer and then across depth. This probability trajectory serves as a head-sensitivity control, and the deployed detector uses the probe-logit trajectory. If the head readouts are largely redundant, Top-5 averaging should add little to Top-1 selection, and random selection should remain close to the ranked choice. Table 4 reports that Top-1 and Top-5 differ by at most 0.1 percentage point, and Random-1 lies within 0.7 percentage point of Top-1 across the nine conditions. Bottom-1 remains above chance with a larger, more variable gap.

The small Top-5 gain and the narrow Random-1/Top-1 gap indicate overlap among the tested within-layer readouts. To quantify this redundancy, we compute the pairwise Pearson correlation of within-layer probe probabilities and the effective rank of the resulting correlation matrix (Table 3). Across the three reported models, the mean probability correlation is at least 0.65 and the effective rank is at most 18% of H.

The two analyses report a modest head-choice difference in the evaluated probability depth means and a larger depth-support difference (Figure 3).

Residual-Stream Readout. We fit layer-wise probes to the answer-onset residual-stream states and aggregate their logits with the same depth mean. This comparison holds the depth support and aggregation rule fixed while changing the layer representation (Table 5).

Table 4: Head selection with probability averaging within and across layers. Top-K and Bottom-1 follow the per-layer training AUROC ranking. Random-1 samples one head per layer uniformly (mean $\pm$ std over 50 draws). AUROC means are in percent; Δ is Random-1 minus Top-1 in AUROC points.

Model	Dataset	Top-1	Top-5	Bottom-1	Random-1	Δ(pp)
Llama-2-7B	HaluEval2	80.8	80.9	78.7	80.4$\pm$.013	-0.4
	TrueFalse	95.4	95.4	91.7	94.7$\pm$.004	-0.7
	HELM	88.3	88.4	87.0	88.0$\pm$.016	-0.3
Llama-3.1-8B	HaluEval2	83.9	83.9	82.0	83.3$\pm$.012	-0.5
	TrueFalse	97.5	97.5	96.2	97.2$\pm$.004	-0.3
	HELM	89.2	89.3	88.1	88.9$\pm$.016	-0.4
Qwen2.5-7B	HaluEval2	82.2	82.2	81.1	81.9$\pm$.014	-0.3
	TrueFalse	97.4	97.4	96.2	97.1$\pm$.004	-0.3
	HELM	88.4	88.4	87.9	88.0$\pm$.015	-0.4

Table 5: Readout mechanism ablation (AUROC): selected-head readout ($d{=}d_h$) vs. residual-stream ($d{=}d_{\text{model}}$, no head selection), both evaluated via the depth mean $\bar{L}$. Δ: residual-stream minus selected-head. The two readouts differ by at most 0.005 on HaluEval2 and TrueFalse, and the residual-stream value is higher on HELM.

Model	Dataset	Selected Head	Residual Stream	Δ
Qwen2.5-32B	HaluEval2	0.852	0.856	$+.005$
	TrueFalse	0.985	0.986	$+.001$
	HELM	0.887	0.907	$+.020$
Qwen2.5-14B	HaluEval2	0.847	0.847	.000
	TrueFalse	0.983	0.984	$+.001$
	HELM	0.884	0.907	$+.023$
Qwen2.5-7B	HaluEval2	0.832	0.829	$-.003$
	TrueFalse	0.974	0.975	$+.001$
	HELM	0.877	0.897	$+.021$
Llama-3.1-8B	HaluEval2	0.846	0.848	$+.002$
	TrueFalse	0.973	0.976	$+.003$
	HELM	0.888	0.906	$+.018$
Llama-2-7B	HaluEval2	0.821	0.822	$+.001$
	TrueFalse	0.952	0.956	$+.004$
	HELM	0.877	0.896	$+.019$

C Complete Feature Definitions

This section defines the statistics used in the ablations: the depth mean $\bar{L}$, four additional trajectory statistics, and 13 statistics derived from the W/K increment decomposition. Table 6 reports the comparison of § 4.5 on all 18 conditions. All feature ablations use the same training and evaluation settings as HalluTracer.

C.1 Trajectory Statistics

Given the trajectory $\boldsymbol{\tau} = [L_0, \ldots, L_{m-1}]^\top$, the ablations use a fixed middle window from one third to three quarters of depth and a late window containing the remaining layers. The deployed detector uses the depth mean $\bar{L}$, independently of these windows. The five statistics are:

1. **Mean** ($\bar{L}$): $m^{-1} \sum_{l=0}^{m-1} L_l$.
2. **OLS slope** (β): the ordinary least-squares coefficient of L_l on the depth index l.
3. **Window mean difference** (Δ_{window}): $\bar{L}_{\text{late}} - \bar{L}_{\text{mid}}$.

Table 6: Aggregation inputs on a fixed probe-logit trajectory under one shared full-support, five-fold run (AUROC, %). Mean: depth mean (1D); Traj.: trajectory statistics (5D); Incr.: increment statistics (13D); Exp.: expanded trajectory ($2m$+5D). The depth mean stays within 1.1 AUROC points of every tested richer representation. Bold marks the largest displayed point estimate within each benchmark.

	HaluEval2				TrueFalse				HELM			
Model	Mean	Traj.	Incr.	Exp.	Mean	Traj.	Incr.	Exp.	Mean	Traj.	Incr.	Exp.
Mistral-Small-24B	**84.7**	**84.7**	84.2	83.7	98.2	**98.4**	**98.4**	**98.4**	**89.5**	**89.5**	89.0	89.2
Qwen2.5-32B	84.0	**84.6**	84.3	83.8	98.5	**98.6**	**98.6**	98.3	88.5	**88.8**	88.4	88.7
Qwen2.5-14B	83.6	**83.8**	**83.8**	83.1	98.3	**98.4**	98.3	98.0	88.3	88.4	**88.5**	88.3
Qwen2.5-7B	82.0	**82.4**	82.2	**82.4**	97.4	**97.5**	**97.5**	97.2	**87.9**	87.7	87.6	87.6
Llama-3.1-8B	83.6	**83.8**	83.3	**83.8**	97.3	97.4	**97.5**	97.1	88.8	88.7	88.3	**88.9**
Llama-2-7B	**80.9**	80.8	80.7	80.1	95.2	**95.3**	**95.3**	**95.3**	87.7	87.7	87.3	**87.8**

4. **Total variation** (TV): $\sum_{l=0}^{m-2} |\Delta L_l|$, where $\Delta L_l := L_{l+1} - L_l$.
5. **Range**: $\max_l L_l - \min_l L_l$.

C.2 Increment-Feature Definition

Within each bootstrap member, the frozen cross-layer readout $\widetilde{\ell}_l^{+} := \phi_l(\mathbf{a}_{l+1}^{(k_l^\star)})$ separates each probe-logit increment into $W_l := \widetilde{\ell}_l^{+} - \ell_l$ and $K_l := \ell_{l+1} - \widetilde{\ell}_l^{+}$. The normalised probe alignment is $\psi_l := \mathbf{v}_l^\top \mathbf{a}_l / \|\mathbf{a}_l\|_2$. The update summaries use the change between consecutive normalised residual-update directions and the excess kurtosis of each block-output activation. These layer-wise quantities are averaged across bootstrap members before forming the window statistics below.

The increment features use the same two windows. A bar denotes the mean over the indicated window, TV denotes total variation, and E_W and E_K denote mean squared magnitudes in the middle window. The 13 features are grouped as follows:

- **Layer level (3):** $\bar{\psi}_{\text{mid}}$, $\bar{L}_{\text{mid}}$, and $L_{m-1} - L_0$.
- **Increments (5):** $\bar{W}_{\text{mid}}$, $\text{TV}(W_{\text{mid}})$, $\bar{K}_{\text{mid}}$, $\text{TV}(K_{\text{mid}})$, and $\log(E_W/E_K)$.
- **Updates (2):** mean update-direction change and mean activation excess kurtosis in the middle window.
- **Late layer (3):** $\bar{\psi}_{\text{late}}$, $\bar{K}_{\text{late}}$, and ψ_{m-1}.

Table 11 evaluates the same four groups.

C.3 Expanded Trajectory

The expanded trajectory concatenates:

1. the raw probe-logit trajectory $[L_0, \ldots, L_{m-1}]$ (m dimensions),
2. the first differences $[\Delta L_0, \ldots, \Delta L_{m-2}]$ ($m-1$ dimensions),
3. six summary statistics: mean, standard deviation, minimum, maximum, final value, and total change ($L_{m-1} - L_0$).

This yields $2m + 5$ dimensions.

D Auxiliary SNR Analysis and Fisher Ratio

D.1 Random-Intercept Model for Probe Logits

Theorem 3.1 uses the full cross-coordinate covariance. The following random-intercept model is a narrower HalluTracer approximation used only to compare the depth mean with a centred slope statistic.

Remark D.1 (Random-Intercept Model for Probe Logits). For sample i and layer l, the random-intercept idealisation writes

$$L_l^{(i)} = \mu_{l,Y_i} + \alpha_i + \xi_{i,l}, \qquad \mu_{l,c} := \mathbb{E}[L_l \mid Y = c], \tag{25}$$

where α_i is a centred sample-level offset shared across depth and $\xi_{i,l}$ is a centred layer-specific innovation. The tractable homoscedastic version assumes $\mathrm{Var}(\alpha_i) = \sigma_\alpha^2$, $\mathrm{Var}(\xi_{i,l}) = \sigma_\xi^2$, and mutual independence of α_i and $\{\xi_{i,l}\}_l$ conditional on the class.

The empirical variance decomposition first estimates $\widehat{\alpha}_i := m^{-1}\sum_l (L_l^{(i)} - \mu_{l,Y_i})$ and then forms $\epsilon_{i,l} := L_l^{(i)} - \mu_{l,Y_i} - \widehat{\alpha}_i$, which satisfies $\sum_l \epsilon_{i,l} = 0$ by construction. This projected residual is used for the empirical variance accounting, while the innovations $\xi_{i,l}$ define the auxiliary calculation below.

Proposition D.2 (Mean Versus the OLS Depth Slope)**.** *Assume the independent, homoscedastic model in Equation 25, with $m \geq 2$ and $\sigma_\xi^2 > 0$. Let $d_l := \mu_{l,1} - \mu_{l,0}$, $\bar{d} := m^{-1}\sum_l d_l$, and $\bar{l} := (m-1)/2$. Define*

$$S_{xx} := \sum_{l=0}^{m-1} (l - \bar{l})^2, \qquad q_l := \frac{l - \bar{l}}{S_{xx}}, \qquad \widehat{\beta} := \sum_l q_l L_l, \tag{26}$$

so that $\widehat{\beta}$ is the ordinary least-squares depth slope. For a scalar statistic T, let $\mathrm{SNR}(T)$ denote its absolute class-conditional mean gap divided by its common within-class standard deviation under the homoscedastic model. If $\bar{d} \neq 0$, set

$$\rho := \frac{m|\sum_l q_l d_l|}{|\bar{d}|}, \qquad \gamma := \sqrt{1 + \frac{m\sigma_\alpha^2}{\sigma_\xi^2}}. \tag{27}$$

Then

$$\frac{\mathrm{SNR}(\widehat{\beta})}{\mathrm{SNR}(m^{-1}\sum_l L_l)} = \frac{\rho\gamma}{\sqrt{12}}\sqrt{1 - \frac{1}{m^2}}. \tag{28}$$

Thus a slowly varying class-gap profile makes a centred depth filter weak relative to the mean in this idealised model. This proposition concerns one comparator and one covariance model. Theorem 3.1 gives the comparison with unrestricted linear weighting.

D.2 Proof of Proposition D.2

Proof. Under Equation 25, write

$$L_l = \mu_{l,Y} + \alpha + \xi_l, \tag{29}$$

where, conditional on either class, the centred random variables satisfy

$$\mathrm{Var}(\alpha) = \sigma_\alpha^2, \qquad \mathrm{Var}(\xi_l) = \sigma_\xi^2, \tag{30}$$

and $\alpha, \xi_0, \ldots, \xi_{m-1}$ are mutually independent. For a scalar statistic T, define its homoscedastic signal-to-noise ratio as the absolute class-conditional mean gap divided by its common within-class standard deviation.

The depth mean has class gap $\bar{d}$ and variance

$$\mathrm{Var}\left(\frac{1}{m}\sum_l L_l \,\middle|\, Y = c\right) = \sigma_\alpha^2 + \frac{\sigma_\xi^2}{m}. \tag{31}$$

For the OLS weights $q_l = (l - \bar{l})/S_{xx}$, we have $\sum_l q_l = 0$ and

$$\sum_l q_l^2 = \frac{1}{S_{xx}}, \qquad S_{xx} = \sum_{l=0}^{m-1}(l - \bar{l})^2 = \frac{m(m^2-1)}{12}. \tag{32}$$

The common offset therefore cancels from $\widehat{\beta} = \sum_l q_l L_l$, whose class gap and variance are

$$\left|\sum_l q_l d_l\right|, \qquad \mathrm{Var}(\widehat{\beta} \mid Y = c) = \frac{12\sigma_\xi^2}{m(m^2-1)}. \tag{33}$$

Combining Equations 31 to 33 gives

$$\frac{\mathrm{SNR}(\widehat{\beta})}{\mathrm{SNR}(m^{-1}\sum_l L_l)} = \frac{|\sum_l q_l d_l|}{|\bar{d}|}\sqrt{\frac{m(m^2-1)(\sigma_\alpha^2+\sigma_\xi^2/m)}{12\sigma_\xi^2}} = \frac{\rho\gamma}{\sqrt{12}}\sqrt{1-\frac{1}{m^2}}, \tag{34}$$

because $\rho = m|\sum_l q_l d_l|/|\bar{d}|$ and $\gamma^2 = 1 + m\sigma_\alpha^2/\sigma_\xi^2$. This is Equation 28. □

Remark D.3 (Finite-depth factor). The simplified expression $\rho\gamma/\sqrt{12}$ omits the multiplicative factor $\sqrt{1-m^{-2}}$. For $m \geq 2$,

$$\frac{1}{2m^2} < 1 - \sqrt{1-\frac{1}{m^2}} < \frac{1}{m^2}, \tag{35}$$

which follows from $1-\sqrt{1-x} = x/(1+\sqrt{1-x})$ with $x = m^{-2}$. Thus the omitted factor and its direction are explicit.

Zero-sum linear filters. The same working model gives a useful extension of Proposition D.2. Write $\mathbf{d} = \bar{d}\,\mathbf{1} + \mathbf{r}$, where $\mathbf{1}^\top\mathbf{r} = 0$, and define

$$\rho_\perp := \frac{\|\mathbf{r}\|_2}{|\bar{d}|\sqrt{m}}. \tag{36}$$

For every unit vector $\mathbf{q}$ satisfying $\mathbf{1}^\top\mathbf{q} = 0$, the common offset cancels and Cauchy-Schwarz gives

$$|\mathbf{q}^\top\mathbf{d}| = |\mathbf{q}^\top\mathbf{r}| \leq \|\mathbf{r}\|_2. \tag{37}$$

Since $\mathrm{Var}(\mathbf{q}^\top\boldsymbol{\tau} \mid Y = c) = \sigma_\xi^2$, it follows that

$$\sup_{\substack{\mathbf{q}\neq\mathbf{0}\\ \mathbf{1}^\top\mathbf{q}=0}} \frac{\mathrm{SNR}(\mathbf{q}^\top\boldsymbol{\tau})}{\mathrm{SNR}(m^{-1}\mathbf{1}^\top\boldsymbol{\tau})} \leq \gamma\rho_\perp. \tag{38}$$

The ratio is scale-invariant, so restricting the supremum to unit vectors loses nothing. If $\mathbf{r} \neq \mathbf{0}$, equality is attained by $\mathbf{q} = \mathbf{r}/\|\mathbf{r}\|_2$.

D.3 Proof of Theorem 3.1

Proof. We take $I = \mathcal{K}_m$ and follow the definitions in § 3.3. The argument has four parts.

1. Fisher-optimal direction. Because $\Sigma \succ 0$, the bilinear form $\langle \mathbf{x}, \mathbf{y}\rangle_\Sigma := \mathbf{x}^\top\Sigma\mathbf{y}$ is an inner product. With $\mathbf{w}_\star = \Sigma^{-1}\mathbf{d}$, we have $\mathbf{d} = \Sigma\mathbf{w}_\star$, and therefore, for every nonzero $\mathbf{w}$,

$$(\mathbf{w}^\top\mathbf{d})^2 = \langle \mathbf{w}, \mathbf{w}_\star\rangle_\Sigma^2 \leq (\mathbf{w}^\top\Sigma\mathbf{w})(\mathbf{w}_\star^\top\Sigma\mathbf{w}_\star). \tag{39}$$

Dividing by the positive first factor yields

$$J(\mathbf{w}) \leq \mathbf{w}_\star^\top\Sigma\mathbf{w}_\star = \mathbf{d}^\top\Sigma^{-1}\mathbf{d} = J(\mathbf{w}_\star). \tag{40}$$

Since $\mathbf{d} \neq \mathbf{0}$, positive definiteness also gives $\mathbf{w}_\star \neq \mathbf{0}$ and

$$\mathbf{w}_\star^\top\mathbf{d} = \mathbf{d}^\top\Sigma^{-1}\mathbf{d} > 0. \tag{41}$$

Thus $\mathbf{w}_\star$ is the oriented Fisher-optimal direction. Every nonzero rescaling attains the same Fisher criterion, while positive rescaling preserves its class orientation.

2. Orthogonal decomposition around the uniform direction. For the full support, the vector $\mathbf{u}_I = m^{-1/2}\mathbf{1}$ has unit norm and can be completed to an orthogonal matrix $U_I = [\mathbf{u}_I, Q_I]$. In the coordinates of Equation 4, positive definiteness of Σ gives $a > 0$, $C \succ 0$, and the positive Schur complement

$$\widetilde{a} := a - \mathbf{b}^\top C^{-1}\mathbf{b} > 0. \tag{42}$$

Consequently

$$0 \le \theta_I = \frac{\mathbf{b}^\top C^{-1}\mathbf{b}}{a} < 1. \tag{43}$$

The quantities are independent of the chosen basis of the orthogonal complement. Indeed, replacing Q_I by $Q_I R$, with R orthogonal, maps

$$\mathbf{z} \mapsto R^\top \mathbf{z}, \qquad \mathbf{b} \mapsto R^\top \mathbf{b}, \qquad C \mapsto R^\top C R, \tag{44}$$

and leaves each quadratic or bilinear scalar in Equation 5 unchanged.

3. Exact Fisher ratio. Let

$$M := U_I^\top \Sigma U_I = \begin{pmatrix} a & \mathbf{b}^\top \\ \mathbf{b} & C \end{pmatrix}. \tag{45}$$

The block inverse associated with Equation 42 is

$$M^{-1} = \begin{pmatrix} \widetilde{a}^{-1} & -\widetilde{a}^{-1}\mathbf{b}^\top C^{-1} \\ -C^{-1}\mathbf{b}\widetilde{a}^{-1} & C^{-1} + C^{-1}\mathbf{b}\widetilde{a}^{-1}\mathbf{b}^\top C^{-1} \end{pmatrix}. \tag{46}$$

Hence

$$\begin{aligned} J(\mathbf{w}_\star) &= \mathbf{d}^\top \Sigma^{-1}\mathbf{d} \\ &= \begin{pmatrix} s \\ \mathbf{z} \end{pmatrix}^\top M^{-1} \begin{pmatrix} s \\ \mathbf{z} \end{pmatrix} \\ &= \frac{(s - \mathbf{b}^\top C^{-1}\mathbf{z})^2}{a - \mathbf{b}^\top C^{-1}\mathbf{b}} + \mathbf{z}^\top C^{-1}\mathbf{z}. \end{aligned} \tag{47}$$

The uniform direction has

$$J(\mathbf{u}_I) = \frac{s^2}{a} > 0. \tag{48}$$

Dividing Equation 47 by s^2/a gives

$$\begin{aligned} R_I^2 &= \frac{a}{s^2}\left[\frac{(s - \mathbf{b}^\top C^{-1}\mathbf{z})^2}{a - \mathbf{b}^\top C^{-1}\mathbf{b}} + \mathbf{z}^\top C^{-1}\mathbf{z}\right] \\ &= \chi_I^2 + \frac{(1 - t_I)^2}{1 - \theta_I}, \end{aligned} \tag{49}$$

which is the additive identity in Equation 7. Part 1 gives $R_I^2 \ge 1$.

4. Bounds for the cross term. Cauchy-Schwarz in the C^{-1}-inner product gives

$$(\mathbf{b}^\top C^{-1}\mathbf{z})^2 \le (\mathbf{b}^\top C^{-1}\mathbf{b})(\mathbf{z}^\top C^{-1}\mathbf{z}). \tag{50}$$

After division by s^2, this becomes

$$t_I^2 \le \theta_I \chi_I^2, \qquad |t_I| \le \sqrt{\theta_I}\,\chi_I. \tag{51}$$

□

Corollary D.4 (Weak Bound on the Fisher Ratio)**.** *Under the assumptions of Theorem 3.1, if $\varepsilon \ge 0$, $\eta \ge 0$, $\chi_I \le \varepsilon$, and $\theta_I \le \eta^2 < 1$, then*

$$1 \le R_I^2 \le B(\varepsilon, \eta), \qquad B(\varepsilon, \eta) := \varepsilon^2 + \frac{(1 + \eta\varepsilon)^2}{1 - \eta^2}. \tag{52}$$

If the additional assumptions of Corollary 3.2 hold, the AUROC difference in Equation 11 is at most $(\sqrt{B(\varepsilon, \eta)} - 1)\Delta_I/(2\sqrt{\pi})$.

Proof. The constraint $|t_I| \le \sqrt{\theta_I}\,\chi_I$ gives $|t_I| \le \eta\varepsilon$, and the exact identity in Equation 7 yields

$$\begin{aligned} R_I^2 &\le \varepsilon^2 + \frac{(1 + |t_I|)^2}{1 - \theta_I} \\ &\le \varepsilon^2 + \frac{(1 + \eta\varepsilon)^2}{1 - \eta^2} = B(\varepsilon, \eta). \end{aligned} \tag{53}$$

Together with $R_I^2 \ge 1$ from Theorem 3.1, this proves Equation 52. Under the additional assumptions of Corollary 3.2, the AUROC statement follows from Equation 11 and $R_I \le \sqrt{B(\varepsilon, \eta)}$. □

Remark D.5 (Limiting Cases of the Fisher Decomposition). If the class gap lies in the uniform direction and that direction is covariance-decoupled from its orthogonal complement, then $\chi_I = \theta_I = t_I = 0$ and Equation 7 gives $R_I = 1$. Away from this case, the same equation records the orthogonal mean difference and cross-covariance as separate terms in the ratio.

D.4 Proof of Corollary 3.2

Proof. Let $\mathbf{w} \in \{\mathbf{u}_I, \mathbf{w}_\star\}$ satisfy the Gaussian difference condition. Its nondegenerate Gaussian difference has no atoms, so the tie probability is zero. If $\mathbf{w}^\top \mathbf{d} > 0$, standardising $D_{\mathbf{w}}$ gives

$$\begin{aligned} \mathrm{AUROC}(\mathbf{w}) &= \mathbb{P}(D_{\mathbf{w}} > 0) \\ &= \Phi\left(\frac{\mathbf{w}^\top \mathbf{d}}{\sqrt{2\mathbf{w}^\top \Sigma \mathbf{w}}}\right) \\ &= \Phi\left(\sqrt{\frac{J(\mathbf{w})}{2}}\right). \end{aligned} \tag{54}$$

The assumption $s > 0$ gives the required sign for $\mathbf{u}_I$, while $\mathbf{w}_\star^\top \mathbf{d} = \mathbf{d}^\top \Sigma^{-1} \mathbf{d} > 0$. Since $\sqrt{J(\mathbf{w}_\star)} = R_I \Delta_I$,

$$\mathrm{AUROC}(\mathbf{w}_\star) = \Phi\left(\frac{R_I \Delta_I}{\sqrt{2}}\right), \qquad \mathrm{AUROC}(\mathbf{u}_I) = \Phi\left(\frac{\Delta_I}{\sqrt{2}}\right). \tag{55}$$

Subtracting the two identities gives Equation 10. The monotonicity of Φ, together with $R_I \geq 1$, gives the lower bound in Equation 11. The standard normal density is bounded above by $1/\sqrt{2\pi}$, so Φ is $1/\sqrt{2\pi}$-Lipschitz. Therefore

$$\begin{aligned} \mathrm{AUROC}(\mathbf{w}_\star) - \mathrm{AUROC}(\mathbf{u}_I) &\leq \frac{1}{\sqrt{2\pi}} \frac{(R_I - 1)\Delta_I}{\sqrt{2}} \\ &= \frac{(R_I - 1)\Delta_I}{2\sqrt{\pi}}. \end{aligned} \tag{56}$$

□

D.5 Proof of Corollary 3.3

Proof. Write

$$A_\star := \mathrm{AUROC}(\mathbf{w}_\star), \qquad A_I := \mathrm{AUROC}(\mathbf{u}_I), \tag{57}$$

and

$$P_\star := \Phi\left(\frac{R_I \Delta_I}{\sqrt{2}}\right), \qquad P_I := \Phi\left(\frac{\Delta_I}{\sqrt{2}}\right). \tag{58}$$

The two approximation assumptions imply

$$A_\star - P_\star \leq \kappa_\star, \qquad P_I - A_I \leq \kappa_I. \tag{59}$$

Because Theorem 3.1 gives $R_I \geq 1$ and $\Delta_I \geq 0$, the standard-normal Lipschitz bound used above gives

$$P_\star - P_I \leq \frac{(R_I - 1)\Delta_I}{2\sqrt{\pi}}. \tag{60}$$

Therefore

$$\begin{aligned} A_\star - A_I &= (A_\star - P_\star) + (P_\star - P_I) + (P_I - A_I) \\ &\leq \frac{(R_I - 1)\Delta_I}{2\sqrt{\pi}} + \kappa_\star + \kappa_I, \end{aligned} \tag{61}$$

which is Equation 13. □

Remark D.6 (Scope of the Gaussian assumption). Theorem 3.1 uses only class-conditional second moments and positive definiteness of Σ. Corollary 3.2 adds a one-dimensional Gaussian difference assumption for the two displayed weights. Independent Gaussian class projections are sufficient. The result also allows a non-Gaussian joint score vector, unequal class covariances, and dependent coordinates. Corollary 3.3 replaces the exact Gaussian law by the two explicit AUROC approximation errors.

Table 7: SNR measurements across six models and three dataset families. ρ: depth-gap ratio; γ: offset penalty; R_{plugin}: covariance-aware predicted ratio $\mathrm{SNR}(\hat{\beta})/\mathrm{SNR}(\bar{L})$; R_{emp}: empirically measured ratio. The final column reports $R_{\text{emp}}/R_{\text{plugin}}$.

Model	Dataset	ρ	γ	R_{plugin}	R_{emp}	$R_{\text{emp}}/R_{\text{plugin}}$
Mistral-Small-24B	HaluEval2	0.666	8.50	0.568	0.567	0.999
	TrueFalse	1.212	6.07	0.823	0.824	1.000
	HELM	0.197	6.83	0.188	0.188	1.002
Qwen2.5-32B	HaluEval2	0.945	8.67	0.700	0.706	1.008
	TrueFalse	1.777	6.25	1.017	1.017	1.000
	HELM	0.144	7.90	0.148	0.148	0.999
Qwen2.5-14B	HaluEval2	0.902	8.23	0.669	0.672	1.005
	TrueFalse	1.882	5.87	1.012	1.013	1.000
	HELM	0.203	7.79	0.205	0.206	1.004
Qwen2.5-7B	HaluEval2	0.730	6.70	0.568	0.572	1.006
	TrueFalse	1.982	4.41	1.006	1.006	1.000
	HELM	0.105	6.16	0.102	0.103	1.001
Llama-3.1-8B	HaluEval2	0.525	6.86	0.487	0.486	0.997
	TrueFalse	1.089	5.53	0.827	0.827	1.000
	HELM	0.125	6.68	0.146	0.146	1.002
Llama-2-7B	HaluEval2	0.499	7.02	0.507	0.509	1.003
	TrueFalse	1.368	5.21	0.846	0.847	1.001
	HELM	0.096	6.23	0.113	0.113	1.002

D.6 Empirical Diagnostics

The auxiliary random-intercept expression in Proposition D.2 separates the depth-gap ratio ρ from the offset penalty γ. Observed trajectories can contain dependent innovations, so we also report a covariance plug-in slope-to-mean ratio using the sample covariance of the projected residual trajectories. This auxiliary quantity concerns Proposition D.2 and is distinct from the equal-class pooled covariance used by Theorem 3.1. The displayed values are descriptive full-sample diagnostics. The held-out plug-in Fisher comparison appears in Table 1.

D.7 Estimation Cost of Fisher-Optimal Aggregation

Fisher weighting depends on estimated class means and cross-layer covariance. Covariance inversion can amplify estimation error along directions with small eigenvalues, motivating regularisation (Ledoit and Wolf, 2004). For m layers, the pooled covariance contains $m(m+1)/2$ distinct parameters, or 528 when $m = 32$. Greater depth increases this number, while limited minority-class data constrain covariance estimation.

We examine whether the gain characterised by Theorem 3.1 translates into better held-out prediction by comparing the training-fold diagnostics and held-out performance across training sizes. The depth mean uses the same fitted probes with fixed aggregation weights. Once fitted, both aggregation rules require $\mathcal{O}(m)$ work per example.

Fitting and evaluation. On each outer-training fold, we estimate the two class means and covariances separately, form the equal-class pooled covariance $\widehat{\Sigma} = (\widehat{\Sigma}_0 + \widehat{\Sigma}_1)/2$, and use

$$\widehat{\Sigma}_{\text{reg}} := \widehat{\Sigma} + 10^{-6}\frac{\mathrm{tr}(\widehat{\Sigma})}{m}\mathbf{I}_m, \qquad \widehat{\mathbf{w}}_\star := \widehat{\Sigma}_{\text{reg}}^{-1}\widehat{\mathbf{d}}. \tag{62}$$

The fixed ridge stabilises inversion and is not tuned on AUROC. After the weight vector is fixed, it is oriented from the training fold so that larger values point toward class 1. Let $\widehat{\mu}_{c,\mathbf{w}}$ and $\widehat{\sigma}^2_{c,\mathbf{w}}$ denote the mean and variance of its scalar scores on the held-out fold. When their summed variance

Table 8: Fisher-ratio diagnostics for Table 1. m: number of layers; $N_{\min}$: minority-class training count. $m(m+1)/(2N_{\min})$ measures covariance parameters per minority-class example. $\widehat{R}_I$, $\widehat{\theta}_I$, $\widehat{\chi}_I$, and $\widehat{t}_I$ are training-fold diagnostics in Equation 7. The held-out SNR ratio is Fisher SNR divided by depth-mean SNR. $\widehat{\kappa}$ is the held-out Gaussian approximation error in AUROC points (Equation 63). Values are five-fold means; $\pm$ denotes standard deviation.

Model	Dataset	m	$\frac{m(m+1)}{2N_{\min}}$	$\widehat{R}_I$	Held-out SNR ratio	$\widehat{\kappa}_{\text{mean}}$	$\widehat{\kappa}_{\text{Fisher}}$	$\widehat{\theta}_I$	$\widehat{\chi}_I$	$\widehat{t}_I$
Llama-2-7B	HaluEval2	32	0.41	$1.024_{\pm 0.004}$	0.986	0.25	0.36	0.479	0.696	0.458
	TrueFalse	32	0.21	$1.048_{\pm 0.006}$	1.028	0.83	1.19	0.944	0.998	0.924
	HELM	32	0.44	$1.044_{\pm 0.004}$	1.014	0.33	0.46	0.357	0.573	0.301
Llama-3.1-8B	HaluEval2	32	0.41	$1.043_{\pm 0.001}$	1.011	0.55	0.59	0.469	0.719	0.450
	TrueFalse	32	0.21	$1.050_{\pm 0.002}$	1.036	1.12	1.51	0.957	1.021	0.950
	HELM	32	0.44	$1.040_{\pm 0.004}$	1.001	0.44	0.60	0.366	0.611	0.329
Qwen2.5-7B	HaluEval2	28	0.32	$1.039_{\pm 0.006}$	1.028	0.55	0.65	0.340	0.687	0.368
	TrueFalse	28	0.16	$1.103_{\pm 0.004}$	1.082	1.07	1.59	0.929	1.071	0.930
	HELM	28	0.34	$1.038_{\pm 0.005}$	1.001	0.23	0.41	0.311	0.558	0.273
Qwen2.5-14B	HaluEval2	48	0.92	$1.058_{\pm 0.005}$	0.990	0.67	0.69	0.556	0.861	0.591
	TrueFalse	48	0.47	$1.089_{\pm 0.008}$	1.070	1.01	1.36	0.949	1.052	0.936
	HELM	48	0.98	$1.051_{\pm 0.004}$	1.004	0.37	0.65	0.507	0.729	0.469
Qwen2.5-32B	HaluEval2	64	1.62	$1.071_{\pm 0.010}$	1.012	0.71	0.79	0.646	0.883	0.639
	TrueFalse	64	0.83	$1.118_{\pm 0.004}$	1.091	0.91	1.22	0.956	1.085	0.943
	HELM	64	1.74	$1.062_{\pm 0.006}$	0.993	0.56	0.58	0.498	0.715	0.443
Mistral-Small-24B	HaluEval2	40	0.64	$1.039_{\pm 0.005}$	0.967	0.80	0.72	0.559	0.803	0.562
	TrueFalse	40	0.33	$1.053_{\pm 0.002}$	1.037	1.22	1.45	0.941	1.027	0.944
	HELM	40	0.69	$1.049_{\pm 0.004}$	1.003	0.51	0.81	0.489	0.765	0.487

is positive, we compute

$$\widehat{\Gamma}(\mathbf{w}) := \frac{\widehat{\mu}_{1,\mathbf{w}} - \widehat{\mu}_{0,\mathbf{w}}}{\sqrt{\widehat{\sigma}^2_{1,\mathbf{w}} + \widehat{\sigma}^2_{0,\mathbf{w}}}}, \qquad \widehat{\kappa}(\mathbf{w}) := \left| \widehat{\mathrm{AUROC}}(\mathbf{w}) - \Phi(\widehat{\Gamma}(\mathbf{w})) \right|. \tag{63}$$

The first quantity is the signed pairwise SNR on held-out scores, and the second is the Gaussian AUROC approximation error. Table 8 reports $\widehat{R}_I$, the held-out SNR ratio, $\widehat{\kappa}$, and the decomposition estimates $\widehat{\theta}_I$, $\widehat{\chi}_I$, and $\widehat{t}_I$ on the 18 conditions of Table 1, together with m and the ratio of covariance parameters to minority-class training examples. $\widehat{R}_I$ exceeds the held-out SNR ratio in every condition.

Decomposition diagnostics. All training-fold decomposition diagnostics use the same $\widehat{\Sigma}_{\text{reg}}$ as the fitted Fisher direction, so the identity $\widehat{R}_I^2 = \widehat{\chi}_I^2 + (1 - \widehat{t}_I)^2/(1 - \widehat{\theta}_I)$ holds exactly within each fold. Table 8 reports rounded fold means, which need not preserve this nonlinear identity exactly. $\widehat{t}_I$ reaches between 87 and 96 percent of its Cauchy-Schwarz bound $\sqrt{\widehat{\theta}_I}\,\widehat{\chi}_I$, so the class separation carried by the orthogonal complement is largely redundant with the uniform direction. Within each fold, setting $\widehat{t}_I = 0$ while retaining the measured $\widehat{\theta}_I$ and $\widehat{\chi}_I$ gives condition-level fold-averaged ratios between 1.33 and 4.93, largest on TrueFalse where $\widehat{\theta}_I$ reaches 0.96. This coupling keeps the measured $\widehat{R}_I$ close to one while the orthogonal complement still carries class separation. In the decoupled limit of the remark following the proof of Theorem 3.1, all three quantities vanish, and the observed values place the trajectories away from that limit, while the uniform direction still retains at least 89 percent of the Fisher-optimal pooled SNR at the condition-level training-fold means.

Covariance-shrinkage path. The shrinkage path retains the same fixed ridge and uses $\widehat{\Sigma}_\alpha = (1-\alpha)\widehat{\Sigma} + (\alpha + 10^{-6})\,\mathrm{tr}(\widehat{\Sigma})/m \cdot \mathbf{I}_m$ in place of $\widehat{\Sigma}_{\text{reg}}$ in $\widehat{\mathbf{w}}_\star$ for $\alpha \in \{0, 0.05, \ldots, 1\}$, with the probe logits and outer folds of Table 1. The path starts at the plug-in Fisher score ($\alpha = 0$) and ends at the mean-gap direction $\widehat{\mathbf{d}}$ ($\alpha = 1$). The same path is evaluated at each training fraction in Appendix D.7. Table 9 reports the largest held-out AUROC gain over the depth mean and its selected α at full training size and at one sixteenth, together with the unshrunk $\alpha = 0$ difference at one sixteenth. Across all tested fractions, the largest gain is 0.79 points, and at full training size the gain averages 0.24 points over the 18 conditions. At one sixteenth, the selected α lies between 0.55 and 1.00 and the largest condition-level gain is 0.58 points.

Table 10: Probe-logit scale control. Five-fold mean AUROC in percent. Raw is the deployed depth mean. Standardised first uses the training-fold mean and standard deviation of each layer-wise logit. Probability averages the sigmoid-transformed logits.

	HaluEval2			TrueFalse			HELM		
Model	Raw	Standardised	Probability	Raw	Standardised	Probability	Raw	Standardised	Probability
Llama-2-7B	**80.85**	80.75	80.84	**95.19**	94.43	95.05	87.70	**87.71**	87.64
Llama-3.1-8B	**83.60**	83.42	83.53	**97.30**	96.96	97.15	88.80	88.79	**88.83**
Qwen2.5-7B	**82.03**	81.70	81.92	**97.43**	96.29	97.07	**87.85**	87.69	87.66
Qwen2.5-14B	**83.56**	83.18	83.53	**98.34**	97.89	98.13	88.34	88.40	**88.44**
Qwen2.5-32B	**84.03**	83.58	83.94	**98.51**	98.07	98.31	88.51	**88.69**	88.66
Mistral-Small-24B	84.71	84.59	**84.74**	**98.22**	98.03	98.08	**89.50**	89.49	89.41

Table 9: Covariance shrinkage across training sizes. AUROC of the shrunk Fisher score minus the depth mean, in AUROC points. For each condition and fraction, $\alpha^\star$ maximises the mean difference over $\alpha \in \{0, 0.05, \ldots, 1\}$ on the held-out folds. The final column reports the unshrunk score at one sixteenth of the training fold.

		Full training fold		One sixteenth		
Model	Dataset	Max. gain	$\alpha^\star$	Max. gain	$\alpha^\star$	$\alpha = 0$
Llama-2-7B	HaluEval2	$+0.08$	0.80	$+0.08$	0.90	-4.70
	TrueFalse	$+0.18$	0.30	$+0.14$	0.55	-0.88
	HELM	$+0.34$	0.25	$+0.02$	0.75	-4.06
Llama-3.1-8B	HaluEval2	$+0.44$	0.15	$+0.15$	0.85	-4.01
	TrueFalse	$+0.04$	1.00	$+0.04$	1.00	-0.49
	HELM	$+0.09$	0.40	$+0.01$	0.95	-4.56
Qwen2.5-7B	HaluEval2	$+0.76$	0.15	$+0.49$	0.70	-2.52
	TrueFalse	$+0.19$	0.85	$+0.19$	0.85	-0.16
	HELM	$+0.08$	0.50	$+0.02$	0.90	-3.73
Qwen2.5-14B	HaluEval2	$+0.44$	0.75	$+0.38$	0.90	-6.43
	TrueFalse	$+0.04$	1.00	$+0.04$	1.00	-0.52
	HELM	$+0.25$	0.30	$+0.05$	0.90	-5.82
Qwen2.5-32B	HaluEval2	$+0.79$	0.50	$+0.58$	0.85	-8.50
	TrueFalse	$+0.06$	0.95	$+0.06$	0.90	-0.56
	HELM	$+0.33$	0.30	$+0.08$	0.85	-9.23
Mistral-Small-24B	HaluEval2	$+0.09$	0.90	$+0.08$	1.00	-5.40
	TrueFalse	$+0.04$	0.95	$+0.04$	0.95	-0.36
	HELM	$+0.08$	0.50	$+0.01$	0.95	-5.16

Training-size sweep. For each outer fold, the probes are fitted on the full training fold and the trajectories are held fixed. The training trajectories are subsampled per class without replacement at fractions $1, \frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}$, with five draws per fraction averaged within the fold and then across the five folds. Only $\widehat{\mathbf{d}}$ and $\widehat{\Sigma}$ are re-estimated, with the ridge of Table 1 and no shrinkage, and the depth mean is unchanged. Figure 2 reports the held-out AUROC difference and the held-out SNR ratio against the parameter ratio $m(m+1)/(2N_{\min})$. The difference is non-increasing along the sweep in all 18 conditions, and the held-out ratio decreases and the training-fold $\widehat{R}_I$ increases in every condition. At full training size, the estimated coupling saturation $|\widehat{t}_I|/(\sqrt{\widehat{\theta}_I\,\widehat{\chi}_I})$ lies between 0.87 and 0.96. It decreases after every successive subsampling step in all 18 conditions, so subsampling moves the plug-in decomposition away from the Cauchy-Schwarz bound.

D.8 Location Estimator Ablation: Mean vs. Median

Theorem 3.1 concerns linear scores. A complementary question compares the depth-wise median with the depth mean under possible outlying layers.

Gaussian reference. Consider the centred scores $X_l := L_l - \mu_{l,Y} = \alpha + \xi_l$ in Equation 25. Conditional on either class, assume $\xi_0, \ldots, \xi_{m-1}$ are iid $\mathcal{N}(0, \sigma_\xi^2)$ with $\sigma_\xi^2 > 0$ and independent of α. Let $\bar{X}$ and $\widetilde{X}$ denote the mean and median of these centred scores. Both estimate the common

Table 11: Increment-feature group ablation on Llama-3.1-8B (AUROC $\times 100$, averaged over five runs). Full uses all 13 statistics; Layer-level, Increment, Update, and Late layer retain only the indicated groups defined in Appendix C.

	Full	Layer-level	Increment	Update	Late layer
HaluEval2	**83.4**	83.3	82.3	65.3	82.0
TrueFalse	97.1	**97.2**	96.7	88.5	96.0
HELM	**88.2**	87.7	86.0	62.4	86.8

residual location α, with

$$\mathrm{Var}(\bar{X} \mid \alpha, Y) = \frac{\sigma_\xi^2}{m}, \qquad \mathrm{Var}(\widetilde{X} \mid \alpha, Y) \sim \frac{\pi \sigma_\xi^2}{2m} \quad (m \to \infty). \tag{64}$$

The median therefore has asymptotic relative efficiency $2/\pi \approx 0.637$ for estimating this location under Gaussian residuals. The empirical ablation below tests whether replacing the mean with the median improves ranking on the original probe logits.

Empirical comparison. Across three models and three benchmarks, the median and 10%-trimmed mean remain within 0.3 AUROC points of the depth mean. The depth mean matches or exceeds the median in all nine conditions, and their score correlation exceeds 0.99 in eight conditions (0.989 in the remaining condition).

D.9 Trajectory and Increment-Feature Controls

Table 11 separates the 13 increment-based statistics into four groups. The layer-level group lies within 0.5 percentage points of the full representation in the three displayed conditions, while the update-distribution group has lower AUROC in isolation.

D.10 Depth Subsampling Ablation

The random-intercept model (Remark D.1) associates depth averaging with an innovation-variance term of order $\sigma_\xi^2/m_{\mathrm{eff}}$, where m_{eff} is the support size under the independent-innovation idealisation. The ablation examines the corresponding performance pattern through uniformly spaced supports and two contiguous half-depth controls.

Subsampling Protocol. For $m_{\mathrm{eff}} \in \{2, 4, 8, 16, m\}$, the uniformly spaced support is

$$I^{\mathrm{unif}}_{m_{\mathrm{eff}}} := \left\{ \left\lfloor \frac{j(m-1)}{m_{\mathrm{eff}} - 1} \right\rfloor : j = 0, \ldots, m_{\mathrm{eff}} - 1 \right\} \subseteq \mathcal{K}_m. \tag{65}$$

The two location controls use $I^{\mathrm{first}} := \{0, \ldots, \lfloor m/2 \rfloor - 1\}$ and $I^{\mathrm{second}} := \{\lfloor m/2 \rfloor, \ldots, m-1\}$. Each support mean is fitted with the same scalar logistic calibration used by the depth-mean detector and evaluated under five-fold cross-validation. The figure displays $m_{\mathrm{eff}} \in \{2, 4, 8, m\}$ together with the two half-depth controls.

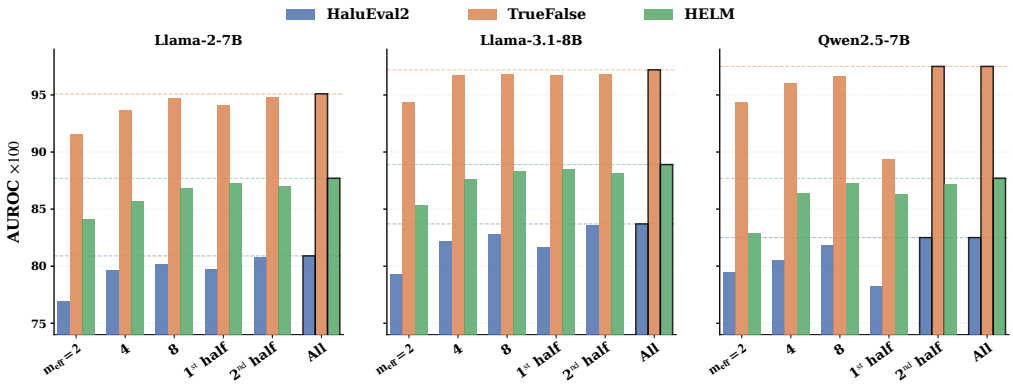


Figure 3: Depth-support ablation. AUROC for uniformly spaced supports, the first and second depth halves, and the full support across three models and three datasets.

Key Observations. Figure 3 reports the results across three models (Llama-2-7B, Llama-3.1-8B, Qwen2.5-7B) and three datasets (HaluEval2, TrueFalse, HELM). Among the half-depth controls, the second half is stronger for Qwen2.5-7B, while the two Llama models have smaller half-to-half gaps.

Consistency with the Variance Reduction Prediction. Under the independent-innovation idealisation, the innovation standard deviation decreases as $\sigma_\xi/\sqrt{m_{\text{eff}}}$. The observed diminishing change after eight uniformly spaced layers is compatible with that directional prediction.

E Scope and Societal Considerations

E.1 Scope of Pre-Decoding Prediction

Prediction unit and empirical scope. We evaluate truthfulness prediction from answer-onset states on the four benchmarks. Token-level error accumulation, attention drift during long generation, and cross-turn context degradation call for signals from the generated sequence or later interaction states.

E.2 Length Control

A classifier using statement length alone reaches 55.5 AUROC on HaluEval2, 52.3 on TrueFalse, and 52.8 on HELM, at least 25 points below the depth mean of every evaluated model. Lengths use a fixed tokenizer on the raw statement, and the classifier is a balanced L_2 logistic fit with training-fold standardisation. For TruthfulQA, the score is computed before the first answer token, so response length is unavailable to the detector by construction.

E.3 Societal Impacts

Potential Benefits. The design offers two practical benefits.

1. **Early risk scoring.** An answer-onset score can support gating, fallback, or additional checking before generation begins. Its use is most direct in systems that already expose internal activations.
2. **Low aggregation overhead.** After the probes at each layer are fitted, the depth mean adds linear time and constant streaming memory in the number of layers. Probe construction and logistic calibration remain supervised offline steps.

Potential Risks. The same scope and access assumptions create three deployment risks.

1. **Error localisation.** An answer-onset score estimates risk for the response as a whole. Locating and correcting individual factual errors requires checking the generated claims.
2. **Distribution and shortcut sensitivity.** Training-fitted probes can change under topic, model, or dataset shift and can use correlates of the benchmark labels. Deployment therefore requires calibration on the intended data distribution and monitoring after shifts.
3. **Adaptive evasion.** White-box disclosure of internal scoring rules may enable prompts or model modifications that reduce the measured trajectory signature. Adaptive attacks remain an unevaluated deployment condition.

HalluTracer supplies an early risk signal within a broader reliability pipeline.