Reading Calibrated Uncertainty from Language Model Trajectories

Aliai Eusebi [1] **Alexander Herzog** [1 2] **Xiaoyu Liang** [1] **Marie Vasek** [1] **Enrico Mariconti** [1] **Lorenzo Cavallaro** [1]

Abstract

The maximum softmax probability (MSP) represents a default approach when evaluating uncertainty quantification for language model generation with structured output. Although cheap, it is often miscalibrated. Methods that probe the model's internal activations feed raw hidden states into opaque classifiers, reading activations as static snapshots and leaving implicit the layer-wise trajectory by which a representation is formed. Yet, similar endpoints can arise from very different paths, and how evidence accumulates, reinforces, or reverses across depth might reveal uncertainty that final probabilities obscure. We extract eleven scale-invariant geometric features, tracing the cumulative path of per-layer MLP updates, and feed them to a sparse linear probe. The probe outperforms MSP under selective abstention, with gains scaling with baseline miscalibration up to 21 AURC points. Because every feature has a closed-form geometric meaning, the probe's coefficients trace *how* and *where* along depth errors take shape – which layers commit prematurely, which contradict the running state, where trajectories drift away from their endpoint.

1. Introduction

A trustworthy language model should know when it might be wrong. Calibrated uncertainty quantification (UQ) captures this requirement: a model's confidence should reflect its likelihood of being correct. Two equally accurate models can nevertheless differ substantially in reliability. If a model assigns low confidence to most of its errors, uncertain predictions can be deferred, abstained from, or escalated for human review. By contrast, when confidence is poorly aligned with correctness, high-confidence errors become indistinguishable from reliable predictions. In clinical triage, the first model would set its uncertain cases aside for a clinician to review; the second would queue a life-threatening case behind a routine one, both stamped high-risk with equally strong confidence. Calibrated uncertainty therefore serves a triage function, enabling uncertainty to guide which predictions warrant intervention or verification.

The default approach to UQ in discrete choice settings is the Maximum Softmax Probability (MSP) (Vashurin et al., 2025; Dakhmouche et al., 2025), which uses the predicted token's softmax probability as a confidence score. MSP incurs no additional computational cost and is often surprisingly competitive. However, it may inherit the well-known pathology of miscalibration: confidence scores that fail to reflect the true likelihood of correctness, often remaining high even when the prediction is wrong (Guo et al., 2017). A parallel line of work reads from the model's activations directly, on the premise that the internal computation leading to a generation carries information about its truthfulness that the output distribution alone does not. Azaria & Mitchell (2023) demonstrated that a simple classifier trained on hidden activations can predict whether an LLM's answer is truthful, and a growing body of work has since traced the contours of this "geometry of truth" (Li et al., 2023b; Marks & Tegmark, 2023; Dakhmouche et al., 2025; Liu et al., 2024; Beigi et al., 2024; Azizian et al., 2025).

Yet across this literature, activations are typically read as static snapshots – a hidden state extracted from one layer, or averaged across layers, and analyzed for the information it contains. This discards the layer-wise trajectory by which the representation is formed. A final hidden state is the endpoint of a path through representation space, and similar endpoints may arise from qualitatively different trajectories. Some representations may develop steadily, others may emerge late, fluctuate, or be partially reversed before settling. These trajectories are not incidental to the prediction as they encode how evidence is accumulated, reinforced, attenuated, or revised across depth. We therefore read activations not only as frozen states, but as representational trajectories; we show that their geometry reveals uncertainty that final probabilities alone may obscure.

We calibrate by tracing the answer-position residual stream

[1]University College London, London, United Kingdom [2]CORE64, London, United Kingdom. Correspondence to: Aliai Eusebi <aliai.eusebi.16@ucl.ac.uk>, Alexander Herzog <alexander.herzog.23@ucl.ac.uk>.

Proceedings of the 43rd International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026. Copyright 2026 by the author(s).

as the cumulative path induced by per-layer MLP write-vectors during the forward pass. We summarize this trajectory by computing scale-invariant geometric features and feed them to a sparse linear probe. Confidence-calibration is evaluated under selective abstention via AURC (Geifman et al., 2018). This leads to an interpretable UQ method that goes beyond single-point evaluation (MSP) while avoiding the opacity of dense probes on raw activations.

To summarize, our main contributions are as follows:

1. We propose a compact set of geometric features that describe how representations evolve across network-depth, and feed them to a sparse linear probe that outperforms MSP under selective abstention, with gains scaling with the baseline's miscalibration.
2. We consider end to end interpretability of the probe: each feature has a closed-form geometric meaning, and its coefficients reveal not only *whether* the model is likely to err but *also how*, which layers commit prematurely, which contradict the running state, and where trajectories drift away from their endpoint. We show that correct and wrong predictions sharing the same MSP score leave different trajectory signatures, exposing information otherwise flattened by the output distribution.
3. We perform a comprehensive set of empirical experiments across 9 instruction-tuned LLMs from three model families (Qwen, Llama, DeepSeek) spanning 3B to 72B parameters on five representative natural language processing tasks.

2. Background

2.1. Uncertainty Quantification

Predictive uncertainty is commonly decomposed into two components: epistemic uncertainty, which stems from limited data or model misspecification and in principle reducible, and aleatoric uncertainty, which reflects the intrinsic stochasticity of the data-generating process and therefore irreducible (Kendall & Gal, 2017). A related dichotomy has emerged in the LLM literature: factual uncertainty pertains to the correctness of generated content with respect to ground-truth knowledge, whereas semantic uncertainty arises from the multiplicity of valid continuations admitted by a prompt; the former is epistemic in nature, the latter aleatoric (Liu et al., 2025). As factual uncertainty is the primary concern in automated decision-making (Dakhmouche et al., 2025), we restrict our analysis to this component. We employ multiple-choice questions, which act as a noise-reduction mechanism that isolates factual gaps by neutralizing the semantic uncertainty found in open-ended text (Li et al., 2026).

UQ methods aim to associate each prediction with a scalar score indicative of its correctness, enabling downstream decisions such as abstention, deferral, or selective generation. Existing approaches for LLM uncertainty estimation fall into three groups that differ primarily in computational cost. Single-sample methods derive uncertainty from a single forward pass, using signals such as maximum token log-probability (Manakul et al., 2023), perplexity (Margatina et al., 2023), and entropy (Kadavath et al., 2022; Kuhn et al., 2023). Multi-sample methods aggregate signals across multiple generations, scoring uncertainty by their consistency, similarity, or variability. Representative examples include semantic (Farquhar et al., 2024) and predictive (Kadavath et al., 2022) entropy, conformal prediction (Kumar et al., 2023), and pairwise similarity methods (Lin et al., 2023). Probing-based methods instead train lightweight predictors over internal activations to infer uncertainty directly from the model's hidden representations (Azaria & Mitchell, 2023; Dakhmouche et al., 2025; Liu et al., 2024).

2.2. Selective Classification

Selective classification (Chow, 1970; Geifman & El-Yaniv, 2017) augments a predictor with the option to abstain, trading coverage for reduced error on the retained inputs. Formally, a selective classifier is a pair (f, g), where $f : \mathcal{X} \to \mathcal{Y}$ is a predictor and $g : \mathcal{X} \to \{0, 1\}$ is a gating function: the prediction $f(x)$ is returned when $g(x) = 1$ and withheld otherwise. In practice, g is induced by thresholding a confidence score $\kappa : \mathcal{X} \to \mathbb{R}$, so that performance reduces to the quality of κ as a ranking of inputs by likely correctness.

Two quantities characterize such a classifier: *coverage*, the fraction of inputs on which f commits to a prediction, and *selective risk*, the average loss on those inputs. Varying the threshold on κ traces the risk-coverage curve (Geifman et al., 2019), whose area (AURC) summarizes performance across all operating points. Hence, selective classification provides a natural testbed for uncertainty estimators: a reliable confidence signal should report low selective risk across coverage levels (Ding et al., 2020).

3. Methodology

We capture a language model's uncertainty by considering trajectory-information during sequential layer-wise processing. At each layer in a transformer architecture, the MLP contributes to the residual stream. The cumulative sum of these contributions forms a trajectory through representation space. We hypothesize that the geometry of these trajectories, including the distribution of update magnitudes across depth, local changes in direction, and the efficiency with which the path approaches its endpoint, contains information about uncertainty (Figure 1). We summarize trajectory geometry using eleven scalar descriptors, which we combine

FEATURE	FORMULA	INTERPRETATION
DEPTH DISTRIBUTION		
Relative update magnitude	$\|m_\ell\|/\mu$	Per-layer update size vs. the mean
Cumulative path fraction	$T^{-1}\sum_{k\le\ell}\|m_k\|$	Front- vs. back-loading of computation
LOCAL TRAJECTORY SHAPE		
Consecutive cosine	$\cos(m_{\ell-1},\, m_\ell)$	Directional consistency of updates
Curvature	$1-\cos(m_{\ell-1},\, m_\ell)$	Abrupt changes of direction
Update-state alignment	$\cos(m_\ell,\, s_{\ell-1})$	Reinforcement vs. contradiction with prior state
RELATION TO ENDPOINT $\hat{s}_L = s_L/\|s_L\|$		
Direction to final	$\cos(s_\ell,\, s_L)$	State convergence to endpoint
Update to final	$\cos(m_\ell,\, s_L)$	Update alignment with endpoint direction
Signed final support[†]	$\langle m_\ell,\, \hat{s}_L\rangle/\|m_\ell\|$	Cosine alignment, magnitude-aware framing
Contradictory support	$\max\big(0,\, -\langle m_\ell,\, \hat{s}_L\rangle/\|m_\ell\|\big)$	Magnitude of opposition to endpoint
Orthogonal mass fraction	$\sqrt{1-\cos^2(m_\ell,\, s_L)}$	Fraction of update orthogonal to endpoint
EFFICIENCY		
Cumulative coherence	$\|s_\ell\|/\sum_{k\le\ell}\|m_k\|$	Net displacement vs. path length

Table 1. Eleven per-layer trajectory features grouped by what they measure. Symbols are defined in the text. Features involving $m_{\ell-1}$ or $s_{\ell-1}$ are undefined at $\ell = 1$ and assigned conventional values (1 for *Consecutive cosine*, 0 for *Curvature* and *Update-state alignment*). [†] *Signed final support* is numerically identical to *Update to final*; both rows are retained because they correspond to distinct geometric formulations (the L1 probe is invariant to this).

with the model's MSP in a sparse linear model.

The remainder of this section describes the representational substrate (§3.1), the extracted trajectory features (§3.2), and the sparse linear probe (§3.3). Code to reproduce the experiments is available at `https://anonymous.4open.science/r/uq-motion-66CC/`. All experiments were conducted on a server equipped with an NVIDIA H100 NVL GPU.

3.1. Setup and Substrate

We consider finite-choice classification with a candidate set $\mathcal{Y}$, $|\mathcal{Y}| = K$. Given a prompt $x \in \mathcal{X}$, the model f_θ induces a distribution $p_\theta(y \mid x)$ over $\mathcal{Y}$ by applying a softmax to the next-token logits of the K candidate-identifying tokens. From a single forward pass, we estimate an uncertainty score $u : \mathcal{X} \to [0, 1]$ corresponding to the probability of error. We use the canonical MSP-based score $1 - p_{\mathrm{msp}}(x)$, where $p_{\mathrm{msp}}(x) = \max_{y\in\mathcal{Y}} p_\theta(y \mid x)$, as the baseline (Hendrycks & Gimpel, 2016). We write $\hat{y}(x) = \arg\max_{y\in\mathcal{Y}} p_\theta(y \mid x)$ for the prediction and $e(x) = \mathbf{1}\big[\hat{y}(x) \neq y\big]$ for the error indicator, where y denotes the ground-truth answer.

As the basis for u we extract a layer-indexed sequence of MLP write-vectors at the final prompt position, the readout position used to compute the next-token logits over $\mathcal{Y}$. Each transformer block ℓ contains a multi-layer perceptron (MLP) sub-layer that writes a contribution $m_\ell(x) \in \mathbb{R}^H$ to the residual stream, where H is the model's hidden dimension; we capture $m_\ell(x)$ for each block $\ell = 1, \ldots, L$ via forward hooks placed on the MLP sub-module, recording its output at the final token of the prompt.

We focus on MLP write-vectors because prior mechanistic interpretability work identifies transformer MLPs as important sites for factual knowledge storage and recall (Geva et al., 2021; Meng et al., 2022; Yu et al., 2024). This motivates using MLP write-vectors as the unit of analysis for studying how predictions are assembled across layers. We also consider the partial sums

$$s_\ell(x) = \sum_{k\le\ell} m_k(x), \tag{1}$$

where $m_k(x)$ is the layer-k MLP write-vector. The sequence $\{s_\ell(x)\}_{\ell=1}^{L}$ traces a discrete trajectory in residual-stream space, with $s_L(x)$ the total MLP-driven displacement over the forward pass and $\hat{s}_L(x) = s_L(x)/\|s_L(x)\|$ its unit direction. We write $\mu = L^{-1}\sum_k \|m_k\|$ for the mean update norm and $T = \sum_k \|m_k\|$ for the total path length.

3.2. Trajectory Features

We describe each trajectory $(m_\ell, s_\ell)_{\ell=1}^{L}$ via eleven layer-wise scalar features (Table 1). The features are scale-invariant, enabling comparison across models with different hidden dimensions, and fall into four geometric groups:

(G1) Depth allocation: how computational effort is distributed across layers;

(G2) Local shape: the local geometry of the trajectory;

(G3) Endpoint alignment: how each layer relates to the trajectory's endpoint; and

(G4) Trajectory efficiency: how directly the trajectory reaches its endpoint.

G1 captures the depth distribution of update magnitude. The relative update magnitude flags layers whose contribution is disproportionately large relative to the layer-wise average, while the cumulative path fraction reveals whether the model front- or back-loads its computation.

G2 characterizes local trajectory shape. The consecutive cosine measures the directional consistency of successive updates, high values indicate a smooth path through representation space, while the curvature highlights abrupt changes of direction. The update-state alignment distinguishes layers that reinforce prior computation from those that contradict it.

G3 relates each layer to the trajectory's endpoint. Two cosines track convergence: the direction to final measures whether the cumulative state aligns with the endpoint, and the update to final measures whether individual updates point toward it. The signed final support quantifies an update's alignment with the final direction. The negative part of the *update to final*, the contradictory support, isolates layers that actively oppose the endpoint, while the orthogonal mass fraction captures the share of an update's magnitude orthogonal to the final direction.

Finally, G4 measures efficiency via the cumulative coherence, the ratio of net displacement to total path length: it equals one for a perfectly straight trajectory and approaches zero when updates cancel rather than accumulate.

3.3. Sparse Linear Probe

We model the uncertainty score via sparse logistic regression:

$$u(x) = \sigma\big(w^\top z(x) + b\big)\,, \tag{2}$$

$$z(x) = \big[\,\varphi(x),\ p_{\mathrm{msp}}(x)\,\varphi(x)\,\big], \tag{3}$$

where $\varphi(x) \in \mathbb{R}^{L\times 10}$ stacks the layer-wise features of Section 3.2 and $p_{\mathrm{msp}}(x)$ is the MSP of the model's prediction. The interaction term lets the probe weight each feature by confidence: curvature under a high-confidence prediction is a stronger error signal than the same curvature when the model is already uncertain. We minimize class-balanced binary cross-entropy against the error indicator $e(x)$ of Section 3.1 under an elastic-net penalty (Zou & Hastie, 2005); sparsity in the fitted w then identifies the predictive features directly, with no separate selection stage.

Each dataset is split 65/15/20 into train, validation, and test folds, stratified by the error indicator so that the error rate is matched across folds. We sweep a 64-point grid over the regularization strength C and ℓ_1 mixing ratio ρ (Appendix B), selecting the setting with lowest validation AURC. The selected hyperparameters are refit on train+validation before evaluation on the held-out test fold.

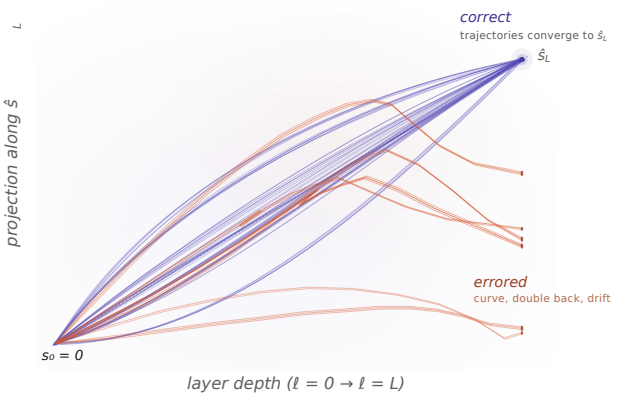

Figure 1. Cumulative MLP write-vectors traced layer-wise. The trajectory-geometry separates the two populations. Correct trajectories converge to $\hat{s}_L$; errored trajectories curve, double back, and drift.

4. Experimental Setup

Models. We conduct the experiments using 9 instruction-tuned LLMs from three model families, spanning scales from 3B to 72B parameters: Qwen (Qwen2.5-7B-Instruct, Qwen2.5-14B-Instruct, Qwen2-72B-Instruct, Qwen2.5-72B-Instruct) (Yang et al., 2024), Llama (Llama-3.2-3B-Instruct, Llama-3.1-8B-Instruct, Llama-3.3-70B-Instruct) (Grattafiori et al., 2024), and DeepSeek (deepseek-llm-7b-chat, deepseek-llm-67b-chat) (Bi et al., 2024).

Datasets. We use five benchmark datasets adapted from Ye et al. (2024) spanning question answering (MMLU; Hendrycks et al., 2020), reading comprehension (CosmosQA; Huang et al., 2019), commonsense inference (HellaSwag; Zellers et al., 2019), dialogue response selection (HaluDial; Li et al., 2023a), and document summarization (HaluSum; Li et al., 2023a). Each dataset contains 10,000 instances formatted as four-option (A–D) multiple-choice questions in a zero-shot setting, using a base prompt that presents the question and options directly with the prefix `Answer:`. The predicted answer is the option with maximum softmax probability (MSP), which also serves as our primary baseline for uncertainty estimation.

Metrics. We evaluate the quality of our probe's uncertainty profile using AURC (Geifman et al., 2019). In essence, a reliable uncertainty profile should allow the model to abstain on observations it is likely to get wrong, so that errors decrease as we restrict predictions to the most

	MMLU		CosmosQA		HellaSwag		HaluDial		HaluSum	
Model	Ours	Δ	Ours	Δ	Ours	Δ	Ours	Δ	Ours	Δ
Llama-3.2-3B	28.68	▲21.83	6.83	▲0.83	31.99	▲1.61	38.17	▲11.45	34.38	▲16.54
Llama-3.1-8B	22.65	▲3.82	3.58	▲0.27	26.87	▲3.27	22.48	▲6.59	21.13	▲21.06
Llama-3.3-70B	14.86	▲8.74	1.41	▲0.37	3.00	▲1.71	7.99	▲5.41	10.10	▲20.17
Qwen2.5-7B	14.87	▲1.65	3.94	▼0.18	6.59	▲2.07	19.74	▲4.28	20.67	▲9.65
Qwen2.5-14B	10.86	▲0.23	2.24	▲0.16	2.75	▲0.38	12.05	▲3.94	9.87	▲16.25
Qwen2-72B	6.10	▲2.21	1.27	▲0.42	1.75	▲1.15	7.29	▲4.45	10.84	▲16.42
Qwen2.5-72B	5.29	▲1.96	1.06	▲0.79	1.82	▲2.75	9.46	▲7.50	8.82	▲16.74
DeepSeek-7B	33.15	▲1.55	7.47	▼0.16	20.65	▲5.32	27.21	▲6.58	29.30	▲14.29
DeepSeek-67B	11.83	▼0.20	1.50	▼0.01	2.75	▲0.28	9.38	▲5.35	11.26	▲20.18

Table 2. We report AURC (×100; lower is better) for our method, together with the absolute reduction $\Delta = \mathrm{MSP} - \mathrm{Ours}$ relative to the MSP baseline. Positive Δ (▲) denotes improvement; negative Δ (▼) denotes regression. Full comparisons against the Ceiling and Ablation experiments are reported in Appendix A.

confident observations.

Let f be a predictor with a confidence function κ, evaluated on $\{(x_i, y_i)\}_{i=1}^{n}$ under the $0/1$ loss, and re-index the samples so that $\kappa(x_1) \geq \cdots \geq \kappa(x_n)$. The selective risk at coverage $c \in (0, 1]$ is the average loss on the top $\lceil cn \rceil$ samples, defined as

$$R(c) \;=\; \frac{1}{\lceil cn \rceil} \sum_{i=1}^{\lceil cn \rceil} \mathbf{1}\big[f(x_i) \neq y_i\big], \tag{4}$$

where AURC averages $R(c)$ over all coverage levels:

$$\mathrm{AURC}(f, \kappa) \;=\; \frac{1}{n} \sum_{k=1}^{n} R\big(\tfrac{k}{n}\big)\,. \tag{5}$$

Lower AURC indicates that the confidence function more effectively separates correct from incorrect predictions, assigning higher confidence to the former. In confidence-stratified analyses (Figure 3) we additionally report AUROC, the area under the ROC curve for discriminating correct from incorrect predictions within each confidence bin.

Relation to calibration. Expected Calibration Error (ECE) (Naeini et al., 2015) assesses whether confidence values match empirical accuracy, but calibration and selective performance are distinct (Ding et al., 2020). A constant score equal to the empirical accuracy can achieve ECE $= 0$ while providing no useful ranking for abstention. Conversely, a score can induce a near-optimal ranking for selective prediction while being numerically miscalibrated. Since AURC directly evaluates the risk-coverage trade-off determined by the ordering induced by κ, we use it as our primary metric.

Baselines. We compare our probe with two baselines and one ablation. The first baseline is maximum softmax probability (MSP), the probability $p_{\mathrm{msp}}(x)$ assigned to the predicted token, which we convert to the uncertainty score $1 - p_{\mathrm{msp}}(x)$. MSP requires no auxiliary computation beyond the model forward pass and is a strong standard baseline for confidence estimation (Vashurin et al., 2025; Dakhmouche et al., 2025). The second baseline is a raw-activation probe in the spirit of Azaria & Mitchell (2023), extended to all layers: we replace our eleven trajectory features with the per-layer hidden states and train a layer-weighted MLP on binary error labels. This provides a high-capacity reference for the error signal recoverable from activations, without the interpretability constraints of our feature-based probe. We refer to this baseline as the *ceiling*. The *trajectory-only* ablation restricts the input to $\varphi(x)$, removing MSP and its interaction terms; its gap to the full probe quantifies the complementarity between trajectory geometry and MSP.

5. Results

We compare the probe with the baselines and ablation from Section 4, reporting headline results in Table 2 and full results in Appendix Table 3. We then assess its uncertainty profile and the geometric signals it uses across depth.

Probe performance. Our probe improves over MSP on 41 of 45 model-dataset pairs, with gains largest where MSP performs worst (Spearman $r_s = 0.78$). For the five configurations with MSP AURC above 40, it reduces AURC by 11.45–21.83 points, with the largest reductions on Llama-3.2-3B MMLU, Llama-3.1-8B HaluSum, and Llama-3.2-3B HaluSum. The probe also matches or outperforms the ceiling on 24 configurations and is within 2 AURC points on another 6, despite using only eleven scalar features instead of full hidden states. Finally, the *trajectory-only* ablation outperforms MSP on 35 configurations, showing that trajectory geometry is independently informative, while the full probe further benefits from MSP as a complementary signal.

Figure 2 dissects these AURC gains into risk-coverage

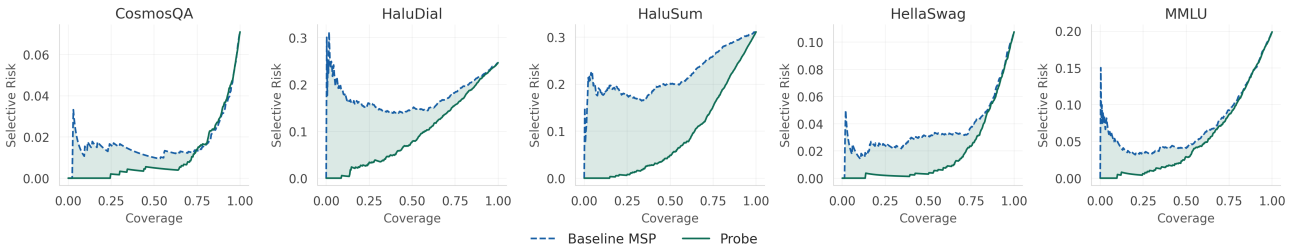

Figure 2. Risk-coverage curves for Qwen2.5-72B across the five evaluation datasets. An ideal curve is monotonically non-decreasing as predictions are rejected in order of decreasing confidence, approaching zero risk at low coverage and the base error rate at full coverage. The MSP baseline is shown in dashed blue and our probe in solid green; the shaded region between them indicates the probe's gain.

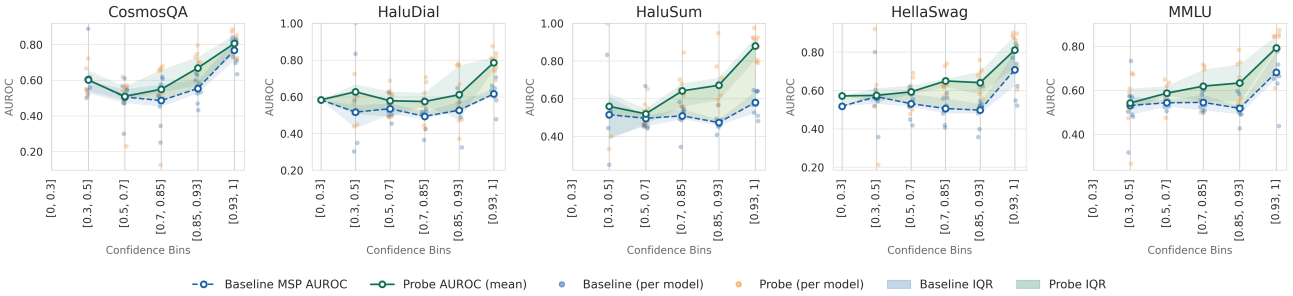

Figure 3. Probe and MSP AUROC across confidence bins, aggregated over models. Lines show mean AUROC, bands the interquartile range, and dots per-model values. Green shading indicates bins where the probe outperforms MSP.

curves for Qwen2.5-72B. The probe (solid green) lies at or below MSP (dashed blue) throughout the coverage range, with the largest gap on HaluDial and HaluSum. MSP exhibits an unstable low-coverage spike on every dataset, whereas the probe rises smoothly from zero and sustains a low-risk regime up to roughly 50–60% coverage.

Figure 3 stratifies AUROC by MSP confidence. Across all five datasets, MSP performs close to random-chance levels (AUROC ≈ 0.5–0.6) outside its highest-confidence bin. In contrast, the probe consistently outperforms MSP across nearly all bins, with the largest improvements emerging in the mid and high-confidence regimes for HaluSum, HaluDial, HellaSwag, and MMLU.

To examine whether the probe finds geometric signal at fixed MSP, Figure 4 compares two Qwen2.5-14B predictions on MMLU with MSP ≈ 0.97 – one correct, one incorrect – via the eleven trajectory features z-scored against the per-layer population. In early layers (depth 0–0.3), the incorrect prediction has already aligned with its final direction (*Direction to final* and *Update to final* saturated at $z \approx +3$), while the correct prediction shows none of this early alignment. The correct prediction also front-loads its computation (*Cumulative path fraction* above the population mean at 0–0.5), whereas the incorrect prediction back-loads its path length and traces an unusually straight cumulative path in the second half (*Cumulative coherence* positive at depths 0.55–0.85).

Figure 4. Two Qwen2.5-14B predictions on MMLU at MSP ≈ 0.97, one correct and one incorrect. Layer-wise z-scores of the eleven trajectory features across normalized depth for the correct prediction (top) and incorrect (bottom).

Probe interpretability. We examine what signal the probe uses by decomposing its total coefficient mass across the four feature groups in Table 1, separating mass assigned to raw trajectory features $\varphi(x)$ from that assigned to their MSP interactions, $p_{\mathrm{msp}} \cdot \varphi(x)$. Figure 5 reports this decomposition by model, aggregated across datasets since the pattern is stable; solid bars indicate trajectory features and hatched bars indicate interactions. The two blocks contribute roughly equally overall. Trajectory features carry a larger share for Llama-3.3-70B, Qwen2.5-14B, Qwen2-72B, and Qwen2.5-72B, whereas DeepSeek-7B and Llama-3.1-8B assign relatively less mass to depth-distribution features. Depth distribution is also the only feature group whose mass lies mainly on the raw features rather than on MSP interactions; the remaining groups split their mass more evenly.

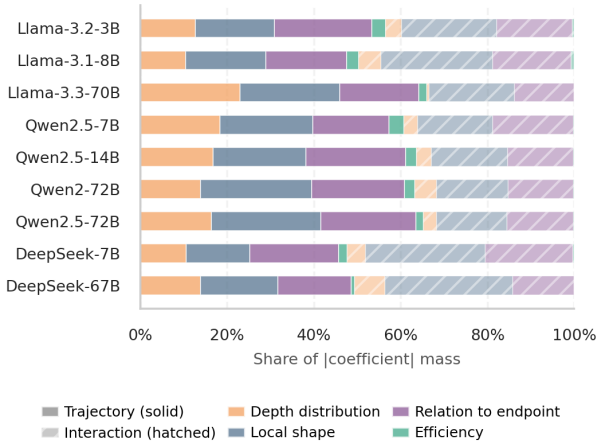

Figure 5. Probe coefficient composition across models, showing coefficient mass by geometric feature family (color), split into main effects (solid) and MSP×trajectory interactions (hatched).

Figure 6 maps the median probe coefficient at each (feature, depth) cell for Qwen2.5-14B on HaluSum and CosmosQA, with direct effects $\varphi(x)$ on the left and MSP interactions on the right. The two datasets show distinct signatures. On HaluSum, signal concentrates in the final quarter (depth $\gtrsim 0.75$): errors carry a large, smooth late update-positive weight on *Relative update magnitude*, *Consecutive cosine*, and *Cumulative path fraction* near the output – counterbalanced by negative *Update-state alignment* and *Cumulative coherence*, indicating that updates aligned with the running state and efficient cumulative paths predict correctness rather than error. *Direction to final* is also negative at depth 0.5–0.6, so trajectories that align with their endpoint earlier are less error-prone. HaluSum errors reflect premature commitment followed by a late, state-breaking correction. CosmosQA shows no such localization: coefficients are smaller and spread across depth, the strongest cell at mid-depth (*Orthogonal mass fraction* at 0.5–0.6), so error trajectories drift sideways relative to the endpoint. Endpoint-alignment features (*Update to final*, *Signed final support*, *Direction to final*) light up in the MSP-interaction panel near the output, predicting error only when weighted by confidence.

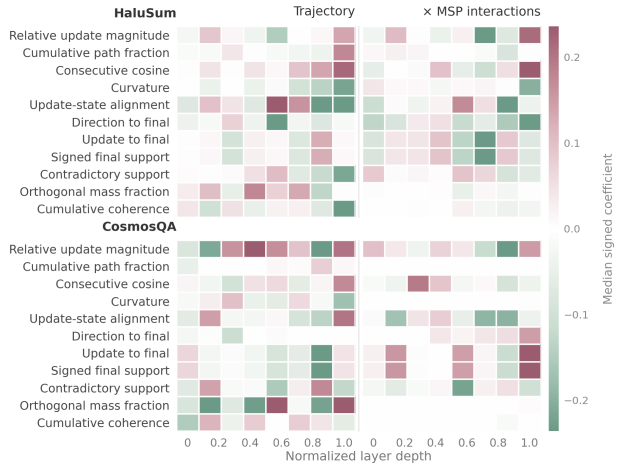

Figure 6. Median probe coefficients per (feature, depth-bin) cell for Qwen2.5-14B on HaluSum (top) and CosmosQA (bottom). Left: direct effects; right: MSP interactions. Pink increases predicted error probability, green decreases it. Values are normalized by the dataset-specific 95th-percentile magnitude.

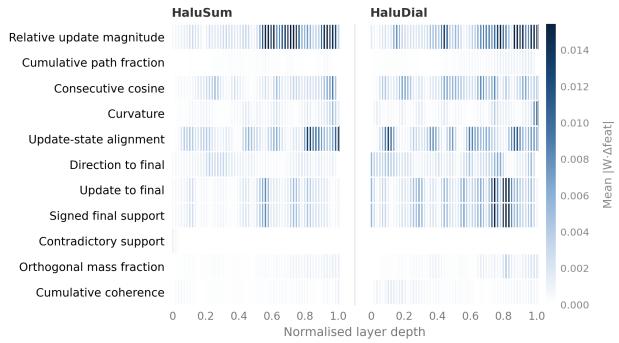

Figure 7. Aggregate attribution maps for Llama-3.2-3B-Instruct on HaluSum (left) and HaluDial (right), averaged over the 100 MSP-matched probe-correct pairs with the largest error-probability gaps. Maps share a normalized-depth axis and color scale.

The coefficient maps in Figure 6 describe the probe's learned structure; to see it deployed, we turn to the pairs on which MSP confidence is silent (Figure 7). Within the top 30% most-confident Llama-3.2-3B-Instruct predictions, we match each probe-flagged error to a probe-cleared non-error with MSP agreeing within 0.02. For each (feature, depth) cell, we multiply the feature-value difference between the paired examples by the probe's effective coefficient at the shared confidence (so both trajectory and MSP-interaction blocks contribute), then average absolute values across pairs. A cell is bright only when trajectories diverge there *and* the

probe weights it, localizing signal MSP alone misses. On both datasets the signal concentrates in the second half of the network, with *Relative update magnitude* dominant – engaging from depth ~ 0.5 on HaluSum but only near depth 1.0 on HaluDial. HaluSum additionally leans on *Update-state alignment* near the output, while HaluDial shifts to the endpoint-alignment channel (*Update to final*, *Signed final support*) at depth 0.7–0.85.

6. Discussion

We find that the trajectory of a model's layer-wise computation reveals when its answer should be trusted. Correct predictions and confident errors can be indistinguishable in the output distribution yet diverge in the geometry that produced them, and a few interpretable descriptors of that geometry suffice to recover the distinction. A sparse linear probe trained on them improves over MSP across nearly all configurations, with gains largest where MSP is most miscalibrated; on more than half of those configurations, the eleven scalar features match a probe trained on the the full hidden states. The resulting selective prediction function is well-behaved: the probe traces a smooth, monotone risk-coverage curve, sustaining low risk where MSP tends to spike at low coverage. Because the features are interpretable by construction, the probe's coefficients localize *how* a model fails, not just *whether* it does. On HaluSum, errors take the form of premature endpoint commitment in the first half of the network, followed by a late, oversized MLP write that breaks the running state. On CosmosQA, errors instead manifest as mid-depth drift orthogonal to the eventual answer direction. These mechanisms remain legible even when MSP fails to discriminate: among predictions matched in confidence, correct and erroneous trajectories diverge at specific (feature, depth) cells the probe weights, localizing discriminative structure that the output distribution has flattened away. The signatures, however, are task-specific, in line with the recent finding that "geometries of truth" are largely orthogonal across tasks (Azizian et al., 2025). Together, these results point to a practical path to trustworthy selective prediction: the features add negligible cost at inference, and their interpretability makes failures auditable rather than opaque. **Limitations.** Our analysis is restricted to discrete-choice settings, where the prediction is localized to a single answer position. The probe is also fit per (model, task), and preliminary leave-one-dataset-out experiments show gains over MSP on only a subset of configurations. Identifying trajectory features that transfer across models and tasks remains open.

7. Related Work

Probing assesses what information is linearly encoded in neural network representations (Alain & Bengio, 2016) and language models (Petroni et al., 2019; Azaria & Mitchell, 2023): a classifier is trained on activations extracted from a frozen model to predict a target property. Azaria & Mitchell (2023) introduced this approach for UQ in LLMs, showing that internal representations contain reliable signals of output correctness. Subsequent work on the "geometry of truth" suggests a direction in model internals along which truthful and erroneous outputs are approximately linearly separable (Azizian et al., 2025; Li et al., 2023b; Marks & Tegmark, 2023; Kossen et al., 2024; Dakhmouche et al., 2025; Liu et al., 2024; Beigi et al., 2024; Burns et al., 2022), differing mainly in how this direction is recovered. Li et al. (2023b) train linear probes on individual attention heads for detection and inference-time steering. Burns et al. (2022) propose Contrast Consistent Search (CCS), recovering the direction without labels by enforcing consistency between a statement and its negation. Dakhmouche et al. (2025) take a Bayesian route, fitting layer-to-layer linear maps and combining their posterior log-likelihoods via sparse regression. In contrast, we extract trajectory geometry rather than classifying individual hidden states: eleven interpretable scalar descriptors of how MLP contributions accumulate across depth, fed to a sparse linear probe whose inputs and decision rule are interpretable end to end.

8. Conclusion

We introduced a method for reading uncertainty from the motion of language model computation. By tracing the answer-position residual stream as a cumulative path of per-layer MLP write-vectors and condensing it through eleven scale-invariant geometric descriptors, we obtain an uncertainty signal whose inputs and decision rule are inspectable end to end. Each feature has a direct geometric reading, computed in closed form from the activations with no auxiliary model, and the sparse probe trained on them localizes not only *whether* a model is likely to err but *how* – which layers commit prematurely, which contradict the running state, where trajectories drift away from their endpoint. Treating activations as motion rather than as snapshots offers a tractable middle path between the opacity of dense probes and the single scalar of MSP, recovering signal from how a prediction is assembled, not only from where it lands. We hope this work encourages the development of simpler, more interpretable representations for uncertainty quantification in LLMs.

Impact Statement

UQ is a core component of trustworthy AI, enabling LLMs to flag predictions that should be deferred, reviewed, or abstained from in high-stakes settings. Our method reads uncertainty from the geometry of MLP write-vectors to the residual stream within a single forward pass, offering an

efficient and interpretable alternative to costly sampling-based or ensemble approaches. Because each feature has a closed-form geometric meaning, the resulting confidence signal is auditable end-to-end, allowing practitioners to inspect not only whether a model is uncertain but where in its computation that uncertainty arises.

References

Alain, G. and Bengio, Y. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2016.

Azaria, A. and Mitchell, T. The internal state of an llm knows when it's lying. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 967–976, 2023.

Azizian, W., Kirchhof, M., Ndiaye, E., Bethune, L., Klein, M., Ablin, P., and Cuturi, M. The geometries of truth are orthogonal across tasks. *arXiv preprint arXiv:2506.08572*, 2025.

Beigi, M., Shen, Y., Yang, R., Lin, Z., Wang, Q., Mohan, A., He, J., Jin, M., Lu, C.-T., and Huang, L. Internalinspector i2: Robust confidence estimation in llms through internal states. In *Findings of the association for computational linguistics: EMNLP 2024*, pp. 12847–12865, 2024.

Bi, X., Chen, D., Chen, G., Chen, S., Dai, D., Deng, C., Ding, H., Dong, K., Du, Q., Fu, Z., et al. Deepseek llm: Scaling open-source language models with longtermism. *arXiv preprint arXiv:2401.02954*, 2024.

Burns, C., Ye, H., Klein, D., and Steinhardt, J. Discovering latent knowledge in language models without supervision. *arXiv preprint arXiv:2212.03827*, 2022.

Chow, C. On optimum recognition error and reject tradeoff. *IEEE Transactions on information theory*, 16(1):41–46, 1970.

Dakhmouche, R., Letellier, A., and Gorji, H. Can linear probes measure llm uncertainty? *arXiv preprint arXiv:2510.04108*, 2025.

Ding, Y., Liu, J., Xiong, J., and Shi, Y. Revisiting the evaluation of uncertainty estimation and its application to explore model complexity-uncertainty trade-off. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 4–5, 2020.

Farquhar, S., Kossen, J., Kuhn, L., and Gal, Y. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630, 2024.

Geifman, Y. and El-Yaniv, R. Selective classification for deep neural networks. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pp. 4885–4894, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

Geifman, Y., Uziel, G., and El-Yaniv, R. Bias-reduced uncertainty estimation for deep neural classifiers. *arXiv preprint arXiv:1805.08206*, 2018.

Geifman, Y., Uziel, G., and El-Yaniv, R. Bias-reduced uncertainty estimation for deep neural classifiers. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=SJfb5jCqKm.

Geva, M., Schuster, R., Berant, J., and Levy, O. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 5484–5495, 2021.

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, pp. 1321–1330. JMLR.org, 2017.

Hendrycks, D. and Gimpel, K. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*, 2016.

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., and Steinhardt, J. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

Huang, L., Le Bras, R., Bhagavatula, C., and Choi, Y. Cosmos qa: Machine reading comprehension with contextual commonsense reasoning. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pp. 2391–2401, 2019.

Kadavath, S., Conerly, T., Askell, A., Henighan, T., Drain, D., Perez, E., Schiefer, N., Hatfield-Dodds, Z., DasSarma, N., Tran-Johnson, E., et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

Kendall, A. and Gal, Y. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems*, 30, 2017.

Kossen, J., Han, J., Razzak, M., Schut, L., Malik, S., and Gal, Y. Semantic entropy probes: Robust and cheap hallucination detection in llms. *arXiv preprint arXiv:2406.15927*, 2024.

Kuhn, L., Gal, Y., and Farquhar, S. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. *arXiv preprint arXiv:2302.09664*, 2023.

Kumar, B., Lu, C., Gupta, G., Palepu, A., Bellamy, D., Raskar, R., and Beam, A. Conformal prediction with large language models for multi-choice question answering. *arXiv preprint arXiv:2305.18404*, 2023.

Li, J., Cheng, X., Zhao, W. X., Nie, J.-Y., and Wen, J.-R. Halueval: A large-scale hallucination evaluation benchmark for large language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 6449–6464, 2023a.

Li, K., Patel, O., Viégas, F., Pfister, H., and Wattenberg, M. Inference-time intervention: Eliciting truthful answers from a language model. *Advances in Neural Information Processing Systems*, 36:41451–41530, 2023b.

Li, X., Yu, Z., Zhang, Z., Zhuang, Y., Shah, S., Sadagopan, N., and Beniwal, A. Semantic volume: Quantifying and detecting both external and internal uncertainty in llms. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 31751–31759, 2026.

Lin, Z., Trivedi, S., and Sun, J. Generating with confidence: Uncertainty quantification for black-box large language models. *arXiv preprint arXiv:2305.19187*, 2023.

Liu, L., Pan, Y., Li, X., and Chen, G. Uncertainty estimation and quantification for llms: A simple supervised approach. *arXiv preprint arXiv:2404.15993*, 2024.

Liu, X., Chen, T., Da, L., Chen, C., Lin, Z., and Wei, H. Uncertainty quantification and confidence calibration in large language models: A survey. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pp. 6107–6117, 2025.

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

Margatina, K., Schick, T., Aletras, N., and Dwivedi-Yu, J. Active learning principles for in-context learning with large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 5011–5034, 2023.

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

Meng, K., Bau, D., Andonian, A., and Belinkov, Y. Locating and editing factual associations in gpt. *Advances in neural information processing systems*, 35:17359–17372, 2022.

Naeini, M. P., Cooper, G., and Hauskrecht, M. Obtaining well calibrated probabilities using bayesian binning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 29, 2015.

Petroni, F., Rocktäschel, T., Riedel, S., Lewis, P., Bakhtin, A., Wu, Y., and Miller, A. Language models as knowledge bases? In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pp. 2463–2473, 2019.

Vashurin, R., Fadeeva, E., Vazhentsev, A., Rvanova, L., Vasilev, D., Tsvigun, A., Petrakov, S., Xing, R., Sadallah, A., Grishchenkov, K., et al. Benchmarking uncertainty quantification methods for large language models with lm-polygraph. *Transactions of the Association for Computational Linguistics*, 13:220–248, 2025.

Yang, Q. A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Li, C., Liu, D., Huang, F., Dong, G., Wei, H., Lin, H., Yang, J., Tu, J., Zhang, J., Yang, J., Yang, J., Zhou, J., Lin, J., Dang, K., Lu, K., Bao, K., Yang, K., Yu, L., Li, M., Xue, M., Zhang, P., Zhu, Q., Men, R., Lin, R., Li, T., Xia, T., Ren, X., Ren, X., Fan, Y., Su, Y., Zhang, Y.-C., Wan, Y., Liu, Y., Cui, Z., Zhang, Z., Qiu, Z., Quan, S., and Wang, Z. Qwen2.5 technical report. *ArXiv*, abs/2412.15115, 2024. URL https://api.semanticscholar.org/CorpusID:274859421.

Ye, F., Yang, M., Pang, J., Wang, L., Wong, D. F., Yilmaz, E., Shi, S., and Tu, Z. Benchmarking llms via uncertainty quantification. *Advances in Neural Information Processing Systems*, 37:15356–15385, 2024.

Yu, L., Cao, M., Cheung, J. C., and Dong, Y. Mechanistic understanding and mitigation of language model non-factual hallucinations. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 7943–7956, 2024.

Zellers, R., Holtzman, A., Bisk, Y., Farhadi, A., and Choi, Y. Hellaswag: Can a machine really finish your sentence? In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pp. 4791–4800, 2019.

Zou, H. and Hastie, T. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 67(2):301–320, 2005.

A. Full AURC Results

Table 2 in the main paper reports our method's AURC alongside the absolute improvement over the MSP baseline. For completeness, Table 3 below provides the full breakdown across all four methods we consider.

		AURC ↓			
Model	**Dataset**	MSP	Ceiling	Ablation	Ours
Llama-3.2-3B	MMLU	50.51	29.58	30.22	28.68
	CosmosQA	7.66	7.14	8.24	6.83
	HellaSwag	33.60	32.86	36.20	31.99
	HaluDial	49.62	35.31	39.94	38.17
	HaluSum	50.92	23.12	34.97	34.38
Llama-3.1-8B	MMLU	26.47	24.64	25.62	22.65
	CosmosQA	3.85	3.73	4.46	3.58
	HellaSwag	30.14	25.51	30.45	26.87
	HaluDial	29.07	19.93	23.98	22.48
	HaluSum	42.19	13.02	21.77	21.13
Qwen2.5-7B	MMLU	16.52	15.17	16.06	14.87
	CosmosQA	3.76	3.62	4.71	3.94
	HellaSwag	8.66	5.96	7.16	6.59
	HaluDial	24.02	13.65	19.73	19.74
	HaluSum	30.32	10.60	20.94	20.67
DeepSeek-7B	MMLU	34.70	31.57	33.83	33.15
	CosmosQA	7.31	6.86	8.04	7.47
	HellaSwag	25.97	18.42	20.94	20.65
	HaluDial	33.79	22.38	27.91	27.21
	HaluSum	43.59	14.84	29.85	29.30
Qwen2.5-14B	MMLU	11.09	10.74	11.54	10.86
	CosmosQA	2.40	1.91	2.56	2.24
	HellaSwag	3.13	2.72	2.94	2.75
	HaluDial	15.99	9.74	12.58	12.05
	HaluSum	26.12	6.47	9.87	9.87
DeepSeek-67B	MMLU	11.63	11.51	12.35	11.83
	CosmosQA	1.49	1.53	1.78	1.50
	HellaSwag	3.03	2.71	2.95	2.75
	HaluDial	14.73	7.81	9.69	9.38
	HaluSum	31.44	8.49	11.54	11.26
Llama-3.3-70B	MMLU	23.60	14.51	15.53	14.86
	CosmosQA	1.78	1.27	1.66	1.41
	HellaSwag	4.71	3.00	3.30	3.00
	HaluDial	13.40	6.80	8.30	7.99
	HaluSum	30.27	7.18	10.15	10.10
Qwen2-72B	MMLU	8.31	5.78	6.30	6.10
	CosmosQA	1.69	1.11	1.39	1.27
	HellaSwag	2.90	1.86	1.80	1.75
	HaluDial	11.74	6.26	7.36	7.29
	HaluSum	27.26	7.40	10.88	10.84
Qwen2.5-72B	MMLU	7.25	5.91	5.53	5.29
	CosmosQA	1.85	0.99	1.09	1.06
	HellaSwag	4.57	1.76	1.82	1.82
	HaluDial	16.96	7.83	9.64	9.46
	HaluSum	25.56	5.74	8.81	8.82

Table 3. Full AURC results ($\times 100$; lower is better). We compare our full probe (**ours**) against the MSP baseline, a raw-activations **ceiling** probe that operates on per-layer hidden states without capacity constraints, and a trajectory-only **ablation** that removes MSP and its interactions from the probe input.

B. Probe Hyperparameters

We fit the sparse linear probe of Section 3.3 using scikit-learn's `LogisticRegression` with the `saga` solver, class-balanced sample weighting, a convergence tolerance of 10^{-3}, and a maximum of 8000 iterations. The 64 configurations swept during model selection comprise 10 pure-ℓ_1 candidates with $C \in \{3{\times}10^{-4},\ 10^{-3},\ 3{\times}10^{-3},\ 10^{-2},\ 3{\times}10^{-2},\ 10^{-1},\ 3{\times}10^{-1},\ 1,\ 3,\ 10\}$, and 54 elastic-net candidates over $C \in \{10^{-3},\ 3\times10^{-3},\ 10^{-2},\ 3\times10^{-2},\ 10^{-1},\ 3\times10^{-1},\ 1,\ 3,\ 10\}$ and ℓ_1 mixing ratio $\rho \in \{0.10,\ 0.25,\ 0.50,\ 0.75,\ 0.90,\ 0.95\}$. All randomness in fitting and splitting is controlled by a fixed seed of 42.