BEYOND HIDDEN STATES: COMPONENT-AWARE CONFIDENCE ESTIMATION FOR LLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models can produce fluent but incorrect answers, with confidence signals that often fail to reflect empirical correctness. Existing estimators rely on output-level signals sensitive to decoding and prompts, or probe hidden states as generic high-dimensional features, underutilizing the functional structure of Transformer components. We propose a component-aware confidence estimator that predicts the correctness of a frozen model's answer from three mechanistic descriptors extracted in a single forward pass: Attention Sink Signature, capturing selected heads' sink behavior; Knowledge Response Alignment, measuring directional agreement between attention and feed forward residual contributions; and Residual Trajectory Geometry, summarizing residual-stream dynamics through normalized distance and angular statistics. A lightweight branch-attentive probe fuses these descriptors against correctness labels with the backbone frozen. Across five question answering benchmarks and four models from the Qwen2.5 and LLaMA-3.1 families, our method achieves the best AUROC and ECE on most combinations and transfers more reliably than internal-state baselines under cross-domain evaluation. These results show that organizing confidence estimation around component-level structure yields more discriminative and transferable signals than treating internal states as a single high-dimensional feature.

1 INTRODUCTION

Large language models (LLMs), such as GPT-4 (Achiam et al., 2023), LLaMA-3 (Dubey et al., 2024), and Qwen2.5 (Yang et al., 2024), have achieved strong performance across a wide range of language understanding and generation tasks. However, they can still produce fluent yet incorrect outputs, often with confidence signals that do not reflect empirical correctness (Kadavath et al., 2022; Tian et al., 2023; Xiong et al., 2024). This mismatch between apparent confidence and actual correctness undermines the reliability of LLM predictions, especially in downstream applications where erroneous outputs may incur real costs. Therefore, a confidence estimator that faithfully estimates the likelihood that an LLM's output is correct is essential for the safe and trustworthy deployment of LLMs.

Prior work on confidence estimation in LLMs generally falls into two categories: output-based methods and internal-state-based methods. Output-based methods use signals from the generated output, including token probabilities (Kadavath et al., 2022), agreement across sampled responses (Manakul et al., 2023; Farquhar et al., 2024), and verbalized self-evaluation (Lin et al., 2022; Tian et al., 2023). However, these signals are sensitive to decoding strategies, prompt formats, and the number of sampled responses. This motivates recent work that examines LLM internal states, showing that hidden representations contain correctness-related information beyond output-level signals (Azaria & Mitchell, 2023; Burns et al., 2023; Chen et al., 2024; Kossen et al., 2024; Beigi et al., 2024; Xiao et al., 2026). However, most of these methods use the hidden state as a single high-dimensional feature, rather than separating signals from different Transformer components. Consequently, finer-grained evidence from attention and feed-forward modules remains underused.

Recent advances in mechanistic interpretability provide a complementary perspective on this gap. Multi-head attention and feed-forward networks play distinct functional roles: attention routes contextual information across token positions, while feed-forward modules retrieve and integrate parametric knowledge stored in the model parameters (Elhage et al., 2021; Geva et al., 2021; Dai et al.,

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.

2022; Ferrando et al., 2024). Building on this view, recent studies extract reliability-related signals from component-level behaviors, either through attention patterns (Chuang et al., 2024; Yuksekgonul et al., 2024; Vazhentsev et al., 2026) or through decomposed attention and FFN contributions to the residual stream (Sun et al., 2025; Zhang et al., 2025). However, these approaches mainly target hallucination detection in retrieval-grounded or open-ended generation settings, and typically focus on a single component family in isolation. How to combine complementary component-level signals into a unified confidence estimator for general-purpose QA remains largely underexplored.

To address this gap, we extract three compact, complementary descriptors from a single forward pass of the frozen LLM. We summarize attention behavior using per-head attention sink values (Xiao et al., 2024), measure the alignment between MHA and FFN updates using per-layer cosine similarity, and characterize residual-stream evolution using two trajectory statistics that are stable under rescaling and rotations of the representation space. A lightweight branch-attentive probe then adaptively combines these descriptors to produce the final confidence score.

In this work, our contributions are summarized as follows:

- We propose a component-aware confidence estimator for frozen LLMs, which predicts answer correctness from structured Transformer signals extracted in a single prompt forward pass.
- We introduce three mechanistic descriptors, namely Attention Sink Signature, Knowledge Response Alignment, and Residual Trajectory Geometry, fused by a lightweight branch-attentive probe.
- Across five QA benchmarks and four LLM backbones, our estimator outperforms output-level and internal-state baselines on most settings and transfers more reliably across domains.

2 Related Work

Output-Level Confidence Estimation for LLMs. Existing confidence estimation methods for LLMs often rely on output-level signals, such as token probabilities, verbalized confidence, self-evaluation prompts, or consistency across sampled generations (Kadavath et al., 2022; Lin et al., 2022; Tian et al., 2023; Manakul et al., 2023; Farquhar et al., 2024). These approaches are broadly applicable because they require limited or no access to model internals. However, their estimates are tied to generated outputs and can be sensitive to prompting formats, decoding strategies, sampling budgets, or output-distribution shifts. This motivates confidence estimation methods that examine the internal computation underlying a prediction.

Representation-Level Internal Confidence Estimation. A growing line of work probes internal representations for truthfulness, hallucination, and answer correctness. Hidden-state-based methods show that LLM internals encode reliability-related information (Azaria & Mitchell, 2023; Burns et al., 2023; Chen et al., 2024; Kossen et al., 2024), and EAGLE (Xiao et al., 2026) aggregates internal beliefs across intermediate layers. InternalInspector (Beigi et al., 2024) pushes this direction further by jointly using activations, attention states, and feed-forward states across all layers as the predictive feature for QA confidence estimation. These studies confirm that internal computation carries useful confidence signals. However, they consume internal states as a single high-dimensional input, with feature size scaling with hidden dimension, layer count, or even total parameter count, and without distinguishing the functional roles of different Transformer components. As a result, correctness-related evidence is mixed with a large amount of unrelated variation in the input, and the resulting probes inherit a strong dependence on the backbone-specific representation space.

Mechanistic and Component-Level Reliability Signals. Mechanistic interpretability analyzes how Transformer components contribute to model predictions. Under the residual-stream view, multi-head attention and feed-forward networks write separable updates into the hidden representation (Elhage et al., 2021; Ferrando et al., 2024), with attention routing contextual information and feed-forward modules storing parametric knowledge (Geva et al., 2021; Dai et al., 2022). Recent studies derive uncertainty or hallucination signals from attention patterns (Chuang et al., 2024; Yuksekgonul et al., 2024; Vazhentsev et al., 2026) or decomposed component contributions to the residual stream (Sun et al., 2025; Zhang et al., 2025), but mainly target retrieval-augmented generation. Inspired by this structural view, we extend the component-aware perspective to QA confidence estimation through an estimator capturing attention behavior, sublayer interaction, and residual-stream dynamics.

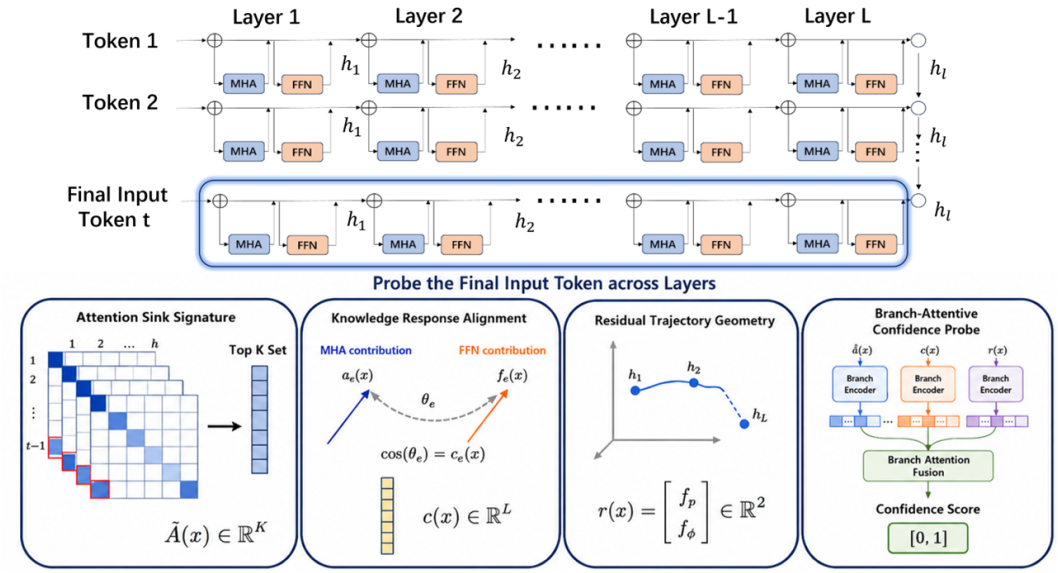

Figure 1: Overview of the proposed component-aware confidence estimator.

3 Preliminaries and Problem Formulation

We consider a frozen decoder-only language model M with L layers, H attention heads per layer, and hidden dimension d. Each layer contains multi-head attention (MHA) and a feed-forward network (FFN). Let h_i^ℓ denote the hidden state of token i at layer ℓ. For simplicity, we leave out normalization details. The hidden state update is:

$$h_i^\ell = h_i^{\ell-1} + \underbrace{\mathrm{MHA}^\ell(h_i^{\ell-1})}_{a_i^\ell} + \underbrace{\mathrm{FFN}^\ell(h_i^{\ell-1} + a_i^\ell)}_{f_i^\ell}. \tag{1}$$

MHA combines contextual information across tokens (Ferrando et al., 2024; Elhage et al., 2021). FFNs apply nonlinear transformations to individual tokens and are viewed as parametric memory (Geva et al., 2021; Dai et al., 2022). This structure allows separate analysis of the two components.

Confidence Estimation. For each input x_j, the frozen model M produces an output y_j under a fixed decoding protocol. The benchmark's official evaluation rule assigns a correctness label $c_j \in \{0, 1\}$. Given $\mathcal{D} = \{(x_j, y_j, c_j)\}_{j=1}^N$, we train an estimator to predict a correctness probability $\hat{c}_j \in [0, 1]$ from internal signals θ_j. We extract these signals at the final input-token position across layers. This requires one forward pass over x_j before generation. Only the estimator parameters ϕ are optimized. Prior work supports predicting response correctness from prompt representations (Dong et al., 2025).

4 Methodology

4.1 Attention Sink Signature

The attention sublayer is the primary mechanism by which a transformer block gathers contextual information into the current residual stream, and prior work has shown that its behavior is informative of model reliability (Chuang et al., 2024; Yuksekgonul et al., 2024). A natural summary is the entropy of the attention distribution, but its value is structurally coupled to the input length through the softmax support, making it not directly comparable across prompts of different lengths. We instead build on the attention sink phenomenon (Xiao et al., 2024), in which a large fraction of attention mass is consistently routed to a fixed initial position regardless of the prompt, providing a length-robust, model-intrinsic summary of attention behavior. We thus use the attention sublayer as our first probing

site and extract a correctness signal in two steps: per-head sink extraction followed by head selection based on class separability, yielding a feature we call the *Attention Sink Signature*.

Per-Head Sink Extraction. Let $\alpha_{i,j}^{(l,h)}(x) \in [0,1]$ denote the softmax attention weight assigned by head h of layer l from query position i to key position j on input x. At the last input token position t and sink target position 0 (Xiao et al., 2024), the per-head sink value is defined as:

$$A^{(l,h)}(x) \;=\; \alpha_{t,\,0}^{(l,h)}(x). \tag{2}$$

Stacking these values over all layers and heads yields the full per-head sink descriptor:

$$\mathbf{A}(x) \;=\; \big[\, A^{(l,h)}(x) \,\big]_{l\in[L],\, h\in[H]} \;\in\; \mathbb{R}^{L\times H}. \tag{3}$$

Class-Separability Scoring. Directly using all $L \cdot H$ entries of $\mathbf{A}(x)$ dilutes the signal, since most heads do not carry correctness-discriminative information. To localize the informative subset, we score each head by the Cohen's d effect size of its sink value between correct and wrong samples on the training set:

$$d^{(l,h)} \;=\; \frac{\mu_c^{(l,h)} - \mu_w^{(l,h)}}{\sqrt{\big(\sigma_c^{(l,h)\,2} + \sigma_w^{(l,h)\,2}\big)/2}}, \tag{4}$$

where μ_c, σ_c^2 and μ_w, σ_w^2 are the mean and variance of $A^{(l,h)}$ computed over the correct and wrong subset of $\mathcal{D}_{\text{train}}$. Cohen's d provides a scale-normalized measure of class separability, making head scores comparable across layers and heads.

Top-K Head Selection. We retain the top-K heads with the largest $|d^{(l,h)}|$ and index the selected set as:

$$\mathcal{S}_K \;=\; \text{Top-}K\Big(\big\{\, |d^{(l,h)}| \,\big\}_{l\in[L],\, h\in[H]}\Big). \tag{5}$$

The final Attention Sink Signature is then the compact descriptor: $\tilde{\mathbf{A}}(x) \;=\; \big[\, A^{(l,h)}(x) \,\big]_{(l,h)\in\mathcal{S}_K} \;\in\; \mathbb{R}^K$. $\mathcal{S}_K$ is computed once on $\mathcal{D}_{\text{train}}$ and kept fixed during both the training and evaluation of g_ϕ, so it receives no gradient from test samples.

4.2 Knowledge Response Alignment

The FFN has been associated with parametric knowledge retrieval and, together with the attention sublayer, jointly writes into the residual stream at each transformer block (Elhage et al., 2021). We probe the interaction between these two sublayers by measuring the directional alignment of their residual contributions at each layer, and refer to the resulting feature as the *Knowledge Response Alignment*. Following the residual-stream view, the output of transformer layer l at the last input token position t decomposes as:

$$h_l(x) \;=\; h_{l-1}(x) \;+\; \mathbf{a}_l(x) \;+\; \mathbf{f}_l(x), \tag{6}$$

where $\mathbf{a}_l(x) = \text{Attn}_l\big(h_{l-1}(x)\big)_t$ is the attention contribution and $\mathbf{f}_l(x) = \text{FFN}_l\big(h_{l-1}(x) + \mathbf{a}_l(x)\big)_t$ is the FFN contribution, both taken at position t. We define the per-layer alignment as the cosine similarity between these two contributions:

$$c_l(x) \;=\; \frac{\langle \mathbf{a}_l(x),\, \mathbf{f}_l(x)\rangle}{\|\mathbf{a}_l(x)\|\,\|\mathbf{f}_l(x)\|} \;\in\; [-1,1]. \tag{7}$$

Concatenating c_l across all L layers yields the Knowledge Response Alignment descriptor: $\mathbf{c}(x) = \big[\, c_1(x),\, c_2(x),\, \ldots,\, c_L(x) \,\big] \;\in\; \mathbb{R}^L$. $\mathbf{c}(x)$ captures, for each layer, whether the context retrieved by the attention sublayer and the parametric knowledge produced by the FFN sublayer push the residual stream in a consistent direction.

4.3 Residual Trajectory Geometry

The preceding signals focus on individual components within a transformer block. Here we take a coarser, model-level view by examining the geometry of the entire residual stream trajectory across all blocks. A common choice is to directly use the raw hidden states (Beigi et al., 2024), but h_l is high-dimensional and model-specific, with scale and orientation not comparable across models

or layers. Even aggregating hidden states updates yields an L-dimensional descriptor still coupled to model depth. We instead propose *Residual Trajectory Geometry*, a descriptor that characterizes the residual-stream evolution using two normalized geometric statistics capturing magnitude and directional changes across layers.

At the final input token position t, let $h_l(x) \in \mathbb{R}^d$ denote the residual state after the first l transformer blocks, with $h_0(x)$ representing the state before the first block and $h_L(x)$ representing the state after the final block. This gives a layerwise trajectory $H(x) = \{h_0(x), h_1(x), \ldots, h_L(x)\}$. We define the state curvature as the average adjacent-state displacement normalized by the overall displacement between the initial and final states:

$$f_\rho(x) = \frac{1}{L} \sum_{l=0}^{L-1} \frac{\|h_{l+1}(x) - h_l(x)\|_2}{\|h_L(x) - h_0(x)\|_2 + \epsilon}. \tag{8}$$

Similarly, we define the angle curvature as the average adjacent-state angular change normalized by the overall angular change between the initial and final states:

$$f_\phi(x) = \frac{1}{L} \sum_{l=0}^{L-1} \frac{\arccos\left(\frac{\langle h_l(x), h_{l+1}(x)\rangle}{\|h_l(x)\|_2 \|h_{l+1}(x)\|_2 + \epsilon}\right)}{\arccos\left(\frac{\langle h_0(x), h_L(x)\rangle}{\|h_0(x)\|_2 \|h_L(x)\|_2 + \epsilon}\right) + \epsilon}. \tag{9}$$

The state curvature reflects how strongly the residual representation moves between adjacent layers relative to its endpoint displacement, while the angle curvature reflects how strongly its direction changes relative to the overall directional shift of the trajectory. Together, they provide a compact trajectory-level signal that complements the Attention Sink Signature and the alignment between MHA and FFN residual updates. We stack them into $r(x) = [f_\rho(x), f_\phi(x)] \in \mathbb{R}^2$.

4.4 Branch-Attentive Confidence Head

The three descriptors $\tilde{\mathbf{A}}(x) \in \mathbb{R}^K$, $\mathbf{c}(x) \in \mathbb{R}^L$, and $\mathbf{r}(x) \in \mathbb{R}^2$ are heterogeneous in scale, meaning, and dimensionality, and we therefore process each in its own subspace rather than collapsing them into a shared feature space at the input. The confidence head g_ϕ consists of three branch encoders $\{e_b\}_{b \in \{\tilde{A}, c, r\}}$, each mapping its input to a common hidden dimension h. The branch representations $u_b(x) = e_b(\cdot)$ are fused through a learned attention over branches,

$$\alpha_b(x) \;=\; \frac{\exp(w^\top u_b(x))}{\sum_{b'} \exp(w^\top u_{b'}(x))}, \qquad u(x) \;=\; \sum_b \alpha_b(x)\, u_b(x), \tag{10}$$

with a single scoring vector $w \in \mathbb{R}^h$. The fused representation $u(x)$ is passed to a two-layer classifier producing the confidence score $\hat{c}_j = \sigma(g_\phi(\theta_j))$. The probe is trained with binary cross-entropy against the correctness labels in $\mathcal{D}_{\text{train}}$ while the LLM remains frozen.

4.5 Theoretical Analysis

We state four elementary properties that support the design choices of our component-aware estimator. Proofs of all propositions are given in Appendix A. **Setup.** Let $(x, C) \sim \mathcal{D}$ with correctness label $C \in \{0, 1\}$. At the final input-token position, the layer-l residual state is written as $h_l(x) = h_{l-1}(x) + a_l(x) + f_l(x)$, where $a_l(x)$ and $f_l(x)$ are the MHA and FFN contributions, respectively. We use $\tilde{A}(x)$, $\mathbf{c}(x) = (c_1(x), \ldots, c_L(x))$, and $r(x) = (f_\rho(x), f_\phi(x))$ for the three descriptors.

Proposition 1 (Component separation can be strictly more informative). There exists a distribution over (a, f, C) such that any estimator observing only the collapsed update $\Delta = a + f$ has Bayes error $1/2$, whereas an estimator observing the component-separated pair (a, f) has Bayes error 0. *Implication:* collapsed residual probes can be information-theoretically insufficient when correctness depends on which sublayer produced the update. *Proof: Appendix A.1.*

Proposition 2 (Cohen's d ranks heads by AUROC). For a head (l, h) with $A^{(l,h)} \mid C{=}k \sim \mathcal{N}(\mu_k, \sigma_k^2)$ and population effect size $d = \frac{\mu_1 - \mu_0}{\sqrt{(\sigma_1^2 + \sigma_0^2)/2}}$, the best orientation of this single feature has $\mathrm{AUROC}^\star\big(A^{(l,h)}\big) = \Phi\big(|d|/\sqrt{2}\big)$. Hence ranking heads by $|d|$ is equivalent to ranking them by best AUROC under this Gaussian class-conditional model. The criterion is single-feature and does

not imply joint optimality of the top-K subset; this is consistent with the non-monotone trend in K observed in Table 5. *Proof: Appendix A.2.*

Proposition 3 ($c_l(x)$ **is the residual interference term).** At any layer l,

$$\|a_l(x) + f_l(x)\|_2^2 = \|a_l(x)\|_2^2 + \|f_l(x)\|_2^2 + 2\|a_l(x)\|_2\|f_l(x)\|_2\, c_l(x). \tag{11}$$

At fixed component norms, $c_l(x)$ is the scalar interaction term that determines whether the two sublayer updates reinforce or cancel each other. This distinguishes $\mathbf{c}(x)$ from an arbitrary similarity feature. *Proof: Appendix A.3.*

Proposition 4 (Invariance of $r(x)$**).** With $\epsilon = 0$ and nonzero denominators in Eqs. 8 and Eqs. 9, $f_\rho(x)$ and $f_\phi(x)$ are invariant under any uniform rescaling $h \mapsto \alpha h$ with $\alpha > 0$, or any orthogonal transformation $h \mapsto Qh$, applied to all residual states $\{h_l(x)\}_{l=0}^{L}$. Thus, $r(x)$ removes two basic sources of representation-level incomparability—global scale and orthogonal coordinate choice—that affect raw hidden-state probes. *Proof: Appendix A.4.*

5 Experiments

5.1 Experiment Settings

Datasets and Models. We evaluate on five benchmarks: TriviaQA (Joshi et al., 2017) for factual recall, MMLU (Hendrycks et al., 2021) for knowledge across 57 subjects, GSM8K (Cobbe et al., 2021) for mathematical reasoning, CommonsenseQA (Talmor et al., 2019) for commonsense reasoning, and OpenBookQA (Mihaylov et al., 2018) for science question answering. These benchmarks span factual knowledge, mathematical reasoning, and commonsense reasoning tasks. We use four frozen LLMs spanning two model families and two scales: Qwen2.5-7B and Qwen2.5-72B (Qwen Team, 2024), and LLaMA-3.1-8B and LLaMA-3.1-70B (Grattafiori et al., 2024).

Baselines. We compare our method with representative confidence estimation baselines from both training-free and training-based methods. The training-free baselines include Self-Consistency(Chen & Mueller, 2024), Semantic Entropy (SE)(Kuhn et al., 2023), and EAGLE (Xiao et al., 2026). Self-Consistency estimates confidence from the agreement among multiple sampled responses, while Semantic Entropy measures uncertainty by grouping sampled responses into semantic equivalence classes. EAGLE is a self-evaluation based method that aggregates layer-wise internal beliefs without training a confidence model. We also compare with training-based internal-state methods, including SEP (Kossen et al., 2024) and InternalInspector (I^2) (Beigi et al., 2024). SEP trains probes on hidden representations to detect uncertainty-related signals, while I^2 uses activation, attention, and feed-forward states across layers to train an internal-state confidence estimator.

Implementation Details. For each benchmark, all training-based methods use the same training splits, with final evaluation on the publicly available labeled validation or test split. For the Attention Sink Signature, we select K separately for each backbone using internal validation subsets of the training data and fix it before final evaluation: $K = 50$ for Qwen2.5-7B and Llama-3.1-8B, $K = 100$ for Qwen2.5-72B, and $K = 300$ for Llama-3.1-70B. Cohen's d-based head rankings and feature normalization statistics are computed only on the probe-training subset. Full implementation details are provided in Appendix B.

Evaluation Metrics. We evaluate confidence estimation along two complementary axes. AUROC (Area Under the Receiver Operating Characteristic Curve) measures how well a confidence score ranks correct predictions above incorrect ones, capturing correctness discrimination. ECE (Expected Calibration Error) measures how closely predicted confidence matches empirical accuracy across confidence bins, capturing calibration quality. We report ECE with 10 equal-width bins, following standard practice. All AUROC and ECE values reported in this paper are scaled by 100.

5.2 Main Results

Table 1 reports AUROC and ECE across five benchmarks and four frozen backbone models, covering two model families, Qwen2.5 and LLaMA 3.1, and two model scales. Overall, our method achieves the strongest balance between correctness discrimination and calibration. It obtains the best AUROC on 17 of 20 model and benchmark combinations and the best ECE on 17 of 20 combinations.

Table 1: Performance comparison of confidence estimation methods across models and benchmarks. The best performing results are in **bold**, and the second best results are underlined.

Model	Method	TriviaQA		MMLU		GSM8K		CommonsenseQA		OpenBookQA	
		AUROC↑	ECE↓	AUROC↑	ECE↓	AUROC↑	ECE↓	AUROC↑	ECE↓	AUROC↑	ECE↓
Qwen2.5-7B	Self-Cons.	68.27	33.42	69.93	22.86	62.74	20.18	68.71	25.36	76.13	18.74
	SE	71.46	30.73	70.82	19.27	67.18	15.53	71.94	22.81	79.38	15.82
	EAGLE	75.54	28.08	64.51	9.19	60.65	64.08	64.17	5.54	72.57	9.58
	SEP	72.46	35.43	75.82	28.93	68.32	16.88	75.33	30.08	75.27	29.00
	I^2	77.30	7.93	78.35	13.82	73.13	20.55	76.38	8.88	83.85	3.90
	Ours	**80.55**	**1.26**	**80.75**	**2.39**	**74.24**	**2.87**	**80.06**	**3.09**	**87.11**	**3.01**
Qwen2.5-72B	Self-Cons.	74.53	18.37	72.94	21.83	78.16	14.92	77.48	17.42	88.23	11.74
	SE	77.82	17.13	75.21	20.34	77.63	15.78	80.17	15.92	90.36	10.27
	EAGLE	70.52	8.19	73.05	51.41	61.40	17.89	68.69	3.21	90.09	6.02
	SEP	78.23	24.49	75.78	31.06	69.68	22.01	79.65	22.23	88.01	14.11
	I^2	80.11	10.60	80.52	9.45	**82.58**	**14.49**	81.76	18.84	91.61	9.72
	Ours	**82.46**	**6.44**	**81.72**	**2.71**	81.97	23.20	**85.40**	**1.29**	**93.42**	**2.76**
LLaMA-3.1-8B	Self-Cons.	65.83	28.17	58.64	11.76	61.42	24.96	64.38	23.27	65.92	19.13
	SE	69.62	16.74	61.18	8.93	67.27	18.24	68.03	20.78	68.57	16.63
	EAGLE	58.46	16.14	65.70	9.06	49.35	46.25	59.46	12.82	55.77	**1.99**
	SEP	74.09	31.85	54.33	53.47	60.29	17.06	69.51	32.37	69.09	36.70
	I^2	75.83	11.75	76.33	10.99	76.31	4.01	**75.66**	5.09	70.44	17.88
	Ours	**79.93**	**1.40**	**78.04**	**2.01**	**77.56**	**3.30**	74.88	**1.91**	**71.89**	6.00
LLaMA-3.1-70B	Self-Cons.	76.11	18.45	62.25	7.21	80.36	8.43	68.67	16.31	80.29	13.53
	SE	80.55	18.62	65.19	9.13	82.47	15.73	71.83	17.41	82.73	12.02
	EAGLE	59.84	3.95	68.69	9.34	55.02	44.43	66.51	5.76	69.15	4.70
	SEP	78.89	27.00	57.34	61.97	77.60	18.11	69.63	37.25	81.24	24.93
	I^2	81.42	16.43	76.01	**2.42**	85.73	7.48	78.38	5.47	**87.37**	17.44
	Ours	**84.00**	**3.15**	**77.28**	2.45	**87.32**	**2.29**	**79.87**	**2.68**	84.16	**2.14**

Table 2: Ablation results on Qwen2.5-7B with $K = 50$ across benchmarks.

Dataset	Descriptor ablation						Architecture ablation	
	$\tilde{A}$ only	c only	r only	$\tilde{A} + \mathbf{c}$	$\tilde{A} + r$	$\mathbf{c} + r$	Ours	Plain MLP
MMLU	76.41 6.22	78.24 4.27	69.98 6.65	79.58 4.92	78.01 3.32	78.90 2.24	**80.75** 2.39	80.02 2.85
TriviaQA	78.65 2.00	72.55 1.90	58.18 1.52	79.26 1.44	79.87 1.83	73.25 1.74	**80.55** 1.26	79.12 1.87
GSM8K	70.20 5.66	67.19 8.05	70.01 5.37	71.43 7.47	73.48 3.71	73.88 1.95	**74.24** 2.87	73.99 6.18
CommonsenseQA	76.99 3.90	75.26 4.20	74.48 1.90	78.12 3.74	79.13 2.80	75.97 3.00	**80.06** 3.09	79.64 3.13
OpenBookQA	82.89 3.02	81.56 1.72	79.05 1.67	85.42 2.04	84.83 1.49	83.11 1.93	**87.11** 3.01	86.12 1.45

Averaged over the five benchmarks, our method improves AUROC over the strongest internal state baseline, I^2, by 0.74 to 2.74 points across different backbones, while reducing ECE by 5.34 to 8.49 points. These results show that our mechanistic descriptors provide confidence signals that are both more discriminative and better calibrated.

Compared with methods without probe training, our method shows more stable behavior across tasks and models. Self-Consistency and Semantic Entropy estimate uncertainty from sampled generations, while EAGLE aggregates layer-wise internal beliefs through self-evaluation. Although these can provide useful signals in some settings, their estimates remain tied to decoding behavior, sampling variation, or prompt-based introspection. In contrast, our method extracts structured internal signals from a single forward pass, yielding competitive AUROC and lower calibration error across diverse benchmarks. Compared with trained internal-state baselines, our method also shows a stronger balance between discrimination and calibration. SEP and I^2 demonstrate that hidden representations carry correctness-related information, but largely treat internal states as high-dimensional predictive features. Our method instead uses compact mechanistic descriptors that summarize how the frozen model forms its prediction, leading to stronger calibration across both model families and scales.

5.3 Ablation Study

Component Ablation. We evaluate the contribution of each mechanistic descriptor on Qwen2.5-7B by comparing single-branch, two-branch, and full three-branch variants. As shown in Table 2, each descriptor captures useful correctness-related information, but no single component consistently performs best across all benchmarks. The full model generally obtains the strongest AUROC, indicating that attention sink behavior, FFN response alignment, and residual trajectory geometry provide complementary signals. Although the best ECE is not always achieved by the full model, the

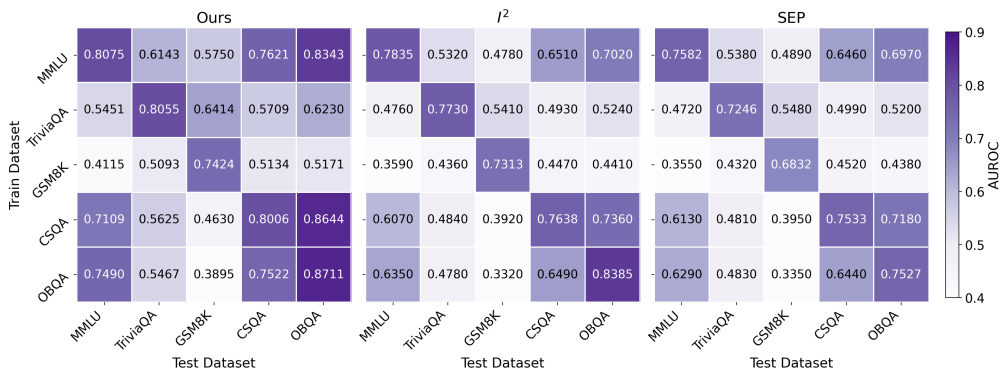

Figure 2: Cross-domain AUROC comparison across benchmarks. Each heatmap corresponds to one confidence estimation method: Ours, I^2, and SEP.

three-branch estimator provides the most stable trade-off between discrimination and calibration. We further compare our branch-attentive design against a plain MLP variant that concatenates the three descriptors as input. As shown in the rightmost column of Table 2, the branch-attentive estimator consistently outperforms the MLP on AUROC across all benchmarks, supporting our choice of processing heterogeneous descriptors in separate branches before fusion.

Top-K Head Selection. We examine how the number of selected heads affects the Attention Sink Signature. We select K for each backbone by considering both AUROC and ECE on internal validation data across the five benchmarks. The selected values are $K = 50$ for Qwen2.5-7B and Llama-3.1-8B, $K = 100$ for Qwen2.5-72B, and $K = 300$ for Llama-3.1-70B. Appendix D presents the internal validation results for each backbones. For all four backbones, the selected settings achieve higher AUROC and lower ECE than using all heads across the five benchmarks. This suggests that including more heads does not necessarily improve confidence estimation.

Effect of Training Data Size. We examine how training data size affects the confidence estimator on Qwen2.5-7B. As reported in Table 9 in Appendix E, AUROC improves as the training fraction increases from 60% to 100%, while the method remains competitive with reduced data, suggesting that the low-dimensional descriptors are data-efficient. The gains are visible on challenging benchmarks, indicating that additional supervision helps when correctness signals are harder to separate.

5.4 Cross-Domain Generalization

We further evaluate robustness to distribution shifts through transfer across benchmarks. For each source benchmark, we train the confidence estimator on its training split and evaluate it on each target benchmark without further training. Figure 2 shows the AUROC heatmaps. Rows denote source benchmarks, and columns denote target benchmarks. Diagonal entries represent matched settings, while the remaining entries measure transfer. Our method outperforms both baselines in all 20 transfer pairs. Its average transfer AUROC is **60.78**, compared with **51.97** for I^2 and **51.92** for SEP, with absolute gains of 8.81 and 8.86 points. All methods degrade under distribution shift on average. However, our method has a smaller drop from the diagonal mean to the transfer mean: **19.76** points, compared with 25.84 for I^2 and 21.52 for SEP. I^2 uses raw internal states as predictive features. Our method instead uses three compact descriptors to summarize attention behavior, interactions between components, and residual dynamics. It processes these descriptors in separate branches before fusion. This structure may reduce reliance on correlations specific to the source benchmark. The results suggest that our structured design helps mitigate performance degradation during transfer.

5.5 Analysis and Discussion

We analyze whether the learned branch weights reflect the functional separation behind our three descriptors. Rather than treating these weights as causal evidence, we use them as a diagnostic signal for how the estimator allocates correctness-related evidence across $\tilde{A}(x)$, $c(x)$, and $r(x)$. As shown in Table 3, the branch weights exhibit a clear task specific pattern: GSM8K is primarily

Table 3: Branch attention statistics on each benchmark (Qwen2.5-7B, K=50).

Benchmark	Mean ± per-sample std			Argmax %	Dominant Pattern
	$\tilde{A}$	**c**	r	$\tilde{A}$ / **c** / r	
MMLU	$\mathbf{0.352 \pm 0.074}$	0.338 ± 0.066	0.310 ± 0.052	**46** / 39 / 16	near-uniform
TriviaQA	$\mathbf{0.364 \pm 0.079}$	0.331 ± 0.082	0.305 ± 0.084	**45** / 31 / 24	near-uniform
GSM8K	0.343 ± 0.095	0.289 ± 0.086	$\mathbf{0.368 \pm 0.111}$	33 / 16 / **51**	r-led
CommonsenseQA	$\mathbf{0.405 \pm 0.103}$	0.324 ± 0.104	0.271 ± 0.073	**62** / 29 / 9	$\tilde{A}$-led
OpenBookQA	$\mathbf{0.387 \pm 0.168}$	0.304 ± 0.115	0.309 ± 0.142	**52** / 19 / 28	$\tilde{A}$-led

$r(x)$-led, suggesting stronger reliance on residual trajectory dynamics for multi-step numerical reasoning; CommonsenseQA and OpenBookQA are $\tilde{A}(x)$-led, indicating greater dependence on attention routing and concentration cues; while MMLU and TriviaQA show more balanced weights.

The balanced pattern on MMLU becomes more informative when analyzed at the subject level. As shown in Figure 3 and further details can be found in Table 10 in Appendix G, the 57 MMLU subjects separate into 37 $\tilde{A}(x)$-dominant, 16 $c(x)$-dominant, and 4 $r(x)$-dominant subjects. The dominant branch aligns systematically with the cognitive demand of each subject: $\tilde{A}(x)$-led subjects center on factual recall or localized evidence selection, $c(x)$-led subjects involve integrating context with domain knowledge, and $r(x)$-led subjects rely on symbolic or quantitative reasoning. Representative examples in Table 11 in Appendix G.2 illustrate this correspondence at the question level. MMLU's near-uniform branch weights at the aggregate level therefore reflect averaging over a heterogeneous mixture of subject types, rather than the absence of branch specialization.

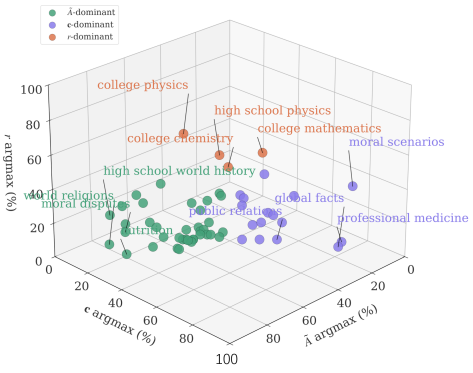

Figure 3: Subject-level branch dominance on MMLU across the 57 subjects.

As a head-level analysis, we further examine the overlap of selected top-K attention heads across benchmarks, with the full matrix reported in Figure 4in the appendix F. This analysis complements the branch-level weights by asking whether similar branch profiles are supported by the same underlying attention heads. Although MMLU and TriviaQA both exhibit near-uniform branch weights at the benchmark level, they share only a small number of selected heads. This suggests that the attention-sink descriptor may capture different head-level evidence in the two benchmarks, even when their aggregate branch allocations appear similar.

6 Conclusion and Limitations

We presented a component-aware confidence estimator for frozen LLMs. The method extracts three compact mechanistic descriptors from a single forward pass, *Attention Sink Signature*, *Knowledge Response Alignment*, and *Residual Trajectory Geometry*, fused through a lightweight branch-attentive probe. Across five QA benchmarks and four decoder-only LLMs from the Qwen2.5 and LLaMA-3.1 families, it achieves the best AUROC on 17 of 20 model-benchmark combinations and the best ECE on 17 of 20, and remains more robust than internal-state baselines under cross-domain evaluation. These results suggest that organizing confidence estimation around the functional structure of Transformer components yields signals more discriminative and transferable than treating internal states as a single high-dimensional feature. Our study has two limitations. The descriptors are read from internal computation, so the method applies to open-weight LLMs but not to closed-source models accessible only through output APIs. The probe is also trained against correctness labels, requiring labeled model outputs for probe training. Extending the framework to closed-source backbones and to label-efficient or unsupervised regimes is a natural direction for future work.

AI USE STATEMENT

In this work, we used generative AI tools to assist with implementing our proposed method. We did not use generative AI tools for research ideation. The authors manually reviewed the AI-assisted implementation for consistency with the intended method and checked its behavior in our experiments.

We also used generative AI tools to aid and polish writing. This included suggestions for preliminary drafts, alternative word choices, and sentence formulations to express our ideas more clearly, as well as grammar checks. We treated AI-suggested drafts as references and rewrote the manuscript sentence by sentence. The authors determined the final wording.

We reviewed all AI assisted work and take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We use publicly available Large Language Models and dataset for training and evaluation, and we have included all training hyperparameters in Appendix B. We will release our implementation code upon acceptance.

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

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*, pp. 967–976, 2023.

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*, pp. 12847–12865, 2024.

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.

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.

Jiuhai Chen and Jonas Mueller. Quantifying uncertainty in answers from any language model and enhancing their trustworthiness. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5186–5200, 2024.

Yung-Sung Chuang, Linlu Qiu, Cheng-Yu Hsieh, Ranjay Krishna, Yoon Kim, and James R. Glass. Lookback lens: Detecting and mitigating contextual hallucinations in large language models using only attention maps. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1419–1436, 2024.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. In *arXiv preprint arXiv:2110.14168*, 2021.

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. Knowledge neurons in pretrained transformers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 8493–8502, 2022.

Zhichen Dong, Zhanhui Zhou, Zhixuan Liu, Chao Yang, and Chaochao Lu. Emergent response planning in LLMs. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, 2025.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

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

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630, 2024.

Javier Ferrando, Gabriele Sarti, Arianna Bisazza, and Marta R. Costa-jussà. A primer on the inner workings of transformer-based language models. *arXiv preprint arXiv:2405.00208*, 2024.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 5484–5495, 2021.

Aaron Grattafiori et al. The Llama 3 herd of models. Technical report, Meta AI, 2024. URL https://arxiv.org/abs/2407.21783.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021. URL https://openreview.net/forum?id=d7KBjmI3GmQ.

Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pp. 1601–1611, 2017.

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

Jannik Kossen, Jiatong Han, Muhammed Razzak, Lisa Schut, Shreshth Malik, and Yarin Gal. Semantic entropy probes: Robust and cheap hallucination detection in LLMs. In *arXiv preprint arXiv:2406.15927*, 2024.

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

Stephanie Lin, Jacob Hilton, and Owain Evans. Teaching models to express their uncertainty in words. In *Transactions on Machine Learning Research (TMLR)*, 2022.

Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 9004–9017, 2023.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2381–2391, Brussels, Belgium, 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1260. URL https://aclanthology.org/D18-1260/.

Qwen Team. Qwen2.5 technical report. Technical report, Alibaba Group, 2024. URL https://arxiv.org/abs/2412.15115.

Zhongxiang Sun, Xiaoxue Zang, Kai Zheng, Jun Xu, Xiao Zhang, Weijie Yu, Yang Song, and Han Li. ReDeEP: Detecting hallucination in retrieval-augmented generation via mechanistic interpretability. In *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4149–4158, Minneapolis, Minnesota, 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1421. URL https://aclanthology.org/N19-1421/.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D. Manning. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 5433–5442, 2023.

Artem Vazhentsev, Lyudmila Rvanova, Gleb Kuzmin, Ekaterina Fadeeva, Ivan Lazichny, Alexander Panchenko, Maxim Panov, Mrinmaya Sachan, Preslav Nakov, Timothy Baldwin, and Artem Shelmanov. Efficient hallucination detection for LLMs using uncertainty-aware attention heads. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2505.20045v3.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://arxiv.org/abs/2309.17453.

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 (AAAI)*, 2026.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. Can LLMs express their uncertainty? an empirical evaluation of confidence elicitation in LLMs. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Mert Yuksekgonul, Varun Chandrasekaran, Erik Jones, Suriya Gunasekar, Ranjita Naik, Hamid Palangi, Ece Kamar, and Besmira Nushi. Attention satisfies: A constraint-satisfaction lens on factual errors of language models. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024.

Zhenliang Zhang et al. ICR probe: Tracking hidden state dynamics for reliable hallucination detection in LLMs. *arXiv preprint arXiv:2507.16488*, 2025.

A Proofs of Theoretical Properties

This appendix contains the proofs of Propositions 1–4 in Section 4.5.

A.1 Proof of Proposition 1

Proof. Let $q \in \mathbb{R}^d$ be any nonzero vector and define

$$(a, f, C) = \begin{cases} (q, 0, 1), & \text{with probability } 1/2, \\ (0, q, 0), & \text{with probability } 1/2. \end{cases} \tag{12}$$

In both cases, the collapsed update is $\Delta = a + f = q$. Thus Δ is constant and contains no information about C:

$$\mathbb{P}(C = 1 \mid \Delta = q) = \mathbb{P}(C = 0 \mid \Delta = q) = 1/2. \tag{13}$$

Hence the Bayes-optimal estimator based only on Δ has error $1/2$. In contrast, the component-separated pair (a, f) identifies the label exactly: $C = 1$ iff $(a, f) = (q, 0)$, and $C = 0$ iff $(a, f) = (0, q)$. Therefore the Bayes error based on (a, f) is 0. □

A.2 Proof of Proposition 2

Proof. Let $Z = A^{(l,h)}(x)$ denote the per-head sink value, and let $Z_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $Z_0 \sim \mathcal{N}(\mu_0, \sigma_0^2)$ be independent draws from the two class-conditional distributions. For threshold-based scoring with Z,

$$\mathrm{AUROC}(Z) = \mathbb{P}(Z_1 > Z_0). \tag{14}$$

Since

$$Z_1 - Z_0 \sim \mathcal{N}(\mu_1 - \mu_0, \sigma_1^2 + \sigma_0^2), \tag{15}$$

we have

$$\mathrm{AUROC}(Z) = \Phi\left(\frac{\mu_1 - \mu_0}{\sqrt{\sigma_1^2 + \sigma_0^2}}\right). \tag{16}$$

Taking the better orientation of the score gives

$$\mathrm{AUROC}^\star(Z) = \Phi\left(\frac{|\mu_1 - \mu_0|}{\sqrt{\sigma_1^2 + \sigma_0^2}}\right). \tag{17}$$

Substituting

$$|d| = \frac{|\mu_1 - \mu_0|}{\sqrt{(\sigma_1^2 + \sigma_0^2)/2}} \tag{18}$$

yields

$$\mathrm{AUROC}^\star(Z) = \Phi\left(\frac{|d|}{\sqrt{2}}\right). \tag{19}$$

Since Φ is strictly increasing, ranking heads by $|d|$ is equivalent to ranking them by best single-feature AUROC. □

A.3 Proof of Proposition 3

Proof. Expanding the squared norm,

$$\|a_l + f_l\|_2^2 = \langle a_l + f_l, a_l + f_l\rangle = \|a_l\|_2^2 + \|f_l\|_2^2 + 2\langle a_l, f_l\rangle. \tag{20}$$

Using

$$\langle a_l, f_l\rangle = \|a_l\|_2 \|f_l\|_2 \, c_l(x) \tag{21}$$

gives the claimed identity. For fixed nonzero component norms, the right-hand side is an affine, strictly increasing function of $c_l(x)$, so the map $c_l(x) \mapsto \|a_l + f_l\|_2^2$ is a bijection on $[-1, 1]$. Hence $c_l(x)$ is the unique scalar determining the joint update energy at fixed component norms. □

A.4 Proof of Proposition 4

Proof. Let $\tilde{h}_l = T(h_l)$ and set $\epsilon = 0$.

For $T(h) = \alpha h$ with $\alpha > 0$,

$$\|\tilde{h}_{l+1} - \tilde{h}_l\|_2 = \alpha\|h_{l+1} - h_l\|_2, \qquad \|\tilde{h}_L - \tilde{h}_0\|_2 = \alpha\|h_L - h_0\|_2, \tag{22}$$

so the factor α cancels in every distance ratio of f_ρ. Moreover,

$$\frac{\langle \tilde{h}_l, \tilde{h}_{l+1}\rangle}{\|\tilde{h}_l\|_2\|\tilde{h}_{l+1}\|_2} = \frac{\langle h_l, h_{l+1}\rangle}{\|h_l\|_2\|h_{l+1}\|_2}, \tag{23}$$

and the same holds for the endpoint cosine. Since f_ϕ depends on these cosines only through $\arccos$, f_ϕ is unchanged.

For $T(h) = Qh$ with $Q^\top Q = I$, orthogonal transformations preserve distances and inner products:

$$\|Qh - Qh'\|_2 = \|h - h'\|_2, \qquad \langle Qh, Qh' \rangle = \langle h, h' \rangle. \tag{24}$$

Therefore all distance ratios in f_ρ and all cosine arguments in f_ϕ are unchanged. This proves the claim. □

B Implementation Details

The branch-attentive confidence head g_ϕ takes the three component-aware descriptors $\tilde{\mathbf{A}}(x) \in \mathbb{R}^K$, $\mathbf{c}(x) \in \mathbb{R}^L$, and $\mathbf{r}(x) \in \mathbb{R}^2$ as separate branch inputs. Each branch is processed by a single linear layer followed by batch normalization, ReLU, and dropout, mapping its input to a shared hidden dimension $h = 32$. The three branch representations are fused by a learned attention with a single scoring vector $\mathbf{w} \in \mathbb{R}^h$ producing softmax weights over branches, and the fused vector is passed to a two-layer classifier ($h \to h/2 \to 1$) that outputs the confidence logit. Dropout is set to 0.3 throughout, and we use Adam with learning rate 10^{-3}, batch size 512, and binary cross-entropy loss. We train for up to 200 epochs with `ReduceLROnPlateau` (factor 0.5, patience 10) and early stopping (patience 20), and select the checkpoint with the lowest internal validation loss. The full architecture and hyperparameters are summarized in Table 4.

For each benchmark, we split the training data into 80% for probe training and 20% for internal validation. We select K separately for each backbone based on validation AUROC and ECE across the five benchmarks. Head rankings and feature normalization statistics are computed only on the probe training subset. Appendix D reports the internal validation results for different values of K. We fix K before final evaluation. All training-based baselines use the same train/evaluation splits as our method. We report results averaged over three seeds (12, 123, 1234). Experiments run on two NVIDIA H200 GPUs. All baseline methods were reproduced using their publicly released implementations and the configurations recommended by the original authors.

Table 4: Architecture and training hyperparameters of the branch-attentive confidence head g_ϕ.

Component	**Specification**
Branch 1 input ($\tilde{\mathbf{A}}$)	$\mathbb{R}^K$, Backbone-Specific K
Branch 2 input ($\mathbf{c}$)	$\mathbb{R}^L$
Branch 3 input ($\mathbf{r}$)	$\mathbb{R}^2$
Branch encoder	Linear → BatchNorm → ReLU → Dropout
Branch hidden dimension h	32
Branch fusion	Softmax attention with scoring vector $\mathbf{w} \in \mathbb{R}^h$
Classifier head	Linear(h, $h/2$) → BN → ReLU → Dropout → Linear($h/2$, 1)
Dropout rate	0.3
Loss function	Binary cross-entropy
Optimizer	Adam
Learning rate	1×10^{-3}
Batch size	512
Max epochs	200
LR scheduler	`ReduceLROnPlateau` (factor 0.5, patience 10)
Early stopping patience	20 epochs
Internal validation ratio	20% (stratified)
Model selection	Lowest internal validation loss

C Prompt Templates

We provide the exact zero-shot prompt templates used to query the frozen language model on each of the five benchmarks. All prompts follow a unified structure across multiple-choice and generation-based tasks.

MMLU and OpenBookQA (4-choice).

```
Question: {question}
A. {choice_A}
B. {choice_B}
C. {choice_C}
D. {choice_D}
Answer:
```

CommonsenseQA (5-choice).

```
Question: {question}
A. {choice_A}
B. {choice_B}
C. {choice_C}
D. {choice_D}
E. {choice_E}
Answer:
```

TriviaQA.

```
Answer the following trivia question with only the short answer.
Question: {question}
Short answer:
```

GSM8K.

```
Answer the following math problem. Return only the final numeric
answer. Do not include reasoning, words, units, or explanation.
Question: {question}
Answer:
```

D TOP-K ABLATION RESULTS

Table 5: Top K ablation results across benchmarks on Qwen2.5-7B.

Top K	**MMLU**		**TriviaQA**		**GSM8K**		**CommonsenseQA**		**OpenBookQA**	
	AUROC↑	ECE↓	AUROC↑	ECE↓	AUROC↑	ECE↓	AUROC↑	ECE↓	AUROC↑	ECE↓
10	80.54	2.47	79.81	1.58	73.09	4.30	78.49	3.23	85.76	3.46
30	81.11	2.61	80.31	1.64	74.32	4.26	78.78	3.69	**88.88**	**1.97**
50	**81.34**	2.45	**81.64**	**1.25**	**76.94**	**3.80**	**79.40**	**3.02**	87.80	3.10
100	80.56	3.29	80.43	3.50	75.73	5.72	78.03	4.03	85.51	3.44
200	80.36	**2.36**	81.15	1.59	76.07	5.76	77.14	3.39	85.60	3.57
300	78.98	5.14	80.20	2.94	75.10	7.46	78.25	3.98	84.63	4.18
all	79.06	6.01	80.73	1.62	75.57	8.94	77.66	5.47	83.98	5.32

Table 6: Top K ablation results across benchmarks on Qwen2.5-72B.

Top K	**MMLU**		**TriviaQA**		**GSM8K**		**CommonsenseQA**		**OpenBookQA**	
	AUROC↑	ECE↓	AUROC↑	ECE↓	AUROC↑	ECE↓	AUROC↑	ECE↓	AUROC↑	ECE↓
10	82.28	**3.29**	81.72	5.23	79.81	21.98	83.14	3.97	85.75	2.34
100	**83.18**	3.82	**82.06**	**4.62**	**82.68**	**20.01**	**84.29**	3.31	90.16	**2.14**
300	82.63	3.46	81.67	5.46	81.09	22.57	83.94	3.73	**90.73**	3.85
500	82.28	5.86	81.77	6.85	81.15	22.61	84.09	3.95	90.01	3.14
800	81.39	4.16	81.35	6.96	80.87	22.55	82.46	**3.05**	86.36	3.70
1200	82.35	8.36	80.33	8.31	81.17	20.32	83.18	3.85	88.11	4.06
2400	81.06	7.92	81.18	7.15	80.21	21.71	81.94	6.00	87.53	3.56
all	79.32	5.34	80.00	9.03	80.38	21.00	82.28	4.38	86.35	4.45

Table 7: Top K ablation results across benchmarks on Llama-3.1-8B.

Top K	MMLU		TriviaQA		GSM8K		CommonsenseQA		OpenBookQA	
	AUROC↑	ECE↓	AUROC↑	ECE↓	AUROC↑	ECE↓	AUROC↑	ECE↓	AUROC↑	ECE↓
10	78.69	4.06	78.45	2.83	75.30	4.20	72.60	3.03	74.22	3.21
30	80.21	4.81	78.73	1.92	**77.64**	**3.37**	73.22	3.45	74.30	4.37
50	**81.25**	**3.15**	**79.05**	**1.78**	77.32	4.03	73.60	**2.55**	**74.89**	**2.78**
100	79.90	3.93	78.57	2.36	74.89	4.25	**74.09**	2.80	73.80	3.61
200	78.83	5.03	77.94	3.41	76.13	4.67	73.37	2.75	73.08	3.18
300	79.16	4.20	78.90	2.95	75.08	3.53	72.64	3.95	72.34	6.90
all	78.24	4.63	77.31	4.01	74.29	4.83	71.03	5.11	72.56	6.93

Table 8: Top K ablation results across benchmarks on Llama-3.1-70B.

Top K	MMLU		TriviaQA		GSM8K		CommonsenseQA		OpenBookQA	
	AUROC↑	ECE↓	AUROC↑	ECE↓	AUROC↑	ECE↓	AUROC↑	ECE↓	AUROC↑	ECE↓
10	81.21	6.63	80.10	3.98	86.59	3.18	78.97	3.34	84.59	2.97
100	80.91	4.86	81.42	3.32	87.79	1.93	79.80	3.67	86.31	3.57
300	**81.66**	4.94	82.99	**3.01**	**88.76**	**1.90**	80.22	3.00	**86.71**	**2.92**
500	80.82	**4.51**	**83.53**	3.15	88.04	2.92	80.21	**2.56**	83.98	3.74
800	78.43	10.23	82.10	4.03	86.46	3.31	**80.47**	2.84	83.84	6.31
1200	78.93	11.55	81.84	3.99	86.26	6.75	80.14	2.99	81.34	4.96
2400	77.59	8.47	79.06	5.33	87.01	3.78	79.49	4.59	80.80	4.47
all	77.58	8.13	80.19	4.72	85.34	7.26	79.12	4.53	80.16	6.93

E Effect of Training Size

Table 9: Effects of training size on AUROC of our method across benchmarks.

Size (%)	TriviaQA	MMLU	GSM8K	CommonsenseQA	OpenBookQA
60	78.18±0.18	77.03±0.78	69.21±1.24	77.09±0.39	85.59±1.25
70	79.45±0.22	78.55±0.28	70.23±0.39	77.85±0.23	86.79±0.82
80	80.38±0.15	79.28±0.63	71.78±0.14	78.69±0.20	86.84±0.85
90	80.42±0.16	80.01±0.16	73.41±1.07	78.77±0.24	86.72±0.57
100	80.55±0.08	80.75±0.24	74.24±0.81	80.06±0.32	87.11±0.42

F Top-K Head Overlap

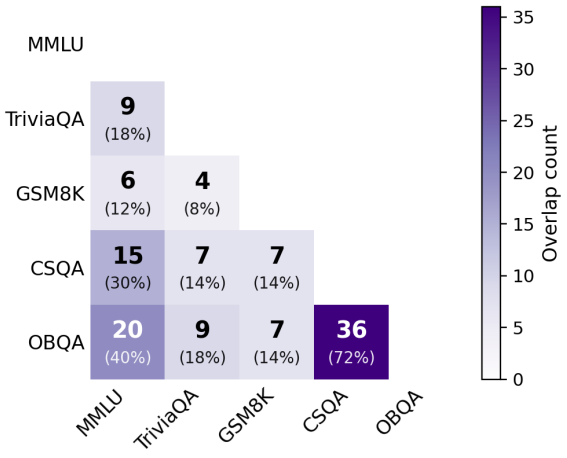

Figure 4: Pairwise overlap of selected top-K attention heads across benchmarks with $K = 50$.

G MMLU PER-SUBJECT BRANCH ATTENTION

G.1 DISTRIBUTION BY QWEN2.5-7B

To verify the heterogeneity-based interpretation of MMLU's near-uniform aggregate attention, we compute branch attention statistics for each of the 57 MMLU subjects. Table 10 reports the mean attention weight assigned to each branch on each subject, the per-sample argmax distribution, and the dominant branch. Subjects are grouped by their dominant branch and sorted within each group by deviation from uniform.

Table 10: Per-subject branch attention on MMLU (Qwen2.5-7B, K=50).

Subject	**N**	$\tilde{A}$	**c**	r	**Argmax (%)** $\tilde{A}$ / **c** / r
$\tilde{A}$**-dominant subjects** (37 / 57)					
nutrition	306	**0.442**	0.322	0.236	**79** / 20 / 1
formal_logic	126	**0.429**	0.285	0.286	**73** / 13 / 13
world_religions	171	**0.407**	0.300	0.293	**82** / 13 / 5
moral_disputes	346	**0.407**	0.296	0.297	**75** / 15 / 10
high_school_world_history	237	**0.399**	0.282	0.320	**76** / 7 / 17
high_school_us_history	204	**0.393**	0.289	0.317	**71** / 9 / 21
high_school_government_and_politics	193	**0.389**	0.323	0.288	**70** / 26 / 4
elementary_mathematics	378	**0.388**	0.290	0.323	**64** / 14 / 22
security_studies	245	**0.385**	0.336	0.279	**63** / 33 / 3
anatomy	135	**0.385**	0.332	0.283	**65** / 27 / 8
high_school_european_history	165	**0.384**	0.285	0.331	**64** / 10 / 25
abstract_algebra	100	**0.382**	0.306	0.312	**55** / 15 / 30
astronomy	152	**0.377**	0.311	0.312	**65** / 23 / 12
high_school_geography	198	**0.377**	0.332	0.291	**61** / 32 / 8
business_ethics	100	**0.375**	0.320	0.305	**66** / 20 / 14
high_school_macroeconomics	390	**0.374**	0.325	0.301	**57** / 32 / 11
us_foreign_policy	100	**0.372**	0.340	0.287	**59** / 34 / 7
high_school_biology	310	**0.372**	0.330	0.298	**60** / 33 / 7
medical_genetics	100	**0.371**	0.339	0.291	**53** / 35 / 12
college_medicine	173	**0.371**	0.340	0.288	**51** / 40 / 9
college_biology	144	**0.370**	0.337	0.293	**58** / 36 / 6
high_school_statistics	216	**0.369**	0.310	0.321	**50** / 28 / 22
miscellaneous	783	**0.368**	0.341	0.291	**59** / 35 / 7
international_law	121	**0.368**	0.338	0.293	**63** / 34 / 3
high_school_microeconomics	238	**0.368**	0.339	0.293	**57** / 36 / 7
management	103	**0.365**	0.347	0.288	**57** / 36 / 7
prehistory	324	**0.364**	0.337	0.300	**54** / 37 / 9
clinical_knowledge	265	**0.361**	0.337	0.303	**52** / 37 / 11
electrical_engineering	145	**0.359**	0.336	0.305	**49** / 37 / 14
marketing	234	**0.358**	0.331	0.311	**53** / 36 / 10
high_school_psychology	545	**0.357**	0.347	0.295	**49** / 43 / 8
conceptual_physics	235	**0.357**	0.332	0.311	**49** / 32 / 19
human_sexuality	131	**0.356**	0.345	0.300	**52** / 39 / 9
high_school_computer_science	100	**0.355**	0.335	0.310	**45** / 32 / 23
college_computer_science	100	**0.349**	0.331	0.320	**40** / 34 / 26
philosophy	311	**0.348**	0.346	0.305	**47** / 43 / 10
sociology	201	**0.342**	0.337	0.320	**40** / 35 / 25
c-dominant subjects (16 / 57)					
global_facts	100	0.307	**0.418**	0.275	17 / **79** / 4
professional_medicine	272	0.299	**0.401**	0.300	15 / **79** / 6
public_relations	110	0.324	**0.381**	0.295	33 / **60** / 7

continued on next page

Table 10 continued from previous page

Subject	N	$\tilde{A}$	c	r	Argmax (%) $\tilde{A}$ / c / r
moral_scenarios	895	0.263	**0.381**	0.355	0 / **71** / 29
econometrics	114	0.331	**0.377**	0.292	30 / **52** / 18
professional_accounting	282	0.303	**0.375**	0.322	19 / **56** / 25
human_aging	223	0.339	**0.367**	0.294	38 / **55** / 7
professional_psychology	612	0.318	**0.365**	0.317	29 / **54** / 17
machine_learning	112	0.319	**0.364**	0.317	28 / **58** / 14
computer_security	100	0.334	**0.361**	0.305	37 / **50** / 13
virology	166	0.346	**0.356**	0.298	43 / **50** / 7
jurisprudence	108	0.339	**0.354**	0.308	34 / **52** / 14
logical_fallacies	163	0.328	**0.347**	0.325	36 / **43** / 21
high_school_mathematics	270	0.316	**0.347**	0.337	23 / **43** / 34
high_school_chemistry	203	0.335	**0.343**	0.322	34 / **40** / 25
professional_law	1534	0.337	**0.337**	0.325	34 / **42** / 24
r-dominant subjects (4 / 57)					
college_physics	102	0.346	0.285	**0.369**	38 / 11 / **51**
high_school_physics	151	0.342	0.312	**0.346**	32 / 26 / **42**
college_mathematics	100	0.325	0.332	**0.343**	19 / 38 / **43**
college_chemistry	100	0.332	0.325	**0.343**	32 / 31 / **37**

G.2 MMLU Subjects Examples

Table 11: Representative MMLU items for the three dominant-branch classes. For each class, we report four representative subjects and one corresponding MMLU item. The correct answer is underlined. These examples are intended as qualitative illustrations of the subject-level branch-attention patterns.

Subject	Representative MMLU item
$\tilde{A}$-dominant subjects	
`nutrition`	**Q:** Supplementation with long-chain fatty acids EPA and DHA during pregnancy and infancy will likely benefit the following: **A.** Visual performance **B.** Motor performance **C.** Behaviour disorders **D.** Memory
`formal_logic`	**Q:** Select the best translation into predicate logic: Ms. Jackson lent me some sugar, but neither Janet nor Latoya did. **A.** $Lm \vee \neg(Jl \vee Ll)$ **B.** $Lm \cdot \neg(Lj \cdot Ll)$ **C.** $Lm \cdot \neg(Lj \vee Ll)$ **D.** $Lm \vee \neg(Lj \cdot Ll)$
`world_religions`	**Q:** Mani referred to God by which of the following names? **A.** Mother of Light **B.** Father of Goodness **C.** Ineffable Lover **D.** Eternal Spirit
`moral_disputes`	**Q:** Leopold claims that history has shown that "the conqueror role is eventually self-defeating." It is self-defeating, in this view, because **A.** the conqueror does not understand what makes the conquered community tick. **B.** the conqueror does not know what and who within the community is valuable. **C.** the conqueror does not know what and who is worthless in community life. **D.** all of the above
c-*dominant subjects*	
`global_facts`	**Q:** As of 2015, the global fertility rate was about **A.** 1.5 children per woman **B.** 2.5 children per woman **C.** 3.5 children per woman **D.** 4.5 children per woman

continued on next page

Table 11 continued from previous page

Subject	Representative MMLU item
professional_ medicine	**Q:** A new severe respiratory illness caused by a newly identified virus is discovered. Which of the following properties of a killed vaccine relative to a live vaccine is the most appropriate rationale for developing a killed vaccine for this illness? **A.** Avoids the concern for reversion to virulence **B.** Develops more rapid protective immunity **C.** Is less likely to require subsequent boosters for lifelong immunity **D.** Is most likely to generate mucosal immunity
public_ relations	**Q:** According to Macnamara's (2005) pyramid of evaluation, at which stage is media coverage most likely to be measured? **A.** Inputs **B.** Outputs **C.** Outtakes **D.** Outcomes
moral_ scenarios	**Q:** For which of these two scenarios does the main character do something clearly morally wrong, according to ordinary moral standards in the US as of 2020? Scenario 1: I entered an ancient pyramid with a tour guide. Scenario 2: I pulled out her eyes and left them dangling so she was forced to watch me. **A.** Wrong, Wrong **B.** Wrong, Not wrong **C.** Not wrong, Wrong **D.** Not wrong, Not wrong
r-dominant subjects	
college_ physics	**Q:** The eigenvalues of a Hermitian operator are always **A.** real **B.** imaginary **C.** degenerate **D.** linear
high_school_ physics	**Q:** A stone is thrown horizontally with an initial speed of 10 m/s from a bridge. Assuming that air resistance is negligible, how long would it take the stone to strike the water 80 m below the bridge? **A.** 1 s **B.** 2 s **C.** 4 s **D.** 8 s
college_ mathematics	**Q:** How many real roots does the polynomial $2x^5 + 8x - 7$ have? **A.** None **B.** One **C.** Two **D.** Three
college_ chemistry	**Q:** Which of the following is a primary standard for use in standardizing bases? **A.** Ammonium hydroxide **B.** Sulfuric acid **C.** Acetic acid **D.** Potassium hydrogen phthalate