Predicting LLM Hallucinations Pre-Generation

Internal Hidden-State Probing vs. Output-Confidence Baselines Across Scale

Syed Sarim Ahsan
sarim.ahsan101@gmail.com

Abstract

Large Language Models (LLMs) frequently generate plausible yet factually incorrect outputs, a phenomenon known as hallucination. This study investigates whether an LLM's internal representations – captured at the final prompt token, prior to generating any response tokens – contain a linear signal capable of predicting downstream factual accuracy. Evaluating the Qwen2.5 model family (1.5B, 3B, and 7B parameters) on the TriviaQA benchmark ($N = 3{,}000$), layer-wise logistic regression probes were trained and compared against standard output-token confidence baselines (entropy and negative log-probability of the first generated token).

Two key findings emerged. First, internal hidden-state probes achieve strong predictive performance pre-generation, with peak Area Under the ROC Curve (AUC) of 0.784 at 1.5B, 0.759 at 3B, and 0.754 at 7B, concentrated primarily in the mid-to-late transformer layers. Second, as parameter scale grows, output-confidence baselines significantly degrade in predictive quality (AUC $= 0.743 \rightarrow 0.691 \rightarrow 0.653$), whereas internal hidden-state probes remain remarkably robust. Consequently, the performance advantage (ΔAUC) of internal hidden-state probing over output confidence widens dramatically with scale ($+0.041 \rightarrow +0.068 \rightarrow +0.101$). These results demonstrate that larger LLMs become increasingly overconfident or noisy in their surface logit probability distributions while retaining internal, linearly separable awareness of factual uncertainty.

1 Introduction

Despite remarkable capabilities in reasoning and natural language processing, LLMs continue to suffer from hallucinations – generating statements that are factually inaccurate or unsubstantiated by underlying knowledge bases. Reliable detection and mitigation of hallucinations remain critical bottlenecks for deploying LLMs in high-stakes domains such as medicine, law, and automated factual retrieval.

Prior approaches to uncertainty estimation predominantly rely on surface-level output logits, such as sequence perplexity, token entropy, or self-consistency via repeated sampling. However, surface confidence metrics suffer from severe limitations: they can only be evaluated after generation has commenced, and models often exhibit high confidence even when outputting false information.

Recent advances in mechanistic interpretability suggest that transformer networks construct rich internal world models and factual representations within their intermediate hidden states. This raises a fundamental research question: does an LLM internalize its own factual boundaries before emitting a single answer token, and can this signal be probed to anticipate hallucinations prior to generation? Furthermore, how does the fidelity of internal factual representations compare to surface output-confidence signals as model parameter size scales?

This work presents a systematic empirical study probing the internal representations of the Qwen2.5 instruction-tuned model series (1.5B, 3B, and 7B parameters) on the TriviaQA open-domain factual

question-answering benchmark. Hidden state vectors are extracted across all transformer layers at the final prompt token position – strictly before any candidate answer tokens are sampled – and linear logistic regression probes are trained to predict binary factual correctness.

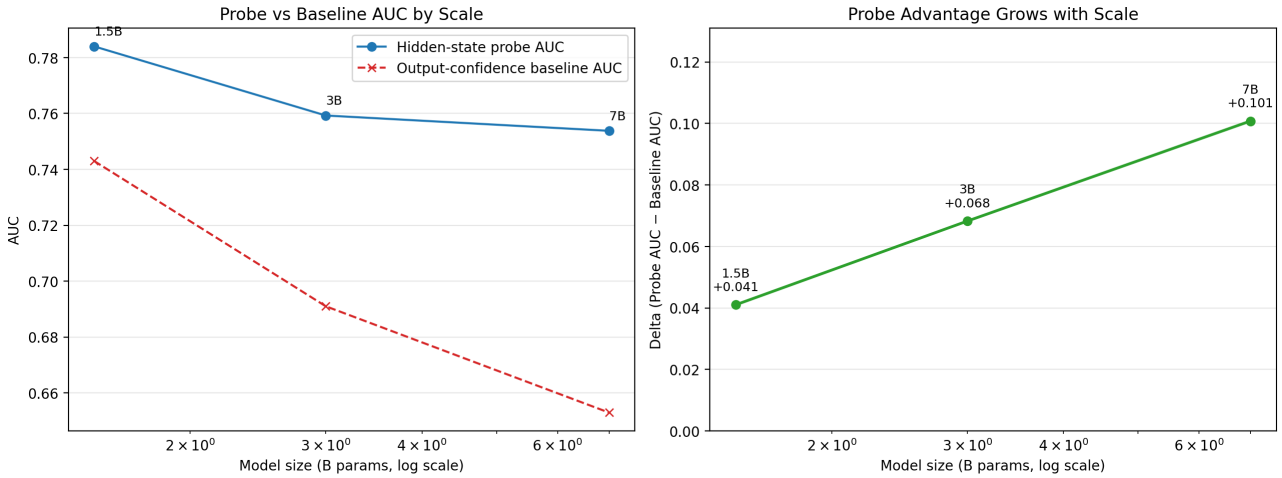


Figure 1: **Probe AUC vs. Baseline AUC across Model Scale.** As Qwen2.5 parameter scale increases (1.5B → 3B → 7B), the surface output-confidence baseline AUC deteriorates (0.743 → 0.653), while the internal hidden-state linear probe AUC remains stable (0.784 → 0.754). The probe advantage (ΔAUC) expands monotonically from +0.041 to +0.101.

Key Contributions

1. **Pre-Generation Hallucination Detection:** Establishing that intermediate hidden states at the prompt boundary contain a robust, linearly separable signal for predicting factual accuracy prior to answer generation.

2. **Layer-Wise Dynamics:** Mapping the trajectory of factual representation across transformer depth, discovering that predictive signal emerges rapidly after early embedding layers and peaks consistently within the mid-to-late network layers (depth ratios 0.44–0.68).

3. **Divergent Scaling Trend:** Demonstrating a critical scaling divergence: while surface token confidence becomes increasingly uncalibrated and noisy at larger parameter scales, internal hidden-state representations retain high predictive fidelity. The probe-to-baseline delta increases by +146% as scale moves from 1.5B to 7B.

2 Methodology

2.1 Task & Dataset Formulation

Factual question answering is evaluated using the TriviaQA dataset (`rc.nocontext` validation split, $N = 3{,}000$). Questions are presented to the model in a closed-book, zero-shot format using a standardized chat template ("Answer in a few words: {question}"). Greedy decoding (temperature = 0) is executed for up to 32 maximum new tokens. For ground-truth evaluation, generated strings are normalized by lowercasing, stripping punctuation, and removing leading articles. A generation is marked correct if any normalized gold answer string or alias appears as a substring within the normalized output; otherwise it is labeled a hallucination.

2.2 Hidden State Extraction

For a transformer model with L layers and hidden dimension d, given a prompt token sequence where the final prompt token precedes answer generation, the activation vector at each layer $l \in \{0, \ldots, L-1\}$ is extracted at that final position:

$$\mathbf{h}_T^{(l)} \in \mathbb{R}^d \tag{1}$$

This produces a complete hidden-state matrix $\mathbf{X} \in \mathbb{R}^{N \times L \times d}$ across all validation examples – recorded prior to emitting the first generation token.

2.3 Linear Probe Architecture

For each layer, an independent linear logistic regression classifier parameterized by weight vector $\mathbf{w}_l \in \mathbb{R}^d$ and bias b_l is trained on the extracted hidden-state vectors:

$$P(y = 1 \mid \mathbf{h}_T^{(l)}) = \sigma\left(\mathbf{w}_l^\top \mathbf{h}_T^{(l)} + b_l\right) \tag{2}$$

To account for potential class imbalance between correct outputs and hallucinations, balanced class weighting is applied. Optimization is performed using L-BFGS with a maximum of 2,000 iterations.

2.4 Output-Confidence Baselines

Two baseline features are computed from the first generated token:

- **Shannon Entropy (H):** the entropy over the full vocabulary logit distribution at the first generated token position.
- **Negative Log-Probability ($-\log p$):** the negative log-likelihood of the greedy-selected token.

Single-variable classifiers are evaluated on H and $-\log p$, as well as a combined bivariate logistic regression baseline using both metrics.

2.5 Experimental Protocol & Data Split

All three models (Qwen2.5-1.5B-Instruct, Qwen2.5-3B, Qwen2.5-7B) are evaluated on an identical 80/20 train/test stratified split ($N_{\text{train}} = 2{,}400$, $N_{\text{test}} = 600$, `random_state`=42). The same split is preserved across all layer probes and baseline classifiers to guarantee rigorous comparability. Model performance is primarily quantified using the Area Under the ROC Curve (ROC-AUC).

3 Experimental Results

3.1 Main Performance Summary

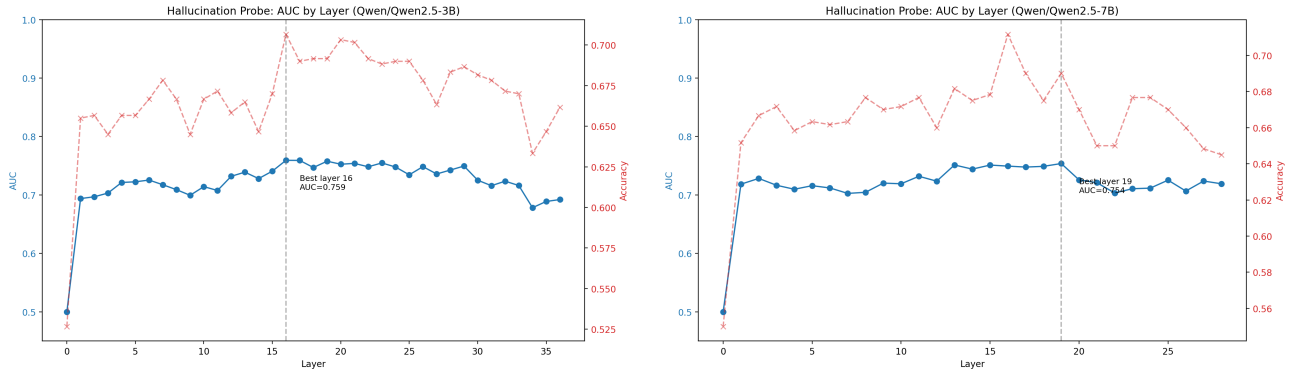
Table 1 provides a comprehensive quantitative comparison of linear hidden-state probes and output-confidence baselines across the three evaluated model sizes.

Table 1: Hallucination Prediction Performance Across Model Scale

Model	L	Peak Layer	Ratio	Probe AUC	Δ
Qwen2.5-1.5B	29	15	0.536	**0.784**	+0.041
Qwen2.5-3B	37	16	0.444	**0.759**	+0.068
Qwen2.5-7B	29	19	0.679	**0.754**	+0.101

Across all three model scales, linear probes operating solely on prompt hidden states outperform the strongest surface confidence baselines. At 1.5B parameters, the best layer probe achieves an AUC of 0.784 compared to the baseline's 0.743 ($\Delta = +0.041$). At 3B parameters, the probe reaches 0.759 AUC against a 0.691 baseline ($\Delta = +0.068$). At 7B parameters, the probe achieves 0.754 AUC compared to a degraded 0.653 baseline AUC, establishing a peak margin of $\Delta = +0.101$.

3.2 Layer-Wise Internal Signal Dynamics



(a) Qwen2.5-3B ($L = 37$ layers)

(b) Qwen2.5-7B ($L = 29$ layers)

Figure 2: **Per-Layer Hallucination Probing AUC & Accuracy Curves.** (a) Qwen2.5-3B achieves peak AUC of 0.759 at layer 16 (depth ratio 0.44). (b) Qwen2.5-7B achieves peak AUC of 0.754 at layer 19 (depth ratio 0.68). Both architectures exhibit near-chance performance at layer 0, a rapid rise through early layers, a broad mid-network plateau, and a slight decay near output projection layers.

Per-layer ROC-AUC and classification accuracy trajectories reveal a clear three-phase evolution of factual representations across network depth:

1. **Early Layer Initialization (Layers 0–4):** Hidden states at the embedding and initial self-attention layers contain negligible predictive signal (AUC $\approx$ 0.50–0.55), reflecting low-level token identity rather than factual confidence.

2. **Mid-Network Factual Emergence (Layers 5–20):** Predictive AUC rises sharply, entering a broad high-performance plateau. Factual truthfulness representations become highly linearly separable in this mid-network regime. Peak performance occurs at depth ratios between 44% and 68% of total model depth.

3. **Late Layer Tapering (Final 5 Layers):** As representation approaches final unembedding projection heads, probe AUC exhibits a modest decline, suggesting that final layers specialize in surface vocabulary selection rather than maintaining abstract semantic truth representations.

3.3 Divergent Scaling Behavior

The central finding of this investigation is the stark divergence between internal probe stability and output confidence degradation under parameter scaling. As model scale increases from 1.5B to 7B parameters:

- **Output Confidence Degradation:** Baseline AUC drops sharply from 0.743 (1.5B) to 0.691 (3B) and further to 0.653 (7B) – a total decay of -0.090 AUC points (-12.1%). Larger models generate surface output tokens with uncalibrated confidence distributions, rendering first-token entropy a poor proxy for actual accuracy.

- **Internal Probe Resilience:** Probe AUC decreases only marginally from 0.784 to 0.754 (-0.030 AUC points), remaining largely robust across scales.
- **Expanding Probe Advantage:** Because surface confidence deteriorates faster than internal representations, the probe advantage ΔAUC increases monotonically with model scale ($+0.041 \rightarrow +0.068 \rightarrow +0.101$).

4 Discussion & Implications

4.1 Mechanistic Interpretation

Why does surface output confidence become noisier at scale while internal states remain predictive? A plausible explanation is that larger instruction-tuned LLMs develop sharper token-level predictive distributions due to RLHF or instruction tuning, leading to overconfident top-token logits even when the factual content is wrong. Intermediate transformer layers, however, maintain an uncorrupted internal representation of factual familiarity or uncertainty. The linear probe successfully extracts this internal signal before it is mapped onto the output token distribution.

4.2 Practical Applications

The ability to detect potential hallucinations prior to generation offers major computational and practical advantages:

- **Early Refusal / Fallback:** If the prompt hidden-state probe signals high hallucination probability, system pipelines can immediately route the query to a web search retrieval module (RAG) or refuse generation without executing expensive decoding loops.
- **Activation Steering:** Linear probe weight vectors define a specific "hallucination direction" in activation space. Subtracting this direction during forward passes could dynamically reduce hallucination rates.

5 Limitations & Future Work

1. **Model & Domain Scope:** Experiments were conducted on the Qwen2.5 model family and TriviaQA dataset. Generalization across architectures (e.g., Llama-3, Mistral) and non-factual reasoning tasks requires further validation.
2. **Decoding Regimes:** Probing was evaluated under greedy decoding; sampling-based decoding methods (e.g., nucleus sampling) warrant future investigation.
3. **Causality:** Linear probing establishes correlation rather than direct causation. Future work will employ causal intervention techniques (activation addition/subtraction) to test whether shifting intermediate activations directly mitigates hallucinations.

6 Conclusion

This study demonstrates that linear probes trained on intermediate transformer hidden states at the prompt boundary can effectively predict LLM hallucinations before generation begins. Crucially, as parameter scale increases from 1.5B to 7B, surface output token confidence becomes an increasingly noisy signal for factual correctness, whereas internal hidden-state representations remain resilient. Consequently, the probing advantage expands significantly with scale ($\Delta\text{AUC} = +0.101$ at 7B). This

work highlights internal hidden states as a crucial resource for real-time hallucination detection and model alignment.