Do Internal Representations Reveal Hallucinations? Predictive, Geometric, and Intervention Evidence Across LLMs

Anonymous ACL submission

Abstract

Large language models (LLMs) can produce fluent answers that are factually unsupported, creating safety and reliability risks in scientific, educational, and decision-support settings. Existing hallucination detectors often depend on retrieval, repeated sampling, or external verification, increasing latency and deployment complexity. This paper tests whether factuality-relevant signals are already encoded in internal activations and can be read out with lightweight methods. We present a three-pillar multi-faceted empirical investigation — predictive probing, exploratory neuron intervention analysis, and representational geometry — using LLaMA-2-7B as a shared backbone across all three pillars. The predictive pillar trains compact MLP probes on final-layer summary features from LLaMA-2-7B (primary) and Qwen2.5-7B (secondary) on a combined 2,317-sample QA corpus. The LLaMA-2-7B detector reaches AUROC 0.8462 ± 0.0082 (5-seed mean $\pm$ std) under semantic-similarity pseudo-labels and retains AUROC 0.7731 on a held-out TruthfulQA native-label test split (full-subset pseudo-to-native gap 0.07), suggesting factuality-related signal beyond the single-reference pseudo-labeling oracle. The intervention pillar applies targeted LLaMA-2-7B neuron ablations, yielding exploratory evidence for depth-nonuniform ablation sensitivity concentrated at early and final layers. The geometric pillar finds a statistically above-chance but weak-to-moderate directional structure in LLaMA-2-7B final-layer embeddings (AUROC 0.624 ± 0.032, permutation $p < 0.001$). Because all three pillars share LLaMA-2-7B, the observed cross-pillar alignment constitutes same-model convergent evidence under related but non-identical operationalizations, though dataset and label differences prevent a fully mechanistic interpretation.

1 Introduction

Large language models have demonstrated broad utility across open-domain question answering, technical writing, scientific assistance, and decision-support workflows. Yet the reliability of these systems remains fundamentally constrained by hallucination: the confident production of unsupported, fabricated, or misleading content. In high-stakes environments such failures can be more damaging than explicit uncertainty because they create the appearance of competence while obscuring error.

Existing mitigation strategies often trade reliability against efficiency. Retrieval-based verification improves factual grounding but introduces infrastructure complexity and latency. Sampling-based self-consistency requires multiple generations (Manakul et al., 2023). Confidence-based methods are lightweight yet frequently fail when models hallucinate with high local confidence (Kadavath et al., 2022; Desai and Durrett, 2020; Farquhar et al., 2024). These limitations motivate a different line of inquiry: whether hallucination-relevant information is already encoded in internal activations.

Prior internal-state studies motivate this direction but leave important connections unresolved. Azaria and Mitchell (2023) probe a single model without neuron-level causal analysis. Chen et al. (2024) and Du et al. (2024) use internal representations without jointly evaluating geometry or causal attribution. To our knowledge, no prior work combines targeted neuron intervention, lightweight hidden-state classification, and directional geometry in a multi-faceted investigation with a shared backbone model across all three pillars.

The present work uses LLaMA-2-7B as the shared backbone across three complementary pillars, with Qwen2.5-7B as a secondary comparison:

- **Intervention (exploratory):** Structured neu-

ron ablation scan in LLaMA-2-7B identifying depth-nonuniform ablation sensitivity patterns.

- **Predictive:** Lightweight hidden-state probes with 5-seed evaluation and TruthfulQA native-label validation.
- **Geometric:** Directional structure of hallucination-related activation shifts in LLaMA-2-7B final-layer embeddings.

Using LLaMA-2-7B across all three pillars enables, to our knowledge, a same-model cross-pillar comparison in this framework. The intervention analysis identifies layers 0, 1, and 31 as showing elevated ablation-estimated influence; predictive probing confirms substantial signal in final-layer states (AUROC 0.8462 ± 0.0082); and geometric analysis independently detects a statistically above-chance directional shift (AUROC 0.624 ± 0.032). Dataset and labelling differences across pillars mean this alignment is convergent but not mechanistically conclusive.

Contributions.

1. A multi-faceted empirical investigation combining exploratory neuron ablation, hidden-state classification, and direction-based geometry, with LLaMA-2-7B shared across all three pillars enabling same-model cross-pillar comparison.
2. A lightweight LLaMA-2-7B hidden-state detector reaching AUROC 0.8462 ± 0.0082 under pseudo-labels (5-seed mean) and retaining AUROC 0.7731 on TruthfulQA's dataset-provided native labels (held-out test split). A Qwen2.5-7B replication achieves AUROC 0.8221 ± 0.0117.
3. Exploratory intervention and geometric analyses with explicit uncertainty quantification, providing preliminary evidence for depth-nonuniform ablation sensitivity and weak-to-moderate directional structure in LLaMA-2-7B final-layer embeddings.

2 Related Work

Hallucination detection approaches fall into four main categories. Retrieval-based verification compares generated claims with external corpora at the cost of latency and infrastructure. Sampling-based consistency methods assess agreement across multiple generations but incur multiplicative inference cost. Uncertainty-based approaches use token-level confidence statistics (Kadavath et al., 2022; Desai and Durrett, 2020; Li et al., 2023b; Farquhar et al., 2024), yet frequently fail when models hallucinate with high confidence. Internal-representation methods directly probe hidden states and latent features, aligning most closely with the present study (Azaria and Mitchell, 2023; Chen et al., 2024; Du et al., 2024; Su et al., 2024; Zhang et al., 2025).

The direction-analysis component is motivated by the linear representation hypothesis: semantic and factual distinctions tend to occupy low-dimensional linear subspaces in transformer hidden states (Mikolov et al., 2013; Park et al., 2023; Marks and Tegmark, 2023; Burns et al., 2022). Zou et al. (2023) report that high-level concepts including honesty occupy identifiable linear directions; unlike their behavioural-steering focus, we evaluate discriminative value for detection. Li et al. (2023a) show inference-time interventions on attention-head outputs can reduce hallucination; Chuang et al. (2024) identify attention-based hallucination signatures. Recent studies report gains from multi-layer feature fusion (Zhang et al., 2025); in contrast, we evaluate deliberately compact final-layer summary statistics.

Work	States	Det.	Interv.	Geom.	Multi-m.
SAPLMA (Azaria and Mitchell, 2023)	✓	✓	–	–	–
INSIDE (Chen et al., 2024)	✓	✓	–	ltd	–
HaloScope (Du et al., 2024)	✓	✓	–	–	part
MHAD (Zhang et al., 2025)	✓	✓	–	–	–
Ours	✓	✓	✓	✓	✓

Table 1: Positioning relative to prior internal-state methods.

Term	Definition
Pseudo-label	$\ell_i = \mathbf{1}[s_i < \tau]$, cosine sim. threshold τ=0.4
TQA_{nat}	TruthfulQA dataset-provided native labels
Held-out split	20% of corpus unseen during MLP training (seed 42)
Full TQA subset	All 817 matched TruthfulQA examples
Retrievable subset	1,505 samples with Wikipedia overlap available
Paired direction set	371 factual/hallucinated LLaMA-2-7B generation pairs
LOO	Leave-one-out cross-dataset evaluation

Table 2: Notation and dataset partitions used throughout.

Positioning. Table 1 summarises how the present work relates to prior methods. Unlike SAPLMA, INSIDE, HaloScope, or MHAD — which each emphasise one detection dimension — this framework explicitly compares intervention, predictive, and geometric evidence on a shared backbone model.

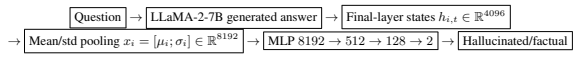


Figure 1: Pillar II hidden-state detection pipeline. The detector uses activations produced during ordinary decoding, without retrieval, repeated sampling, or reference answers at inference time.

3 Methodology

3.1 Overview

All three pillars use LLaMA-2-7B as the shared backbone. Figure 1 illustrates the Pillar II hidden-state detection pipeline. The intervention pillar uses targeted neuron ablations in LLaMA-2-7B to estimate the influence of individual MLP channels on last-token predictions. The predictive pillar uses final-layer hidden states from LLaMA-2-7B (primary) and Qwen2.5-7B (secondary) to train lightweight hallucination detectors. The geometric pillar studies whether hallucination-relevant shifts in LLaMA-2-7B activations produce measurable directional structure. Because the pillars use different datasets and labelling procedures, cross-pillar observations should be treated as convergent but not mechanistically conclusive.

3.2 Pillar I: Targeted Neuron Intervention (LLaMA-2-7B)

Using SQuAD-v2 validation samples (Rajpurkar et al., 2018; Touvron et al., 2023), we construct a balanced 200-prompt corpus of 100 answerable and 100 unanswerable examples. SQuAD-v2 unanswerable prompts represent model refusal failure, not the broader hallucination scenario; findings should be interpreted as evidence about context-sensitivity to unanswerable inputs. The pipeline:

1. Run a batched hidden-state scan; retain top-$k = 15$ neurons at the final prompt token per layer by absolute activation magnitude.
2. Form layer-wise candidate sets as the union across all prompts.
3. Cache unablated last-token logit vectors.
4. For each prompt and layer, zero one MLP-output channel per replica via a forward hook (batches of 128).
5. Measure estimated importance via Equation 1.

$$\Delta_{\ell,n}(x) = \left\| z^{\text{orig}}(x) - z^{\text{abl}(\ell,n)}(x) \right\|_2 \quad (1)$$

Zero-ablation is used for tractability across 1.1M intervention records; it may overestimate the importance of high-activation neurons. Mean-ablation replication is identified as future work.

Random-neuron baseline. To assess whether the top-k activation filter identifies genuinely influential neurons or merely high-magnitude ones, we pair each targeted neuron set with a matched random baseline: for each layer and prompt, k=15 neurons are sampled uniformly at random from the remaining neurons (those not selected by the top-k filter), and the same $\Delta_{\ell,n}$ measure is computed. We repeat random sampling across three seeds and report mean importance with Welch's t-test against the targeted set.

3.3 Pillar II: Hidden-State Classifier

Dataset. The 2,317-sample corpus combines: full TruthfulQA validation (Lin et al., 2022); 500-sample TriviaQA (Joshi et al., 2017); 500-sample Natural Questions (Kwiatkowski et al., 2019); 500-sample SQuAD (Rajpurkar et al., 2016).

Feature extraction. For LLaMA-2-7B (d=4096) and Qwen2.5-7B (d=3584), let final-layer generated-token states be $h_{i,1}, \ldots, h_{i,T_i} \in \mathbb{R}^d$. We intentionally evaluate compact final-layer mean and standard-deviation statistics to test whether lightweight activation summaries alone carry useful factuality-related signal, without requiring multi-layer fusion or high-capacity probes. Features:

$$\mu_i = \frac{1}{T_i} \sum_{t=1}^{T_i} h_{i,t}, \quad \sigma_i = \sqrt{\frac{1}{T_i} \sum_t (h_{i,t} - \mu_i)^2} \quad (2)$$

$$x_i = [\mu_i;\, \sigma_i] \in \mathbb{R}^{2d} \quad (3)$$

This yields $x_i \in \mathbb{R}^{8192}$ for LLaMA-2-7B and $x_i \in \mathbb{R}^{7168}$ for Qwen2.5-7B.

Pseudo-labels.

$$s_i = \frac{e(\hat{y}_i) \cdot e(y_i)}{\|e(\hat{y}_i)\|\, \|e(y_i)\|}, \quad (4)$$

$$\ell_i = \mathbf{1}[s_i < \tau], \quad \tau = 0.4. \quad (5)$$

Labels use a separate sentence-embedding model (`all-MiniLM-L6-v2`) to partially mitigate representational circularity.

TruthfulQA native-label validation. We evaluate trained detectors on the TruthfulQA subset (817 matched samples, 172 in the held-out test split) using TruthfulQA's dataset-provided correct/incorrect answer lists (Lin et al., 2022) as ground truth, independently of the cosine-similarity training oracle.

Exact native-label construction. For each matched TruthfulQA question, let $C=\{c_1, \ldots, c_m\}$ be the dataset-provided correct answers and $W=\{w_1, \ldots, w_k\}$ be the incorrect answers. We embed the generated answer $\hat{y}$ and all entries in C and W using `all-MiniLM-L6-v2` (the same embedder used for pseudo-labeling, applied to a different reference set). We then compute:

$$s_C = \max_{c \in C} \cos(e(\hat{y}), e(c)), \tag{6}$$

$$s_W = \max_{w \in W} \cos(e(\hat{y}), e(w)). \tag{7}$$

and assign a factual native label if $s_C \geq \tau$ and $s_C > s_W$, and hallucinated otherwise (i.e., $s_C < \tau$ or $s_W \geq s_C$). This uses the incorrect answer set as a tiebreaker, giving priority to correct-answer proximity. Pseudo/native agreement is the fraction of all 817 matched samples where the cosine-threshold pseudo-label agrees with this native binary label. The held-out test split (172 samples) is used for AUROC evaluation.

Classifier. Three-layer MLP with identical architecture for both models:

$$2d \to 512 \to 128 \to 2 \tag{8}$$

ReLU, dropout 0.3, Adam lr $= 10^{-4}$, up to 15 epochs, early-stopping patience 3. Evaluated across 5 seeds $\{42, 7, 123, 256, 999\}$ affecting both train/test split (80/20 stratified) and weight initialisation; mean $\pm$ std reported.

3.4 Secondary Oracle Validation

Wikipedia word-overlap (threshold 0.2, retrievable samples) is used post-hoc only — it does not affect training or model selection.

3.5 Pillar III: Direction Analysis (LLaMA-2-7B)

A 371-sample paired dataset from LLaMA-2-7B outputs across SQuAD, TriviaQA, and BoolQ. For each of 300 questions per dataset, LLaMA-2-7B samples 8 candidate answers (nucleus sampling, T=1.3, top-p=0.95, 50 tokens). Candidates are scored by token-F1 and sentence-embedding similarity; a candidate is hallucinated only when both fall below 0.4. Questions with both candidate types retained (371 pairs, 742 answer embeddings).

Mean-pooled final-layer embeddings ($\mathbb{R}^{4096}$) are extracted per answer. Normalised class-mean direction:

$$d = \frac{\mu_{\text{hall}} - \mu_{\text{fact}}}{\|\mu_{\text{hall}} - \mu_{\text{fact}}\|_2}, \quad s_i = x_i^\top d \tag{9}$$

Validation: 80/20 splits, 5 seeds, 1000-bootstrap 95% CI, 1000-permutation test.

4 Experimental Setup

4.1 Predictive Procedures

(1) Build 2,317-example corpus. (2) Generate greedy answers with LLaMA-2-7B and Qwen2.5-7B; collect final-layer hidden states. (3) Pool to $x_i=[\mu_i; \sigma_i]$. (4) Compute semantic-similarity labels (τ=0.4). (5) Run 5-seed evaluation; train MLP on training split only. (6) Report mean $\pm$ std AUROC, Macro-F1, Accuracy on held-out split. (7) Separately evaluate on TruthfulQA subset with native labels.

4.2 Direction-Analysis Procedures

(1) Select first 300 validation questions from SQuAD, TriviaQA, BoolQ. (2) Sample 8 LLaMA-2-7B candidate answers per question. (3) Filter to 371 pairs. (4) Extract mean-pooled final-layer embeddings. (5) Estimate class-mean direction on training split only. (6) Report held-out AUROC, 5-seed variability, bootstrap CI, permutation test, and baseline comparisons.

4.3 Intervention Procedures

(1) Select 100 answerable and 100 unanswerable SQuAD-v2 prompts. (2) Tokenise individually with truncation; collate with dynamic padding. (3) Retain top-15 neurons per layer per prompt by activation magnitude. (4) Cache original last-token logits. (5) Ablate in batches of 128 per prompt and layer. (6) Compute $\Delta_{\ell,n}$ and aggregate layer-wise means and differential rankings.

4.4 Metrics

Classifiers: per-class precision, recall, F1; macro-F1; accuracy; ROC-AUC (mean $\pm$ std across 5 seeds). **Direction:** held-out AUROC, bootstrap 95% CI, permutation p-value, baseline comparisons. **Ablation:** layer-wise mean importance; neuron-level differential importance.

Model	Acc.	F1	AUROC	TQA_{nat}
Qwen2.5-7B	$0.755_{\pm.013}$	$0.751_{\pm.014}$	$0.822_{\pm.012}$	0.640
LLaMA-2-7B	$0.769_{\pm.008}$	$0.741_{\pm.013}$	$0.846_{\pm.008}$	0.773

Table 3: Held-out hidden-state detector performance (5-seed mean ± std). TQA_{nat} uses TruthfulQA's dataset-provided native labels. ECE and Brier score reported in Section 5.9.

Model	s42	s7	s123	s256	s999
Qwen	.840	.820	.822	.825	.803
LLaMA	.840	.843	.861	.850	.838

Table 4: Per-seed AUROC across 5 random seeds (mean ± std: Qwen 0.822 ± 0.012; LLaMA 0.846 ± 0.008).

4.5 Reproducibility

HuggingFace identifiers: `meta-llama/Llama-2-7b-hf`, `Qwen/Qwen2.5-7B`. Seeds: $\{42, 7, 123, 256, 999\}$. Adam lr $= 10^{-4}$, dropout 0.3, 15 epochs, patience 3. Code and features released upon acceptance.

5 Results

5.1 Hidden-State Classifier Performance

Table 3 reports held-out performance across 5 seeds. LLaMA-2-7B is the primary model; Qwen2.5-7B is included as a secondary cross-architecture comparison only — it is not a co-primary result. The Qwen2.5-7B pseudo-to-native AUROC gap (0.21) is 3× larger than LLaMA-2-7B's (0.07), indicating substantially greater sensitivity to the pseudo-labelling scheme; its native-label performance (0.640) should be interpreted with corresponding caution.

On the aggregate held-out split, LLaMA-2-7B outperforms Qwen2.5-7B on all reported metrics with tighter variance. Three TruthfulQA-related AUROC values appear in this paper and should be distinguished clearly: (1) TruthfulQA pseudo-label AUROC (0.9193, Table 7) — the per-source slice evaluated under the same cosine-similarity oracle used for training, which is partially circular; (2) TruthfulQA native-label AUROC, full subset (0.8452, Table 5) — all 817 matched TruthfulQA samples evaluated against TruthfulQA's dataset-provided native labels; and (3) TruthfulQA native-label AUROC, held-out test split only (0.7731, Table 3) — the cleanest estimate, using only the 172 test-split samples unseen during training. The held-out pseudo-to-native gap is 0.031 (LLaMA-2-7B, Table 5); the full-subset gap is 0.07.

Model	$AUROC_{pse}$	$AUROC_{nat}$	Label agree.
Qwen2.5-7B	0.715	0.640	85.9%
LLaMA-2-7B	0.804	0.773	90.0%

Table 5: TruthfulQA native-label evaluation (817 matched samples, 172 in held-out test split). $AUROC_{pse}$ is the detector score under pseudo-labels on the TruthfulQA corpus slice; $AUROC_{nat}$ uses TruthfulQA's dataset-provided native labels.

5.2 Multi-Seed Stability

Table 4 reports per-seed AUROC. Both models are stable across seeds (std < 0.015). LLaMA-2-7B's tighter variance (0.0082) is particularly notable given its larger feature dimensionality (8192 vs. 7168).

5.3 TruthfulQA Native-Label Evaluation

LLaMA-2-7B's higher native-label AUROC (0.773 vs. 0.640) and higher label agreement (90.0% vs. 85.9%) indicate that its representations encode factuality-related signal that is more robust to the labeling oracle than Qwen2.5-7B's.

5.4 Secondary Oracle Validation

A secondary Wikipedia word-overlap oracle (threshold 0.2, retrievable subset of 1,505 samples) was used as a post-hoc robustness check. Under semantic-similarity labels, both models perform well (LLaMA: 0.844, Qwen: 0.749). Under Wikipedia overlap, LLaMA-2-7B falls to 0.464 (below chance) while Qwen2.5-7B reaches 0.610. The below-chance LLaMA result most likely reflects fluency-driven lexical divergence rather than a pseudo-label artefact — the TruthfulQA native-label result (0.7731, 90% agreement) argues against the artefact interpretation.

5.5 Surface-Level Baseline Comparison

Under pseudo-labels, Token F1 (0.917) and Jaccard overlap (0.917) achieve higher AUROC than the MLP detectors — but both are partially correlated with the similarity oracle used for labeling. Under TruthfulQA dataset-provided native labels, Token F1 drops to 0.806 (a fall of 0.111) while the LLaMA-2-7B MLP drops only to 0.773 (a fall of 0.067), suggesting the MLP captures signal that generalises better beyond the training oracle. Critically, Token F1 and Jaccard require a ground-truth reference answer at inference time, while the MLP operates solely on activations produced during ordinary decoding.

Method	$\text{AUROC}_{\text{pseudo}}$	$\text{AUROC}_{\text{native}}$
Token F1†	0.917	0.806
Jaccard overlap†	0.917	0.806
Answer length	0.506	0.543
MLP (Qwen2.5-7B)	$0.822_{\pm.012}$	0.640
MLP (LLaMA-2-7B)	$0.846_{\pm.008}$	0.773

Table 6: Independent surface-level baselines vs. hidden-state MLP detectors, evaluated under both pseudo-labels (full test split, n=464) and TruthfulQA dataset-provided native labels (TQA test subset, n=172). Similarity-based scores are excluded from the pseudo-label column as they are circular with the labeling oracle (τ=0.4). †Token F1 and Jaccard require a reference answer at inference time; the MLP requires only model activations.

Source	N	LLaMA AUROC	Qwen AUROC
TruthfulQA	817	0.9193	0.8252
TriviaQA	500	0.8411	0.8992
SQuAD	500	0.8979	0.8204
TQA_{nat}	817	0.8452	0.7118

Table 7: Per-source AUROC breakdown. TQA_{nat} uses TruthfulQA dataset-provided native labels. All sources evaluated on the corresponding corpus slice.

τ	Hall.%	LLaMA	Qwen	LLaMA F1
0.30	51.5	0.829	0.793	0.733
0.35	56.5	0.838	0.840	0.733
0.40*	62.1	0.839	0.841	0.722
0.45	67.6	0.877	0.848	0.764
0.50	72.0	0.846	0.831	0.751
Range	–	0.048	0.055	–

Table 8: Threshold sensitivity: MLP detector AUROC across $\tau \in [0.30, 0.50]$. LLaMA-2-7B and Qwen2.5-7B retrained at each τ. * = setting used throughout.

Train	Test	Within-src	LOO AUROC
TriviaQA + SQuAD	TruthfulQA	$0.843_{\pm.033}$	$0.595_{\pm.018}$
TruthfulQA + SQuAD	TriviaQA	$0.677_{\pm.065}$	$0.652_{\pm.008}$
TruthfulQA + TriviaQA	SQuAD	$0.805_{\pm.016}$	$0.587_{\pm.006}$
All three (mixed)	Mixed 20%	0.841 (reference)	

Table 9: Leave-one-out cross-dataset transfer (LLaMA-2-7B, 3 seeds mean ± std). Scaler fitted on training data only. Mixed-source trains on all three sources.

5.6 Per-Source AUROC Breakdown

Table 7 reports per-source AUROC for both models evaluated separately on each constituent dataset. LLaMA-2-7B is strongest on TruthfulQA (0.919) and SQuAD (0.898), while Qwen2.5-7B peaks on TriviaQA (0.899). The reversal on TriviaQA suggests model-specific sensitivity to question style. Both detectors exceed AUROC 0.82 on all three evaluated sources, confirming that aggregate performance is not driven by a single dominant dataset.

5.7 Pseudo-Label Threshold Sensitivity

Table 8 reports detector performance across $\tau \in [0.30, 0.50]$. LLaMA-2-7B AUROC ranges from 0.829 to 0.877 (spread 0.048); Qwen2.5-7B ranges from 0.793 to 0.848 (spread 0.055). Both models remain well above chance across all thresholds, supporting τ=0.4 as a principled mid-range value that balances class representation (62.1% hallucinated for LLaMA, 56.7% for Qwen).

5.8 Cross-Dataset Transfer

Table 9 reports leave-one-out (LOO) cross-dataset transfer. Training on two sources and testing on the held-out third drops AUROC to 0.587–0.652 vs. within-source 0.677–0.843, indicating dataset-specific components in the probe. Mixed-source training recovers full performance (0.841), suggesting training diversity rather than signal absence as the limiting factor. Appendix Figure 5 illustrates this within-source vs. cross-dataset gap.

5.9 Calibration and Error Analysis

The LLaMA-2-7B detector is well-calibrated (ECE = 0.064, Brier = 0.156), indicating predicted probabilities reliably reflect empirical hallucination rates. Of 464 test predictions, 109 are errors: 33 false negatives (missed hallucinations, mean model confidence 0.344) and 76 false positives (false alarms, mean confidence 0.319). Notably, 78% of all errors (85/109) involve moderate-similarity boundary cases (sim ≈ 0.4) where the pseudo-label itself is ambiguous. False positives show higher mean reference similarity (0.531) than false negatives (0.186), indicating the main confusion source is factual answers that are lexically distant from the reference string.

Appendix Figures 4 and 3 show the reliability diagram, confidence distribution, error analysis, and threshold robustness.

5.10 Comparison with Prior Work

Appendix Table 12 provides a per-dataset comparison against prior methods. Our per-source results (Table 7) provide contextual comparison on TruthfulQA and TriviaQA, though they are not a controlled benchmark. LLaMA-2-7B achieves numerically higher AUROC than previously reported values for HaloScope on TruthfulQA and TriviaQA under our non-matched evaluation conditions.

Layer	Targeted	Random	Ratio	Sig.
0	2.850	0.904	3.15×	***
1	3.198	0.988	3.24×	***
4	2.286	1.217	1.88×	***
15	3.029	1.287	2.35×	***
20	4.786	1.359	3.52×	***
30	8.908	1.338	6.66×	***
31	43.171	2.609	16.54×	***
Mean (all 32)	–	–	2.82×	***

Table 10: Targeted (top-k activation) vs. random (matched-k, same layer) neuron ablation importance. All 32 layers show targeted $>$ random ($p<0.001$, Welch's t-test, 3 seeds). Values are mean $\Delta_{\ell,n}$ (L2) over 200 prompts. *** $p<0.001$.

5.11 Naturalistic Direction Evaluation

The class-mean projection exceeds random directions and PCA, approaching the supervised logistic probe (0.648) without label-aware optimisation. This suggests factuality-relevant information occupies a direction in activation space discoverable from class means alone (Appendix Table 14).

5.12 Random-Neuron Ablation Baseline

Targeted neurons are substantially more intervention-sensitive than randomly sampled neurons from the same layer across all 32 layers (mean ratio 2.82×, all $p<0.001$). Layer 31 shows the largest effect (16.54×), consistent with its dominant position in the layer-wise importance analysis. Layers 0 and 1 also show clear differentiation (3.15× and 3.24×). This result supports the validity of the top-k activation filter as a targeted selection procedure (Table 10). This baseline should be interpreted as an exploratory robustness check rather than definitive mechanistic localisation.

Appendix Figure 5 shows the targeted vs. random neuron ablation results.

5.13 Layer-Wise Ablation Importance

Globally influential neurons concentrate at layers 0, 1, and 31 rather than being distributed uniformly. Neurons most selective for unanswerable (hallucination-adjacent) contexts are more front-loaded than the absolute-importance ranking (Appendix Table 15).

6 Discussion

6.1 Interpretation

Predictive results (primary contribution). The predictive pillar is the primary contribution of this paper and provides the strongest and most practically significant evidence. LLaMA-2-7B achieves AUROC 0.8462 ± 0.0082 across 5 seeds under pseudo-labels, and retains AUROC 0.7731 on the held-out TruthfulQA native-label test split (held-out gap 0.031; full-subset gap 0.07). The 90% pseudo/native label agreement provides evidence consistent with the view that the pseudo-labelling scheme was a reliable approximation and that the learned detector generalises to TruthfulQA's dataset-provided native labels beyond the single-reference pseudo-labeling oracle. Per-source evaluation (Table 7) further confirms that strong performance holds across TruthfulQA (0.919), TriviaQA (0.841), and SQuAD (0.898), with Qwen2.5-7B showing a complementary pattern (TriviaQA 0.899). In contrast, the internal-state detector operates solely on activations produced during ordinary decoding — making it deployable in settings where references are unavailable.

Intervention results (exploratory). The intervention analysis is consistent with depth-structured ablation sensitivity: layer-wise means peak at 31, 1, and 0, with a secondary rise at layer 4. Crucially, targeted neurons are substantially more intervention-sensitive than randomly sampled neurons across all 32 layers (mean 2.82×, all $p<0.001$), with the largest effect at layer 31 (16.54×). This validates the activation-based neuron selection procedure and supports the depth-nonuniform influence pattern as a non-trivial finding.

Geometric results (partial negative). AUROC 0.624 ± 0.032 is statistically above chance ($p < 0.001$, bootstrap CI 0.573–0.757; Appendix Table 13) but practically limited. This is best read as a partial negative result: a simple class-mean direction is not sufficient as a standalone detector, which is itself informative about LLaMA-2-7B's representational geometry. The weak but detectable signal motivates future work with larger paired datasets and stronger factuality labels rather than implying deployment readiness.

6.2 Cross-Pillar Alignment on a Shared Model

All three pillars share LLaMA-2-7B, enabling same-model cross-pillar comparison:

- **Pillar I (intervention):** Layers 0, 1, 31 show elevated ablation sensitivity; targeted neurons at these layers are $3.15\times$, $3.24\times$, and $16.54\times$ more influential than matched random neurons ($p<0.001$), validating the top-k selection procedure.

- **Pillar II (predictive):** Final-layer states achieve AUROC 0.8462 ± 0.0082 under pseudo-labels and 0.7731 under native labels, with a smaller oracle-dependence drop than surface-level signals (0.07 vs. 0.11 for Token F1).

- **Pillar III (geometric):** Final-layer embeddings show a statistically above-chance but weak-to-moderate directional signal (AUROC 0.624 ± 0.032).

This constitutes cross-pillar convergent evidence on a shared model, stronger than convergence across different architectures. Because datasets and labelling procedures still differ across pillars, this alignment is convergent but not mechanistically conclusive. Confirming any mechanistic link requires matched prompts and stronger factuality labels across all three pillars.

6.3 Efficiency and Practicality

The detector operates on activations already produced during decoding, avoiding external retrieval and repeated sampling. The tight seed variance (LLaMA: ± 0.0082) further supports deployment reliability.

6.4 Cautions

The pseudo-label concern is substantially mitigated for LLaMA-2-7B (90% agreement, 0.07 gap) but remains valid for Qwen2.5-7B (85.9%, 0.21 gap). The Wikipedia oracle below-chance result (AUROC 0.464) does not contradict the native-label result and is most likely attributable to fluency-driven lexical divergence. Intervention analysis conclusions are tentative: zero-ablation may overestimate ablation importance, and the top-15 activation filter may miss influential low-activation neurons.

7 Conclusion

We study whether internal representations contain usable hallucination-related signals under three complementary operationalizations — lightweight probing, neuron-level intervention, and directional geometry — with LLaMA-2-7B as a shared backbone. Our strongest evidence comes from final-layer LLaMA-2-7B probes: AUROC 0.8462 ± 0.0082 under pseudo-labels and 0.7731 on a held-out TruthfulQA native-label test split (full-subset pseudo-to-native gap 0.07; 90% label agreement). A Qwen2.5-7B cross-architecture replication achieves AUROC 0.8221 ± 0.0117. Intervention and geometry results provide exploratory supporting evidence for depth-nonuniform ablation sensitivity, validated by a random-neuron baseline showing targeted neurons are $2.82\times$ more influential than unselected neurons across all layers ($p<0.001$), and weak-to-moderate directional structure, with explicit uncertainty quantification. Sharing LLaMA-2-7B across all three pillars enables, to our knowledge, same-model cross-pillar comparison within this type of multi-pillar hallucination study. Per-source evaluation shows strong within-source performance across three evaluated sources (TruthfulQA, TriviaQA, SQuAD), while NQ exposes a limitation of fixed-threshold semantic labeling. The detector is well-calibrated (ECE = 0.064), and error analysis shows failures concentrate at ambiguous similarity-boundary cases rather than on clearly correct or incorrect generations. Leave-one-out cross-dataset evaluation reveals LOO AUROC of 0.595–0.652 vs. within-source 0.677–0.843, indicating dataset-specific components; mixed-source training recovers full performance (0.841). These results support internal representations as a practical low-latency substrate for hallucination monitoring.

Limitations

- **Pseudo-label validity:** Affects the interpretation of predictive AUROC as hallucination detection rather than semantic-similarity detection. Substantially mitigated for LLaMA-2-7B by TruthfulQA native-label validation (90% agreement, full-subset gap 0.07); Qwen2.5-7B's larger gap (0.21) indicates greater oracle sensitivity and limits confidence in its per-source results. Threshold sensitivity across $\tau \in [0.3, 0.5]$ (Table 8) shows AUROC spread of 0.048 for LLaMA-2-7B and 0.055

for Qwen2.5-7B, confirming stability.

- **Ablation coverage:** 200 SQuAD-v2 prompts; restricted to answerability failure, not the full diversity of hallucination types.
- **Candidate restriction:** Top-15 activation filter may miss influential low-activation neurons. The matched random-neuron baseline (Table 10) confirms selected neurons are $2.82\times$ more influential on average.
- **Zero-ablation:** May overestimate ablation importance for high-activation neurons by producing out-of-distribution representations. This affects the absolute magnitude of reported importance values but not necessarily the relative layer rankings; mean-ablation replication is reserved for future work.
- **Direction sample size:** 371 pairs; small per-dataset analysis would be underpowered.
- **Cross-pillar non-equivalence:** Shared backbone but different datasets and labels; alignment is convergent, not mechanistically conclusive.
- **Per-source evaluation:** LLaMA-2-7B NQ is excluded because pseudo-labeling under τ=0.4 assigns 100% hallucination labels to all 500 NQ samples: NQ reference answers are long and structured, producing systematically low cosine similarity against LLaMA-2-7B's shorter greedy generations regardless of factual correctness. This is a labeling-scheme limitation specific to NQ's reference format, not a model failure.
- **Model specificity:** Results are specific to LLaMA-2-7B and Qwen2.5-7B; generalisation to other architectures and scales is unknown. Cross-dataset LOO evaluation (Table 9) suggests dataset-specific components in the learned signal.
- **Wikipedia oracle:** Below-chance LLaMA-2-7B result (AUROC 0.464) under Wikipedia word-overlap is discussed in Appendix A. The native-label result (0.7731, 90% agreement) argues against a pseudo-label artefact interpretation.

References

Amos Azaria and Tom Mitchell. 2023. The internal state of an LLM knows when it's lying. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 967–976. Association for Computational Linguistics.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. 2022. Discovering latent knowledge in language models without supervision. *arXiv preprint*.

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

Yung-Sung Chuang, Linlu Qiu, Cheng-Yu Hsieh, Ranjay Krishna, Yoon Kim, and James R. Glass. 2024. 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*. Association for Computational Linguistics.

Shrey Desai and Greg Durrett. 2020. Calibration of pre-trained transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

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

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

Arian Hosseini, Xingdi Yuan, Nikolay Malkin, Aaron Courville, Alessandro Sordoni, and Rishabh Agarwal. 2024. V-STaR: Training verifiers for self-taught reasoners. In *Proceedings of the 2024 Conference on Language Modeling*.

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

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, and 17 others. 2022. Language models (mostly) know what they know. *arXiv preprint*.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023a. Inference-time intervention: Eliciting truthful answers from a language model. In *Advances in Neural Information Processing Systems*.

Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori Hashimoto, Luke Zettlemoyer, and Mike Lewis. 2023b. Contrastive decoding: Open-ended text generation as optimization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

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

Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. 2023. 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*. Association for Computational Linguistics.

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

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *International Conference on Learning Representations Workshop*.

Kiho Park, Yo Joong Choe, and Victor Veitch. 2023. The linear representation hypothesis and the geometry of large language models. *arXiv preprint*.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392. Association for Computational Linguistics.

Yusheng Song, Lirong Qiu, Xi Zhang, and Zhihao Tang. 2025. Hallucination detection via internal states and structured reasoning consistency in large language models. *arXiv preprint*.

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

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and 49 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint*.

Luan Zhang, Dandan Song, Zhijing Wu, Yuhang Tian, Changzhi Zhou, Jing Xu, Ziyi Yang, and Shuhao Zhang. 2025. Detecting hallucination in large language models through deep internal representation analysis. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pages 8357–8365.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, and 2 others. 2023. Representation engineering: A top-down approach to AI transparency. *arXiv preprint*.

A Wikipedia Oracle Cross-Validation

Model	Oracle	ROC-AUC
Qwen2.5-7B	Semantic similarity	0.7490
Qwen2.5-7B	Wikipedia overlap	0.6100
LLaMA-2-7B	Semantic similarity	0.8442
LLaMA-2-7B	Wikipedia overlap	0.4644

Table 11: ROC-AUC across two label oracles (retrievable subset, 1,505 of 2,317 samples). Wikipedia overlap is a supplementary diagnostic only and does not affect training or model selection.

The below-chance Wikipedia overlap for LLaMA-2-7B (AUROC 0.464) is a known open question. Two interpretations are plausible: (1) LLaMA-2-7B generates more fluent, lexically varied responses that reduce surface overlap even when factually correct — a known limitation of lexical-overlap metrics applied to capable generators; or (2) the detector has learned to predict

semantic embedding similarity rather than factuality per se. The TruthfulQA native-label result (AUROC 0.7731, 90% label agreement) provides evidence against interpretation (2): a detector that merely predicts cosine similarity would not retain 0.7731 under TruthfulQA's dataset-provided native labels. We treat the below-chance Wikipedia result as most consistent with interpretation (1), acknowledging that independent factuality annotations are needed for a definitive resolution.

B Additional Tables and Figures

LLM	Method	TQA	TriviaQA
Qwen2.5-7B	SAPLMA (Azaria and Mitchell, 2023)	$59.66_{\pm 1.69}$	$62.36_{\pm 1.38}$
	SelfCheckGPT (Manakul et al., 2023)	$55.08_{\pm 1.15}$	$74.65_{\pm 0.92}$
	Sem. Entropy (Farquhar et al., 2024)	$64.72_{\pm 1.26}$	$75.68_{\pm 0.88}$
	V-STaR (Hosseini et al., 2024)	$63.91_{\pm 0.93}$	$71.09_{\pm 1.15}$
	HaloScope (Du et al., 2024)	$79.31_{\pm 2.33}$	$81.52_{\pm 2.08}$
	Ours	**82.52**	**89.92**
LLaMA-2-7B	SAPLMA (Azaria and Mitchell, 2023)	$57.41_{\pm 1.71}$	$60.32_{\pm 1.83}$
	SelfCheckGPT (Manakul et al., 2023)	$52.95_{\pm 1.20}$	$73.22_{\pm 1.01}$
	Sem. Entropy (Farquhar et al., 2024)	$62.17_{\pm 1.32}$	$73.65_{\pm 0.77}$
	V-STaR (Hosseini et al., 2024)	$61.28_{\pm 1.12}$	$68.83_{\pm 1.33}$
	HaloScope (Du et al., 2024)	$78.64_{\pm 2.25}$	$77.40_{\pm 1.98}$
	Ours	**91.93**	**84.11**

Table 12: Per-dataset comparison with prior methods. Best baselines underlined. Baseline numbers from Song et al. (2025) use separate controlled splits; our values come from per-source corpus slices with a different prompting format — not a controlled benchmark. GSM8K not in our corpus.

Metric	Result
AUROC across 5 splits	0.624 ± 0.032
Bootstrap 95% CI	0.573–0.757
Permutation p-value	< 0.001 (0/1000)
Interpretation	Weak-to-moderate signal

Table 13: Naturalistic 371-sample directional projection evaluation (LLaMA-2-7B).

Method	AUROC	Above chance?
Random baseline	$0.503_{\pm 0.037}$	No
PCA baseline	0.500	No
Logistic baseline	0.648	Yes
Cosine-to-mean	0.563	Weakly
Directional proj.	$\mathbf{0.624}_{\pm \mathbf{0.032}}$	**Yes**

Table 14: Baseline comparison for the 371-sample direction task.

Layer	0	1	2	3	4	30	31
Imp.	≈14.5	≈18.6	≈4.4	≈4.1	≈6.1	≈3.8	≈20.8

Table 15: Selected layer-wise mean importance (LLaMA-2-7B, 200-prompt sample). Treat as indicative pending stability validation.

Rk	L	N	Imp.	Rk	L	N	Imp.
1	0	1512	147.5	11	1	339	34.8
2	31	3556	77.7	12	31	3839	33.6
3	31	1415	76.9	13	31	2215	32.0
4	31	1360	73.7	14	1	2298	31.5
5	31	3241	65.0	15	31	3909	30.6
6	31	2084	62.9	16	1	1076	30.2
7	31	1573	61.8	17	1	2789	30.0
8	31	2883	59.6	18	1	3209	28.7
9	0	2298	47.6	19	1	3431	28.0
10	31	1839	38.6	20	31	2021	27.7

Table 16: Top-20 neurons by mean importance (single 200-prompt sample).

Rank	Layer	Neuron	Diff. Imp.
1	0	1512	7.951
2	0	2298	6.315
3	1	2789	2.405
4	31	3591	2.250
5	1	1076	2.234
6	4	1512	1.637
7	1	2393	1.620
8	30	2042	1.531
9	30	2393	1.145
10	28	363	1.106

Table 17: Top differential neurons (unanswerable − answerable).

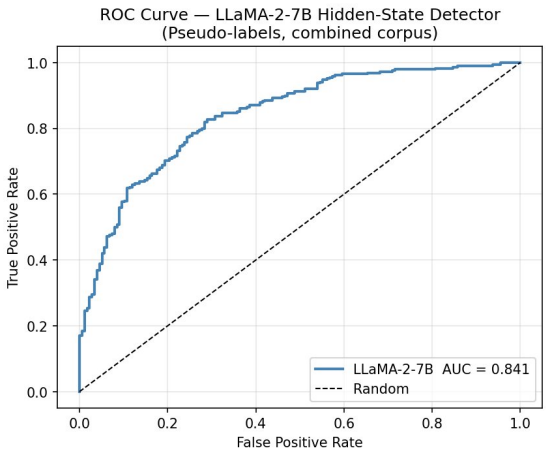


Figure 2: ROC curve for the LLaMA-2-7B hidden-state detector (pseudo-labels, combined corpus). AUC = 0.841.

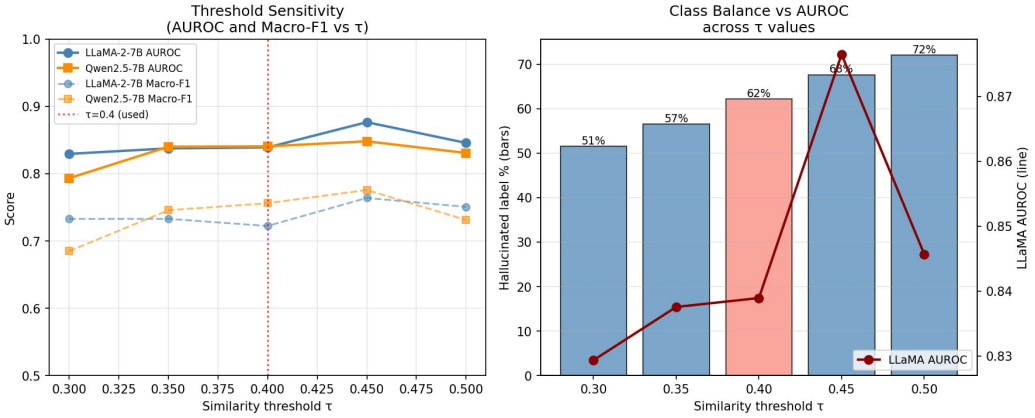


Figure 3: Threshold sensitivity: AUROC and Macro-F1 vs. τ for both models (left); class balance vs. AUROC across τ values (right). LLaMA-2-7B spread 0.048; Qwen2.5-7B spread 0.055. Both models remain well above chance across $\tau \in [0.30, 0.50]$.

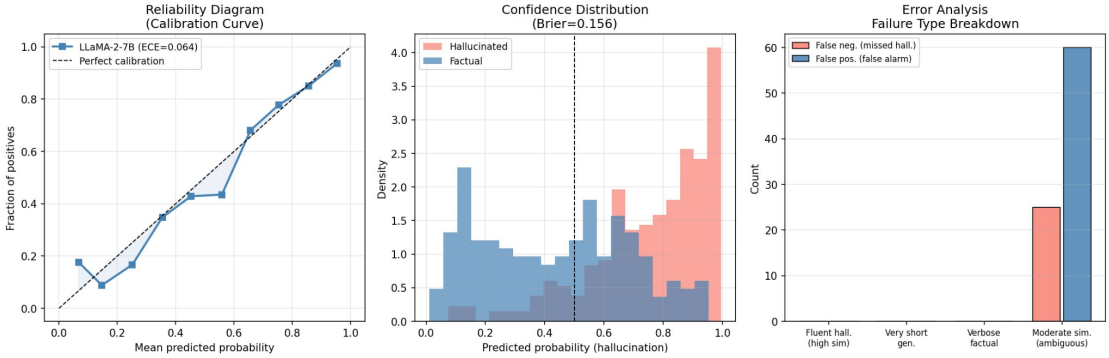
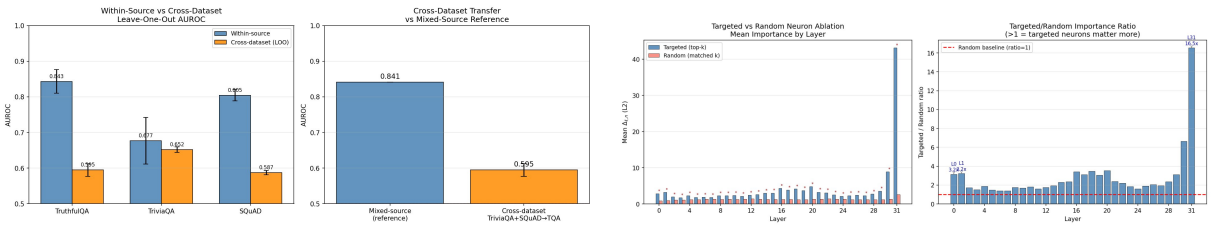


Figure 4: Reliability diagram (left), confidence distribution (centre), and error analysis (right) for the LLaMA-2-7B detector. ECE = 0.064; 78% of errors fall at ambiguous similarity-boundary cases.



(a) Within-source vs. cross-dataset leave-one-out AUROC and mixed-source reference.

(b) Targeted vs. random neuron ablation: mean importance and targeted/random ratio by layer.

Figure 5: Additional robustness analyses. Left: cross-dataset transfer; mixed-source training recovers performance lost under leave-one-out evaluation. Right: targeted vs. random neuron ablation; layer 31 shows a $16.5\times$ ratio, and all 32 layers show targeted $>$ random ($p < 0.001$).