Do Hallucination Signals Travel Across Models? A Cross-Dataset and Cross-Model Study

Anonymous ACL submission

Abstract

Internal-state hallucination probes offer a low-overhead alternative to inference-time verification, but it remains unclear whether they learn *dataset-general* grounding signals or *model-specific* representation artifacts. We study the portability of hallucination probes across three axes: (1) five diverse QA datasets within the same model, (2) multiple variants within the LLaMA and Mistral model families, and (3) direct cross-family weight transfer between LLaMA and Mistral. Our central finding is a two-way split: probes are *dataset-general but representation-space-specific*. Within the source model, probes retain meaningful signal across NQ-Open, TriviaQA, HotpotQA, PubMedQA, and SQuAD v2, with mean AUC of 0.69–0.76 across probe architectures. Same-family transfer preserves strong signal for close variants (LLaMA-2 base/chat; Mistral Instruct v0.2, avg. AUC 0.82), but generation-shift variants degrade severely and trigger widespread score inversion. Cross-family transfer fails in direction-dependent ways: LLaMA probes on Mistral produce calibration drift (avg. AUC 0.54, mean prob 0.81), while Mistral probes on LLaMA show partial sign reversal (avg. AUC 0.43; inverted 0.57). These results suggest that hallucination-related information is broadly present across model families but its geometry is model-specific, requiring calibration or alignment for portable deployment.

1 Introduction

Representation-level hallucination probes are attractive because they replace costly inference-time verification—retrieval, judge model calls, repeated sampling—with a single forward pass over internal activations (Azaria and Mitchell, 2023; Burns et al., 2023; Li et al., 2023). Prior work established that such probes detect hallucination with strong discrimination in single-model, single-dataset settings (Ji et al., 2023).

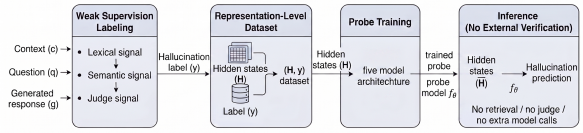


Figure 1: Overview of the representation-level distillation framework. External grounding signals supervise probe training offline; deployment consumes only hidden states. This paper studies whether trained probe weights remain useful when the target model or dataset changes.

The open portability question. Practical deployment is never a single-model, single-dataset setting. A hallucination monitor trained on one QA distribution must survive new domains, fine-tune updates, model version upgrades, and potentially a different model family. The NLP community has studied distributional shift for output-level metrics (Maynez et al., 2020; Min et al., 2023), but probe portability across model geometries is largely unexplored.

This paper. We perform a structured portability study using LLaMA-2-7B and Mistral-7B probes as training sources, addressing three research questions. **RQ1 (cross-dataset):** Do probes generalize beyond the training dataset to other QA domains? **RQ2 (same-family):** Do trained probe weights survive model version updates within the same family? **RQ3 (cross-family):** Can weights transfer directly when applied to a different model family's hidden states?

Summary of findings. Dataset generalization works: probes retain above-chance AUC across all five evaluation datasets, indicating learned signals reflect grounding risk patterns rather than SQuAD-specific artifacts. Same-family transfer is locally reliable for close variants but degrades sharply across generation-level architecture changes. Cross-family transfer fails without alignment, in structurally distinct ways for each direc-

tion.

Contributions. (1) A cross-dataset portability benchmark evaluating probes on five QA datasets, reporting both AUC and F1 to expose class-imbalance effects. (2) A same-family model transfer study across nine LLaMA and Mistral variants, revealing that instruction-tuning variants preserve signal while generation shifts cause inversion. (3) A cross-family transfer analysis showing asymmetric failure: calibration drift versus score-direction reversal, characterized both at the ensemble and per-probe level. (4) An architecture-specific robustness analysis of M0–M4, showing no universally portable architecture—structured probes excel close-in, depth-weighted M1 is most stable under geometric shift.

2 Related Work

2.1 Hallucination Detection

Hallucination in LLMs (Ji et al., 2023) manifests as fluent but factually unsupported generation. Existing detectors use lexical overlap (Maynez et al., 2020), retrieval-backed factuality estimation (Min et al., 2023), LLM-as-a-judge evaluation (Zheng et al., 2023), and sampling-based consistency checking via SelfCheckGPT (Manakul et al., 2023) or ConSens (Vankov et al., 2025). All require additional inference-time computation and cannot flag signals before generation completes.

2.2 Probing and Representation Analysis

Probing classifiers (Alain and Bengio, 2017) test what information is linearly encoded in hidden representations. Transformer layers distribute different syntactic and semantic abstractions across depth (Tenney et al., 2019; Ethayarajh, 2019). Most directly relevant: Azaria and Mitchell (2023) show internal states predict output falsehood; Burns et al. (2023) recover latent truth signals without direct supervision; Li et al. (2023) use probe directions for inference-time intervention. These works establish that internal states *contain* hallucination signals. Our paper asks whether those signals *transfer*.

2.3 Generalization and Calibration Under Shift

Transfer learning work shows that representations are model-specific (Devlin et al., 2019). Calibration under covariate shift (Guo et al., 2017) is particularly relevant: even when ranking signal survives, probability magnitude often breaks. We bring this calibration lens to probe portability, distinguishing inversion (sign reversal) from drift (miscalibrated magnitude).

2.4 Positioning

Unlike prior detection work, we do not ask *whether* hallucination is detectable from internal states. We ask *what transfers*: which signals survive dataset shift, model version updates, and family changes.

Relation to concurrent work. A concurrent anonymized submission (Anonymous, 2026) studies whether hallucination can be detected from hidden states in a single source setting: LLaMA-2-7B generations on a SQuAD v2-derived representation-level dataset. That work establishes source-model detectability under a fixed model–dataset setting. In contrast, the present paper studies portability: whether trained probes generalize across datasets, model variants, and model families. Our contributions are therefore the cross-dataset evaluation, same-family transfer study, cross-family transfer analysis, calibration/inversion diagnostics, and architecture-specific robustness analysis under distribution and representation shift.

3 Background: Framework and Probe Family

3.1 Representation-Level Distillation

Figure 1 illustrates the pipeline. During dataset construction, external grounding signals label model generations. For context c, question q, gold-answer set $\mathcal{A}$, and response g, an offline procedure assigns a hallucination label $y \in \{0, 1\}$. Simultaneously, the generator produces a hidden-state tensor $\tilde{H} \in \mathbb{R}^{L \times T_{\text{fixed}} \times D}$, padded or truncated to L=32, T_{fixed}=96, D=4096 (the common shape across all evaluated models). We fit a probe $\hat{y} = f_\theta(\tilde{H})$: grounding checks are paid once offline; inference uses only internal representations.

Source datasets. We construct two representation-level datasets on SQuAD v2 (Rajpurkar et al., 2018): one with LLaMA-2-7B (Touvron et al., 2023) as generator, one with Mistral-7B-v0.1 (Jiang et al., 2023). Each uses a hybrid labeling cascade (substring match, MiniLM semantic similarity (Reimers and Gurevych, 2019; Wang et al., 2020), abstention heuristics) followed by a Mistral-7B-Instruct-v0.2 judge; judge labels (y_{judge}) serve as the primary evaluation target throughout.

3.2 Probe Family M0–M4

We study five probes ordered by inductive bias over $\tilde{H}$ (Figures 2 and 3). All are trained identically: class-weighted binary cross-entropy on y_{judge}, AdamW (Loshchilov and Hutter, 2019) with cosine scheduling, batch size 8, $\eta{=}2{\times}10^{-4}$, early stopping, label smoothing $\epsilon{=}0.05$.

M0 (global pooled MLP) mean-pools over all $L{\times}T$ positions and classifies with an MLP. Tests whether hallucination leaves a first-order footprint under aggressive compression.

M1 (depth-aware weighted MLP) pools within each layer, then uses learned layer-attention for an adaptive depth summary. Tests whether signal is unevenly distributed across depth; also provides interpretable per-layer importance weights useful for layer analysis.

M2 (full layer–token transformer) flattens the full $L{\times}T$ grid, prepends a CLS token, and applies global self-attention. The most expressive probe; reference for unrestricted cross-layer, cross-token interactions.

M3 (hierarchical transformer) applies a local transformer within each layer's token sequence, then a global transformer across layer summaries. Tests whether signals emerge within layers before integrating across depth.

M4 (token-wise cross-layer fusion) fuses layer representations per token position via multi-head attention before a temporal transformer. Tests depth-first fusion versus M3's token-first factorization.

Source performance. On the held-out 5,000-sample SQuAD v2 test set (LLaMA-2-7B source model), probes achieve AUC 0.744 (M0), 0.807 (M1), 0.849 (M2), 0.858 (M3), and 0.829 (M4), with ECE 0.209, 0.143, 0.175, 0.103, 0.114 respectively. Probe latency is 0.15–5.62 ms batched—negligible relative to generation. These establish strong baselines; the paper then asks what survives when the model or dataset changes.

4 Experimental Design

4.1 Evaluation Protocol

All portability experiments use 1,000 generated samples per dataset. Generations use greedy decoding under the same prompt format as training; hidden states are extracted identically (final-token position, all L layers, $T{=}96$). Trained probe weights are applied *without modification*—no retraining, no adapter, no calibration. A Mistral-7B-Instruct-v0.2 judge provides consistent evaluation labels across all target generators.

Shape compatibility. All LLaMA-2/3/3.1 and Mistral-7B variants evaluated here share the $[32{\times}4096]$ hidden-state shape enabling direct weight transfer.

4.2 RQ1: Cross-Dataset Generalization

LLaMA probes (trained on SQuAD v2) are evaluated against the LLaMA-2-7B generator across five datasets representing diverse QA styles: **SQuAD v2** (Rajpurkar et al., 2018) (context-grounded, incl. unanswerable); **NQ-Open** (Kwiatkowski et al., 2019) (open-domain, short answers); **TriviaQA** (Joshi et al., 2017) (trivia-style, alias variation); **HotpotQA** (Yang et al., 2018) (multi-hop, context-heavy); **PubMedQA** (Jin et al., 2019) (biomedical domain shift).

4.3 RQ2: Same-Family Transfer

LLaMA family: LLaMA-2-7B base, LLaMA-2-7B chat (Touvron et al., 2023), LLaMA-3-8B base, LLaMA-3.1-8B base, LLaMA-3.1-8B Instruct (Meta AI, 2024). **Mistral family:** Mistral-7B v0.1 base, Mistral-7B Instruct v0.1, Mistral-7B Instruct v0.2, Mistral-7B v0.3 base (Jiang et al., 2023). Each run uses 100 samples per dataset across all five datasets.

4.4 RQ3: Cross-Family Transfer

We apply trained LLaMA probe weights to Mistral-7B-v0.1 generated hidden states and Mistral probe weights to LLaMA-2-7B hidden states, without any adaptation (1,000 samples per dataset).

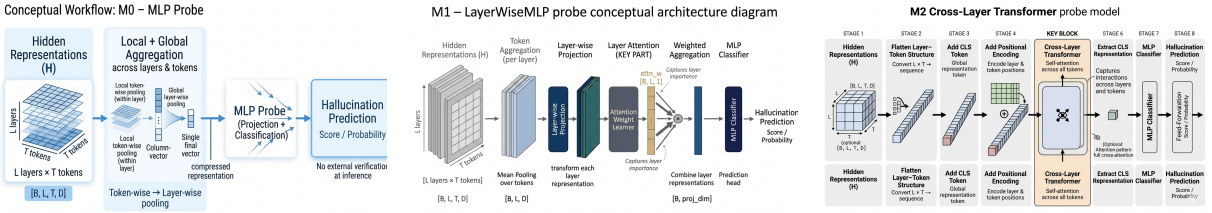
4.5 Metrics

AUC-ROC (primary; threshold-free ranking), **PR-AUC** (imbalanced positive class), **F1@0.5** (fixed-threshold deployment—sensitive to class imbalance, diagnostic of calibration), **mean predicted probability** (to diagnose overfiring), and **inversion mode** (whether score reversal recovers AUC above 0.5). Individual probe results are reported throughout; ensemble (mean probability across M0–M4) is a secondary comparison.

5 Results

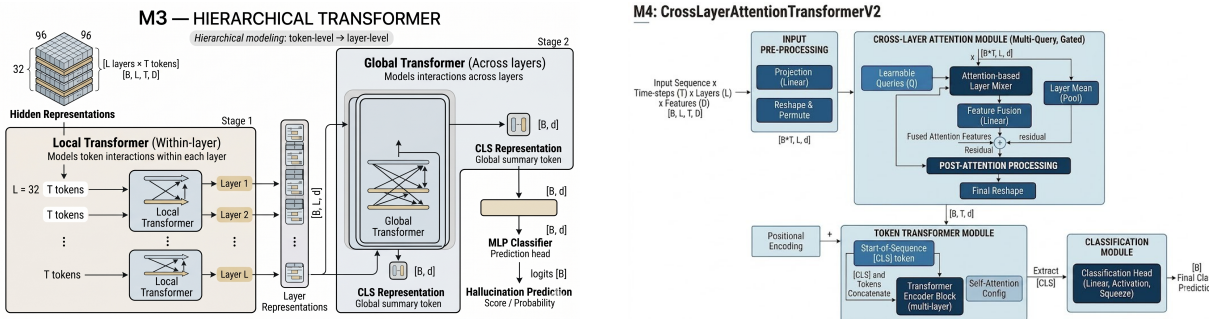
5.1 RQ1: Probes Generalize Across Datasets Within the Source Model

Table 1 reports AUC and F1@0.5 per probe across all five datasets for the LLaMA-2-7B generator.



(a) M0: Global pooled MLP (b) M1: Depth-aware MLP (c) M2: Cross-layer Transformer

Figure 2: Compression and full-interaction probes. M0 tests whether hallucination leaves a global pooled footprint, M1 preserves layer-level weighting, and M2 models unrestricted layer–token interactions.



(a) M3: Hierarchical Transformer (b) M4: Token-wise cross-layer fusion

Figure 3: Structured factorization probes. M3 first summarizes token trajectories within each layer and then reasons across layers; M4 first fuses cross-layer information at each token position and then models the token sequence.

All probes exceed chance uniformly across both metrics.

AUC vs. F1: why both matter. The AUC/F1 split in Table 1 is not incidental. TriviaQA shows AUC 0.73–0.80 but F1@0.5 only 0.22–0.35—a gap driven entirely by the 6% positive rate. At a fixed 0.5 threshold, the probe correctly ranks hallucinated responses above grounded ones (captured by AUC), but the majority-class skew means the default threshold is far too aggressive. Conversely, HotpotQA—with 33.5% positive rate—shows much tighter AUC/F1 alignment. This confirms that threshold calibration is an essential step before any deployment-time use, regardless of how strong the AUC appears.

Dataset-specific patterns. SQuAD v2 and NQ-Open yield the strongest results across both metrics, consistent with clear answerable/unanswerable grounding decisions. HotpotQA is harder (AUC 0.60–0.73), reflecting ambiguity in multi-hop grounding. PubMedQA shows the weakest but still above-chance signal (AUC 0.64–0.67), attributable to biomedical domain shift and high gold–judge disagreement (54.6%, Table 2); we discuss this instability in the Limitations.

Architecture behavior. M2 and M3 lead on datasets with clear grounding structure. On PubMedQA, M1 narrows the gap with M3 slightly leading on F1—depth-weighted pooling generalizes more stably to domain-shifted biomedical language than M2's full cross-layer self-attention. M0 establishes a lower bound: AUC $\geq$0.60 on all datasets confirms hallucination has a first-order footprint even under aggressive pooling.

The key takeaway is that probes are not learning SQuAD-specific artifacts: the learned signal captures grounding risk patterns recurring across diverse QA distributions, answer styles, and reasoning types.

5.2 RQ2: Same-Family Transfer—Close Variants Work, Generation Shifts Break

Table 3 summarizes same-family transfer using the mean best-mode AUC (maximum of normal and inverted-score AUC) averaged over five datasets, plus the count of datasets requiring score inversion.

LLaMA family. LLaMA-2 base to LLaMA-2 chat achieves avg. AUC 0.757, with only PubMedQA requiring inversion and TriviaQA reaching 0.955. Instruction tuning within the same model generation preserves representational geometry for hallucination detection.

	AUC-ROC						F1@0.5					
Dataset	M0	M1	M2	M3	M4	Ens.	M0	M1	M2	M3	M4	Ens.
SQuAD v2	0.752	0.787	**0.851**	0.846	0.821	0.846	0.489	0.500	**0.599**	0.509	0.526	0.558
NQ-Open	0.714	0.724	**0.804**	0.791	0.787	0.803	0.393	0.339	**0.503**	0.449	0.458	0.508
TriviaQA	0.733	0.709	0.780	**0.798**	0.789	0.805	0.249	0.222	**0.348**	0.241	0.349	0.333
HotpotQA	0.598	0.675	**0.728**	0.727	0.660	0.704	0.539	0.553	**0.554**	0.530	0.532	0.548
PubMedQA	0.644	**0.672**	0.659	0.644	0.650	0.682	0.336	0.367	0.263	**0.379**	0.350	0.364
Mean	0.688	0.713	**0.764**	0.761	0.741	0.768	0.401	0.396	**0.453**	0.422	0.443	0.462

Table 1: Cross-dataset AUC and F1@0.5 (LLaMA probes on LLaMA-2-7B, judge labels, n=1,000 per dataset). Bold = best individual probe per row. Comparing AUC and F1 reveals the class-imbalance effect: TriviaQA's AUC of 0.73–0.80 collapses to F1 of 0.22–0.35 at a fixed 0.5 threshold due to only 6% positive rate, confirming threshold calibration is essential for deployment.

Dataset	Judge pos.	Gold pos.	Agree.
TriviaQA	6.0%	3.3%	92.7%
NQ-Open	18.2%	18.2%	79.2%
HotpotQA	33.5%	56.6%	58.9%
PubMedQA	19.3%	44.5%	54.6%
SQuAD v2	21.3%	66.6%	51.1%

Table 2: Label statistics by dataset (LLaMA generator, n=1,000). Gold–judge agreement correlates with cross-dataset probe reliability. TriviaQA's high agreement reflects clean grounding decisions; SQuAD v2 and PubMedQA's low agreement reflects inherently harder cases.

Probe family	Target generator	Avg AUC	Inv./5
LLaMA probes	LLaMA-2-7B base	0.722	0
	LLaMA-2-7B chat	0.757	1
	LLaMA-3-8B base	0.581	4
	LLaMA-3.1-8B base	0.578	4
	LLaMA-3.1-8B Instruct	0.594	2
Mistral probes	Mistral-7B v0.1 base	0.663	0
	Mistral-7B Instruct v0.1	0.733	0
	Mistral-7B Instruct v0.2	**0.816**	1
	Mistral-7B v0.3 base	0.618	1

Table 3: Same-family model transfer (ensemble, judge labels, 100 samples per dataset, 5 datasets). "Avg AUC" = mean best-mode AUC; "Inv./5" = number of datasets requiring score inversion for best discrimination.

Direction	Avg AUC	Avg AUC†	Mean prob	Mode
LLaMA → Mistral	0.544	0.456	0.810	Drift
Mistral → LLaMA	0.427	0.573	0.732	Inversion

Table 4: Cross-family transfer summary (ensemble, judge labels, n=1,000, averaged over 5 datasets). AUC† = inverted-score AUC. LLaMA→Mistral fires near-universally (mean prob 0.81) with partial ranking preserved; Mistral→LLaMA shows partial sign reversal (AUC $<$ 0.5).

The LLaMA-2 to LLaMA-3/3.1 generation shifts are substantially harder: avg. AUC drops to 0.578–0.594 and four of five datasets require score inversion. This is a strong diagnostic signal that the LLaMA-3 manifold has geometrically shifted, not merely rescaled. LLaMA-3.1-8B Instruct recovers modestly (0.594, 2 inversions) relative to its base counterpart, suggesting instruction alignment partially realigns probe-readable geometry.

Mistral family. Mistral shows cleaner intra-family transfer. Mistral Instruct v0.2—the same model used as the labeling judge—achieves the strongest result (avg. AUC 0.816, only PubMedQA inverted). Instruct v0.1 also transfers well (0.733, 0 inversions), indicating Mistral's representation geometry is more stable across versions than LLaMA's generation-level shift. Base variants v0.1 and v0.3 are weaker (0.663, 0.618), consistent with base models having less task-aligned representation structure for context-grounded QA.

The key takeaway: same-family portability is real but *local*. Instruction tuning within a generation preserves geometry; generation-level architecture changes distort the learned boundary severely.

5.3 RQ3: Cross-Family Transfer Fails Without Alignment

Tables 4 and 5 characterize cross-family transfer at the ensemble level and per dataset respectively. The two directions show qualitatively distinct failure modes.

LLaMA probes → Mistral generator. Avg. AUC 0.544 across five datasets—marginally above

Dataset	LLaMA→Mistral AUC	LLaMA→Mistral Mean prob	Mistral→LLaMA AUC	Mistral→LLaMA Mean prob
SQuAD v2	0.567	0.821	0.458	0.730
NQ-Open	0.535	0.793	0.364	0.661
TriviaQA	0.453	0.811	0.367	0.763
HotpotQA	0.595	0.761	0.474	0.753
PubMedQA	0.568	0.863	0.471	0.753
Mean	0.544	0.810	0.427	0.732

Table 5: Cross-family transfer by dataset (ensemble). LLaMA→Mistral: AUC consistently above 0.5 but probabilities severely inflated. Mistral→LLaMA: AUC consistently below 0.5, confirming systematic sign reversal across all datasets. TriviaQA and NQ-Open are most inverted ($\leq$0.37).

Probe	LLaMA→Mistral AUC	LLaMA→Mistral AUC†	LLaMA→Mistral MProb	Mistral→LLaMA AUC	Mistral→LLaMA AUC†	Mistral→LLaMA MProb
M0	0.540	0.460	0.86	0.431	0.569	0.77
M1	0.552	0.448	0.60	0.478	0.522	0.72
M2	0.510	0.490	0.85	0.420	0.580	0.72
M3	0.521	0.479	0.92	0.481	0.519	0.94
M4	0.532	0.468	0.72	0.442	0.558	0.54
Ens.	0.544	0.456	0.81	0.427	0.573	0.73

Table 6: Cross-family per-probe AUC (avg. over 5 datasets, n=1,000 each). AUC† = inverted-score AUC. MProb = mean predicted hallucination probability. Three patterns emerge: M1 is least inverted in Mistral→LLaMA (0.478); M3's high MProb in LLaMA→Mistral (0.92) indicates near-universal overfiring; M4 is most extreme in both directions.

chance. The mean predicted probability is 0.810, far above judge-positive rates of 6–33%, indicating severe overfiring. AUC above 0.5 confirms that rank ordering is partially preserved, but calibration is broken. Notably, TriviaQA is the weakest (AUC 0.453), and PubMedQA shows the most extreme overfiring (mean prob 0.863), suggesting Mistral generates biomedical responses that activate the LLaMA probe's hallucination features at particularly high magnitude.

Mistral probes → LLaMA generator. Avg. AUC 0.427—consistently below 0.5—with inverted scores yielding avg. AUC 0.573. NQ-Open and TriviaQA are the most inverted (AUC 0.364, 0.367), while HotpotQA is the least (0.474), suggesting that multi-hop grounding patterns are the most structurally similar across families. This partial sign reversal indicates that features the Mistral probe learned to associate with hallucination are more strongly activated by *grounded* LLaMA responses.

Asymmetry interpretation. Cross-family transfer is not random failure—it fails in structured, direction-dependent ways. This asymmetry implies the two families encode hallucination-relevant information in geometrically related but directionally opposite regions of representation space.

5.4 Architecture-Specific Transfer Behavior

Table 6 reports per-probe AUC for both cross-family directions, revealing that architecture choice matters significantly under transfer.

Several probe-specific patterns are noteworthy. **M1** is the least inverted in the Mistral→LLaMA direction (normal AUC 0.478, closest to 0.5) and has the lowest mean probability in LLaMA→Mistral (0.60 vs. 0.72–0.92 for others), making it the most calibration-stable probe under cross-family shift. Its depth-weighted pooling does not lock into family-specific cross-layer interaction patterns the way M2 and M4 do.

M2 is the most strongly inverted in Mistral→LLaMA (0.420), consistent with its unrestricted cross-layer self-attention encoding highly model-specific geometry.

M3 shows an interesting split: it resists inversion most strongly among the structured probes (0.481 normal), suggesting its hierarchical factorization—handling tokens within layers first, then integrating across layers—captures coarser depth patterns that are more structurally similar across families. However, M3's mean probability in LLaMA→Mistral is 0.92, nearly universal overfiring, indicating that the hierarchical CLS token activates at maximal magnitude in the Mistral geometry.

M4 shows the most extreme behavior in both directions: near-universal overfiring in LLaMA→Mistral and strong inversion in Mistral→LLaMA. Its depth-first, token-local fusion appears to encode the most model-specific cross-layer patterns.

For cross-dataset transfer within the source model (Table 1), M2 and M3 dominate. For same-family generation-shift (Table 3), M1's simpler depth pooling is most stable. For cross-family settings (Table 6), no architecture dominates cleanly: M1 is most stable on the Mistral→LLaMA direction, but the ensemble averages conceal this. The practical implication is that there is no universally portable architecture—probe selection should be transfer-setting-aware.

6 Discussion

Dataset-general, representation-space-specific. The learned signal is not SQuAD-specific: it captures grounding risk patterns recurring across TriviaQA's alias variation, HotpotQA's multi-hop reasoning, and NQ-Open's open-domain style. This generality is non-trivial—SQuAD v2 and TriviaQA have very different answer distributions, yet probes trained on SQuAD retain strong discrimination on TriviaQA (AUC 0.73–0.80). At the same time, the same probe architectures can be *retrained* effectively on Mistral (up to AUC 0.816), while direct weight transfer from LLaMA to Mistral yields near-random discrimination. The signal exists in Mistral's space; it requires a probe trained to that geometry.

Calibration drift vs. score inversion. The two cross-family failure modes have different remediation paths. Calibration drift (LLaMA→Mistral, mean prob 0.81): rank ordering is partially preserved (AUC above 0.5); temperature scaling or a small calibration split may be sufficient to recover useful signal without retraining. Score inversion (Mistral→LLaMA, AUC 0.427): the decision boundary partially flips sign. Simple recalibration is insufficient; the probe needs either retraining, representation alignment (e.g., learned CKA-based mapping), or at minimum an inversion check on a small calibration set from the target model. The per-probe results in Table 6 further suggest that M1 may be the most recoverable architecture in both directions: lowest inversion severity and lowest overfiring magnitude.

Deployment protocol. These results motivate a tiered deployment approach. *Same-model/close fine-tune:* probes are viable with standard calibration (avg. AUC 0.72–0.82). *Same-family generation shift* (e.g., LLaMA-2 → LLaMA-3): probe weights are unreliable without retraining; widespread inversion makes threshold tuning insufficient. *Cross-family transfer:* direct weight application is not recommended; retraining on the new model or representation alignment is required. In all cases, probes should be advisory signals rather than hard safety gates, and calibration must be verified on held-out data from the target model and domain.

7 Conclusion

We studied the portability of internal-state hallucination probes across datasets, model variants, and model families. The central finding is a two-way split: probe signals are *dataset-general* within a model's representation space, but *representation-space-specific* across model families and large architecture shifts.

Probes trained on SQuAD-style supervision retain useful hallucination signal across five diverse QA datasets (avg. AUC 0.69–0.76), with M2 and M3 leading on structured datasets and M1 showing relative stability under domain shift. The AUC/F1 comparison reveals a critical deployment caveat: strong AUC on imbalanced datasets such as TriviaQA (AUC $\approx$0.80, F1@0.5 $\approx$0.28) does not translate to usable F1 without threshold calibration. Same-family transfer preserves strong signal for close variants (avg. AUC up to 0.82 for Mistral Instruct v0.2), while generation-level shifts cause widespread score inversion. Cross-family transfer fails in direction-dependent ways—calibration drift from LLaMA to Mistral, score inversion from Mistral to LLaMA—with M1 being the most transfer-stable architecture in both cases.

These findings motivate two concrete directions: lightweight representation alignment methods for cross-family portability, and inversion-aware calibration protocols for same-family generation-shift scenarios. More broadly, the results separate the *existence* of internal hallucination signals from the *portability* of learned classifiers over them, suggesting that hallucination-related information is broadly distributed across model families but its geometry is model-specific.

Limitations

PubMedQA instability. PubMedQA consistently shows the weakest cross-dataset transfer and high label disagreement between gold annotations and the Mistral judge (54.6%). This reflects a genuine methodological limitation: the judge is trained on general-domain language and may systematically mislabel biomedical responses, making AUC estimates on this dataset noisy and potentially underestimated. Evaluation with a domain-appropriate biomedical judge or human annotation on a PubMedQA subset is needed for reliable conclusions.

Small sample size for family sweep. The family sweep uses 100 samples per dataset per model, sufficient to identify inversion patterns but potentially unstable for AUC estimates in close-transfer settings where differences are small. Re-running strong configurations at 500–1,000 samples per dataset is planned.

Judge label dependence. All portability evaluation uses Mistral-7B-Instruct-v0.2 as the judge. For Mistral-family targets, this may introduce a favorable bias. Multiple judges or human validation on a subset would provide more balanced evaluation.

Greedy decoding and fixed prompts. All generations use greedy decoding with a fixed prompt format. Sampling-based decoding and prompt variation may produce different hidden-state distributions, changing both probe performance and transfer behavior.

Shape compatibility constraint. Direct weight transfer requires identical $[L \times D]$ hidden-state shape. Models with different hidden dimensions are excluded from this study without a learned adapter.

Cross-family alignment not yet evaluated. We evaluate direct weight transfer and document failure modes, but do not yet implement representation alignment methods (e.g., CKA-based linear maps, Procrustes alignment). These are the natural next step given our finding that cross-family signals exist but are geometrically misaligned.

References

Guillaume Alain and Yoshua Bengio. 2017. Understanding intermediate layers using linear classifier probes. In *International Conference on Learning Representations (ICLR Workshop)*.

Anonymous. 2026. Do LLMs know when they hallucinate? A representation learning approach. Under review.

Amos Azaria and Tom Mitchell. 2023. The internal state of an llm knows when it's lying. *arXiv preprint arXiv:2304.13734*.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. 2023. Discovering latent knowledge in language models without supervision. *arXiv preprint arXiv:2212.03827*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*.

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? comparing the geometry of bert, elmo, and gpt-2 embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tianjie Yu, Dan Su, Yi Xu, Etsuko Ishii, Younghee Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12).

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. PubMedQA: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Mandar Joshi, Eunsol Choi, Daniel 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 (ACL)*.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Lee, Adam Roberts, Noam Shazeer, Alexandre Lacoste, Sheena Rashford, and Kristina Toutanova. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association of Computational Linguistics*, 7:452–466.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023. Inference-time intervention: Eliciting truthful answers from a language model. *arXiv preprint arXiv:2306.03341*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*.

Potsawee Manakul, Adian Liusie, and Mark 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 (EMNLP)*.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings*

of the 58th Annual Meeting of the Association for Computational Linguistics (ACL).

Meta AI. 2024. Meta Llama 3: The most capable openly available LLM to date. https://ai.meta.com/blog/meta-llama-3/.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP).*

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 (ACL).*

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP-IJCNLP).*

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. What do you learn from context? probing for sentence structure in contextualized word representations. In *International Conference on Learning Representations (ICLR).*

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. *Preprint*, arXiv:2307.09288.

Ivan Vankov, Matyo Ivanov, Adriana Correia, and Victor Botev. 2025. Consens: Assessing context grounding in open-book question answering. *arXiv preprint arXiv:2505.00065.*

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. MiniLM: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP).*

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-bench and chatbot arena. *arXiv preprint arXiv:2306.05685.*

A Additional Probe Results

A.1 Per-probe cross-dataset PR-AUC.

Table 7 reports PR-AUC alongside AUC for completeness.

Dataset	Probe	AUC	PR-AUC
SQuAD v2	M0	0.752	0.427
	M1	0.787	0.514
	M2	0.851	0.651
	M3	0.846	0.648
	M4	0.821	0.582
NQ-Open	M0	0.714	0.288
	M1	0.724	0.301
	M2	0.804	0.523
	M3	0.791	0.522
	M4	0.787	0.476
TriviaQA	M0	0.733	0.197
	M1	0.709	0.149
	M2	0.780	0.400
	M3	0.798	0.376
	M4	0.789	0.297
HotpotQA	M0	0.598	0.392
	M1	0.675	0.473
	M2	0.728	0.590
	M3	0.727	0.579
	M4	0.660	0.505
PubMedQA	M0	0.644	0.320
	M1	0.672	0.369
	M2	0.659	0.308
	M3	0.644	0.278
	M4	0.650	0.322

Table 7: Full per-probe AUC and PR-AUC (LLaMA probes on LLaMA-2-7B, n=1,000 per dataset, judge labels).

A.2 Source model performance.

Table 8 reports the held-out source-model baseline from which portability experiments depart.

Probe	AUC	PR-AUC	F1	Best F1	τ^*	ECE
M0	0.744	0.512	0.532	0.535	0.547	0.209
M1	0.807	0.607	0.603	0.608	0.458	0.143
M2	0.849	0.696	0.648	0.660	0.645	0.175
M3	0.858	0.706	0.664	0.665	0.473	0.103
M4	0.829	0.645	0.624	0.628	0.365	0.114

Table 8: Source model performance: LLaMA probes on the 5,000-row held-out SQuAD v2 test set. τ^* = optimal F1 threshold.