WHEN ANSWERS STRAY FROM QUESTIONS: HALLUCINATION DETECTION VIA QUESTION-ANSWER ORTHOGONAL DECOMPOSITION

Anonymous authors
Paper under double-blind review

ABSTRACT

Hallucination detection in large language models (LLMs) requires balancing accuracy, efficiency, and robustness to distribution shift. Black-box consistency methods are effective but demand repeated inference; single-pass white-box probes are efficient yet treat answer representations in isolation, often degrading sharply under domain shift. We propose QAOD (**Q**uestion-**A**nswer **O**rthogonal **D**ecomposition), a single-pass framework that projects away the question-aligned direction from the answer representation to obtain a question-orthogonal component for cross-domain correctness prediction. To identify informative signals, QAOD further selects layers via diversity-penalized Fisher scoring and discriminative neurons via Fisher importance. To address both in-domain detection and cross-domain generalization, we design two complementary probing strategies: pairing the orthogonal component with question context yields a joint probe that maximizes in-domain discriminability, while using the orthogonal component alone supports robust transfer in our cross-domain evaluations. QAOD's joint probe achieves the best in-domain AUROC across all evaluated model-dataset pairs, while the orthogonal-only probe delivers the strongest OOD transfer, surpassing the strongest evaluated white-box baseline for each model by 6.27–14.22 AUROC percentage points on BioASQ at under 2% of generation cost.

1 INTRODUCTION

Despite their remarkable progress in natural language generation and reasoning, large language models (LLMs) remain vulnerable to hallucination, the production of factually unsupported or fabricated content (Huang et al., 2025; Ji et al., 2023; Tonmoy et al., 2024; Zhang et al., 2025c). This limitation continues to hinder the safe use of LLMs in high-stakes settings such as healthcare and legal services (Wang et al., 2024; Azamfirei et al., 2023).

A common strategy for hallucination detection samples multiple candidate answers and compares their consistency. Although effective, such black-box methods typically require repeated inference, incurring substantial computational overhead (Abdaljalil et al., 2025; Manakul et al., 2023; Farquhar et al., 2024). This efficiency bottleneck has motivated a second line of work that aims to detect hallucinations from a single generation by probing internal states of the model (Azaria & Mitchell, 2023; Su et al., 2024; Sriramanan et al., 2024). These single-pass white-box methods are faster, but many of them rely primarily on answer-side representations and may be sensitive to domain shift or to how factuality-related signals are entangled with contextual priors (Kwiatkowski et al., 2019).

We take a third path: rather than comparing across multiple sampled outputs or probing the answer representation alone, QAOD (**Q**uestion-**A**nswer **O**rthogonal **D**ecomposition) exploits the geometric relationship between question and answer representations within a single forward pass. Through this geometric decomposition, we realize the following contributions:

- **Geometric decoupling for robust representations**: We introduce $\mathcal{H}_{V_\perp}$ by removing the question-aligned component from answer representations, yielding a residual representation that empirically improves cross-domain correctness prediction. A joint variant $\mathcal{H}_{Q \oplus V_\perp}$ additionally incorporates question context for improved in-domain performance.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

- **Fisher-based layer and neuron selection**: We propose a Fisher-discriminant criterion to identify informative layers and neurons in a single pass over training features, retaining the most discriminative signals without iterative optimization.
- **Efficient detection for in-domain and cross-domain settings**: The joint probe achieves the best in-domain AUROC on all evaluated model-dataset pairs and the orthogonal-only probe surpasses the strongest evaluated white-box baseline for each model by 6.27–14.22 AUROC percentage points on zero-shot BioASQ. Both operate in a single forward pass at under 2% of generation cost.

2 Related work

2.1 Black-box methods: uncertainty estimation

Hallucination detection has been studied extensively through output probabilities, repeated sampling, and semantic consistency. SelfCheckGPT (Manakul et al., 2023) detects hallucinations by sampling multiple responses and measuring their mutual consistency. Semantic entropy (Farquhar et al., 2024) addresses the mismatch between token diversity and semantic agreement by clustering responses via bidirectional entailment and computing class-level entropy, while SINdex (Abdaljalil et al., 2025) extends this with cosine-similarity-based clustering and coherence-weighted entropy. Semantic Dispersion (Lin et al., 2024), Self-Prompt (Kadavath et al., 2022), direct verbalization (Lin et al., 2022), and ranked-voting self-consistency (Wang et al., 2025) offer further variants. These approaches are effective, but require multiple inference passes, incurring substantial computational overhead (Manakul et al., 2023; Chern et al., 2023; Kadavath et al., 2022).

2.2 White-box methods: internal state-based probing

White-box methods instead probe hidden states for factuality-related signals (Belinkov, 2022). Prior work has shown that factual knowledge can appear as a linear direction in hidden space (Li et al., 2021; Marks & Tegmark, 2024; Zou et al., 2025), and that such internal signals can be exploited for probing even when the model output is hallucinated (Li et al., 2024). Existing methods include lightweight classifiers on hidden states (Azaria & Mitchell, 2023; Su et al., 2024), layer-comparison strategies such as DoLa (Chuang et al., 2024), and other layer-wise or contrastive probes that aggregate intermediate representations across layers (Kim et al., 2025; Sriramanan et al., 2024; Chen et al., 2024b;a). Recent methods further enrich the detection signal: PRISM elicits answer-conditioned hidden-state evidence through a verification prompt (Zhang et al., 2025a); ICR Probe characterizes cross-layer state-update dynamics using hidden states and attention maps (Zhang et al., 2025d); MHAD uses auxiliary probes to select neurons and layers, then concatenates representations at the initial and final generation steps (Zhang et al., 2025b); and SpikeScore quantifies uncertainty fluctuations over multi-turn continuations (Deng et al., 2026). These works make initial progress toward cross-domain generalization, but involve distinct practical costs: PRISM depends on prompt quality and prompt search; ICR Probe's attention maps and divergences substantially increase inference time and memory use; MHAD trains auxiliary probes across layers at two generation stages; and SpikeScore requires repeated multi-turn inference. This motivates detectors that achieve robust OOD performance with low detection overhead.

3 Methodology

The hallucination detection framework QAoD consists of two branches. The offline branch identifies informative layers and neurons during training, while the online branch performs single-pass detection at test time. Figure 1 provides an overview of this offline-online pipeline.

3.1 Hidden state extraction

For question sequence S_Q and answer sequence S_A, we extract the final-token hidden states at each transformer layer l, denoted $h_Q^{(l)}, h_A^{(l)} \in \mathbb{R}^d$. In decoder-only models, causal self-attention makes the final token a natural read-out point because it aggregates preceding information (Vaswani et al.,

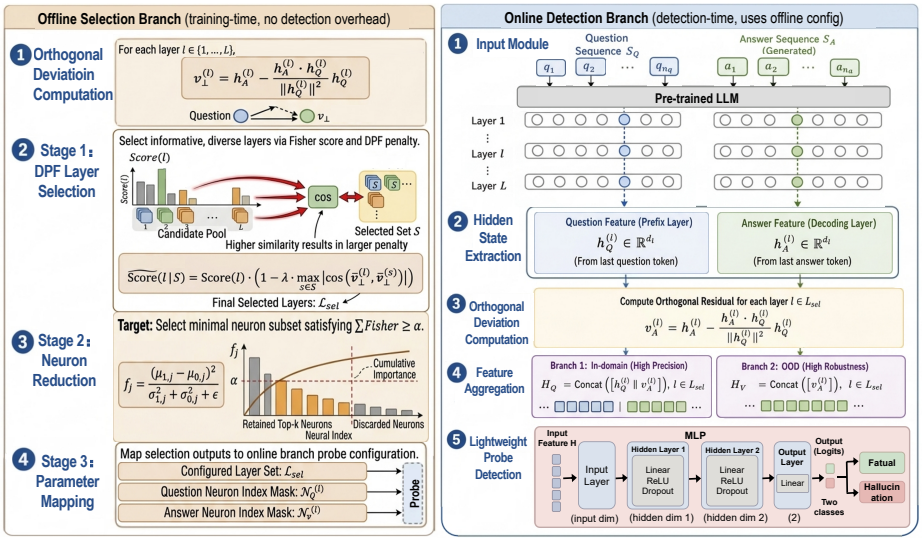


Figure 1: QAoD architecture: an offline selection branch identifies discriminative layers and neurons via Fisher-based scores; the online backbone extracts selected features and computes question-orthogonal components in a single forward pass.

2023); prior probing work also supports its effectiveness for factuality detection (Azaria & Mitchell, 2023; Su et al., 2024; Li et al., 2024).

3.2 Orthogonal Deviation Computation

For each transformer layer $l \in \{1, \ldots, L\}$, we compute the question-orthogonal component $v_\perp^{(l)}$ by removing from the answer representation its component aligned with the question representation:

$$v_\perp^{(l)} = h_A^{(l)} - \mathrm{proj}_{h_Q^{(l)}}(h_A^{(l)}) = h_A^{(l)} - \frac{h_A^{(l)} \cdot h_Q^{(l)}}{\|h_Q^{(l)}\|^2} h_Q^{(l)}. \tag{1}$$

The sample-wise projection removes the component of $h_A^{(l)}$ collinear with the current question direction, leaving a question-orthogonal component expected to attenuate question-conditioned domain variation. Section 4.3 empirically examines whether this orthogonal decomposition improves OOD stability.

3.3 Fisher-discriminant layer and neuron selection

Concatenating hidden states from all layers creates parameter redundancy and overfitting risk in deep LLMs. We therefore use Fisher discriminant scores based on class-conditional statistics to select informative layers and neurons.

3.3.1 Layer scoring via multivariate Fisher discriminant

For each candidate layer l, we compute Fisher scores over the question representations $h_Q^{(l)}$ and the question-orthogonal components $v_\perp^{(l)}$. We use the score

$$F(X, y) = \frac{\|\mu_1 - \mu_0\|^2}{\mathrm{tr}(\Sigma_1) + \mathrm{tr}(\Sigma_0)} \tag{2}$$

where μ_c and Σ_c are the class-conditional mean and covariance estimated per label group from the training set. In our implementation, the denominator uses a per-dimension diagonal variance approximation. The layer score is $\mathrm{Score}(l) = F\left(v_\perp^{(l)}, y\right)$ for $\mathcal{H}_{V_\perp}$, and $\mathrm{Score}(l) = \left(F\left(h_Q^{(l)}, y\right) + F\left(v_\perp^{(l)}, y\right)\right)/2$ for $\mathcal{H}_{Q \oplus V_\perp}$.

Diversity-penalized Fisher (DPF) layer selection. Since adjacent layers share high inter-layer representational similarity, the pure top=K Fisher ranking tends to select layers encode largely redundant information while missing complementary signals from distinct processing stages. We therefore employ a greedy *Diversity-Penalized Fisher* (DPF) procedure: given the set S of already-selected layers, each candidate l is scored as

$$\widetilde{\text{Score}}(l \mid S) = \text{Score}(l) \cdot \left(1 - \lambda \cdot \max_{s \in S} \left|\cos\left(\bar{v}_{\perp}^{(l)}, \bar{v}_{\perp}^{(s)}\right)\right|\right), \tag{3}$$

where $\bar{v}_{\perp}^{(l)}$ is the mean $v_{\perp}$ direction at layer l and $\lambda \in [0, 1]$ controls the diversity-discriminability trade-off. At each greedy step the highest-scoring layer is added to S; the process repeats until $|S| = K$, yielding the final selected set $\mathcal{L}_{sel}$.

3.3.2 Neuron selection via per-dimension Fisher scores

Within each selected layer l, we further compress representations by identifying the most discriminative neuron subsets. For each dimension $j \in \{1, \ldots, d\}$, the 1D Fisher score is:

$$f_j = \frac{(\mu_{1,j} - \mu_{0,j})^2}{\sigma_{1,j}^2 + \sigma_{0,j}^2 + \epsilon}. \tag{4}$$

Neurons are ranked by f_j and selected via an α-threshold cumulative importance criterion: the smallest set of neurons whose cumulative f_j covers $\geq \alpha$ fraction of the total Fisher importance is retained. This selection is applied independently to $h_Q^{(l)}$ and $v_{\perp}^{(l)}$, respecting the orthogonal decomposition and avoiding conflation of contextual priors with deviation signals. The procedure needs only one pass over training features, accumulating class-wise statistics for closed-form ranking.

3.4 Joint feature construction

The Fisher selection procedure (Section 3.3) identifies the most discriminative neuron subsets from both h_Q and $v_{\perp}$. Depending on the deployment scenario, these two feature sources can be combined in different ways: incorporating h_Q provides explicit question context that can enhance in-domain discriminability, while using $v_{\perp}$ alone supports cross-domain transfer in our evaluations. We therefore define two probe inputs that serve these different operational priorities.

Question-orthogonal joint ($\mathcal{H}_{Q \oplus V_{\perp}}$): Concatenates question and question-orthogonal-component features, preserving both query context and deviation signal:

$$\mathcal{H}_{Q \oplus V_{\perp}} = \text{Concat}\left([h_Q^{(l)}[\mathcal{N}_Q^{(l)}], v_{\perp}^{(l)}[\mathcal{N}_v^{(l)}]]\right)_{l \in \mathcal{L}_{sel}}, \tag{5}$$

where $\mathcal{N}_Q^{(l)}$ and $\mathcal{N}_v^{(l)}$ denote the Fisher-selected neuron indices for h_Q and $v_{\perp}$ at layer l, respectively.

Question-orthogonal component ($\mathcal{H}_{V_{\perp}}$): Uses only the question-orthogonal component:

$$\mathcal{H}_{V_{\perp}} = \text{Concat}\left(v_{\perp}^{(l)}[\mathcal{N}_v^{(l)}]\right)_{l \in \mathcal{L}_{sel}}. \tag{6}$$

The concatenated feature vector is standardized and fed into a lightweight MLP (1024 and 128 dimensions, ReLU, dropout, cross-entropy loss).

4 Experiments

4.1 Experimental setup

Benchmarks. We evaluate on four benchmarks: TriviaQA (Joshi et al., 2017), SQuAD (Rajpurkar et al., 2018), NQ (Natural Questions) (Kwiatkowski et al., 2019), and BioASQ (Krithara et al., 2022).

Models. We test on four open-weight LLMs: gemma-2-2b (Team et al., 2024), Llama-2-7B-chat (Touvron et al., 2023), Qwen3-14B (Yang et al., 2025), and Qwen3-30B-A3B (Yang et al., 2025).

Baselines. We compare against output-probability heuristics (PPL, Logit Entropy), black-box sampling consistency methods (SINdex (Abdaljalil et al., 2025), P_{false} (Kadavath et al., 2022)), and internal state-based white-box probing methods (P_{ik} (Kadavath et al., 2022), MIND (Su et al., 2024), SAPLMA (Azaria & Mitchell, 2023), LLM-Check (Sriramanan et al., 2024), and ICR Probe (Zhang et al., 2025d)).

Training configuration. The QAoD probe is a two-hidden-layer MLP with ReLU and dropout. Class-weighted cross-entropy is used to address label imbalance. QAoD probe dimensions, optimization hyperparameters, and SINdex sampling settings are provided in Appendix A.

4.2 In-domain detection performance

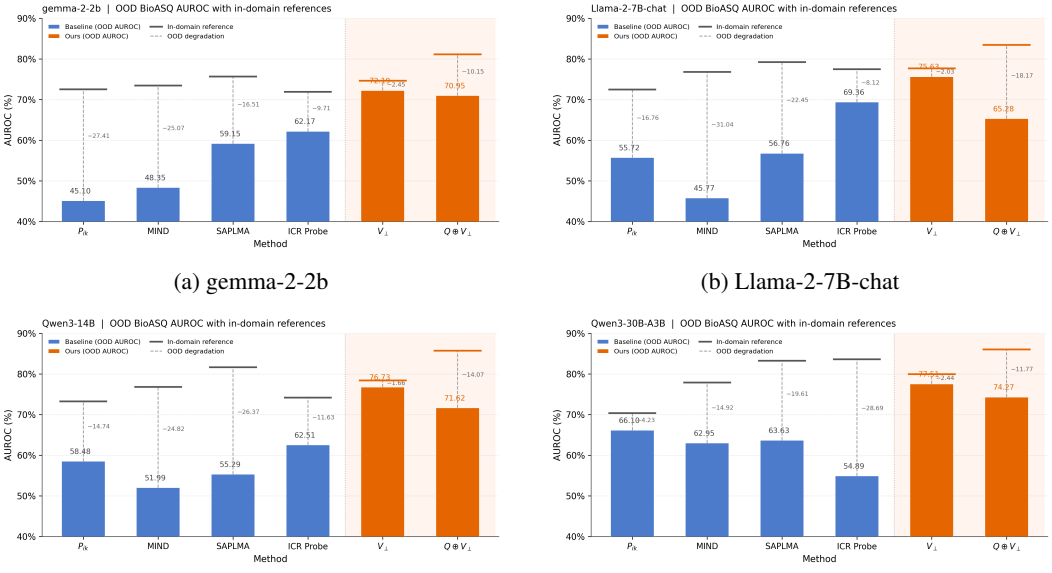
Table 1 reports AUROC and F1 for both QAoD configurations against all baselines.

$\mathcal{H}_{Q\oplus V_\perp}$ **achieves the best in-domain performance**, ranking first across all 16 model-dataset combinations by jointly encoding question context alongside the orthogonal deviation signal. Its advantage over the strongest single-pass baseline (SAPLMA) is consistent across all four model families, indicating that the benefit of incorporating question context is robust to model scale and architecture. $\mathcal{H}_{Q\oplus V_\perp}$ matches or surpasses SINdex across all model-dataset pairs with one forward pass instead of ten, suggesting that internal geometric structure encodes discriminative information at least as rich as that extracted through repeated output sampling.

$\mathcal{H}_{V_\perp}$ **remains competitive in-domain**, and our transfer results suggest that it is particularly effective under substantial domain shift.

4.3 Cross-domain generalization

To evaluate cross-domain transferability, we train detectors on a general-domain mixture of TriviaQA, SQuAD, and NQ, and evaluate them directly on BioASQ without any target-domain fine-tuning. The BioASQ training split is used only for the in-domain results, while the zero-shot OOD setting is reserved for cross-domain evaluation.



(a) gemma-2-2b

(b) Llama-2-7B-chat

(c) Qwen3-14B

(d) Qwen3-30B-A3B

Figure 2: Zero-shot OOD AUROC on BioASQ. Bars show OOD AUROC for detectors trained on {TriviaQA, SQuAD, NQ} and evaluated on BioASQ; ticks indicate in-domain BioASQ AUROC; dashed lines mark degradation gap. QAoD$_{V_\perp}$ achieves best OOD AUROC and smallest gap.

Cross-domain generalization to BioASQ is the primary distinguishing capability of QAoD relative to existing white-box detectors. As shown in Figure 2, $\mathcal{H}_{V_\perp}$ achieves the highest zero-shot BioASQ

Table 1: Hallucination detection performance on four datasets (seed=42), reporting AUROC with F1 in parentheses (%). **Bold**: best; underline: second best.

Model	Method	TriviaQA	SQuAD	NQ	BioASQ
		AUC (F1)	AUC (F1)	AUC (F1)	AUC (F1)
gemma-2-2b	PPL	73.90 (38.28)	76.18 (40.34)	67.48 (84.22)	59.32 (69.82)
	Logit Entropy	75.59 (38.94)	77.35 (40.54)	69.14 (84.44)	64.80 (69.89)
	SINdex	80.19 (42.38)	78.26 (42.00)	75.20 (85.05)	72.91 (71.46)
	P_{false}	69.51 (35.12)	65.29 (33.65)	69.89 (83.73)	65.45 (72.34)
	LLM-Check	59.97 (27.62)	59.18 (28.13)	49.90 (83.63)	63.99 (69.14)
	P_{ik}	77.12 (41.62)	69.31 (35.16)	71.58 (83.72)	72.51 (70.35)
	MIND	76.80 (49.94)	66.76 (34.20)	72.41 (84.00)	73.42 (70.94)
	SAPLMA	83.32 (51.42)	77.23 (42.29)	78.40 (84.96)	75.66 (71.06)
	ICR Probe	64.73(34.93)	66.21 (32.42)	68.06 (82.08)	71.88 (70.44)
	QAoD$_{V_\perp}$	82.26 (52.18)	79.73 (47.55)	77.91 (85.19)	74.64 (70.93)
	QAoD$_{Q\oplus V_\perp}$	**90.10** (63.75)	**82.16** (48.84)	**81.42** (85.68)	**81.10** (74.70)
Llama-2-7B-chat	PPL	52.16 (23.59)	56.19 (22.59)	56.81 (78.94)	53.39 (64.23)
	Logit Entropy	52.23 (23.59)	56.48 (22.59)	56.85 (78.94)	57.10 (64.23)
	SINdex	68.93 (40.65)	75.72 (39.56)	73.78 (79.72)	79.19 (68.44)
	P_{false}	65.12 (31.29)	69.42 (32.00)	68.08 (79.57)	73.76 (68.71)
	LLM-Check	65.06 (30.23)	57.35 (25.50)	56.06 (79.14)	71.40 (71.13)
	P_{ik}	75.40 (43.63)	65.93 (37.26)	62.37 (79.30)	72.48 (67.62)
	MIND	79.31 (47.32)	68.06 (36.10)	67.10 (79.72)	76.81 (72.66)
	SAPLMA	84.33 (52.74)	79.47 (43.25)	74.00 (80.90)	79.21 (69.35)
	ICR Probe	73.67(31.03)	64.88(36.46)	67.29(76.15)	77.48(68.92)
	QAoD$_{V_\perp}$	86.20 (54.26)	80.75 (44.08)	75.08 (81.20)	77.46 (74.32)
	QAoD$_{Q\oplus V_\perp}$	**90.21** (61.28)	**83.09** (47.47)	**77.94** (81.68)	**83.45** (76.08)
Qwen3-14B	PPL	53.25 (13.13)	54.89 (14.75)	61.81 (49.47)	61.33 (46.28)
	Logit Entropy	53.25 (12.94)	54.92 (14.65)	62.06 (49.47)	61.88 (46.52)
	SINdex	65.73 (32.08)	74.02 (20.46)	78.30 (74.65)	80.50 (67.49)
	P_{false}	78.04 (32.71)	58.33 (12.33)	79.05 (77.71)	79.17 (66.97)
	LLM-Check	61.32 (21.81)	62.64 (13.06)	53.16 (77.03)	59.67 (57.42)
	P_{ik}	79.14 (39.61)	69.76 (19.45)	67.73 (70.52)	73.22 (64.13)
	MIND	81.35 (45.99)	72.12 (24.03)	73.69 (78.34)	76.81 (65.41)
	SAPLMA	85.45 (54.26)	83.57 (29.03)	78.93 (79.03)	81.66 (68.59)
	ICR Probe	73.00 (28.53)	73.26 (23.41)	69.01 (74.74)	74.14 (58.60)
	QAoD$_{V_\perp}$	87.42 (58.67)	81.34 (29.39)	79.98 (80.25)	78.39 (67.14)
	QAoD$_{Q\oplus V_\perp}$	**90.13** (60.81)	**85.17** (32.09)	**82.69** (81.25)	**85.69** (73.51)
Qwen3-30B-A3B	PPL	54.51 (16.85)	55.78 (17.74)	60.33 (77.54)	60.07 (52.34)
	Logit Entropy	54.52 (16.85)	55.82 (17.74)	60.47 (77.54)	59.73 (52.34)
	SINdex	68.26 (36.68)	74.07 (35.46)	77.33 (79.43)	81.37 (67.67)
	P_{false}	80.90 (44.52)	70.59 (22.79)	80.30 (81.84)	79.33 (66.81)
	LLM-Check	66.89 (25.95)	58.69 (18.21)	55.53 (77.52)	62.25 (55.22)
	P_{ik}	74.04 (42.11)	71.60 (27.71)	69.99 (77.61)	70.32 (59.20)
	MIND	81.38 (49.57)	72.90 (30.48)	72.69 (78.30)	77.87 (65.78)
	SAPLMA	86.74 (52.23)	83.31 (39.76)	78.78 (80.27)	83.24 (68.22)
	ICR Probe	76.41 (26.72)	76.46 (32.07)	73.22(74.56)	83.58 (65.71)
	QAoD$_{V_\perp}$	87.26 (53.90)	82.52 (36.45)	79.27 (80.48)	79.95 (67.85)
	QAoD$_{Q\oplus V_\perp}$	**91.84** (63.89)	**85.10** (40.81)	**82.76** (81.76)	**86.04** (73.38)

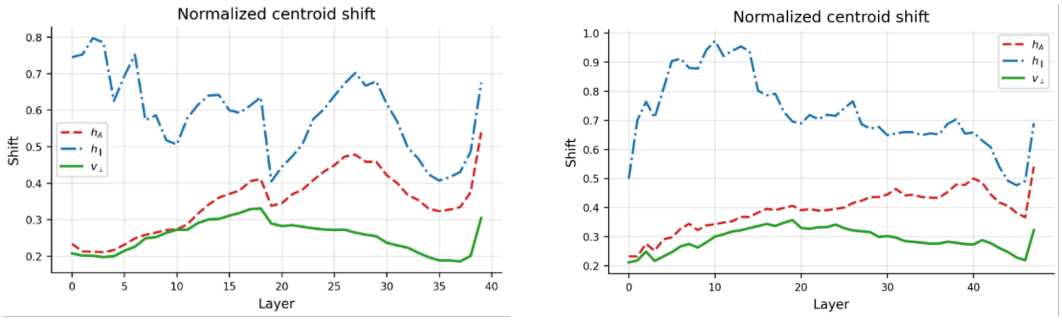
AUROC on all four models, exceeding the strongest evaluated white-box baseline for each model by 6.27–14.22 AUROC percentage points.

To examine whether this result is specific to a held-out target, we conduct leave-one-domain-out evaluation over four benchmarks with Llama-2-7B-chat. For each target, the detector is trained on the remaining three datasets without target-domain fine-tuning. We further test transfer to GSM8K(Cobbe et al., 2021), an unseen mathematical-reasoning target, using two configurations: *GeneralQA* $\rightarrow$ *Math*, trained on TriviaQA, SQuAD, and NQ, and *Biomedical* $\rightarrow$ *Math*, trained only on BioASQ.

Table 2: Cross-domain evaluation on Llama-2-7B-chat (AUROC, %). The first four rows are leave-one-domain-out QA evaluations; the last two transfer to GSM8K.

Training source → target	P_{ik}	MIND	SAPLMA	ICR PROBE	QAOD$_{V_\perp}$	QAOD$_{Q\oplus V_\perp}$
SQuAD + NQ + BioASQ → TriviaQA	60.45	61.50	68.42	60.81	71.28	**71.82**
TriviaQA + NQ + BioASQ → SQuAD	60.22	57.87	64.87	70.38	69.97	**70.86**
TriviaQA + SQuAD + BioASQ → NQ	54.55	56.97	61.70	60.22	63.49	**68.41**
TriviaQA + SQuAD + NQ → BioASQ	55.72	45.77	56.76	69.36	**75.63**	65.28
TriviaQA + SQuAD + NQ → GSM8K	60.69	62.56	60.29	65.32	**71.77**	66.33
BioASQ → GSM8K	60.60	62.84	57.42	41.83	**70.12**	66.00

Table 2 shows complementary transfer behavior for the two QAOD probes. The joint probe performs best when TriviaQA, SQuAD, or NQ is held out, whereas $\mathcal{H}_{V_\perp}$ performs best on BioASQ and in both GSM8K transfer settings. TriviaQA, SQuAD, and NQ are broad-coverage factual-QA benchmarks, in contrast to BioASQ's specialized biomedical setting. When a general factual-QA benchmark is held out, the remaining sources still provide question formats and knowledge-seeking contexts broadly similar to those encountered at test time, which may benefit $\mathcal{H}_{Q\oplus V_\perp}$. By contrast, BioASQ and GSM8K involve specialized biomedical knowledge and mathematical reasoning respectively, and thus present larger domain or task shifts. The results are consistent with $\mathcal{H}_{Q\oplus V_\perp}$ being more effective when source and target question contexts are relatively similar, while $\mathcal{H}_{V_\perp}$ transfers better under larger shifts.



(a) Qwen3-14B (b) Qwen3-30B-A3B

Figure 3: Layer-wise normalized centroid shift ($\|\mu_{\text{src}} - \mu_{\text{tgt}}\|/\bar{\sigma}$) from {TriviaQA, SQuAD, NQ} to BioASQ. The question-orthogonal component $v_\perp$ exhibits the smallest displacement across all layers, while the question-aligned component $h_\parallel$ exhibits the largest displacement.

To investigate a representation-level interpretation of this result, we analyze domain-induced feature changes in the general-QA-to-BioASQ setting, where the source and target domains are explicitly defined and the effect is consistent across all four models. We use Qwen3-14B and Qwen3-30B-A3B as representative dense and MoE backbones, respectively, to examine whether the observed representation patterns hold across two qualitatively different architectures.

Geometric view. Figure 3 plots the normalized centroid shift $\|\mu_{\text{src}} - \mu_{\text{tgt}}\|/\bar{\sigma}$ from the source-domain mixture to BioASQ. For both representative models, the same layer-wise ordering is observed: the question-aligned component $h_\parallel = \frac{h_A^{(l)} \cdot h_Q^{(l)}}{\|h_Q^{(l)}\|^2} h_Q^{(l)}$ undergoes the largest displacement, the original answer state h_A lies in the middle, and the question-orthogonal component $v_\perp$ is the least displaced. This ordering is consistent with the interpretation that the residual is less sensitive to source-target domain variation.

CKA evidence. We use linear CKA (Kornblith et al., 2019) to characterize domain-related variation from a representation-alignment perspective. Specifically, $\text{CKA}_{\text{domain}}$ measures linear alignment with domain identity ({TriviaQA, SQuAD, NQ} vs. BioASQ). Table 3 shows a pattern on both Qwen3-14B and Qwen3-30B-A3B: $v_\perp$ has the lowest domain alignment across three feature-selection

regimes. Together with the centroid-shift analysis, this pattern is consistent with reduced sensitivity to source-target domain variation.

$\mathrm{CKA}_{\text{hall}}$ for $v_\perp$ varies with the backbone and feature-selection regime, and is lower than that of h_A in the all-layer setting. After Fisher layer-and-neuron selection, it exceeds that of h_A for both backbones. The selectivity ratio $\mathrm{CKA}_{\text{hall}}/\mathrm{CKA}_{\text{domain}}$ is highest for $v_\perp$ under every selection regime (Table 3). Together with the detection results, these observations are consistent with the residual retaining useful information for correctness prediction while exhibiting lower domain-label alignment.

Table 3: Linear CKA alignment with hallucination ($\mathrm{CKA}_{\text{hall}}$) and domain-membership ($\mathrm{CKA}_{\text{domain}}$) labels, and selectivity ratio, at three feature-selection operating points; $v_\perp$ achieves the highest ratio under every regime, and Fisher selection widens the advantage.

		$\mathrm{CKA}_{\text{hall}}$			$\mathrm{CKA}_{\text{domain}}$			$\mathrm{CKA}_{\text{hall}}/\mathrm{CKA}_{\text{domain}}$		
Model	**Selection**	h_A	$h_\parallel$	$v_\perp$	h_A	$h_\parallel$	$v_\perp$	h_A	$h_\parallel$	$v_\perp$
Qwen3-14B	All layers	0.099	0.090	**0.068**	0.207	0.155	**0.114**	0.476	0.584	**0.596**
	Fisher layer	0.073	0.068	**0.072**	0.218	0.171	**0.115**	0.334	0.398	**0.621**
	Fisher layer+neuron	0.069	0.065	**0.072**	0.202	0.151	**0.109**	0.344	0.434	**0.663**
Qwen3-30B-A3B	All layers	0.083	0.077	**0.080**	0.269	0.226	**0.149**	0.310	0.343	**0.537**
	Fisher layer	0.059	0.046	**0.076**	0.240	0.196	**0.137**	0.244	0.237	**0.556**
	Fisher layer+neuron	0.056	0.044	**0.077**	0.236	0.184	**0.136**	0.239	0.238	**0.564**

4.4 Ablation studies

We report two ablations on Qwen3-14B and Qwen3-30B-A3B (dense and MoE representatives): a zero-shot OOD feature-construction study (trained on {TriviaQA, SQuAD, NQ}, tested on BioASQ) and a layer-selection comparison on Qwen3-14B / TriviaQA. All other settings follow the main experiments. The neuron-selection threshold ablation is provided in Appendix A.5.

Table 4: Feature-construction ablation: zero-shot OOD AUROC (%) on BioASQ.

Feature	**14B**	**30B**
Random	64.24 ± 1.57	72.58 ± 1.66
Q-only	71.12	73.03
A-only	66.09	72.35
Q+A (no proj.)	69.88	76.11
QAoD ($\mathcal{H}_{V_\perp}$)	**76.73**	**77.51**

Table 5: Layer-selection ablation on Qwen3-14B / TriviaQA (AUROC, %); DPF leads at every budget and outperforms full-layer aggregation.

Method	$K=5$	$K=10$	$K=15$	**Layers** (K=5)
Last-N layers	84.34	85.06	86.64	[35,36,37,38,39]
Pure Fisher	84.92	86.02	86.42	[19,20,21,22,23]
Uniform spacing	86.83	87.03	87.02	[0,10,20,29,39]
Random	86.49±0.43	86.74±0.19	86.86±0.37	—
DPF (ours)	**87.27**	**87.36**	**87.42**	[10,19,26,33,39]

Table 4 holds Fisher selection fixed, varying only the feature combination. The alternative feature constructions are as follows: *Random* subtracts the projection of the answer representation onto a random unit direction u, yielding $h_A - (h_A^\top u)u$ with $\|u\|_2 = 1$, instead of subtracting its projection onto the question direction; *Q-only* and *A-only* apply Fisher selection to h_Q or h_A alone; and *Q+A (no proj.)* follows the same Fisher selection and concatenation pipeline as QAoD but concatenates h_Q and h_A without computing $v_\perp$. *Random* performs worst, confirming OOD gains require geometrically meaningful decoupling; *Q-only*, *A-only*, and *Q+A (no proj.)* all underperform QAoD. These results support the usefulness of the question-orthogonal representation within the evaluated pipeline.

In Table 5, combining discriminability with diversity is the key factor: *Last-N layers* and pure top-K *Fisher* both select largely redundant consecutive layers, limiting coverage of complementary processing stages. *Uniform spacing* and *Random* selection improve over these by increasing layer diversity, but without discriminability guidance. DPF combines both objectives, consistently achieving the best AUROC at every budget and selecting well-spread layers. Using all 40 layers (AUROC 86.95) also falls below DPF at K=15 (87.42), confirming that diversity-aware compact selection outperforms full-depth aggregation.

4.5 EFFICIENCY

Because all methods share the same answer-generation forward pass, we compare only the additional detection overhead. Figure 4 shows the runtime breakdown on TriviaQA with Qwen3-14B (NVIDIA A100, batch size 256), highlighting the substantial computational gap between repeated-sampling methods and single-pass white-box probes in practice. Per-query time is the total elapsed time for a batch of 256 queries divided by 256, i.e., an amortized per-query time under batched execution. Detection overhead includes all additional online costs introduced by each method beyond the shared answer generation: hidden-state capture, memory transfers, feature extraction and processing, probe inference, and any additional model calls required by the detector. Only the shared model computation for generating the original answer is excluded from detection overhead.

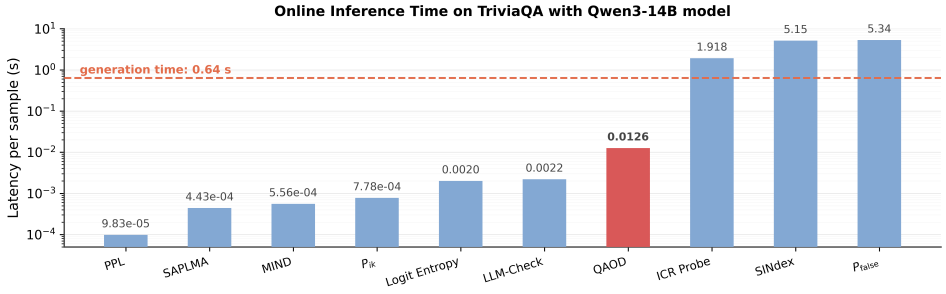


Figure 4: Inference-time comparison (log scale), including shared generation cost and additional detection overhead; QAOD's detection cost is negligible ($\approx 2\%$) relative to generation time.

Detection overhead is orders of magnitude lower than sampling-based methods such as SINdex ($N{=}10$). Among single-pass white-box probes, QAOD's per-query detection overhead of 0.0126s is marginally higher than lighter baselines, but remains negligible relative to the shared generation cost of ~ 0.64s ($\approx 2\%$). Given the substantial OOD robustness advantage demonstrated in Section 4.3, this marginal overhead is well-justified within the single-pass operating regime.

5 LIMITATIONS

Our evaluation uses reference-based correctness as a proxy for hallucination in QA; finer distinctions among response types remain for future work. The evaluation covers knowledge-intensive QA tasks, where factual errors can be traced to specific misapplied entities or relationships. Extending QAOD to tasks such as open-ended generation, where the geometric properties of hidden states may differ substantially, is a promising direction for future work. As a white-box method, QAOD requires intermediate hidden states and therefore cannot serve closed-source models exposed only through inference APIs. Extending it to such settings, for instance by transferring probe knowledge from an accessible surrogate model to a black-box target, is a direction we plan to explore in future work.

6 CONCLUSION

In LLM hallucination detection, black-box methods demand repeated inference while white-box probes degrade sharply under domain shift. We proposed QAOD, which decomposes the answer representation into a component parallel to the question representation and an orthogonal residual, and applies Fisher scoring to select compact probe features. Our experiments show that this decomposition improves cross-domain correctness prediction. The centroid-shift and CKA analyses are consistent with the interpretation that the residual reduces sensitivity to domain variation while retaining useful information for correctness prediction. Evaluated across four LLMs and four benchmarks, $\mathcal{H}_{Q\oplus V_\perp}$ achieves the best in-domain AUROC on all model-dataset pairs, while $\mathcal{H}_{V_\perp}$ surpasses the strongest evaluated white-box baseline for each model by 6.27–14.22 AUROC percentage points on zero-shot BioASQ. These results suggest that explicitly modeling the geometric relationship between question and answer representations is a promising direction toward hallucination detectors that are simultaneously efficient and robust to domain shift, particularly in real-world settings where target-domain labels are scarce.

REFERENCES

Samir Abdaljalil, Hasan Kurban, Parichit Sharma, Erchin Serpedin, and Rachad Atat. Sindex: Semantic inconsistency index for hallucination detection in llms, 2025. URL https://arxiv.org/abs/2503.05980.

Razvan Azamfirei, Sapna Kudchadkar, and James Fackler. Large language models and the perils of their hallucinations. *Critical Care*, 27, 03 2023. doi: 10.1186/s13054-023-04393-x.

Amos Azaria and Tom Mitchell. The internal state of an llm knows when it's lying, 2023. URL https://arxiv.org/abs/2304.13734.

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, March 2022. doi: 10.1162/coli$_{a0}$0422.URLhttps://aclanthology.org/2022.cl-1.7/.

Chao Chen, Kai Liu, Ze Chen, Yi Gu, Yue Wu, Mingyuan Tao, Zhihang Fu, and Jieping Ye. Inside: Llms' internal states retain the power of hallucination detection, 2024a. URL https://arxiv.org/abs/2402.03744.

Yuyan Chen, Qiang Fu, Yichen Yuan, Zhihao Wen, Ge Fan, Dayiheng Liu, Dongmei Zhang, Zhixu Li, and Yanghua Xiao. Hallucination detection: Robustly discerning reliable answers in large language models, 2024b. URL https://arxiv.org/abs/2407.04121.

I-Chun Chern, Steffi Chern, Shiqi Chen, Weizhe Yuan, Kehua Feng, Chunting Zhou, Junxian He, Graham Neubig, and Pengfei Liu. Factool: Factuality detection in generative ai – a tool augmented framework for multi-task and multi-domain scenarios, 2023. URL https://arxiv.org/abs/2307.13528.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James Glass, and Pengcheng He. Dola: Decoding by contrasting layers improves factuality in large language models, 2024. URL https://arxiv.org/abs/2309.03883.

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

Yongxin Deng, Zhen Fang, Sharon Li, and Ling Chen. Beyond in-domain detection: Spikescore for cross-domain hallucination detection, 2026. URL https://arxiv.org/abs/2601.19245.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630, 06 2024. doi: 10.1038/s41586-024-07421-0. URL https://doi.org/10.1038/s41586-024-07421-0.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55, January 2025. ISSN 1558-2868. doi: 10.1145/3703155. URL http://dx.doi.org/10.1145/3703155.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38, March 2023. ISSN 1557-7341. doi: 10.1145/3571730. URL http://dx.doi.org/10.1145/3571730.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In Regina Barzilay and Min-Yen Kan (eds.), *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1601–1611, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1147. URL https://aclanthology.org/P17-1147/.

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, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. Language models (mostly) know what they know, 2022. URL https://arxiv.org/abs/2207.05221.

Hazel Kim, Tom A. Lamb, Adel Bibi, Philip Torr, and Yarin Gal. Detecting llm hallucination through layer-wise information deficiency: Analysis of ambiguous prompts and unanswerable questions, 2025. URL https://arxiv.org/abs/2412.10246.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited, 2019. URL https://arxiv.org/abs/1905.00414.

Anastasia Krithara, Anastasios Nentidis, Konstantinos Bougiatiotis, and Georgios Paliouras. Bioasq-qa: A manually curated corpus for biomedical question answering. *bioRxiv*, 2022. doi: 10.1101/2022.12.14.520213. URL https://www.biorxiv.org/content/early/2022/12/16/2022.12.14.520213.

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. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466, 08 2019. ISSN 2307-387X. doi: 10.1162/tacl$_{a0}$0276.URL

Benjamin A. Levinstein and Daniel A. Herrmann. Still no lie detector for language models: probing empirical and conceptual roadblocks. *Philosophical Studies*, 182(7): 1539–1565, February 2024. ISSN 1573-0883. doi: 10.1007/s11098-023-02094-3. URL http://dx.doi.org/10.1007/s11098-023-02094-3.

Belinda Z. Li, Maxwell Nye, and Jacob Andreas. Implicit representations of meaning in neural language models, 2021. URL https://arxiv.org/abs/2106.00737.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model, 2024. URL https://arxiv.org/abs/2306.03341.

Stephanie Lin, Jacob Hilton, and Owain Evans. Teaching models to express their uncertainty in words, 2022. URL https://arxiv.org/abs/2205.14334.

Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. Generating with confidence: Uncertainty quantification for black-box large language models, 2024. URL https://arxiv.org/abs/2305.19187.

Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models, 2023. URL https://arxiv.org/abs/2303.08896.

Samuel Marks and Max Tegmark. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets, 2024. URL https://arxiv.org/abs/2310.06824.

Pranav Rajpurkar, Robin Jia, and Percy Liang. Know what you don't know: Unanswerable questions for squad, 2018. URL https://arxiv.org/abs/1806.03822.

Gaurang Sriramanan, Siddhant Bharti, Vinu Sankar Sadasivan, Shoumik Saha, Priyatham Kattakinda, and Soheil Feizi. Llm-check: Investigating detection of hallucinations in large language models. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 34188–34216. Curran Associates, Inc., 2024. doi: 10.52202/079017-1077. URL https://proceedings.neurips.cc/paper$_f$iles/paper/2024/$file$/3c1e1fdf305195cd620c118aaa9717$ad-Paper-Conference.pdf$.

Weihang Su, Changyue Wang, Qingyao Ai, Yiran Hu, Zhijing Wu, Yujia Zhou, and Yiqun Liu. Unsupervised real-time hallucination detection based on the internal states of large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 14379–14391, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.854. URL https://aclanthology.org/2024.findings-acl.854/.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, Anton Tsitsulin, Nino Vieillard, Piotr Stanczyk, Sertan Girgin, Nikola Momchev, Matt Hoffman, Shantanu Thakoor, Jean-Bastien Grill, Behnam Neyshabur, Olivier Bachem, Alanna Walton, Aliaksei Severyn, Alicia Parrish, Aliya Ahmad, Allen Hutchison, Alvin Abdagic, Amanda Carl, Amy Shen, Andy Brock, Andy Coenen, Anthony Laforge, Antonia Paterson, Ben Bastian, Bilal Piot, Bo Wu, Brandon Royal, Charlie Chen, Chintu Kumar, Chris Perry, Chris Welty, Christopher A. Choquette-Choo, Danila Sinopalnikov, David Weinberger, Dimple Vijaykumar, Dominika Rogozińska, Dustin Herbison, Elisa Bandy, Emma Wang, Eric Noland, Erica Moreira, Evan Senter, Evgenii Eltyshev, Francesco Visin, Gabriel Rasskin, Gary Wei, Glenn Cameron, Gus Martins, Hadi Hashemi, Hanna Klimczak-Plucińska, Harleen Batra, Harsh Dhand, Ivan Nardini, Jacinda Mein, Jack Zhou, James Svensson, Jeff Stanway, Jetha Chan, Jin Peng Zhou, Joana Carrasqueira, Joana Iljazi, Jocelyn Becker, Joe Fernandez, Joost van Amersfoort, Josh Gordon, Josh Lipschultz, Josh Newlan, Ju yeong Ji, Kareem Mohamed, Kartikeya Badola, Kat Black, Katie Millican, Keelin McDonell, Kelvin Nguyen, Kiranbir Sodhia, Kish Greene, Lars Lowe Sjoesund, Lauren Usui, Laurent Sifre, Lena Heuermann, Leticia Lago, Lilly McNealus, Livio Baldini Soares, Logan Kilpatrick, Lucas Dixon, Luciano Martins, Machel Reid, Manvinder Singh, Mark Iverson, Martin Görner, Mat Velloso, Mateo Wirth, Matt Davidow, Matt Miller, Matthew Rahtz, Matthew Watson, Meg Risdal, Mehran Kazemi, Michael Moynihan, Ming Zhang, Minsuk Kahng, Minwoo Park, Mofi Rahman, Mohit Khatwani, Natalie Dao, Nenshad Bardoliwalla, Nesh Devanathan, Neta Dumai, Nilay Chauhan, Oscar Wahltinez, Pankil Botarda, Parker Barnes, Paul Barham, Paul Michel, Pengchong Jin, Petko Georgiev, Phil Culliton, Pradeep Kuppala, Ramona Comanescu, Ramona Merhej, Reena Jana, Reza Ardeshir Rokni, Rishabh Agarwal, Ryan Mullins, Samaneh Saadat, Sara Mc Carthy, Sarah Cogan, Sarah Perrin, Sébastien M. R. Arnold, Sebastian Krause, Shengyang Dai, Shruti Garg, Shruti Sheth, Sue Ronstrom, Susan Chan, Timothy Jordan, Ting Yu, Tom Eccles, Tom Hennigan, Tomas Kocisky, Tulsee Doshi, Vihan Jain, Vikas Yadav, Vilobh Meshram, Vishal Dharmadhikari, Warren Barkley, Wei Wei, Wenming Ye, Woohyun Han, Woosuk Kwon, Xiang Xu, Zhe Shen, Zhitao Gong, Zichuan Wei, Victor Cotruta, Phoebe Kirk, Anand Rao, Minh Giang, Ludovic Peran, Tris Warkentin, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, D. Sculley, Jeanine Banks, Anca Dragan, Slav Petrov, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Sebastian Borgeaud, Noah Fiedel, Armand Joulin, Kathleen Kenealy, Robert Dadashi, and Alek Andreev. Gemma 2: Improving open language models at a practical size, 2024. URL https://arxiv.org/abs/2408.00118.

S. M Towhidul Islam Tonmoy, S M Mehedi Zaman, Vinija Jain, Anku Rani, Vipula Rawte, Aman Chadha, and Amitava Das. A comprehensive survey of hallucination mitigation techniques in large language models, 2024. URL https://arxiv.org/abs/2401.01313.

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, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models, 2023. URL https://arxiv.org/abs/2307.09288.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023. URL https://arxiv.org/abs/1706.03762.

Weiqin Wang, Yile Wang, and Hui Huang. Ranked voting based self-consistency of large language models. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 14410–14426, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.744. URL https://aclanthology.org/2025.findings-acl.744/.

Yuxia Wang, Minghan Wang, Muhammad Arslan Manzoor, Fei Liu, Georgi Georgiev, Rocktim Jyoti Das, and Preslav Nakov. Factuality of large language models: A survey, 2024. URL https://arxiv.org/abs/2402.02420.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Fujie Zhang, Peiqi Yu, Biao Yi, Baolei Zhang, Tong Li, and Zheli Liu. Prompt-guided internal states for hallucination detection of large language models. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 21806–21818, Vienna, Austria, July 2025a. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1058. URL https://aclanthology.org/2025.acl-long.1058/.

Luan Zhang, Dandan Song, Zhijing Wu, Yuhang Tian, Changzhi Zhou, Jing Xu, Ziyi Yang, and Shuhao Zhang. Detecting hallucination in large language models through deep internal representation analysis. In James Kwok (ed.), *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI-25*, pp. 8357–8365. International Joint Conferences on Artificial Intelligence Organization, 8 2025b. doi: 10.24963/ijcai.2025/929. URL https://doi.org/10.24963/ijcai.2025/929. Main Track.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Chen Xu, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, and Shuming Shi. Siren's song in the ai ocean: A survey on hallucination in large language models, 2025c. URL https://arxiv.org/abs/2309.01219.

Zhenliang Zhang, Xinyu Hu, Huixuan Zhang, Junzhe Zhang, and Xiaojun Wan. ICR probe: Tracking hidden state dynamics for reliable hallucination detection in LLMs. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 17986–18002, Vienna, Austria, July 2025d. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.880. URL https://aclanthology.org/2025.acl-long.880/.

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. Judging llm-as-a-judge with mt-bench and chatbot arena, 2023. URL https://arxiv.org/abs/2306.05685.

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, J. Zico Kolter, and Dan Hendrycks. Representation engineering: A top-down approach to ai transparency, 2025. URL https://arxiv.org/abs/2310.01405.

A IMPLEMENTATION DETAILS

A.1 DATASET SAMPLING

For the four main benchmarks, we use the full training split of BioASQ and randomly sample 10,000 examples from the training split of each of TriviaQA, SQuAD, and NQ. We use the full validation split of each dataset without subsampling. These sampling choices specify the data used in our experiments; the source–target configurations for cross-domain evaluation follow those described in the main text.

A.2 ANSWER-GENERATION PROMPT

The general QA generation template is shown below. The context and question fields are filled with the corresponding example content, and generation begins after `Answer:`.

```
You are a helpful QA assistant.
Read the context and answer the question concisely.
- Output ONLY the short answer text (no explanations, no
restatements, no multiple sentences).

Context: [context]
Question: [question]
Answer:
```

A.3 LABEL CONSTRUCTION AND JUDGE PROTOCOL

We employed different annotation strategies depending on dataset characteristics.

SQuAD. Following the official dataset evaluation protocol, we operationalize correctness using a token-overlap F1 threshold of 50.0; answers above the threshold are treated as correct and those below it are treated as incorrect for label construction. This is a dataset-specific proxy rather than a universal hallucination definition.

TriviaQA, NQ, BioASQ, and GSM8K. We formulate an *LLM-as-a-judge* labeling protocol (Zheng et al., 2023) using `DeepSeek-V3` as the adjudicator for semantic equivalence. When multiple reference answers are available, we ask whether the predicted answer means the same as any of the expected answers; when only one reference answer is available, we ask the same question with the singular form. In both cases, the judge is required to respond only with "yes" or "no," following the protocol used in semantic-entropy-style semantic equivalence evaluation.

Judge prompt template. The judge receives the following text as a single user message, without an additional system prompt. We use the singular form when there is one reference answer and the plural form otherwise; the bracketed fields are replaced with the corresponding example content:

```
We are assessing the quality of answers to the following
question: [QUESTION]
The expected answer is: [EXPECTED_ANSWER].
The proposed answer is: [PROPOSED_ANSWER]
Within the context of the question, does the proposed answer
mean the same as the expected answer? Respond only with
'yes' or 'no'.
Response:
```

For multiple references, we use:

```
We are assessing the quality of answers to the following
question: [QUESTION]
The following are expected answers to this question:
[EXPECTED_ANSWERS].
The proposed answer is: [PROPOSED_ANSWER]
Within the context of the question, does the proposed answer
```

```
mean the same as any of the expected answers?  Respond only
with 'yes' or 'no'.
Response:
```

A "yes" is labeled correct and a "no" incorrect. We first query at temperature 0.01; if the returned text does not begin with "yes" or "no," we retry the same prompt at temperature 1.0.

Separation of answer generation and label adjudication. During answer generation, each evaluated LLM receives the question, the generation instruction, and any context supplied by the QA input; the gold reference answer(s) are withheld. We extract QAoD features from the same question-and-generated-answer sequence, without providing the reference answer(s), the judge prompt, or the judge output as model input or features. Only after generation and feature extraction are complete, a separate `DeepSeek-V3` call receives the question, generated answer, and gold reference answer(s) to assign the reference-equivalence label. The judge output is used only as the supervision/evaluation label, not as an input feature. Thus, the gold answer is not directly exposed to the evaluated model during generation or feature extraction, while the resulting labels remain defined by agreement with the dataset references rather than by an unconstrained factuality assessment.

Blinded audit of reference-based labels To assess the reliability of the reference-based labels on the four main datasets, we randomly sampled 300 validation examples from each dataset and conducted a blinded manual audit. Human reviewers assessed the generated answers against the dataset references to adjudicate reference-based correctness. The original labels agreed with the human-adjudicated reference labels on 93.17% of examples (95% Wilson confidence interval: 91.60%–94.46%; Cohen's $\kappa = 0.843$). Dataset-wise agreement was 95.00% for TriviaQA, 91.67% for SQuAD, 93.00% for NQ, and 93.00% for BioASQ. These results indicate strong agreement in the reference-based label judgments. The audit evaluates agreement with reference-answer matching labels and does not establish coverage of every form of factual error or unsupported content. All methods use the same fixed labels.

A.4 Probe configuration and optimization

F1 threshold selection and reporting. For the reported F1 scores, we select the decision threshold that maximizes F1 on the validation split and report the resulting maximum F1 on that same split. These scores therefore represent validation-set F1 with an optimized threshold, rather than F1 on an independent test set. AUROC does not require threshold selection.

Feature-standardization statistics, Fisher scores, DPF layer selection, and Fisher neuron selection are fitted or computed strictly on the training split. The resulting standardization parameters and selected feature indices are then held fixed when processing validation and evaluation data. In cross-domain experiments, these quantities are obtained only from the source-domain training data.

For any given LLM-dataset pair, we compute multivariate Fisher scores for all layers in a single pass over training data. Unless otherwise stated, we use fixed settings across all experiments: $K = 15$ for the layer budget, $\lambda = 1$ for the diversity weight, $\alpha = 0.9$ for neuron selection, MLP hidden dimensions of 1024 and 128, Adam with learning rate $\eta = 0.001$, weight decay 0.01, and dropout probability 0.1. The rationale for choosing $\alpha = 0.9$ is examined in Appendix A.5; this value is not claimed to be universally optimal and can be adjusted jointly with other budget choices when different deployment constraints are preferred.

For SINdex, we follow the original setting with $N = 10$ samples and `temperature`=1.0. The QAoD probe is a two-hidden-layer MLP ($d_{\text{in}} \to 1024 \to 128 \to 2$, ReLU, dropout 0.1), trained for 30 epochs with Adam ($\eta = 0.001$, weight decay 0.01) using class-weighted cross-entropy. For the other baselines, including MIND and SAPLMA, we use the parameter settings specified in their respective original papers or official implementations, such as the MLP architecture and optimization settings where applicable.

A.5 Neuron-selection threshold ablation

We analyze the cumulative Fisher-importance threshold α with the layer budget fixed at $K = 15$ and diversity weight fixed at $\lambda = 1$. Figure 5 shows that performance stabilizes for $\alpha \in [0.8, 0.9]$. We

therefore use $\alpha = 0.9$ as a practical balance between AUROC and feature dimensionality; it is not intended as a universally optimal value.

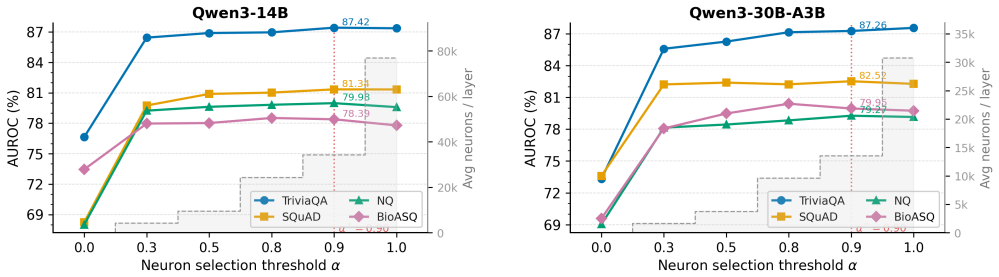


Figure 5: Neuron-selection ablation on the cumulative Fisher threshold α under $K = 15$ and $\lambda = 1$; performance stabilizes at $\alpha \in [0.8, 0.9]$, motivating the default $\alpha = 0.9$.

Compute infrastructure. All experiments were conducted on a single NVIDIA A100 80 GB GPU. Hidden-state extraction and probe training for a given LLM-dataset pair typically complete within a few hours on this hardware; the exact wall-clock time varies with model size (e.g., shorter for Gemma-2-2B and longer for Qwen3-30B-A3B). The efficiency measurements reported in Section 4.5 were collected under the same hardware configuration with batch size 256.

A.6 Multi-seed stability on Qwen models

Table 6 summarizes the multi-seed results for Qwen3-14B and Qwen3-30B-A3B on the main in-domain benchmarks. To keep the presentation consistent with the main text, we report AUROC in percentage form here. Across both models, the joint QAoD probe achieves the highest mean AUROC on every benchmark. Its standard deviation is below 0.4 AUROC points in seven of the eight model-dataset settings, but reaches 1.66 on Qwen3-30B-A3B / NQ, compared with 0.11 for the orthogonal-only probe, 1.84 for SAPLMA, and 2.11 for MIND. SAPLMA shows a notably larger spread on in-domain BioASQ for Qwen3-14B (4.75 AUROC points).

Table 7 reports the corresponding zero-shot OOD results on BioASQ. The orthogonal-only QAoD probe achieves the highest mean AUROC and the lowest standard deviation among the listed methods for both Qwen models (0.36 and 0.24 AUROC points, respectively). The joint probe is more variable, with standard deviations of 2.15 and 0.76. These results support the stability of the cross-domain advantage of the question-orthogonal component across seeds.

Table 6: Multi-seed in-domain AUROC on Qwen3-14B and Qwen3-30B-A3B (seeds 42 to 46, mean $\pm$ std over five runs); the joint QAoD probe achieves the highest mean AUROC in every setting, with variability depending on the model and dataset.

Model	Method	TriviaQA	SQuAD	NQ	BioASQ
Qwen3-14B	MIND	82.89 ± 0.36	72.45 ± 1.36	74.86 ± 1.55	78.38 ± 1.31
	SAPLMA	86.19 ± 0.39	83.68 ± 0.32	79.17 ± 0.61	78.12 ± 4.75
	QAoD ($V_\perp$)	87.25 ± 0.24	81.39 ± 0.37	79.62 ± 0.21	78.15 ± 0.14
	QAoD ($Q \oplus V_\perp$)	90.35 ± 0.15	85.51 ± 0.19	82.53 ± 0.18	85.51 ± 0.07
Qwen3-30B-A3B	MIND	80.38 ± 0.61	71.88 ± 1.00	75.39 ± 2.11	77.08 ± 0.72
	SAPLMA	86.87 ± 0.73	83.89 ± 0.28	81.15 ± 1.84	83.06 ± 0.87
	QAoD ($V_\perp$)	87.01 ± 0.15	82.54 ± 0.14	79.28 ± 0.11	79.97 ± 0.21
	QAoD ($Q \oplus V_\perp$)	91.86 ± 0.08	84.92 ± 0.15	83.90 ± 1.66	85.94 ± 0.06

Table 7: Multi-seed zero-shot OOD AUROC on BioASQ (seeds 42 to 46, mean ± std over five runs); the OOD advantage of $\mathcal{H}_{V_\perp}$ holds consistently across seeds.

Method	**Qwen3-14B**	**Qwen3-30B-A3B**
	BioASQ	BioASQ
MIND	47.61 ± 2.57	64.36 ± 2.61
SAPLMA	51.42 ± 4.55	72.37 ± 4.93
QAoD ($V_\perp$)	**77.68** ± 0.36	**77.21** ± 0.24
QAoD ($Q \oplus V_\perp$)	70.52 ± 2.15	73.93 ± 0.76

B ADDITIONAL COMPARISONS WITH RECENT DETECTION METHODS

Beyond the baselines in the main experiments, we evaluate PRISM's fixed verification prompt and provide a reproduction of MHAD on Llama-2-7B-chat.

Prompt construction and selection. PRISM (Zhang et al., 2025a) uses GPT-4o to generate ten candidate verification templates. The authors rank them by the average hidden-state structural salience score across the six True-False sub-datasets, using the final-token representation from the final layer of Llama-2-7B-chat. The chosen template maximizes this score within the candidate pool under that selection setting; its effectiveness need not carry over to another task or model, where prompt reselection may be necessary. Applying it to a generated QA answer requires an additional forward pass through the verification template.

Fixed-prompt QA evaluation. We evaluate the PRISM-SAPLMA combination on Llama-2-7B-chat using the verification template selected in the original paper. Because our tasks are question answering, we first generate an answer using the QA generation prompt. We then prepend the question to the verification template and insert the generated answer into `[statement]`, producing the following input for an additional forward pass:

> Question: [question]
> Does the statement '[generated answer]' accurately reflect the truth?

The question supplies context for interpreting short answers. We use the resulting prompt-guided hidden-state representation for SAPLMA and retain this construction across all evaluated datasets, without performing a new prompt search. This experiment evaluates the selected verification template with question context in our QA setting; its performance may depend on how the question and answer are incorporated into the prompt. For the BioASQ OOD evaluation, the detector is trained on TriviaQA, SQuAD, and NQ and evaluated on BioASQ, following the source–target configuration of the main experiments.

Table 8: Additional comparison on Llama-2-7B-chat (AUROC, %). PRISM-SAPLMA uses the verification template selected in the original paper, with the question prepended and the generated answer inserted as the statement. The two QAoD columns correspond to the orthogonal-only and joint probes. Bold indicates the best result in each row.

Evaluation	**PRISM-SAPLMA**	**SAPLMA**	$\text{QAoD}_{V_\perp}$	$\text{QAoD}_{Q\oplus V_\perp}$
TriviaQA	77.25	84.33	86.20	**90.21**
SQuAD	65.11	79.47	80.75	**83.09**
NQ	74.35	74.00	75.08	**77.94**
BioASQ	81.98	79.21	77.46	**83.45**
BioASQ OOD	40.81	56.76	**75.63**	65.28

Results and implications. Table 8 shows that the fixed verification prompt reduces SAPLMA's AUROC by 7.08 percentage points on TriviaQA and 14.36 points on SQuAD, while improving it by 0.35 points on NQ and 2.77 points on BioASQ. The joint QAoD probe outperforms PRISM-SAPLMA on all four in-domain benchmarks. PRISM-SAPLMA exceeds the orthogonal-only probe

on in-domain BioASQ, but its AUROC drops to 40.81% under BioASQ OOD, falling below SAPLMA by 15.95 points and below the orthogonal-only QAOD probe by 34.82 points. These results show mixed gains for the evaluated fixed-prompt construction in our QA setting, including degradation under the tested domain shift. They highlight the importance of verification-prompt construction and selection when adapting PRISM to QA; other prompt constructions may yield different results. They highlight a practical limitation of prompt-guided detection: a template selected for one setting does not guarantee reliable transfer, despite requiring an additional verification pass. The comparison concerns this fixed-template configuration; it does not establish the performance of PRISM after a new source-domain prompt-selection procedure.

MHAD reproduction. MHAD (Zhang et al., 2025b) selects neurons and layers with auxiliary probes and combines representations from the initial and final generation steps. Because the authors do not provide official open-source code, we did not include MHAD as a baseline in the main experiments. Instead, we independently reproduced the method based on its paper and ran the corresponding experiments on Llama-2-7B-chat. The in-domain results are reported in Table 9. For zero-shot OOD evaluation, we train on TriviaQA, SQuAD, and NQ, then evaluate on BioASQ, matching the source–target configuration used in the main experiments.

Table 9: Our MHAD reproduction on Llama-2-7B-chat. In-domain and zero-shot OOD results report AUROC. For OOD evaluation, the detector is trained on TriviaQA, SQuAD, and NQ and tested on BioASQ.

Evaluation	**AUROC (%)**
TriviaQA (in-domain)	82.20
SQuAD (in-domain)	71.28
NQ (in-domain)	66.49
BioASQ (in-domain)	82.47
TriviaQA + SQuAD + NQ $\rightarrow$ BioASQ (OOD)	48.46

SpikeScore and computational scope. SpikeScore (Deng et al., 2026) is a relevant cross-domain detector that measures uncertainty fluctuations over multi-turn continuations following an initial answer. These continuations require additional autoregressive generation and per-turn uncertainty computation, increasing detection-time inference costs relative to single-pass probes. Given our computational budget and emphasis on low-overhead detection, we did not include SpikeScore in the experimental comparisons.