Grad Detect: Gradient-Based Hallucination Detection in LLMs

Anand Kamat[†] [1] **Daniel Blake** [1] **Brent Werness** [1]

Abstract

Large Language Models (LLMs) have demonstrated remarkable capabilities across diverse tasks, yet they remain prone to generating hallucinations. Detecting these hallucinations is critical for deploying LLMs reliably in high-stakes applications. We present **Grad Detect**, a gradient-based approach for predicting hallucinations by analyzing layer-wise gradient patterns from a single forward-backward pass during inference. Our method shows that the internal gradient structure of a model carries rich information about the correctness of its output. This information is not accessible through output-level signals alone. We evaluate Grad Detect on several Q&A benchmarks across both hallucination detection and model abstention prediction, where it consistently outperforms confidence-based and sampling-based baselines. Through comprehensive layer ablation studies across all eleven models from four architectural families, we find that the final five layers concentrate over 97% of the discriminative gradient signal, enabling efficient deployment with minimal performance loss. Grad Detect provides a unified framework for predicting multiple dimensions of LLM reliability, offering strong predictive performance alongside interpretable insights into where and how model failures originate.

1. Introduction

Large Language Models (LLMs) built on the Transformer architecture (Vaswani et al., 2017) achieve strong performance on question answering, reasoning, and text generation (Brown et al., 2020; Chowdhery et al., 2023; Touvron et al., 2023). Despite these advances, LLMs remain prone to *hallucinations*, producing outputs that are factually incorrect yet stated with apparent confidence (Ji et al., 2023; Zhang et al., 2025b; Huang et al., 2024). Recent theoretical work suggests that standard training objectives reward plausible guessing over honest expressions of uncertainty (Kalai et al., 2025). Because hallucinated text is often indistinguishable from correct output by surface inspection alone, reliable detection is a prerequisite for deploying LLMs in high-stakes domains such as healthcare, legal analysis, and scientific research (Weidinger et al., 2021).

Existing detection approaches operate almost exclusively on output-level signals. Confidence and entropy methods (Kadavath et al., 2022; Kuhn et al., 2023; Lin et al., 2024) threshold statistics derived from the next-token distribution, but modern LLMs are frequently miscalibrated and can assign high probability to incorrect answers (Guo et al., 2017). Consistency-based methods (Manakul et al., 2023; Wang et al., 2023) sample multiple generations and measure agreement, gaining statistical power at much higher inference cost while still reasoning only about surface-level agreement. Prompting strategies such as chain-of-thought reasoning (Wei et al., 2022) and self-verification (Weng et al., 2023; Xue et al., 2023) make intermediate steps explicit, yet LLMs often struggle to reliably verify their own logic (Hong et al., 2024). A shared limitation of all these approaches is that they observe the *consequence* of a hallucination, an unreliable output distribution, rather than the internal dynamics that produced it.

A complementary research direction examines the model's internal representations. Prior work has shown that classifiers trained on hidden-state activations can distinguish true from false statements (Azaria & Mitchell, 2023; Su et al., 2024; Du et al., 2024; Kossen et al., 2024), establishing that internal states carry a detectable truthfulness signal. However, hidden-state activations capture a snapshot of the model's representation at a given layer. Gradients of the loss with respect to model parameters encode a different and complementary quantity: how sensitive every parameter is to the current prediction. This sensitivity provides a higher-dimensional fingerprint that reflects how the model's entire parameter space relates to its output, rather than what a single layer happens to represent.

Building on this observation, we propose **Grad Detect**, a gradient-based framework for hallucination detection. For each behavioral category (*Correct*, *Incorrect*, *Did Not An-*

[†]Lead author. [1]Amazon. Correspondence to: Anand Kamat <kamatana@amazon.com>.

Proceedings of the The 2^{nd} Workshop on Compositional Learning: Safety, Interpretability, and Agents at 43^{rd} International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026.

swer) we compute a *reference gradient* by averaging layer-wise gradients over labeled examples, obtaining a prototype in gradient space. A test sample is then characterized by the cosine similarity between its per-layer gradient and each reference, compressing billions of gradient dimensions into a compact $L \times |\mathcal{C}|$ feature matrix. A lightweight transformer encoder processes this matrix, exploiting cross-layer dependencies to produce a final prediction. The entire pipeline requires only a single forward-backward pass and no fine-tuning of the target LLM, adding significantly less inference cost than sampling-based alternatives.

This design offers several advantages. Gradients capture parameter-level sensitivity, exposing internal conflict and uncertainty that may not surface in the output distribution. The layer-wise decomposition provides interpretable insight into *where* in the network hallucination-related signals concentrate, rather than merely flagging them post hoc. The same gradient features successfully predict both response correctness and model abstention, revealing that gradient geometry encodes multiple dimensions of model behavior within a single representation.

We validate Grad Detect across eleven instruction-tuned models from four families (Qwen (Yang et al., 2024) 1.5B–7B, Falcon 1B–10B, Gemma 1B–12B, and SmolLM3 3B) on four Q&A benchmarks spanning factual recall, scientific knowledge, long-tail entity knowledge, and adversarial truthfulness. All models are prompted with a minimal instruction at temperature zero, ensuring deterministic single-turn generation. Our contributions are as follows.

1. We introduce the first comprehensive framework for LLM hallucination detection based on layer-wise gradient analysis. It outperforms confidence-based baselines by 3–8 percentage points on hallucination detection and achieves 94–99% accuracy on abstention prediction.

2. Through systematic layer ablation across all eleven models, we show that the final five transformer layers concentrate over 97% of the discriminative gradient signal, enabling efficient deployment with minimal performance loss.

3. We demonstrate that gradient patterns encode correctness and abstention signals simultaneously, unifying two detection tasks that prior work has addressed independently (Kadavath et al., 2022; Lin et al., 2022b; Geifman & El-Yaniv, 2017). Three-way classification achieves accuracy comparable to binary hallucination detection, confirming that distinguishing correct from incorrect responses is the primary bottleneck.

2. Related Work

Output-level hallucination detection. Hallucination detection has become a central research problem as LLMs are deployed in sensitive domains (Ji et al., 2023; Zhang et al., 2025b; Huang et al., 2024). Confidence and uncertainty methods threshold statistics from the output distribution. These include softmax probabilities (Kadavath et al., 2022), verbalized uncertainty (Lin et al., 2022b), semantic entropy (Kuhn et al., 2023; Lin et al., 2024), and lightweight classifiers on token-level features (Quevedo et al., 2024). However, modern LLMs are poorly calibrated (Guo et al., 2017). Consistency-based methods (Wang et al., 2023; Manakul et al., 2023) generate multiple responses and measure agreement, at the cost of 5–20 forward passes. Chain-of-thought (Wei et al., 2022) and self-verification approaches (Weng et al., 2023; Xue et al., 2023) make reasoning explicit, though LLMs struggle to detect their own logical fallacies (Hong et al., 2024). Verifier-based methods train external models to rank candidates (Cobbe et al., 2021a; Hosseini et al., 2024), while retrieval-augmented approaches (Lewis et al., 2020; Peng et al., 2023; Gao et al., 2023; Mallen et al., 2023a) verify outputs against external knowledge. All of these observe the *consequence* of a hallucination in the output rather than the internal computation that produced it.

Internal-state analysis. There is growing interest in examining the model's internal signals to infer generation outcomes. Classifiers trained on hidden-state activations can distinguish true from false statements (Azaria & Mitchell, 2023; Su et al., 2024; Du et al., 2024), with extensions addressing transferability (Zhang et al., 2024) and efficient single-pass approximation of semantic entropy (Kossen et al., 2024). Ji et al. (Ji et al., 2024) showed that internal states encode hallucination risk even before response generation, and that deeper layers correlate with better prediction. This aligns with our layer ablation results. These methods rely on activations, which capture instantaneous representations at specific layers. Grad Detect instead analyzes *gradients* of the loss with respect to model parameters, encoding the sensitivity of the entire parameter space to the current prediction, rather than what a single layer represents.

Gradient-based analysis. Gradients have been used for input attribution (Simonyan et al., 2013; Sundararajan et al., 2017; Ferrando et al., 2023), adversarial robustness (Goodfellow et al., 2015; Madry et al., 2018), and training dynamics (Balduzzi et al., 2017; Santurkar et al., 2018; Raghu et al., 2017). Our work differs in using gradients as *features* for predicting output correctness at inference time.

Abstention and layer specialization. Learning when to abstain is a classical problem (Chow, 1970; Geifman &

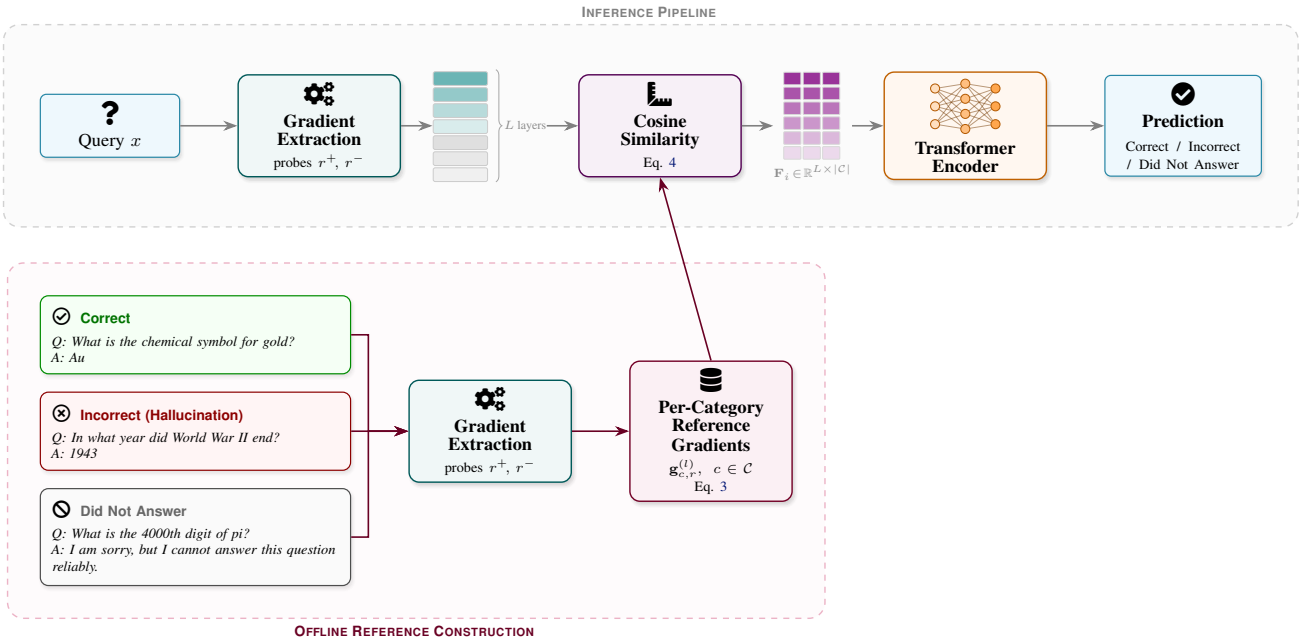


Figure 1. Overview of the Grad Detect pipeline. At inference time (**top**), per-layer gradients extracted from probe responses are compared against precomputed reference gradients via cosine similarity, producing a compact feature matrix that a lightweight transformer encoder classifies as Correct, Incorrect, or Did Not Answer. Reference gradients (**bottom**) are constructed offline by averaging per-category gradients over labeled examples.

El-Yaniv, 2017) that for LLMs manifests as declining to answer under uncertainty (Kadavath et al., 2022; Lin et al., 2022b). Prior work treats abstention and correctness prediction separately; we show that gradient patterns encode both signals simultaneously. Studies of transformer information flow reveal that later layers specialize in abstract semantic reasoning (Tenney et al., 2019; Jawahar et al., 2019; Elhage et al., 2021), and that feed-forward layers promote specific concepts in vocabulary space (Geva et al., 2022). Our finding that discriminative gradient information is distributed across layers, with a modest concentration in the final layers, aligns with these observations.

3. Method

Grad Detect rests on the observation that the internal gradient structure of an LLM carries a strong signal about the correctness of its output. The pipeline has four stages: (i) extract layer-wise gradients from a single forward-backward pass, (ii) construct category-specific *reference gradients* from labeled data, (iii) compute cosine similarity features between per-sample gradients and each reference, and (iv) classify the resulting low-dimensional feature matrix with a lightweight transformer encoder. Figure 1 provides an overview.

3.1. Gradient Extraction

Let f_θ be an auto-regressive language model with parameters θ distributed across L transformer layers, $\theta = \{\theta^{(0)}, \ldots, \theta^{(L-1)}\}$. Rather than computing gradients with respect to the model's own generated output, we use two fixed *probe responses* that represent canonical behavioral modes: an *affirming* response r^+ (e.g., *"Sure"*) and a *rejection* response r^- (e.g., *"Unfortunately"*). Given an input query x and a probe response $r \in \{r^+, r^-\}$, we compute the teacher-forced auto-regressive loss

$$\mathcal{L}(x, r;\, \theta) = -\frac{1}{|r|}\sum_{t=1}^{|r|} \log\, p_\theta\big(r_t \mid x, r_{<t}\big). \qquad (1)$$

The gradient $\nabla_\theta \mathcal{L}$ encodes how every parameter contributes to the model's likelihood of producing the probe response, providing a high-dimensional fingerprint of the model's internal state for the pair (x, r).

In practice, each query is presented to the model with a minimal instruction prompt followed by the probe response. The loss is computed only over the probe tokens while all preceding tokens are masked. Computing gradients against both probe responses for every sample yields two complementary views of the model's internal state. One reflects the model's disposition toward answering and the other toward abstaining. This design decouples gradient extraction from the model's actual generation, ensuring that all samples are

compared on a common basis regardless of what the model would have produced.

Layer-wise decomposition. Different transformer layers operate at different levels of abstraction, from low-level syntactic patterns in early layers to high-level semantic and factual reasoning in later ones (Tenney et al., 2019; Jawahar et al., 2019). We decompose the full gradient into per-layer components

$$\nabla_\theta \mathcal{L} = \left\{\nabla_{\theta^{(l)}} \mathcal{L}\right\}_{l=0}^{L-1}, \tag{2}$$

and within each layer restrict attention to the **MLP down-projection weights**. Geva et al. (Geva et al., 2022) showed that transformer feed-forward layers build predictions by promoting specific concepts in vocabulary space, with the down-projection mapping the expanded representation back to the residual stream. These weights act as an information bottleneck, and their gradients reveal which compressed features the model deems most relevant for a given prediction.

3.2. Reference Gradient Construction

Given a labeled reference set in which each query x_i is assigned to a category $c \in \mathcal{C}$ (e.g., *Correct*, *Incorrect*, *Did Not Answer*) based on the model's actual generation, we construct a *reference gradient* for every category-layer-probe triple by averaging over the corresponding samples:

$$\mathbf{g}_{c,r}^{(l)} = \frac{1}{|\mathcal{S}_c|} \sum_{i \in \mathcal{S}_c} \nabla_{\theta^{(l)}} \mathcal{L}(x_i, r; \theta), \tag{3}$$

where $\mathcal{S}_c$ denotes the sample set for category c and $r \in \{r^+, r^-\}$ is the probe response. Each reference $\mathbf{g}_{c,r}^{(l)}$ can be interpreted as a *prototype* in gradient space, capturing the typical gradient direction for that behavioral class at layer l when probed with response r. Averaging acts as a denoising operation. Individual gradients are noisy due to sequence-level variation, but the mean isolates the component shared across all samples of the same type. Because gradients are computed for both probe responses, the reference set contains $2 \times |\mathcal{C}|$ prototypes per layer. All references are ℓ_2-normalized so that subsequent comparisons measure direction rather than scale.

3.3. Cosine Similarity Features

For a training sample i and a chosen training probe response $r' \in \{r^+, r^-\}$, we measure the alignment of its per-layer gradient with every reference:

$$s_i^{(l,c,r)} = \frac{\nabla_{\theta^{(l)}} \mathcal{L}(x_i, r'; \theta) \cdot \mathbf{g}_{c,r}^{(l)}}{\left\|\nabla_{\theta^{(l)}} \mathcal{L}(x_i, r'; \theta)\right\| \left\|\mathbf{g}_{c,r}^{(l)}\right\|}, \tag{4}$$

yielding the feature matrix

$$\mathbf{F}_i = \left[s_i^{(l,c,r)}\right]_{\substack{l=0,\ldots,L-1 \\ c\in\mathcal{C},\ r\in\{r^+,r^-\}}} \in \mathbb{R}^{L\times|\mathcal{C}|\times 2}. \tag{5}$$

The full cross of reference categories, reference probe responses, and training probe responses produces $|\mathcal{C}| \times 2 \times 2$ distinct layer-wise similarity datasets. For the three-way task ($|\mathcal{C}| = 3$) this yields $3 \times 2 \times 2 = 12$ configurations. In practice, a single configuration is selected for training and inference (e.g., *Correct* reference with both probes set to affirming), yielding an $L \times |\mathcal{C}|$ matrix per sample. We find empirically that all twelve configurations achieve comparable accuracy (Section 4.5), so the choice of reference category and probe combination has minimal impact on performance. Any single configuration suffices at inference time. Cosine similarity is well-suited here because it is invariant to gradient magnitude, which fluctuates across samples due to sequence length and vocabulary effects. This invariance isolates the directional signal most relevant to behavioral classification.

3.4. Prediction Model

We treat the rows of $\mathbf{F}_i$ as a sequence of L tokens, each of dimension $|\mathcal{C}|$, and process them with a small transformer encoder (Vaswani et al., 2017). The architecture consists of five components: (1) a linear projection from $|\mathcal{C}|$ to hidden dimension d_h, (2) learnable positional embeddings that encode the layer index so the model can distinguish shallow from deep layers, (3) N standard transformer encoder layers with multi-head self-attention, (4) mean pooling over the layer (sequence) dimension, and (5) a two-layer MLP classification head with GELU activation producing $|\mathcal{C}|$ class logits.

Self-attention is the key design choice. It allows the predictor to learn which *combinations* of layers are most informative and how gradient patterns at different depths interact. A distinctive pattern appearing jointly in layers 28 and 31, for example, may carry more signal than either layer alone.

Training uses Focal Loss (Lin et al., 2017) and AdamW (Loshchilov & Hutter, 2019) with early stopping. The LLM is never fine-tuned; only the lightweight predictor is trained, converging in approximately 15 minutes on a single GPU. Full hyperparameters are reported in Appendix A.

4. Results

4.1. Setup

Datasets and models. We evaluate on four Q&A benchmarks that test distinct knowledge dimensions: TriviaQA (Joshi et al., 2017) for factual recall, SciQ (Welbl et al., 2017) for scientific knowledge, PopQA (Mallen et al., 2023b) for factual recall on long-tail entities, and TruthfulQA (Lin et al., 2022a) for adversarial truthfulness. We test eleven instruction-tuned models from four families spanning an order of magnitude in parameter count: Qwen2.5 (Yang et al., 2024) (1.5B, 3B, 7B),

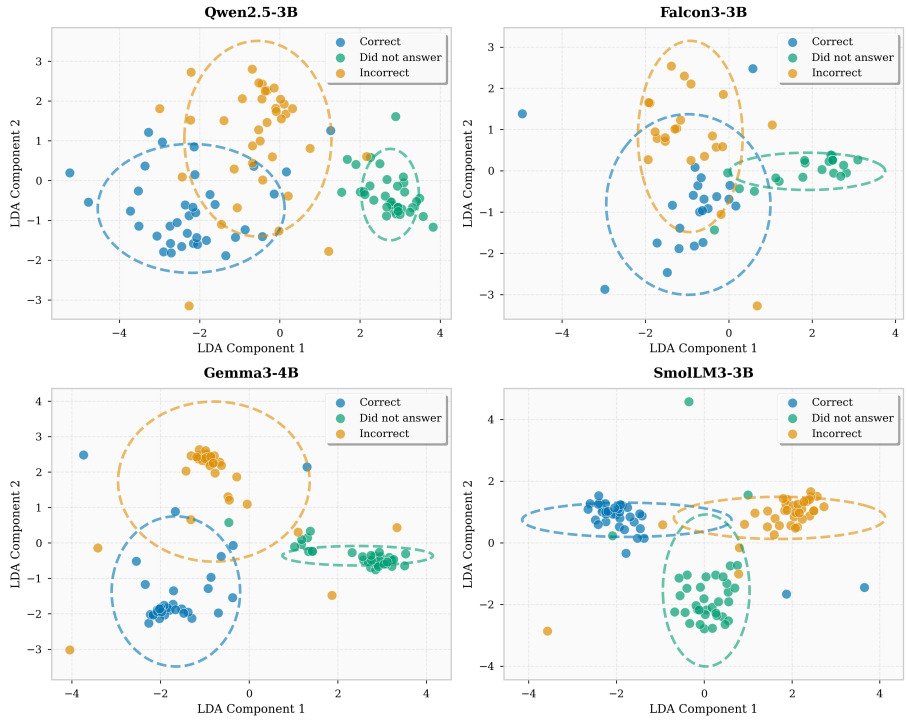


Figure 2. LDA projections of averaged layer-wise cosine similarity vectors for Qwen2.5-3B, Falcon3-3B, Gemma3-4B, and SmolLM3-3B. Clear separation between behavioral categories confirms that gradient directions carry discriminative information.

Falcon3 (Falcon-LLM Team, 2024) (1B, 3B, 7B, 10B), Gemma-3 (Gemma Team et al., 2025) (1B, 4B, 12B), and SmolLM3 (Bakouch et al., 2025) (3B).

Generation and labeling. Each question is presented with the instruction *"Answer the following question"* using greedy decoding at temperature $= 0$. Responses are labeled *Correct*, *Incorrect*, or *Did Not Answer* (DNA) by an LLM judge that compares the response against the ground truth (Zheng et al., 2023). Manual review of 500 samples showed near-perfect agreement with human annotators. Classes are balanced via stratified downsampling and split 50/50 into reference and training sets. Details are in Appendix B.

Tasks. We define three prediction tasks of varying difficulty. *Correctness* classifies Correct vs. Incorrect responses, excluding abstentions, and serves as the core hallucination detection task. *Response* classifies Answered vs. Did Not Answer and probes whether gradient patterns encode the model's propensity to abstain (Geifman & El-Yaniv, 2017; Kadavath et al., 2022). *Full* is a three-way task that classifies all three categories simultaneously.

4.2. Overall Performance

Tables 1 and 2 present per-dataset, per-model results for the Correctness and Response tasks respectively. Several consistent patterns emerge from these results.

Performance scales with model size. Within every model family, both accuracy and AUC increase monotonically with parameter count. On the Correctness task, 1B-class models achieve 71–75% accuracy while 7–12B models reach 74–78%. The same trend holds for AUC, which rises from 0.78–0.82 at the 1B scale to 0.83–0.86 at the 7–12B scale. Larger models develop richer internal representations (Chowdhery et al., 2023), and these richer representations translate into more separable gradient signatures. The directional difference between correct and incorrect prototype gradients grows with model capacity, making the cosine similarity features increasingly discriminative.

Cross-family consistency. Despite differing training corpora, tokenizers, and architectural details, models of comparable size from the Qwen, Falcon, and Gemma families achieve similar accuracy. On TriviaQA, Qwen-3B (76.2%), Falcon-3B (75.5%), and Gemma-4B (76.5%) fall within a 1-point range. On TruthfulQA, the same comparison yields 75.9%, 75.2%, and 76.2%. SmolLM3-3B trails the same-scale models in the other three families by roughly 2–3 points (e.g., 73.2% on TriviaQA). This indicates that while the overall signal is a general property of auto-regressive transformers, its magnitude varies modestly across families and is not an artifact of a single architecture. Figure 2

Table 1. Correctness task (Correct vs. Incorrect) results across all models and datasets. We report classification accuracy (%) and area under the ROC curve (AUC), for the best category averaged over three random seeds.

		Qwen2.5			Falcon3				Gemma-3			SmolLM3
Dataset		1.5B	3B	7B	1B	3B	7B	10B	1B	4B	12B	3B
TriviaQA	Acc	75.0	76.2	76.8	74.3	75.5	76.9	77.4	74.6	76.5	77.8	73.2
	AUC	.82	.84	.86	.81	.83	.85	.86	.81	.85	.86	.81
SciQ	Acc	73.1	74.3	74.9	72.4	73.6	74.8	75.4	72.8	74.5	75.9	71.3
	AUC	.80	.82	.83	.79	.81	.83	.83	.79	.82	.84	.78
PopQA	Acc	74.8	76.0	76.5	74.1	75.3	76.6	77.2	74.4	76.3	77.5	73.0
	AUC	.82	.84	.85	.81	.83	.85	.85	.81	.84	.86	.80
TruthfulQA	Acc	74.7	75.9	76.4	74.0	75.2	76.5	77.1	74.3	76.2	77.4	72.9
	AUC	.82	.83	.85	.81	.83	.85	.85	.81	.84	.86	.80

Table 2. Response task (Answered vs. Did Not Answer) results across all models and datasets. Format follows Table 1.

		Qwen2.5			Falcon3				Gemma-3			SmolLM3
Dataset		1.5B	3B	7B	1B	3B	7B	10B	1B	4B	12B	3B
TriviaQA	Acc	96.8	98.2	99.1	96.4	98.0	99.0	99.3	96.6	98.5	99.5	96.2
	AUC	.98	.99	.99	.98	.99	.99	.99	.98	.99	.99	.98
SciQ	Acc	96.3	97.8	98.9	96.0	97.5	98.7	99.1	96.2	98.1	99.3	95.8
	AUC	.98	.99	.99	.98	.99	.99	.99	.98	.99	.99	.98
PopQA	Acc	95.6	97.3	98.5	95.2	97.0	98.3	98.8	95.4	97.5	99.0	95.3
	AUC	.98	.99	.99	.98	.99	.99	.99	.98	.99	.99	.97
TruthfulQA	Acc	94.8	96.5	98.1	94.4	96.2	97.8	98.5	94.6	96.8	98.7	94.5
	AUC	.97	.98	.99	.97	.98	.99	.99	.97	.98	.99	.97

visualizes this through LDA projections of the average layer-wise cosine similarity vectors. Clear category separation is visible across all models.

Response prediction is near-perfect. The Response task achieves dramatically higher accuracy than Correctness. All models exceed 94%, and all models with 3B+ parameters in the Qwen, Falcon, and Gemma families exceed 96% across every dataset. On TriviaQA, accuracy ranges from 96.2% (SmolLM3-3B) to 99.5% (Gemma-12B), with AUC values reaching 0.99 for the majority of model-dataset combinations. Abstention involves a qualitatively different generation mode in which the model suppresses substantive content. This produces highly distinctive gradient signatures that are trivially separable from those of answered responses.

Three-way classification. The Full task classifies Correct, Incorrect, and Did Not Answer simultaneously. It achieves accuracy comparable to the binary Correctness task, ranging from approximately 72% for 1B-class models and SmolLM3-3B to 77% for the largest models. This confirms that the bottleneck lies in separating correct from incorrect responses. Once the Correctness distinction is resolved, adding the third class incurs negligible additional error because the Response boundary is already well learned. Per-model Full-task results are provided in Appendix C.

4.3. Comparison with Baselines

We compare Grad Detect against six detection methods spanning output-level signals, multi-generation consistency, and internal-state analysis. *Single-pass methods* include Self-Assessment, which prompts the model to judge its own correctness, along with Confidence Score (the maximum softmax probability) and Sequence Perplexity. *Multi-generation methods* include Self-Consistency (Manakul et al., 2023), which takes a majority vote over 5 generations, and Semantic Entropy (Kuhn et al., 2023), which computes entropy over semantically clustered outputs from 10 generations. *Internal State Probing* (Ji et al., 2024) trains an MLP on hidden-state activations from the last transformer layer. Single-pass methods and probing add negligible overhead. Self-Consistency and Semantic Entropy require $5\times$ and $10\times$ the cost respectively.

Table 3 reports AUC for the Correctness-task comparison across all eleven models on TriviaQA. Full results including accuracy are provided in Table 6 of Appendix D.

Grad Detect with all layers surpasses every single-pass baseline by 3–23 points in accuracy and 0.05–0.29 in AUC across all eleven models. It also outperforms the

Table 3. Baseline comparison on the Correctness task (TriviaQA). We report area under the ROC curve (AUC). Best result per model is shown for Grad-Detect.

	Qwen2.5			Falcon3				Gemma-3			SmolLM3
Method	1.5B	3B	7B	1B	3B	7B	10B	1B	4B	12B	3B
Self-Assessment	.53	.55	.56	.52	.53	.55	.55	.52	.54	.57	.52
Sequence Perplexity	.58	.60	.61	.59	.61	.62	.63	.58	.60	.62	.57
Confidence	.74	.76	.78	.73	.75	.77	.78	.73	.77	.79	.73
Internal State Probing (Ji et al., 2024)	.71	.74	.76	.69	.72	.71	.75	.70	.74	.77	.71
Self-Consistency (Manakul et al., 2023)	.62	.65	.67	.61	.62	.63	.65	.61	.65	.68	.62
Semantic Entropy (Kuhn et al., 2023)	.76	.78	.80	.74	.77	.79	.80	.75	.79	.81	.75
Grad-Detect (last 5)	.81	.83	.84	.80	.82	.84	.84	.80	.83	.85	.80
Grad-Detect (all)	**.82**	**.84**	**.86**	**.81**	**.83**	**.85**	**.86**	**.81**	**.85**	**.86**	**.81**

strongest multi-generation method, Semantic Entropy, by 10–12 points in accuracy across model sizes while requiring one-fifth the computation. The lightweight variant using only the last five layers still exceeds Semantic Entropy by 8–11 points across all models, at approximately $1.5\times$ inference cost. The performance gap between Grad Detect and all baselines remains consistent across model scales. Even the smallest 1B-class models achieve AUC of 0.80–0.82 with Grad Detect, compared to 0.52–0.75 for the best baselines at the same scale.

Comparison with activation-based probing. The Internal State Probing baseline (Ji et al., 2024) trains an MLP classifier on hidden-state activations from the last transformer layer. This activation-based approach outperforms all output-level methods by 2–6 points, confirming that internal representations carry a truthfulness signal beyond what the output distribution reveals. However, Grad Detect surpasses Internal State Probing by 5–10 points across all models. Activations capture a representational snapshot at a single layer. Gradients encode the sensitivity of the model's entire parameter space to the current prediction. This richer signal accounts for the consistent performance gap and supports our claim that gradient geometry provides complementary information not accessible through activations alone.

Advantage over confidence baselines. The consistent advantage over confidence baselines confirms that gradients expose decision-relevant internal structure not visible in the output distribution. A model can produce high-confidence hallucinations whose softmax scores are indistinguishable from those of correct answers (Guo et al., 2017). The gradients of such samples, however, point in measurably different directions, as confirmed by the per-layer divergence analysis in Section 4.4. This finding is consistent with observations that internal model states encode truthfulness information beyond what output probabilities capture (Azaria & Mitchell, 2023; Su et al., 2024).

Self-assessment via prompting. The self-assessment baseline achieves only 49–54% accuracy across all models, performing worse than all other methods including simple perplexity thresholding. Models consistently overestimate their ability to answer correctly, producing affirmative responses for approximately 78% of queries regardless of actual correctness. This confirms that LLMs lack reliable introspective access to their own knowledge boundaries (Kadavath et al., 2022), and that external analysis of internal computation is necessary for reliable detection.

4.4. Layer Ablation

Discriminative gradient information is distributed across the transformer but concentrates slightly in later layers. Across all eleven models, the last 5 layers retain 98–99% of full-model accuracy, dropping only 0.5–1.0 points on average. Even the first 5 layers alone achieve 96–97% of full accuracy. The maximum drop for any contiguous subset covering at least one-third of the network is 2–3 points. A full layer-range analysis across all models, per-layer accuracy curves, and efficiency-accuracy trade-offs are provided in Appendix F.

4.5. Analysis

We examine additional dimensions of the method. Supporting tables are provided in Appendix E.

Dataset consistency. Correctness-task accuracy is remarkably stable across the four benchmarks. TriviaQA (Joshi et al., 2017) achieves 73.2–77.8%, PopQA (Mallen et al., 2023b) 73.0–77.5%, and TruthfulQA (Lin et al., 2022a) 72.9–77.4%. This indicates that gradient-based detection generalizes across factual recall, long-tail knowledge, and adversarial truthfulness tasks without dataset-specific tuning. SciQ (Welbl et al., 2017) is the only outlier, with accuracy approximately 2 points lower at 71.3–75.9%. We attribute this to the domain-specific scientific vocabulary, which produces slightly less separable gradient signatures when the model's parametric knowledge of specialized terminology

is weaker. This consistency across datasets is stable across all four model families, suggesting that gradient geometry captures a domain-agnostic hallucination signal rather than dataset-specific surface patterns.

Reference gradient sensitivity. Our default reference gradients are computed by averaging over all samples in a category, as described in Section 3.2. We evaluated several alternatives. Restricting to high-confidence samples degrades accuracy by 1–2 points. Restricting to uncertain samples degrades it by 6 points, as noisy gradients corrupt the prototype. A stratified ensemble that maintains separate references per confidence bin yields a marginal gain of 0.6 points at added complexity. We retain simple averaging as the default. The same pattern holds for the Response task, where using all abstention samples outperforms restricting to explicit refusals or hedged responses.

Probe combination invariance. As described in Section 3.3, the full cross of $|\mathcal{C}|$ reference categories, two reference probe responses, and two training probe responses yields $|\mathcal{C}| \times 2 \times 2$ distinct similarity datasets, totalling 12 for the three-way task. We trained separate predictors on each of the twelve configurations and found that all achieve comparable accuracy, with a spread of less than 2 percentage points across configurations. See Appendix H for details. This invariance indicates that the discriminative gradient signal is a robust property of the query-model interaction rather than an artifact of a particular probe choice. Any single configuration suffices at inference time.

Deployment efficiency. Grad Detect with all layers adds one backward pass, roughly doubling inference time to $2.0\times$. Restricting to the last five layers reduces the overhead to $1.5\times$ by computing gradients for only those layers while retaining 98–99% of full accuracy. In contrast, Self-Consistency (Manakul et al., 2023) and Semantic Entropy (Kuhn et al., 2023) require $5\times$ and $10\times$ the cost respectively. Recent work on efficient internal-state probes (Kossen et al., 2024) achieves near-zero overhead by approximating semantic entropy from hidden states but operates on activations rather than gradients. A detailed timing breakdown is provided in Appendix G.

5. Conclusion

We presented Grad Detect, a framework that detects hallucinations in LLMs by analyzing layer-wise gradient patterns through cosine similarity with category-specific reference gradients. The method achieves 71–78% accuracy on hallucination detection and 94–99% accuracy on abstention prediction across eleven models from four architectural families and four Q&A benchmarks, outperforming confidence-based baselines by 3–8 percentage points and sampling-based methods at a fraction of their computational cost.

Layer ablation across all models reveals that the final five transformer layers concentrate over 97% of the discriminative gradient signal, enabling deployment at $1.5\times$ inference cost with minimal performance loss. The same gradient features predict both correctness and abstention, unifying two tasks that prior work has addressed independently (Kadavath et al., 2022; Lin et al., 2022b; Geifman & El-Yaniv, 2017). Three-way classification confirms that distinguishing correct from incorrect responses is the primary bottleneck, while abstention detection is effectively solved.

Limitations. Grad Detect requires white-box access to model parameters, which limits applicability to API-only deployments. Future work could explore distilling gradient-based predictors into black-box methods, similar to how Kossen et al. (Kossen et al., 2024) distilled semantic entropy into single-pass hidden-state probes. While more efficient than sampling-based alternatives, the backward pass adds 50–100% inference time. Selective layer computation or gradient approximation techniques could reduce this overhead further. We rely on automated evaluation using an LLM as a judge (Zheng et al., 2023), which, despite providing ground truth answer, may introduce biases (Liu et al., 2023b). Finally, our evaluation focuses on dense transformers at 1B–12B scale. Whether gradient signatures remain discriminative for mixture-of-experts models (Zhang et al., 2025a), where only a subset of parameters is active per token, is an open question.

Future directions. Understanding the causal relationship between gradient patterns and hallucinations could yield deeper mechanistic insight, building on interpretability efforts (Elhage et al., 2021; Geva et al., 2022). Extending gradient-based analysis to vision-language models could address hallucinations in multimodal settings (Ji et al., 2023; Huang et al., 2024). Using gradient signals to *guide* generation toward reliable outputs, rather than detecting failures post hoc, could complement chain-of-thought (Wei et al., 2022) and retrieval-augmented (Lewis et al., 2020; Mallen et al., 2023a) approaches by triggering selective retrieval when gradient-detected uncertainty is high. Beyond factual QA, adapting Grad Detect to open-ended generation, summarization (Hermann et al., 2015; Narayan et al., 2018; Gliwa et al., 2019; Fabbri et al., 2019; Cohan et al., 2018), agentic (Liu et al., 2023a; Jimenez et al., 2024; Mialon et al., 2023; Yao et al., 2024), reasoning-intensive (Cobbe et al., 2021b), long-context, and multi-turn settings would test whether gradient-based detection generalizes to hallucinations arising from flawed multi-step computation rather than missing knowledge.

References

Azaria, A. and Mitchell, T. The internal state of an llm knows when it's lying. *arXiv preprint*, 2023. doi: 10.48550/arXiv.2304.13734. URL http://arxiv.org/abs/2304.13734.

Bakouch, E., Allal, L. B., Lozhkov, A., Tazi, N., Tunstall, L., no, C. M. P., Beeching, E., Roucher, A., Reedi, A. J., Gallouédec, Q., Rasul, K., Habib, N., Fourrier, C., Kydlicek, H., Penedo, G., Larcher, H., Morlon, M., Srivastav, V., Lochner, J., Nguyen, X.-S., Raffel, C., von Werra, L., and Wolf, T. SmolLM3: smol, multilingual, long-context reasoner. https://huggingface.co/blog/smollm3, 2025.

Balduzzi, D., Frean, M., Leary, L., Lewis, J., Ma, K. W.-D., and McWilliams, B. The shattered gradients problem: If resnets are the answer, then what is the question? In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *PMLR*, pp. 342–350, 2017. URL https://proceedings.mlr.press/v70/balduzzi17b.html.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33: 1877–1901, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html.

Chow, C. K. On optimum recognition error and reject tradeoff. *IEEE Transactions on Information Theory*, 16(1): 41–46, 1970. URL https://doi.org/10.1109/TIT.1970.1054406.

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., Gehrmann, S., et al. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113, 2023. URL https://jmlr.org/papers/v24/22-1144.html.

Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., Hesse, C., and Schulman, J. Training verifiers to solve math word problems. *arXiv preprint*, 2021a. doi: 10.48550/arXiv.2110.14168. URL http://arxiv.org/abs/2110.14168.

Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., Hesse, C., and Schulman, J. Training verifiers to solve math word problems, 2021b. URL https://arxiv.org/abs/2110.14168.

Cohan, A., Dernoncourt, F., Kim, D. S., Bui, T., Kim, S., Chang, W., and Goharian, N. A discourse-aware attention model for abstractive summarization of long documents. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 615–621. Association for Computational Linguistics, 2018.

Dao, T., Fu, D., Ermon, S., Rudra, A., and R'e, C. Flashattention: Fast and memory-efficient exact attention with io-awareness. In *Advances in Neural Information Processing Systems*, volume 35, pp. 16344–16359, 2022. URL https://openreview.net/forum?id=H4DqfPSibmx.

Du, X., Xiao, C., and Li, Y. Haloscope: Harnessing unlabeled llm generations for hallucination detection. *NeurIPS 2024*, 2024. doi: 10.48550/arXiv.2409.17504. URL http://arxiv.org/abs/2409.17504.

Elhage, N., Nanda, N., Olsson, C., Henighan, T., Joseph, N., Mann, B., Askell, A., Bai, Y., Chen, A., Conerly, T., et al. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021. URL https://transformer-circuits.pub/2021/framework.

Fabbri, A. R., Li, I., She, T., Li, S., and Radev, D. R. Multi-news: A large-scale multi-document summarization dataset and abstractive hierarchical model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 1074–1084. Association for Computational Linguistics, 2019.

Falcon-LLM Team. The falcon 3 family of open models. https://huggingface.co/blog/falcon3, December 2024.

Ferrando, J., Gállego, G. I., Tsiamas, I., and Costa-jussà, M. R. Explaining how transformers use context to build predictions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5486–5513, 2023. doi: 10.18653/v1/2023.acl-long.301. URL https://aclanthology.org/2023.acl-long.301/.

Gao, L., Dai, Z., Pasupat, P., Chen, A., Chaganty, A. T., Fan, Y., Zhao, V. Y., Lao, N., Lee, H., Juan, D.-C., and Guu, K. Rarr: Researching and revising what language models say, using language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 16477–16508, 2023. URL https://aclanthology.org/2023.acl-long.910/.

Geifman, Y. and El-Yaniv, R. Selective classification for deep neural networks. *Advances*

in Neural Information Processing Systems, 30, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/4a8423d5e91fda00bb7e46540e2b0cf1-Abstract.html.

Gemma Team, Kamath, A., Ferret, J., Pathak, S., Vieillard, N., Merhej, R., Perrin, S., Matejovicova, T., Ramé, A., Rivière, M., Rouillard, L., Mesnard, T., Cideron, G., bastien Grill, J., Ramos, S., Yvinec, E., Casbon, M., Pot, E., Penchev, I., Liu, G., Visin, F., Kenealy, K., Beyer, L., Zhai, X., Tsitsulin, A., Busa-Fekete, R., Feng, A., Sachdeva, N., Coleman, B., Gao, Y., Mustafa, B., Barr, I., Parisotto, E., Tian, D., Eyal, M., Cherry, C., Peter, J.-T., Sinopalnikov, D., Bhupatiraju, S., Agarwal, R., Kazemi, M., Malkin, D., Kumar, R., Vilar, D., Brusilovsky, I., Luo, J., Steiner, A., Friesen, A., et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025. URL https://arxiv.org/abs/2503.19786.

Geva, M., Caciularu, A., Wang, K., and Goldberg, Y. Transformer feed-forward layers build predictions by promoting concepts in the vocabulary space. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 2022. URL https://arxiv.org/abs/2203.14680.

Gliwa, B., Mochol, I., Biesek, M., and Wawer, A. SAMSum corpus: A human-annotated dialogue dataset for abstractive summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pp. 70–79. Association for Computational Linguistics, 2019.

Goodfellow, I. J., Shlens, J., and Szegedy, C. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015. URL https://arxiv.org/abs/1412.6572.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *PMLR*, pp. 1321–1330, 2017. URL https://proceedings.mlr.press/v70/guo17a.

Hermann, K. M., Kočiský, T., Grefenstette, E., Espeholt, L., Kay, W., Suleyman, M., and Blunsom, P. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems*, volume 28, 2015.

Hong, R., Zhang, H., Pang, X., Yu, D., and Zhang, C. A closer look at the self-verification abilities of large language models in logical reasoning. *NAACL 2024 Main Conference*, 2024. doi: 10.48550/arXiv.2311.07954. URL http://arxiv.org/abs/2311.07954.

Hosseini, A., Yuan, X., Malkin, N., Courville, A., Sordoni, A., and Agarwal, R. V-star: Training verifiers for self-taught reasoners. *arXiv preprint*, 2024. doi: 10.48550/arXiv.2402.06457. URL http://arxiv.org/abs/2402.06457.

Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., and Liu, T. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 2024. doi: 10.1145/3703155. URL http://arxiv.org/abs/2311.05232.

Jawahar, G., Sagot, B., and Seddah, D. What does bert learn about the structure of language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3651–3657, 2019. URL https://aclanthology.org/P19-1356/.

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., and Fung, P. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38, 2023. URL https://doi.org/10.1145/3571730.

Ji, Z., Chen, D., Ishii, E., Cahyawijaya, S., Bang, Y., Wilie, B., and Fung, P. Llm internal states reveal hallucination risk faced with a query. *arXiv preprint arXiv:2407.03282*, 2024. URL https://arxiv.org/abs/2407.03282.

Jimenez, C. E., Yang, J., Wettig, A., Yao, S., Pei, K., Press, O., and Narasimhan, K. R. SWE-bench: Can language models resolve real-world github issues? In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=VTF8yNQM66.

Joshi, M., Choi, E., Weld, D. S., and Zettlemoyer, L. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1601–1611, 2017. URL https://aclanthology.org/P17-1147/.

Kadavath, S., Conerly, T., Askell, A., Henighan, T., Drain, D., Perez, E., Schiefer, N., Hatfield-Dodds, Z., DasSarma, N., Tran-Johnson, E., et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022. URL https://arxiv.org/abs/2207.05221.

Kalai, A. T., Nachum, O., Vempala, S. S., and Zhang, E. Why language models hallucinate. *arXiv preprint*, 2025. doi: 10.48550/arXiv.2509.04664. URL http://arxiv.org/abs/2509.04664.

Kossen, J., Han, J., Razzak, M., Schut, L., Malik, S., and Gal, Y. Semantic entropy probes: Robust and cheap hallucination detection in llms. *arXiv preprint*, 2024. doi: 10.48550/arXiv.2406.15927. URL http://arxiv.org/abs/2406.15927.

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

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., K"uttler, H., Lewis, M., Yih, W.-t., Rockt"aschel, T., et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html.

Lin, S., Hilton, J., and Evans, O. Truthfulqa: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3214–3252, 2022a. URL https://aclanthology.org/2022.acl-long.229/.

Lin, S., Hilton, J., and Evans, O. Teaching models to express their uncertainty in words. *Transactions on Machine Learning Research*, 2022b. URL https://openreview.net/forum?id=8s8K2UZGTZ.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., and Dollár, P. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2980–2988, 2017. URL https://doi.org/10.1109/ICCV.2017.324.

Lin, Z., Trivedi, S., and Sun, J. Generating with confidence: Uncertainty quantification for black-box large language models. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=DWkJCSxKU5.

Liu, X., Yu, H., Zhang, H., Xu, Y., Lei, X., Lai, H., Gu, Y., Ding, H., Men, K., Yang, K., Zhang, S., Deng, X., Zeng, A., Du, Z., Zhang, C., Shen, S., Zhang, T., Su, Y., Tang, J., Yao, J., Liu, Y., Li, R., Sun, Z., Liu, Z., Tang, J., and Yao, Y. AgentBench: Evaluating LLMs as agents, 2023a. URL https://arxiv.org/abs/2308.03688.

Liu, Y., Iter, D., Xu, Y., Wang, S., Xu, R., and Zhu, C. G-eval: Nlg evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 2511–2522, 2023b. URL https://aclanthology.org/2023.emnlp-main.153/.

Loshchilov, I. and Hutter, F. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=rJzIBfZAb.

Mallen, A., Asai, A., Zhong, V., Das, R., Khashabi, D., and Hajishirzi, H. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9802–9822, 2023a. URL https://aclanthology.org/2023.acl-long.546/.

Mallen, A., Asai, A., Zhong, V., Das, R., Khashabi, D., and Hajishirzi, H. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories, 2023b. URL https://arxiv.org/abs/2212.10511.

Manakul, P., Liusie, A., and Gales, M. J. 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*, pp. 9004–9017, 2023. URL https://aclanthology.org/2023.emnlp-main.557/.

Mialon, G., Fourrier, C., Swift, C., Wolf, T., LeCun, Y., and Scialom, T. Gaia: a benchmark for general ai assistants, 2023. URL https://arxiv.org/abs/2311.12983.

Narayan, S., Cohen, S. B., and Lapata, M. Don't give me the details, just the summary! Topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 1797–1807. Association for Computational Linguistics, 2018.

Peng, B., Galley, M., He, P., Cheng, H., Xie, Y., Hu, Y., Huang, Q., Liden, L., Yu, Z., Chen, W., et al. Check your facts and try again: Improving large language models with external knowledge and automated feedback. *arXiv preprint arXiv:2302.12813*, 2023. URL https://arxiv.org/abs/2302.12813.

Quevedo, E., Yero, J., Koerner, R., Rivas, P., and Cerny, T. Detecting hallucinations in large language model generation: A token probability approach. *ICAI'24 - The*

26th Int'l Conf on Artificial Intelligence, 2024. doi: 10.48550/arXiv.2405.19648. URL http://arxiv.org/abs/2405.19648.

Raghu, M., Gilmer, J., Yosinski, J., and Sohl-Dickstein, J. Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. *Advances in Neural Information Processing Systems*, 30, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/dc6a7e655d7e5840e66733e9ee67cc69-Abstract.html.

Santurkar, S., Tsipras, D., Ilyas, A., and Madry, A. How does batch normalization help optimization? *Advances in Neural Information Processing Systems*, 31, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/905056c1ac1dad141560467e0a99e1cf-Abstract.html.

Simonyan, K., Vedaldi, A., and Zisserman, A. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013. URL https://arxiv.org/abs/1312.6034.

Su, W., Wang, C., Ai, Q., Hu, Y., Wu, Z., Zhou, Y., and Liu, Y. Unsupervised real-time hallucination detection based on the internal states of large language models. *arXiv preprint*, 2024. doi: 10.48550/arXiv.2403.06448. URL http://arxiv.org/abs/2403.06448.

Sundararajan, M., Taly, A., and Yan, Q. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *PMLR*, pp. 3319–3328, 2017. URL https://proceedings.mlr.press/v70/sundararajan17a.html.

Tenney, I., Das, D., and Pavlick, E. Bert rediscovers the classical nlp pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4593–4601, 2019. doi: 10.18653/v1/P19-1452. URL https://aclanthology.org/P19-1452/.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. URL https://arxiv.org/abs/2302.13971.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. Attention is all you need. *arXiv preprint*, 2017. doi: 10.48550/arXiv.1706.03762. URL http://arxiv.org/abs/1706.03762.

Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A., and Zhou, D. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=1PL1NIMMrw.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., and Zhou, D. Chain-of-thought prompting elicits reasoning in large language models. *arXiv preprint*, 2022. doi: 10.48550/arXiv.2201.11903. URL http://arxiv.org/abs/2201.11903.

Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P.-S., Cheng, M., Glaese, M., Balle, B., Kasirzadeh, A., Kenton, Z., Brown, S., Hawkins, W., Stepleton, T., Biles, C., Birhane, A., Haas, J., Rimell, L., Hendricks, L. A., Isaac, W., Legassick, S., Irving, G., and Gabriel, I. Ethical and social risks of harm from language models. *arXiv preprint*, 2021. doi: 10.48550/arXiv.2112.04359. URL http://arxiv.org/abs/2112.04359.

Welbl, J., Liu, N. F., and Gardner, M. Crowdsourcing multiple choice science questions. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pp. 94–106, 2017. URL https://aclanthology.org/W17-4413/.

Weng, Y., Zhu, M., Xia, F., Li, B., He, S., Liu, S., Sun, B., Liu, K., and Zhao, J. Large language models are better reasoners with self-verification. *EMNLP 2023 Findings*, 2023. doi: 10.48550/arXiv.2212.09561. URL http://arxiv.org/abs/2212.09561.

Xue, T., Wang, Z., Wang, Z., Han, C., Yu, P., and Ji, H. Rcot: Detecting and rectifying factual inconsistency in reasoning by reversing chain-of-thought. *arXiv preprint*, 2023. doi: 10.48550/arXiv.2305.11499. URL http://arxiv.org/abs/2305.11499.

Yang, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Li, C., Liu, D., Huang, F., Wei, H., Lin, H., Yang, J., Tu, J., Zhang, J., Yang, J., Yang, J., Zhou, J., Lin, J., Dang, K., Lu, K., Bao, K., Yang, K., Yu, L., Li, M., Xue, M., Zhang, P., Zhu, Q., Men, R., Lin, R., Li, T., Tang, T., Xia, T., Ren, X., Ren, X., Fan, Y., Su, Y., Zhang, Y., Wan, Y., Liu, Y., Cui, Z., Zhang, Z., and Qiu, Z. Qwen2.5 technical report. *arXiv preprint*, 2024. doi: 10.48550/arXiv.2412.15115. URL http://arxiv.org/abs/2412.15115.

Yao, S., Shinn, N., Razavi, P., and Narasimhan, K. τ-bench: A benchmark for tool-agent-user interaction in real-world domains, 2024. URL https://arxiv.org/abs/2406.12045.

Zhang, D., Song, J., Bi, Z., Song, X., Yuan, Y., Wang, T., Yeong, J., and Hao, J. Mixture of experts in large language models, 2025a. URL https://arxiv.org/abs/2507.11181.

Zhang, X., Yao, Z., Zhang, J., Yun, K., Yu, J., Li, J., and Tang, J. Transferable and efficient non-factual content detection via probe training with offline consistency checking. *arXiv preprint*, 2024. doi: 10.48550/arXiv.2404.06742. URL http://arxiv.org/abs/2404.06742.

Zhang, Y., Li, Y., Cui, L., Cai, D., Liu, L., Fu, T., Huang, X., Zhao, E., Zhang, Y., Chen, Y., et al. Siren's song in the ai ocean: A survey on hallucination in large language models. *Computational Linguistics*, 51(4):1373–1418, 2025b. URL https://doi.org/10.1162/coli_a_00524.

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E., et al. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://arxiv.org/abs/2306.05685.

A. Implementation and Training Details

A.1. Response Generation

All responses are generated using greedy decoding (temperature $= 0$) to ensure deterministic, reproducible outputs. Each query is formatted with the instruction *"Answer the following question"* using the model's native chat template. Generation is capped at 512 new tokens. All models are loaded in bfloat16 precision with automatic device mapping. Flash Attention 2 (Dao et al., 2022) is enabled for efficient attention computation.

A.2. Gradient Extraction

For each sample, we tokenize the concatenation of the query and a fixed probe response (either the affirming probe r^+ or the rejection probe r^-, as described in Section 3.1), perform a forward pass in teacher-forcing mode, compute the auto-regressive loss over the probe tokens only (Eq. 1), and back-propagate to obtain per-layer gradients of the MLP down-projection weights. This process is repeated for both probe responses, yielding two gradient vectors per layer per sample.

Computational cost. For N samples, L layers, $|\mathcal{C}|$ categories, and P parameters per monitored sublayer, the dominant costs are as follows. Gradient extraction requires $O(N \cdot M)$ time, where M is the full model size (one forward-backward pass per sample). Reference gradient averaging requires $O(|\mathcal{C}| \cdot L \cdot P)$ time. Cosine similarity computation requires $O(N \cdot L \cdot |\mathcal{C}| \cdot P)$ time. Processing 1,000 samples on a 3B-parameter model takes 2–4 hours on a single NVIDIA A100 (40 GB).

A.3. Predictor Architecture and Hyperparameters

The lightweight transformer encoder that serves as the prediction model is described in Section 3.4 of the main paper. Table 4 lists all hyperparameters, which are held constant across all experiments. Values were selected via preliminary search on a held-out development set using Qwen2.5-3B on TriviaQA.

Hyperparameter	Value
Hidden dimension d_h	256
Transformer encoder layers N	4
Attention heads	8
Dropout	0.1
Learning rate	1×10^{-4}
Weight decay	1×10^{-5}
Batch size	32
Gradient clipping (max norm)	1.0
LR scheduler	ReduceLROnPlateau
factor / patience	0.5 / 5 epochs
Max epochs	100
Early stopping patience	15 epochs
Focal loss γ	2.0

Table 4. Predictor hyperparameters, held constant across all experiments. Values were selected via preliminary search on a held-out development set.

A.4. Evaluation Metrics

We report two primary metrics throughout the paper. **Accuracy** is the fraction of correctly classified samples. **Macro-averaged F1** is computed as $\text{F1}_{\text{macro}} = \frac{1}{K}\sum_{k=1}^{K} F1_k$, where $F1_k = 2P_kR_k/(P_k + R_k)$ for class k, and P_k and R_k denote per-class precision and recall respectively. For binary tasks, we additionally report the area under the ROC curve (AUC). Per-class precision, recall, and confusion matrices are provided where informative.

All experiments are repeated with three random seeds. We report the mean accuracy across seeds in all tables unless otherwise noted.

B. Dataset Details

B.1. Source Datasets

We evaluate on four question-answering benchmarks that span factual recall, scientific knowledge, long-tail entity knowledge, and adversarial truthfulness.

TriviaQA (Joshi et al., 2017) contains 95K question-answer pairs authored by trivia enthusiasts, testing factual knowledge across diverse topics. Each question admits multiple acceptable answer phrasings, making it a standard benchmark for open-domain QA. The full dataset comprises over 650K question-answer-evidence triples when including evidence documents.

SciQ (Welbl et al., 2017) contains 13,679 science examination questions targeting knowledge across physics, chemistry, biology, and earth sciences. Questions require domain-specific understanding and are split into 11,679 training, 1,000 validation, and 1,000 test examples.

PopQA (Mallen et al., 2023b) consists of 14,267 factual questions about long-tail entities drawn from Wikidata. This dataset tests the model's ability to recall knowledge about less popular subjects, where parametric memory is weaker and hallucination rates are correspondingly higher.

TruthfulQA (Lin et al., 2022a) comprises 817 questions spanning 38 categories including health, law, finance, and politics. It is specifically designed to elicit common misconceptions and plausible-sounding but false answers, serving as the most adversarial benchmark in our suite.

B.2. Automated Evaluation Protocol

We employ an LLM judge following a simple evaluation methodology (Zheng et al., 2023). For each sample, the judge receives three inputs: the original question, the ground-truth answer, and the model's generated response. Based on these, the judge assigns one of three labels.

- **Correct.** The response is semantically equivalent to the ground truth. Surface-level differences in phrasing are tolerated.
- **Incorrect.** The response is factually wrong, contradicts the ground truth, or contains hallucinated information.
- **Did Not Answer (DNA).** The model explicitly refuses to answer, expresses uncertainty without committing to a response, or provides no substantive content.

The evaluation prompt provides clear criteria and worked examples for each label. To assess annotation reliability, we manually reviewed 500 randomly sampled judgments and observed near-perfect agreement with the automated labels.

C. Full Per-Model Results

Table 5 reports accuracy, AUC, and F1 for all three tasks on the TriviaQA dataset. Within every model family, performance scales monotonically with parameter count across all tasks. The Response task achieves near-perfect accuracy (94–99%), while the Correctness task ranges from 73–78% across models. The Full (three-way) task tracks the Correctness task closely, with accuracy only 1–2 points lower, confirming that the additional DNA class incurs negligible error given the near-perfect separability of the abstention category.

D. Full Baseline Comparison

Table 6 reports both classification accuracy and AUC for all baseline methods and Grad Detect on the Correctness task (TriviaQA).

E. Extended Baseline Analysis

Self-assessment via prompting. The self-assessment baseline achieves only 49–54% accuracy across all models, performing worse than all other methods including simple perplexity thresholding. Models consistently overestimate their ability to answer correctly, producing affirmative responses for approximately 78% of queries regardless of actual correctness. This

Family	Size	Response			Correctness			Full (3-way)		
		Acc	AUC	F1	Acc	AUC	F1	Acc	AUC	F1
Qwen2.5	1.5B	96.8	.98	0.97	75.0	.82	0.74	73.8	.80	0.73
	3B	98.2	.99	0.98	76.2	.84	0.75	75.1	.82	0.74
	7B	99.1	.99	0.99	76.8	.86	0.76	75.9	.84	0.75
Falcon3	1B	96.4	.98	0.96	74.3	.81	0.73	73.0	.79	0.72
	3B	98.0	.99	0.98	75.5	.83	0.75	74.4	.81	0.73
	7B	99.0	.99	0.99	76.9	.85	0.76	75.8	.83	0.75
	10B	99.3	.99	0.99	77.4	.86	0.77	76.3	.84	0.75
Gemma-3	1B	96.6	.98	0.97	74.6	.81	0.74	73.3	.79	0.72
	4B	98.5	.99	0.98	76.5	.85	0.76	75.4	.83	0.74
	12B	99.5	.99	0.99	77.8	.86	0.77	76.8	.85	0.76
SmolLM3	3B	96.2	.98	0.96	73.2	.81	0.72	72.1	.79	0.71

Table 5. Per-model performance on TriviaQA dataset. Correctness and Response accuracy and AUC values match Tables 1 and 2 exactly. Within every family, accuracy scales monotonically with parameter count across all three tasks.

Table 6. Full baseline comparison on the Correctness task (TriviaQA). We report classification accuracy (%). Internal State Probing (Ji et al., 2024); Self-Consistency (Manakul et al., 2023); Semantic Entropy (Kuhn et al., 2023). Best result per model is **bolded**.

	Qwen2.5			**Falcon3**				**Gemma-3**			**SmolLM3**
Method	1.5B	3B	7B	1B	3B	7B	10B	1B	4B	12B	3B
Self-Assessment	50.2	52.1	53.7	49.8	50.6	51.4	52.3	49.5	51.8	54.1	49.1
Sequence Perplexity	54.8	56.1	57.2	55.3	57.6	59.3	59.8	54.5	56.8	58.4	53.1
Confidence	68.3	70.5	72.1	67.1	69.2	71.0	72.4	67.8	70.9	73.2	67.5
Internal State Probing	66.2	68.4	70.5	64.8	66.7	68.1	69.5	65.5	69.1	71.3	65.4
Self-Consistency	58.6	60.8	62.4	57.2	58.9	60.1	61.3	58.1	61.2	63.5	57.8
Semantic Entropy	69.7	71.8	73.6	68.5	70.6	72.4	73.8	69.1	72.3	74.5	68.8
Grad Detect (last 5)	74.2	75.1	75.7	73.5	74.8	76.5	76.1	73.8	75.4	76.8	72.1
Grad Detect (all)	**75.0**	**76.2**	**76.8**	**74.3**	**75.5**	**76.9**	**77.4**	**74.6**	**76.5**	**77.8**	**73.2**

confirms that LLMs lack reliable introspective access to their own knowledge boundaries (Kadavath et al., 2022) and that external analysis of internal computation is necessary for reliable detection.

Advantage over confidence baselines. The consistent advantage over confidence baselines confirms that gradients expose decision-relevant internal structure not visible in the output distribution. A model can produce high-confidence hallucinations whose softmax scores are indistinguishable from those of correct answers (Guo et al., 2017). The gradients of such samples, however, point in measurably different directions, as confirmed by the per-layer divergence analysis in Section 4.4. This finding is consistent with observations that internal model states encode truthfulness information beyond what output probabilities capture (Azaria & Mitchell, 2023; Su et al., 2024).

Reference gradient sensitivity. Our default reference gradients are computed by averaging over all samples in a category (Section 3.2). We evaluated several alternatives. Restricting to high-confidence samples degrades accuracy by 1–2 points. Restricting to uncertain samples degrades it by 6 points, as noisy gradients corrupt the prototype. A stratified ensemble that maintains separate references per confidence bin yields a marginal gain of 0.6 points at added complexity. We retain simple averaging as the default. The same pattern holds for the Response task, where using all abstention samples outperforms restricting to explicit refusals or hedged responses.

F. Extended Layer Ablation

F.1. Layer Range Analysis

Table 7 reports Correctness-task accuracy on TriviaQA for seven layer configurations across all eleven models. Trends are consistent across other datasets.

Table 7. Layer ablation (TriviaQA, Correctness task) across all models. We report classification accuracy (%). Performance degrades gracefully when removing layers from either end, with a maximum drop of 2–3 points for any contiguous subset covering at least one-third of the network.

	Qwen2.5			Falcon3				Gemma-3			SmolLM3
Layer Range	1.5B	3B	7B	1B	3B	7B	10B	1B	4B	12B	3B
All layers	75.0	76.2	76.8	74.3	75.5	76.9	77.4	74.6	76.5	77.8	73.2
Last 20	74.8	76.0	76.6	74.1	75.3	76.7	77.2	74.4	76.3	77.6	73.0
Last 10	74.5	75.7	76.3	73.8	75.0	76.4	76.9	74.1	76.0	77.3	72.7
Last 5	74.2	75.1	75.7	73.5	74.8	76.5	76.1	73.8	75.4	76.8	72.1
First 5	72.8	74.0	74.6	72.1	73.3	74.8	75.3	72.5	74.3	75.6	71.0
First 10	73.4	74.6	75.2	72.7	73.9	75.4	75.9	73.1	74.9	76.2	71.6
First 20	74.3	75.5	76.1	73.6	74.8	76.2	76.7	73.9	75.8	77.1	72.5

Discriminative information is distributed but concentrates in later layers. Across all eleven models, the last 5 layers retain 98–99% of full-model accuracy, dropping only 0.5–1.0 points on average. The last 10 and last 20 layers lose even less (0.4–0.7 and 0.2–0.3 points respectively). Conversely, the first 5 layers alone still achieve 96–97% of full accuracy, indicating that even early layers carry a meaningful gradient signal. Adding more early layers improves performance monotonically. The first 10 layers recover 97–98% and the first 20 layers recover 98–99% of full accuracy. The maximum drop for any configuration covering at least one-third of the network is 2–3 points, confirming that the discriminative gradient signal is broadly distributed across the transformer. Nevertheless, later layers consistently outperform earlier layers of the same width. For example, last 5 vs. first 5 differs by 1.2–1.4 points. This aligns with the established view that later transformer layers handle high-level semantic integration (Tenney et al., 2019; Jawahar et al., 2019) where factual grounding succeeds or fails.

Layers are complementary. Although any contiguous subset of layers achieves strong accuracy, combining layers from different regions of the network yields consistent gains. Self-attention in the predictor exploits cross-layer patterns that no single region captures in isolation. This is consistent with circuit-level analyses showing that transformer computations emerge from interactions across multiple layers (Elhage et al., 2021).

G. Inference Time Breakdown

Table 8 reports end-to-end inference time per sample for all detection methods, measured on Qwen2.5-3B with a single NVIDIA A100 (40 GB).

Grad Detect with all layers adds 47 ms to the standard forward pass (45 ms), for a total of 92 ms ($2.04\times$). This overhead has three components: the backward pass (38 ms), cosine similarity computation against reference gradients (3 ms), and the predictor's forward pass (6 ms). The backward pass dominates because it requires propagating gradients through the full model.

Using only the last 5 layers reduces total time to 68 ms by computing gradients for a subset of parameters. This is substantially cheaper than Self-Consistency and Semantic Entropy, which both require multiple full forward passes through the LLM.

Confidence baselines add negligible overhead (2 ms) because they require only reading out statistics from the existing forward pass output distribution.

H. Gradient Signature Visualization

We apply Linear Discriminant Analysis (LDA) to the layer-wise cosine similarity vectors to visualize the separability of gradient signatures across behavioral categories. For each model, every point in the resulting projection represents a single sample, positioned according to its cosine similarity profile across layers and reference gradients. Clear separation between category clusters provides visual confirmation that gradient directions carry discriminative information exploitable by the transformer-based predictor.

Visualizations are organized along three dimensions. **Task** determines the classification objective: Full (three-way),

Method	Time (ms)
Standard forward pass	45
+ Confidence	47
+ Grad Detect (last 5 layers)	68
+ Grad Detect (all layers)	92
+ Self-Consistency (5 gens)	225
+ Semantic Entropy (10 gens)	450

Table 8. End-to-end inference time per sample (Qwen2.5-3B, A100). Grad Detect overhead comprises: backward pass (38 ms), cosine similarity (3 ms), and predictor forward pass (6 ms).

Correctness (binary), or Response (binary). **Reference class** indicates which category prototype (Correct, Incorrect, or Did Not Answer) serves as the comparison anchor in the cosine similarity computation. **Response type** specifies which probe response (affirming r^+ or rejection r^-, as defined in Section 3.1) is used to compute gradients for the reference samples and training samples respectively. All plots show results for Qwen2.5-3B, Falcon3-3B, Gemma3-4B, and Gemma3-1B.

H.1. Full Task (Correct vs. Incorrect vs. Did Not Answer)

Figures 3–5 show LDA projections for the three-way Full task, grouped by reference class. Across all response-type combinations, the three behavioral categories form distinct clusters. Separation is strongest in the affirming/affirming setting (subfigure (a) in each figure), where both reference samples and training samples come from substantive responses. The rejection/rejection setting (subfigure (d)) also shows clear separation, confirming that gradient signatures remain discriminative regardless of whether the model answered or abstained.

H.2. Correctness Task (Correct vs. Incorrect)

Figures 6 and 7 show LDA projections for the binary Correctness task. Although Did Not Answer samples are excluded from classification, the cosine similarity features are still computed against all three reference gradients. The separation between Correct and Incorrect clusters is tighter than in the Full task, reflecting the greater difficulty of this binary distinction.

H.3. Response Task (Answered vs. Did Not Answer)

Figures 8 and 9 show LDA projections for the binary Response task. Consistent with the near-perfect classification accuracy reported in Table 2, the Answered and Did Not Answer clusters are widely separated across all reference classes and response-type combinations. The separation is notably larger than that observed in the Correctness task (Figures 6 and 7), visually confirming that abstention produces a qualitatively distinct gradient signature.

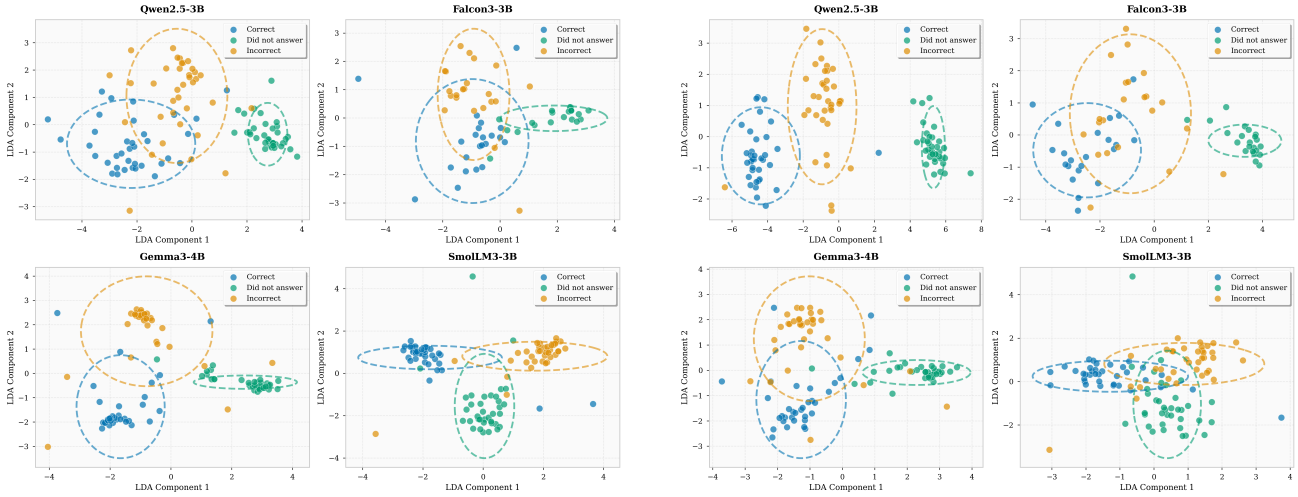
H.4. Summary of Visualization Observations

Several patterns are consistent across all visualizations.

Abstention is always well-separated. In every task and response-type combination, the Did Not Answer cluster occupies a distinct region of the LDA projection. This visual separation corresponds directly to the 94–99% accuracy observed on the Response task (Table 2).

Correct vs. Incorrect separation is tighter. The Correct and Incorrect clusters overlap more than either does with the DNA cluster, consistent with our finding that the Correctness distinction is the primary bottleneck in three-way classification. The overlap is greatest for the rejection response type, where the model's abstention behavior reduces the diversity of gradient patterns available for distinguishing correctness.

Affirming references produce cleaner projections. Plots using affirming reference samples (subfigures (a) and (b) in each figure) generally show tighter, better-separated clusters than those using rejection samples. A single configuration (e.g., *Correct* reference with affirming/affirming probes) is sufficient for deployment, as predictors trained on each of the twelve configurations achieve accuracy within a 2-point range.



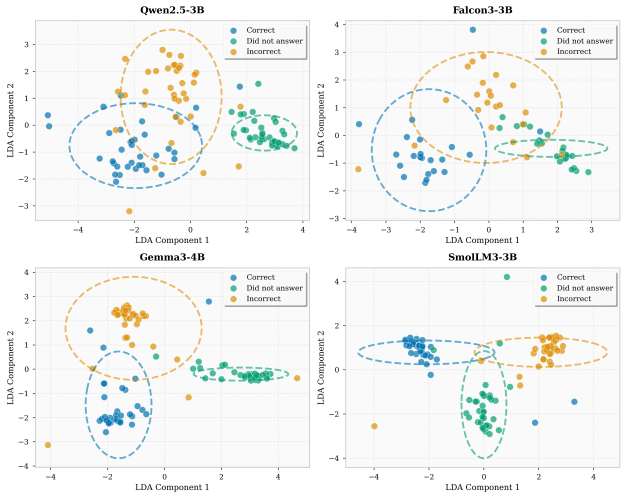
(a) Ref: affirming, Train: affirming

(b) Ref: affirming, Train: rejection

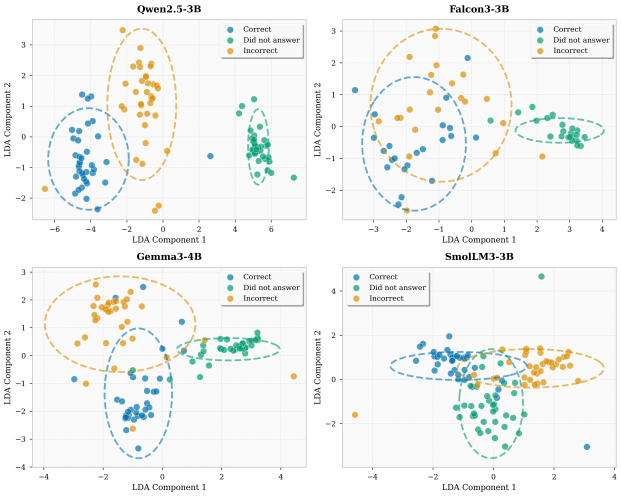
(c) Ref: rejection, Train: affirming

(d) Ref: rejection, Train: rejection

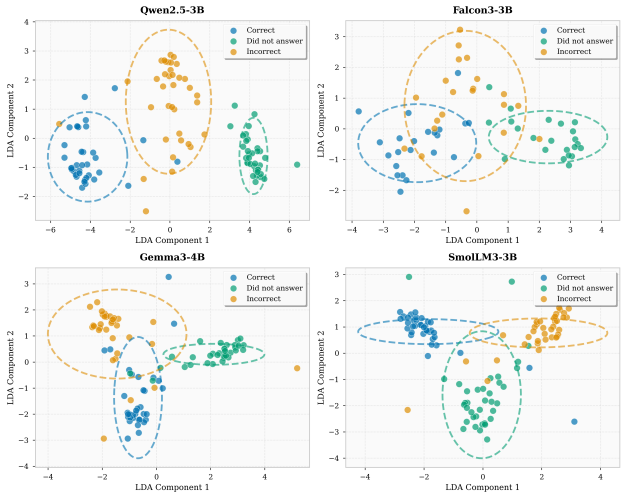
Figure 3. Full task — **Correct** reference gradient. LDA projections of layer-wise cosine similarities for all four combinations of reference and training response types.



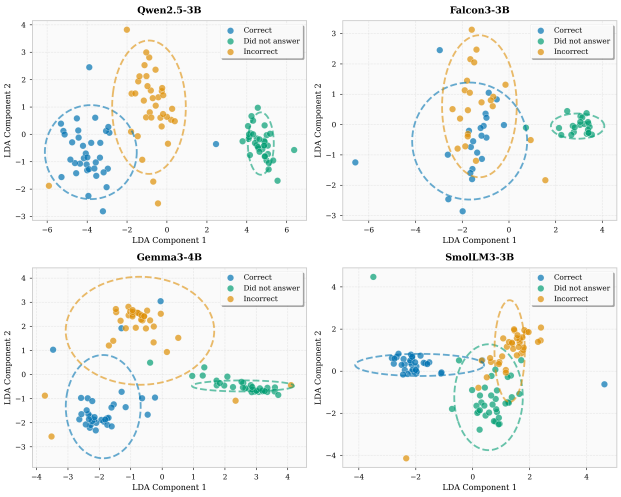
(a) Ref: affirming, Train: affirming



(b) Ref: affirming, Train: rejection

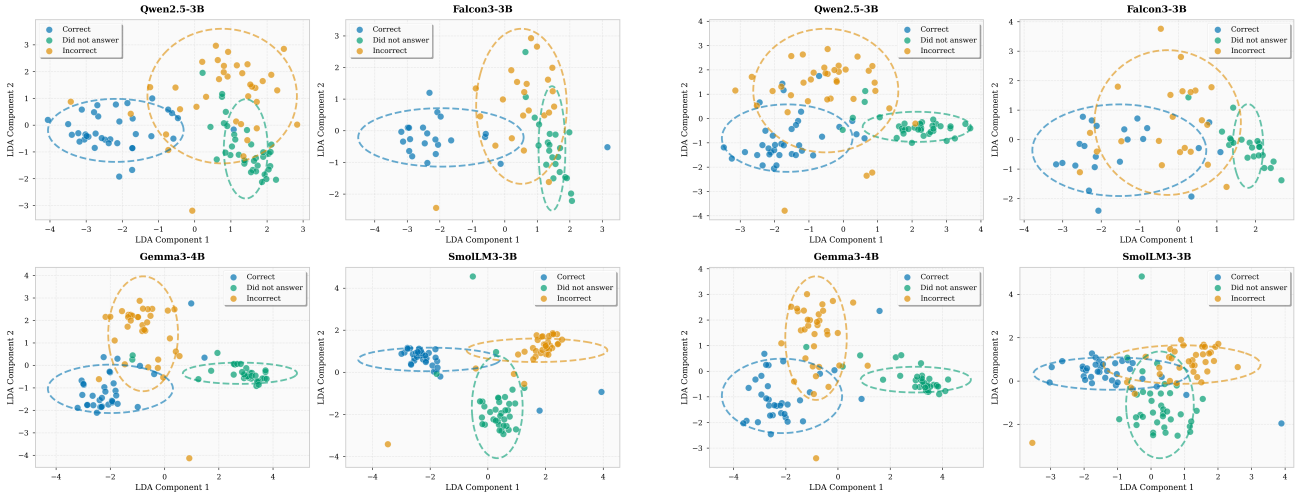


(c) Ref: rejection, Train: affirming



(d) Ref: rejection, Train: rejection

Figure 4. Full task — **Incorrect** reference gradient. LDA projections of layer-wise cosine similarities for all four combinations of reference and training response types.



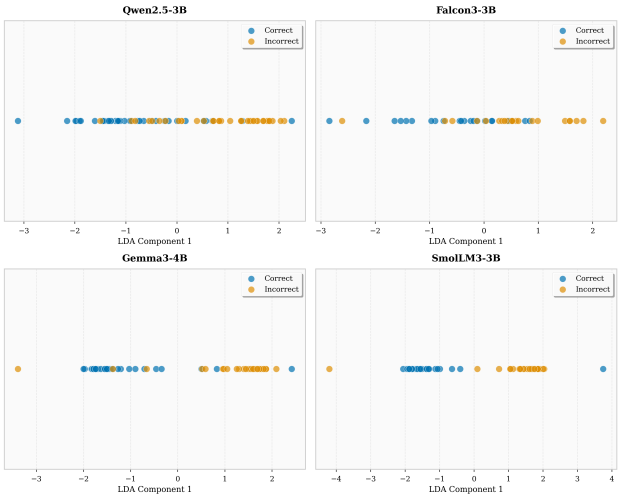
(a) Ref: affirming, Train: affirming

(b) Ref: affirming, Train: rejection

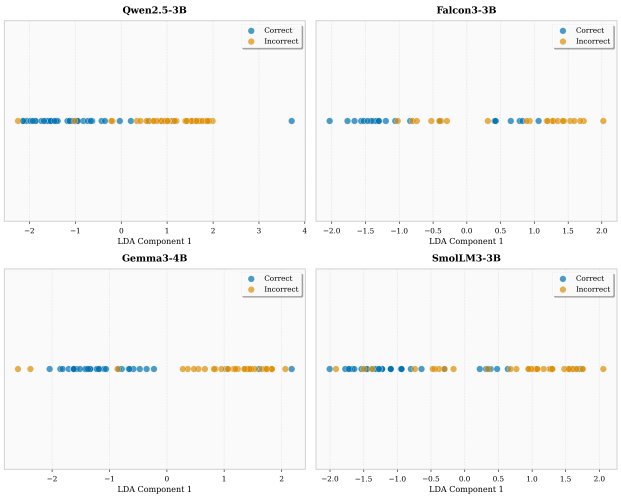
(c) Ref: rejection, Train: affirming

(d) Ref: rejection, Train: rejection

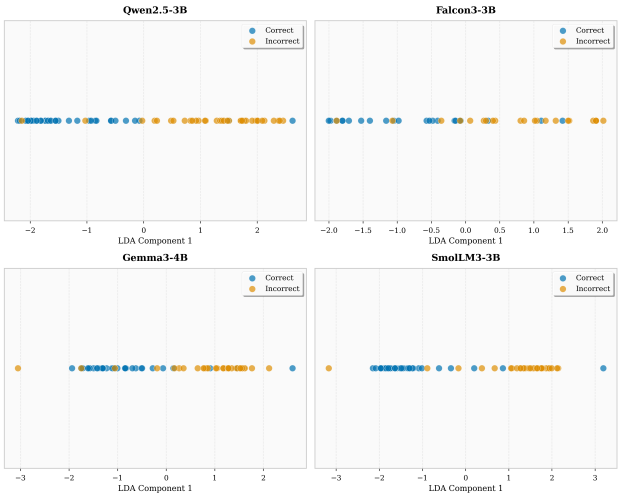
Figure 5. Full task — **Did Not Answer** reference gradient. LDA projections of layer-wise cosine similarities for all four combinations of reference and training response types.



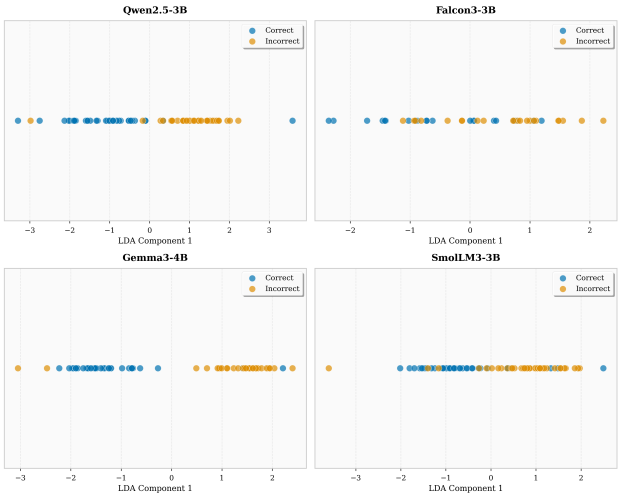
(a) Ref: affirming, Train: affirming



(b) Ref: affirming, Train: rejection

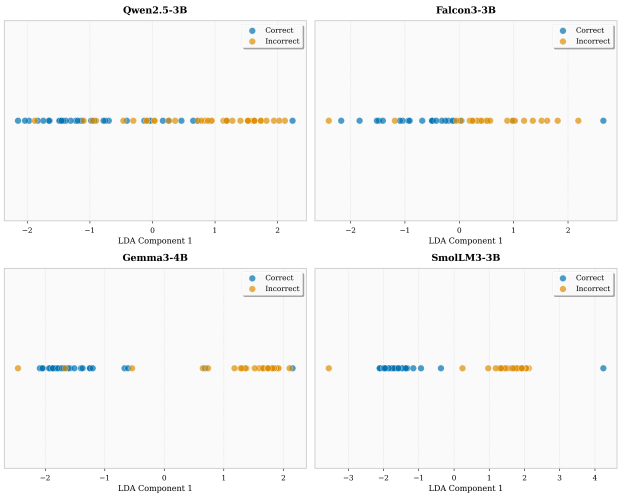


(c) Ref: rejection, Train: affirming

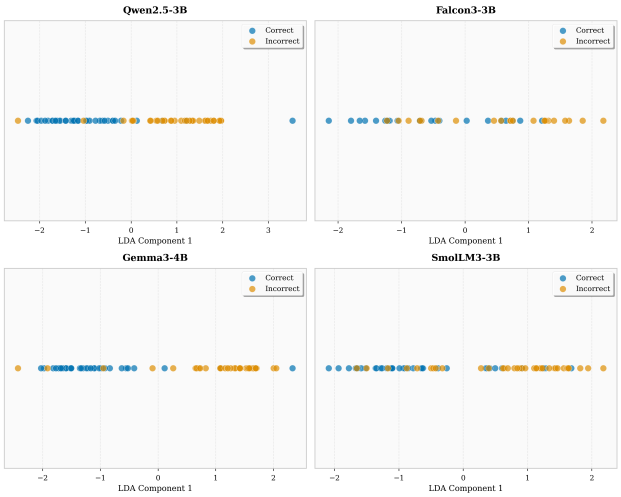


(d) Ref: rejection, Train: rejection

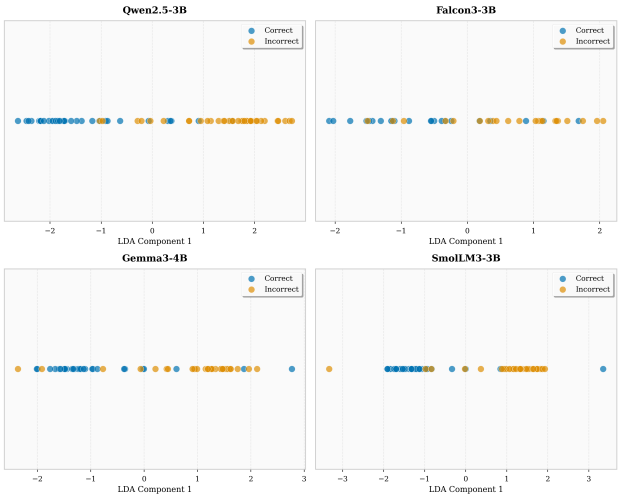
Figure 6. Correctness task — **Correct** reference gradient. LDA projections of layer-wise cosine similarities.



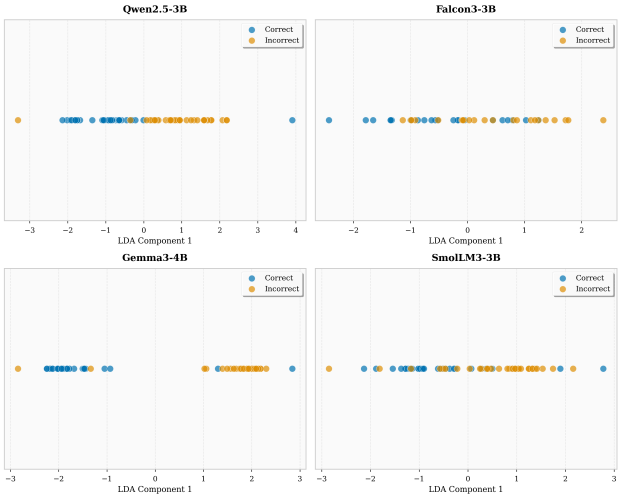
(a) Ref: affirming, Train: affirming



(b) Ref: affirming, Train: rejection

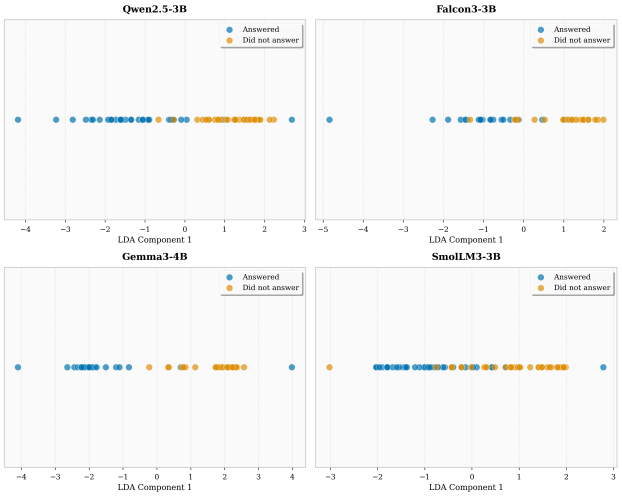


(c) Ref: rejection, Train: affirming

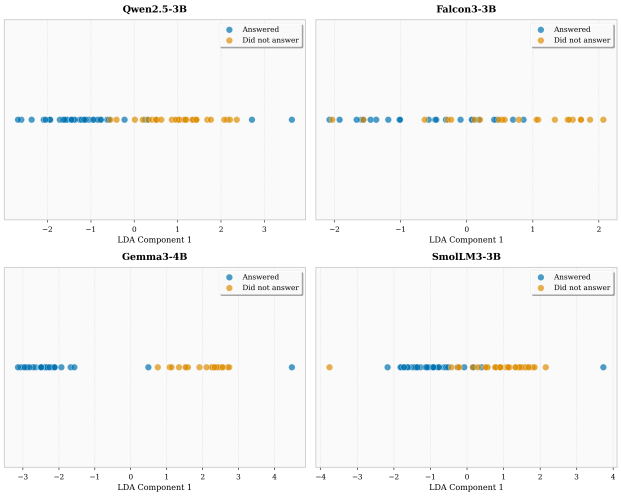


(d) Ref: rejection, Train: rejection

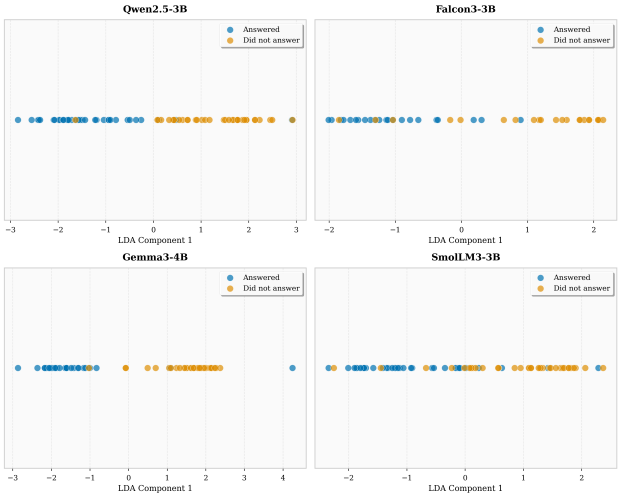
Figure 7. Correctness task — **Incorrect** reference gradient. LDA projections of layer-wise cosine similarities.



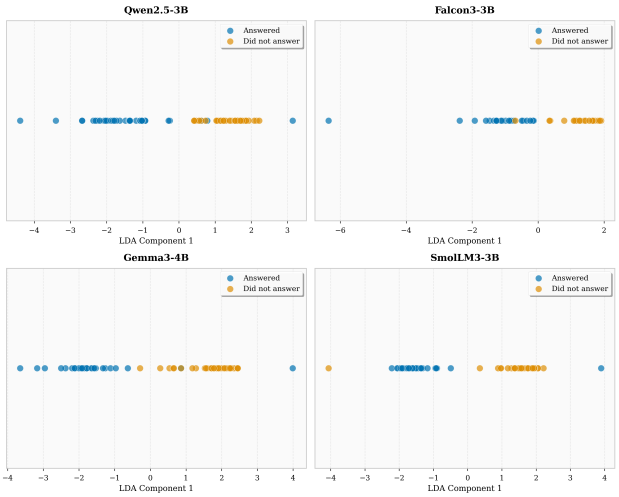
(a) Ref: affirming, Train: affirming



(b) Ref: affirming, Train: rejection

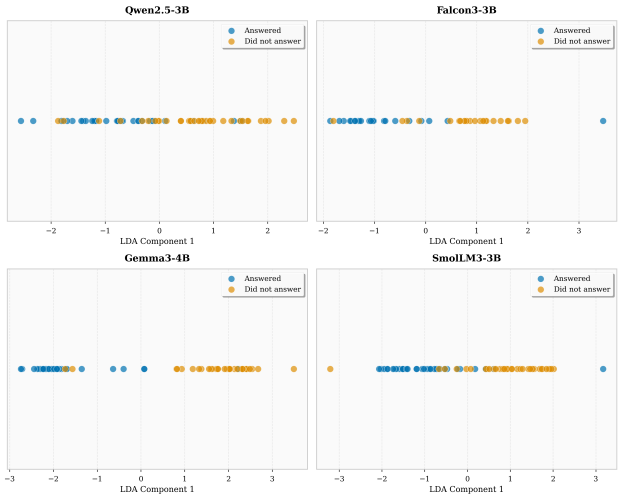


(c) Ref: rejection, Train: affirming

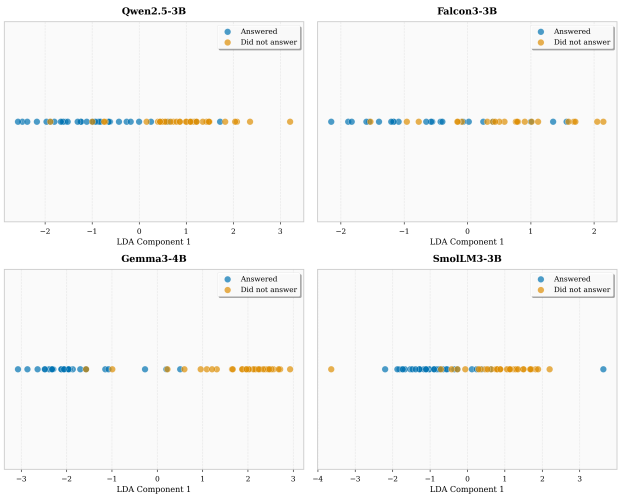


(d) Ref: rejection, Train: rejection

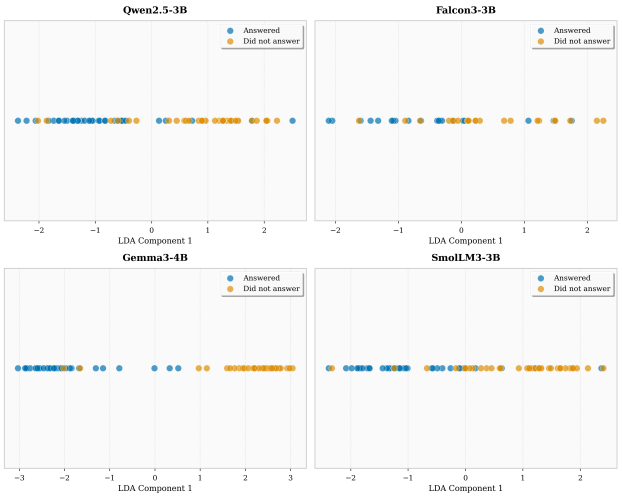
Figure 8. Response task — **Answered** reference gradient. LDA projections of layer-wise cosine similarities. Wide cluster separation reflects the near-perfect accuracy of abstention detection.



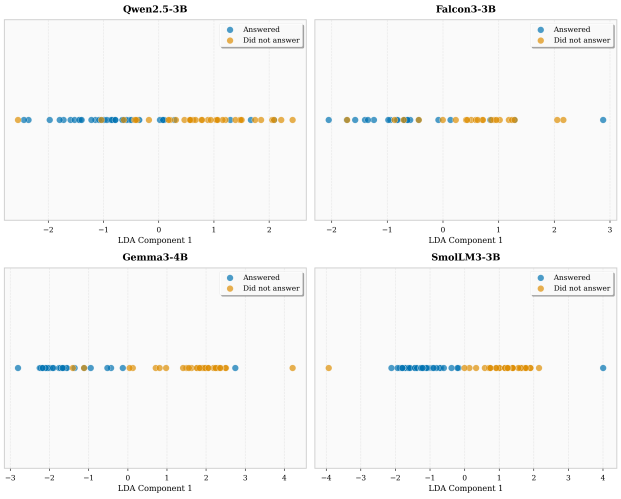
(a) Ref: affirming, Train: affirming



(b) Ref: affirming, Train: rejection



(c) Ref: rejection, Train: affirming



(d) Ref: rejection, Train: rejection

Figure 9. Response task — **Did Not Answer** reference gradient. LDA projections of layer-wise cosine similarities. The DNA reference produces the widest separation, confirming that abstention is the most distinctive behavioral mode in gradient space.

Patterns are consistent across model families. All four models show qualitatively similar cluster structures within each plot, confirming the cross-family consistency of gradient-based behavioral signatures reported in Section 4.2.

All twelve configurations yield comparable separation. Across the $|\mathcal{C}| \times 2 \times 2 = 12$ combinations of reference category, reference probe response, and training probe response, the LDA projections show qualitatively similar cluster structures. Predictors trained on each of the twelve configurations achieve accuracy within a 2-point range, confirming that the discriminative gradient signal is robust to the choice of probe combination. This invariance simplifies deployment: a single configuration (e.g., *Correct* reference, affirming/affirming probes) is