GRADIENTS KNOW BEST: GRADIENT-BASED HALLUCINATION AND ABSTENTION DETECTION IN LLMS

Anonymous authors
Paper under double-blind review

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 signal is distributed across the network: any five layers retain about 97% of full accuracy, and restricting gradients to the final five costs only 0.4–1.3 points, enabling efficient deployment. 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

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.

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 Answer*) 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 a framework for LLM hallucination detection that compares layer-wise gradients with behavior-specific reference gradients. It outperforms the Confidence baseline by 4.6–7.2 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 any five layers retain about 97% of full accuracy, and that restricting gradients to the final five costs only 0.4–1.3 points at $1.5\times$ inference cost.
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., 2021; Hosseini et al., 2024), while retrieval-augmented approaches (Lewis et al., 2020; Peng et al., 2023; Gao et al., 2023; Mallen et al., 2023) 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

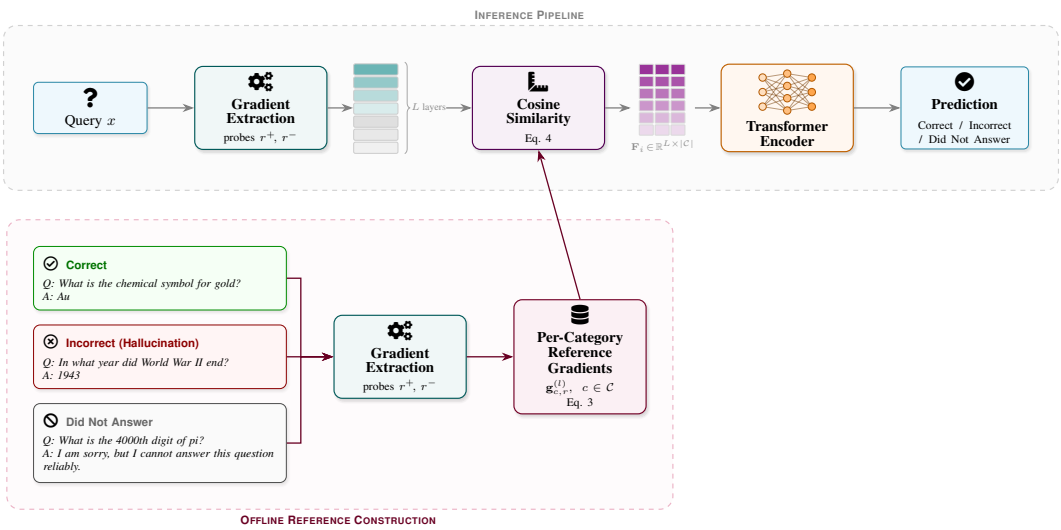


Figure 1: Overview of the Grad Detect pipeline. At inference time (**top**), per-layer gradients extracted from probe responses are compared against precomputed reference gradients via cosine similarity, producing a compact feature matrix that a lightweight transformer encoder classifies as Correct,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

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. (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. GradSafe (Xie et al., 2024) also pairs a prompt with the response "Sure", but it screens prompts for jailbreak *intent*, a binary property of the input that says nothing about whether the model's knowledge supports an answer. Its reference comes from two unsafe and two safe prompts over hand-filtered parameter slices, and it scores a prompt by one averaged similarity against a fixed threshold, discarding where in the network the signal arises. Grad Detect instead builds per-class references from labeled generations, keeps the full layer profile, and learns cross-layer interactions to predict both correctness and abstention. Gradient-based hallucination detectors score scalar statistics of the gradient, such as a Taylor-expanded conditional-output distance (Hu et al., 2024), sensitivity to embedding perturbations (Liew et al., 2026), or, concurrently, weight-update geometry (Zhang et al., 2026), rather than comparing it with behavior-specific prototypes.

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

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

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

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

and within each layer restrict attention to the **MLP down-projection weights**. 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

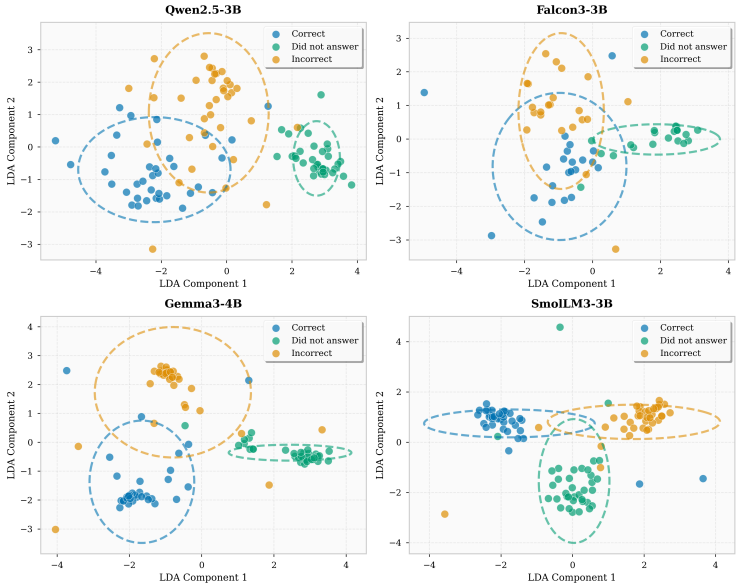


Figure 2: LDA projections of averaged layer-wise cosine similarity vectors for Qwen2.5-3B, Falcon3-3B, Gemma3-4B, and SmolLM3-3B. Clear separation between behavioral categories confirms that gradient directions carry discriminative information.

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. Our main results fit the references and the predictor separately for each model and dataset. Appendix F shows that artifacts fitted on one factual Q&A dataset can be reused on another at a modest cost in accuracy.

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 on a validation split and used 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., 2023) 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), 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). The judge never assesses factuality from its own knowledge; it only checks the response against the benchmark's reference answer. 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. The reference and probe configuration for each model and dataset is selected on a held-out validation split. 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).

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 reference and probe configuration selected on the validation split, 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

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 visualizes this through LDA projections of the average layer-wise cosine similarity vectors. All models separate abstentions from answers, while the overlap between correct and incorrect responses varies across families.

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 distinctive gradient signatures that separate from those of answered responses far more cleanly than correct responses separate from incorrect ones.

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.

Table 3: Baseline comparison on the Correctness task (TriviaQA). We report area under the ROC curve (AUC). Internal State Probing (Ji et al., 2024); Self-Consistency (Wang et al., 2023); Semantic Entropy (Kuhn et al., 2023). Grad Detect uses the configuration selected on the validation split.

	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	.71	.74	.76	.69	.72	.71	.75	.70	.74	.77	.71
Self-Consistency	.62	.65	.67	.61	.62	.63	.65	.61	.65	.68	.62
Semantic Entropy	.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**

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 (Wang 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 4.6–25.5 points in accuracy and 0.07–0.31 in AUC across all eleven models. It also outperforms the strongest multi-generation method, Semantic Entropy, by 3.2–5.8 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 2.1–5.0 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 0.81–0.82 with Grad Detect, compared to 0.74–0.76 for Semantic Entropy, the strongest baseline at that 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 Self-Assessment, Sequence Perplexity, and Self-Consistency, confirming that internal representations carry a truthfulness signal, although it trails the Confidence baseline on every model. Grad Detect surpasses Internal State Probing by 6.3–9.5 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. Appendix E further analyzes the advantage over confidence baselines and the failure of self-assessment prompting.

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 0.4–1.3 points (0.9 on average). Even the first 5 layers alone achieve 97% of full accuracy. No evaluated range covering at least one-third of the network loses more than 1.6 points. A full layer-range analysis across all models, per-layer accuracy curves, and efficiency-accuracy trade-offs are provided in Appendix G.

4.5 Analysis

Alternatives to averaging for building the reference gradients are compared 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., 2023) 73.0–77.5%, and TruthfulQA (Lin et al., 2022a) 72.9–77.4%. The gradient signal is therefore present at comparable strength across factual recall, long-tail knowledge, and adversarial truthfulness. Each figure is obtained with references and a predictor fitted on that same dataset, so comparable in-domain accuracy does not by itself establish that the setup transfers; we measure transfer directly below. 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. The same ordering of datasets holds across all four model families.

Cross-dataset transfer. To test whether the setup itself generalizes, we fit the references and the predictor on one dataset and evaluate on the remaining three using no target labels. Across all eleven models, this transfer lowers accuracy by 2.9–3.8 points on average and by 4.2–5.4 points in the worst case, and the cost shrinks with model size within every family. SciQ is the hardest target for every model. On TriviaQA, the mean transferred accuracy still exceeds the Confidence, Internal State Probing, and Semantic Entropy baselines for every model. A single setup can therefore serve several factual Q&A datasets. Transfer results for all eleven models are reported in Table 7 of Appendix F.2. Transfer across *models* is not supported by the present architecture, for reasons given in Appendix F.3.

Probe combination invariance. Predictors trained separately on each of the twelve reference and probe configurations of Section 3.3 achieve accuracy within 2 percentage points of one another (Appendix I). The discriminative gradient signal is therefore a property of the query-model interaction rather than of a particular probe choice, and 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 gradients to the last five layers reduces this to $1.5\times$ while retaining 98–99% of full accuracy. Self-Consistency and Semantic Entropy require $5\times$ and $10\times$ the cost respectively. Efficient internal-state probes (Kossen et al., 2024) reach near-zero overhead by approximating semantic entropy from hidden states, but operate on activations rather than gradients. Appendix H gives a detailed timing breakdown.

5 Conclusion

We presented Grad Detect, which detects hallucinations from the cosine similarity between layer-wise gradients and category-specific reference gradients. Across eleven models from four families and four Q&A benchmarks, it reaches 71–78% accuracy on hallucination detection and 94–99% on abstention prediction, outperforming the Confidence baseline by 4.6–7.2 points and sampling-based methods at a fraction of their cost. Restricting gradients to the final five layers costs 0.4–1.3 points and allows deployment at $1.5\times$ inference cost, and the same features predict both correctness and abstention, two tasks prior work has addressed separately (Kadavath et al., 2022; Lin et al., 2022b; Geifman & El-Yaniv, 2017).

Future directions. References and predictors transfer across factual Q&A datasets at a cost of 1–5 points (Appendix F), but each new model needs its own setup (Appendix F.3); indexing layers by relative depth could lift this requirement. Distilling the predictor into a black-box method, as Kossen et al. (2024) did for semantic entropy, would reach API-only deployments. Open questions include mixture-of-experts models (Zhang et al., 2025a), the causal link between gradient patterns and hallucinations (Elhage et al., 2021; Geva et al., 2022), and using gradient signals to trigger retrieval (Lewis et al., 2020; Mallen et al., 2023). Summarization (Hermann et al., 2015; Narayan et al., 2018), agentic (Liu et al., 2023a; Jimenez et al., 2024), reasoning-intensive (Cobbe et al., 2021), long-context, and multi-turn settings would test whether the method extends to hallucinations caused by flawed multi-step computation rather than missing knowledge.

AI USE STATEMENT

The original manuscript was human authored. We used large language models only for editing: to shorten or lengthen passages to fit the page limit and to rewrite existing sections for clarity. We did not use generative AI tools for research ideation, method design, implementation, running experiments, or analyzing results. The authors reviewed and revised all AI-edited text. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We share the complete codebase as supplementary material to reproduce all results in this paper, including gradient extraction, predictor training, and evaluation. Section 3 specifies every stage of the method. Appendix A gives the generation settings, the gradient extraction procedure, its computational cost, and all predictor hyperparameters (Table 4), which are held fixed across experiments. Appendix B describes the four public benchmarks and the labeling protocol, and Appendix F.2 defines the cross-dataset transfer protocol. All models evaluated are publicly available checkpoints.

REFERENCES

Amos Azaria and Tom Mitchell. 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.

Elie Bakouch, Loubna Ben Allal, Anton Lozhkov, Nouamane Tazi, Lewis Tunstall, Carlos Miguel Patiño, Edward Beeching, Aymeric Roucher, Aksel Joonas Reedi, Quentin Gallouédec, Kashif Rasul, Nathan Habib, Clémentine Fourrier, Hynek Kydlicek, Guilherme Penedo, Hugo Larcher, Mathieu Morlon, Vaibhav Srivastav, Joshua Lochner, Xuan-Son Nguyen, Colin Raffel, Leandro von Werra, and Thomas Wolf. SmolLM3: smol, multilingual, long-context reasoner. https://huggingface.co/blog/smollm3, 2025.

David Balduzzi, Marcus Frean, Lennox Leary, JP Lewis, Kurt Wan-Duo Ma, and Brian McWilliams. 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.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, 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.

C. K. Chow. 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.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, 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.

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*, 2021. doi: 10.48550/arXiv.2110.14168. URL http://arxiv.org/abs/2110.14168.

Tri Dao, Dan Fu, Stefano Ermon, Atri Rudra, and Christopher R'e. 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.

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

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, et al. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021. URL https://transformer-circuits.pub/2021/framework.

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

Javier Ferrando, Gerard I Gállego, Ioannis Tsiamas, and Marta R Costa-jussà. 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/.

Luyu Gao, Zhuyun Dai, Panupong Pasupat, Anthony Chen, Arun Tejasvi Chaganty, Yicheng Fan, Vincent Y Zhao, Ni Lao, Hongrae Lee, Da-Cheng Juan, and Kelvin Guu. 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/.

Yonatan Geifman and Ran El-Yaniv. 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, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025. URL https://arxiv.org/abs/2503.19786.

Mor Geva, Avi Caciularu, Kevin Wang, and Yoav Goldberg. 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.

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

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 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.

Karl Moritz Hermann, Tomáš Kočiský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems*, volume 28, 2015.

Ruixin Hong, Hongming Zhang, Xinyu Pang, Dong Yu, and Changshui Zhang. 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.

Arian Hosseini, Xingdi Yuan, Nikolay Malkin, Aaron Courville, Alessandro Sordoni, and Rishabh Agarwal. V-star: Training verifiers for self-taught reasoners. *arXiv preprint*, 2024. doi: 10.48550/arXiv.2402.06457. URL http://arxiv.org/abs/2402.06457.

Xiaomeng Hu, Yiming Zhang, Ru Peng, Haozhe Zhang, Chenwei Wu, Gang Chen, and Junbo Zhao. Embedding and gradient say wrong: A white-box method for hallucination detection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 1950–1959, 2024. URL https://aclanthology.org/2024.emnlp-main.116/.

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*, 2024. doi: 10.1145/3703155. URL http://arxiv.org/abs/2311.05232.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 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/.

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, 2023. URL https://doi.org/10.1145/3571730.

Ziwei Ji, Delong Chen, Etsuko Ishii, Samuel Cahyawijaya, Yejin Bang, Bryan Wilie, and Pascale Fung. Llm internal states reveal hallucination risk faced with a query. *arXiv preprint arXiv:2407.03282*, 2024. URL https://arxiv.org/abs/2407.03282.

Carlos E Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik R Narasimhan. 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.

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. 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/.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022. URL https://arxiv.org/abs/2207.05221.

Adam Tauman Kalai, Ofir Nachum, Santosh S. Vempala, and Edwin Zhang. Why language models hallucinate. *arXiv preprint*, 2025. doi: 10.48550/arXiv.2509.04664. URL http://arxiv.org/abs/2509.04664.

Jannik Kossen, Jiatong Han, Muhammed Razzak, Lisa Schut, Shreshth Malik, and Yarin Gal. 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.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 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.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich K"uttler, Mike Lewis, Wen-tau Yih, Tim Rockt"aschel, 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.

Yee Zhing Liew, Andrew Huey Ping Tan, and Anwar P. P. Abdul Majeed. From flat facts to sharp hallucinations: Detecting stubborn errors via gradient sensitivity. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2605.00939.

Stephanie Lin, Jacob Hilton, and Owain Evans. 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/.

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

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 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.

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

Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, Shudan Zhang, Xiang Deng, Aohan Zeng, Zhengxiao Du, Chenhui Zhang, Sheng Shen, Tianjun Zhang, Yu Su, Jie Tang, Juntao Yao, Yankai Liu, Ruobing Li, Zhiyuan Sun, Zhiyuan Liu, Jie Tang, and Yuan Yao. AgentBench: Evaluating LLMs as agents, 2023a. URL https://arxiv.org/abs/2308.03688.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 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/.

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

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

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 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, 2023. URL https://aclanthology.org/2023.acl-long.546/.

Potsawee Manakul, Adian Liusie, and Mark JF Gales. 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/.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 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.

Baolin Peng, Michel Galley, Pengcheng He, Hao Cheng, Yujia Xie, Yu Hu, Qiuyuan Huang, Lars Liden, Zhou Yu, Weizhu Chen, 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.

Ernesto Quevedo, Jorge Yero, Rachel Koerner, Pablo Rivas, and Tomas Cerny. 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.

Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. 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.

Shibani Santurkar, Dimitris Tsipras, Andrew Ilyas, and Aleksander Madry. 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.

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

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. *arXiv preprint*, 2024. doi: 10.48550/arXiv.2403.06448. URL http://arxiv.org/abs/2403.06448.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 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.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 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/.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. URL https://arxiv.org/abs/2302.13971.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *arXiv preprint*, 2017. doi: 10.48550/arXiv.1706.03762. URL http://arxiv.org/abs/1706.03762.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=1PL1NIMMrw.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 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.

Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, Zac Kenton, Sasha Brown, Will Hawkins, Tom Stepleton, Courtney Biles, Abeba Birhane, Julia Haas, Laura Rimell, Lisa Anne Hendricks, William Isaac, Sean Legassick, Geoffrey Irving, and Iason Gabriel. 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.

Johannes Welbl, Nelson F Liu, and Matt Gardner. 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/.

Yixuan Weng, Minjun Zhu, Fei Xia, Bin Li, Shizhu He, Shengping Liu, Bin Sun, Kang Liu, and Jun Zhao. 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.

Yueqi Xie, Minghong Fang, Renjie Pi, and Neil Zhenqiang Gong. GradSafe: Detecting jailbreak prompts for LLMs via safety-critical gradient analysis. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2024. URL https://aclanthology.org/2024.acl-long.30/.

Tianci Xue, Ziqi Wang, Zhenhailong Wang, Chi Han, Pengfei Yu, and Heng Ji. 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.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. *arXiv preprint*, 2024. doi: 10.48550/arXiv.2412.15115. URL http://arxiv.org/abs/2412.15115.

Danyang Zhang, Junhao Song, Ziqian Bi, Xinyuan Song, Yingfang Yuan, Tianyang Wang, Joe Yeong, and Junfeng Hao. Mixture of experts in large language models, 2025a. URL https://arxiv.org/abs/2507.11181.

Xiaokang Zhang, Zijun Yao, Jing Zhang, Kaifeng Yun, Jifan Yu, Juanzi Li, and Jie Tang. 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.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, 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.

Zishuai Zhang, Hainan Zhang, and Zhiming Zheng. AURORA: Asymmetry and update-induced rotation for robust hallucination detection in large language models, 2026. URL https://arxiv.org/abs/2606.29545.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, 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. Measured per-sample inference cost is reported in Appendix H.

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., 2023) 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.

Two properties of this protocol keep the labeling reliable. First, the judge is never asked to assess factuality from its own knowledge. Every question in all four benchmarks comes with a gold reference answer, together with the accepted answer aliases where the benchmark provides them, and these are given to the judge. The task therefore reduces to deciding whether the response matches a known target, a far more constrained decision than the open-ended quality judgments for which LLM-judge biases have been reported (Zheng et al., 2023; Liu et al., 2023b). Second, the resulting labels are checked directly against human annotation on the 500-sample audit above rather than assumed to be correct.

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.

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.

D Full Baseline Comparison

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

Table 6: Full baseline comparison on the Correctness task (TriviaQA). We report classification accuracy (%). Internal State Probing (Ji et al., 2024); Self-Consistency (Wang 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**

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 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 Generalization and Transfer

F.1 What Requires Setup

Grad Detect requires two fitted artifacts: the category reference gradients of Eq. 3 and the predictor of Section 3.4. The in-domain results of Section 4.2 fit both separately for every model and dataset. Setting up a new model does not require human annotation. It requires a set of questions with known answers, one greedy generation pass to obtain the model's responses, one judge pass to assign categories (Appendix B), and one gradient extraction pass (Appendix A). For a public benchmark the labels come with the data, and for a proprietary domain the only manual step is assembling a question-answer list.

F.2 Cross-Dataset Transfer

We test whether a setup fitted on one factual Q&A dataset can be reused on another. Both the reference gradients and the predictor are fitted on a source dataset and applied unchanged to a target dataset, so no target labels are used. All four benchmarks share the same hallucination mechanism: errors arise from missing or unreliable parametric knowledge in short-form factual recall. Hallucinations with a different mechanism, such as flawed multi-step reasoning, are outside the scope of this test (Section 5). We report Correctness-task accuracy averaged over three seeds, with the probe configuration and all hyperparameters held at their in-domain values.

Table 7 reports transfer results for all eleven models. For each target dataset we report the in-domain accuracy and the mean accuracy of the three predictors transferred from the other source datasets. Across models, transfer lowers accuracy by 2.9–3.8 points on average (3.4 over all eleven models), and the worst single source–target pair costs 4.2–5.4 points.

The cost of transfer shrinks with model size. Within every family, Δ_{mean} decreases monotonically with parameter count: from 3.6 to 3.1 points for Qwen2.5, 3.7 to 3.0 for Falcon3, and 3.6 to 2.9 for Gemma-3. SmolLM3-3B shows the largest degradation (3.8 points), in line with its lowest in-domain accuracy (Table 1). This mirrors the in-domain scaling trend of Section 4.2: larger models produce more separable gradient signatures, and these carry over better to a new dataset.

The ordering of targets is the same for every model. SciQ is the hardest target, losing 3.6–4.6 points, roughly one point more than any other target of the same model. This is consistent with SciQ already being the weakest in-domain dataset (Section 4.5) and covering a single technical domain, whereas the other three benchmarks test broader factual recall. TriviaQA, PopQA, and TruthfulQA lose 2.6–3.6 points as targets.

Transferred predictors remain competitive with the baselines. On TriviaQA (Table 6), the mean transferred accuracy exceeds Confidence and Internal State Probing for every model, by 1.9–3.8 and 3.5–6.1 points respectively. It also exceeds Semantic Entropy for every model, by 0.4–2.4 points, at one-fifth of its inference cost.

F.3 Transfer Across Models

The present method does not support transfer across models with different architectures, and we therefore fit a separate setup for each of the eleven models we evaluate. Two properties of the method prevent such transfer. First, the predictor consumes the $L \times |\mathcal{C}|$ feature matrix as a sequence of L tokens indexed by absolute layer position, with learnable positional embeddings over that index (Section 3.4). Two models with different depths induce feature sequences of different lengths, and the fitted positional embeddings of one do not apply to the other. Second, the reference gradients

Table 7: Cross-dataset transfer across all models (Correctness task). For each target dataset, ID is the in-domain accuracy of Table 1 and Tr is the mean accuracy of the three predictors fitted on the other datasets. Δ_{mean} is the mean change relative to in-domain accuracy over all twelve source–target pairs, and Δ_{worst} is the largest drop for a single pair.

Family	Size	TriviaQA		SciQ		PopQA		TruthfulQA		Δ_{mean}	Δ_{worst}
		ID	Tr	ID	Tr	ID	Tr	ID	Tr		
Qwen2.5	1.5B	75.0	71.8	73.1	68.8	74.8	71.5	74.7	71.3	−3.6	−5.0
	3B	76.2	73.2	74.3	70.3	76.0	72.9	75.9	72.7	−3.3	−4.7
	7B	76.8	74.0	74.9	71.1	76.5	73.6	76.4	73.4	−3.1	−4.4
Falcon3	1B	74.3	70.9	72.4	67.9	74.1	70.7	74.0	70.5	−3.7	−5.2
	3B	75.5	72.4	73.6	69.5	75.3	72.1	75.2	71.9	−3.4	−4.8
	7B	76.9	74.0	74.8	70.9	76.6	73.6	76.5	73.4	−3.2	−4.5
	10B	77.4	74.7	75.4	71.7	77.2	74.4	77.1	74.2	−3.0	−4.3
Gemma-3	1B	74.6	71.3	72.8	68.4	74.4	71.0	74.3	70.9	−3.6	−5.1
	4B	76.5	73.5	74.5	70.5	76.3	73.3	76.2	73.1	−3.3	−4.6
	12B	77.8	75.2	75.9	72.3	77.5	74.8	77.4	74.6	−2.9	−4.2
SmolLM3	3B	73.2	69.7	71.3	66.7	73.0	69.4	72.9	69.3	−3.8	−5.4

live in the MLP down-projection parameter space of one architecture, so for a model with different layer widths the cosine similarities of Eq. 4 are not defined. The setup cost of Appendix F.1 is therefore incurred once per model. Checkpoints that share an architecture, such as a base model and its instruction-tuned variant, avoid both obstacles, but whether fitted artifacts remain accurate after fine-tuning shifts the gradient geometry is untested. Re-indexing layers onto a fixed relative-depth grid would make transfer across depths expressible, and we leave it to future work.

G Extended Layer Ablation

G.1 Layer Range Analysis

Table 8 reports Correctness-task accuracy on TriviaQA for seven layer configurations across all eleven models. Trends are consistent across other datasets.

Table 8: Layer ablation (TriviaQA, Correctness task) across all models. We report classification accuracy (%). Performance degrades gracefully when removing layers from either end, and no range covering at least one-third of the network loses more than 1.6 points. Falcon3-1B has 18 layers, so its 20-layer ranges are not applicable.

	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	74.6	75.5	76.0	–	75.1	76.4	77.1	74.2	76.0	77.3	72.4
Last 10	74.5	75.5	76.2	73.7	74.6	76.2	76.5	74.0	76.1	77.2	72.5
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	–	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 0.4–1.3 points (0.9 on average). The last 10 and last 20 layers lose 0.4–0.9 and 0.3–0.8 points respectively, so widening the window beyond five layers does not reliably help: for Falcon3-3B and Falcon3-7B the last 10 layers score below the last 5. Conversely, the first 5 layers alone still achieve 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. No configuration covering at least one-third of the network loses more than 1.6 points, confirming that the discriminative gradient signal is broadly distributed across the transformer. Nevertheless, at widths of 5 and 10 layers, later layers outperform earlier layers of the same width in every model; last 5 vs. first 5 differs by 0.8–1.7 points. At a width of 20 the two ranges overlap heavily and differ by at most 0.4 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).

H Inference Time Breakdown

Table 9 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.

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 9: 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).

I Gradient Signature Visualization

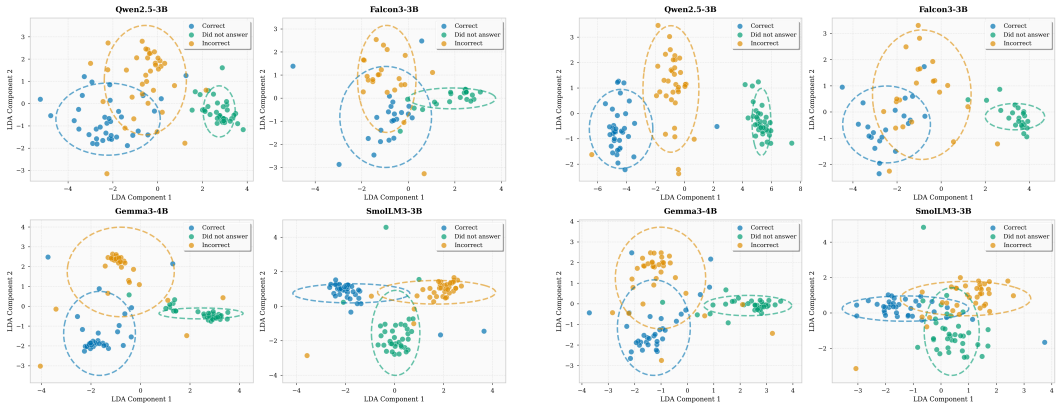
We apply Linear Discriminant Analysis (LDA) to the layer-wise cosine similarity vectors to visualize the separability of gradient signatures across behavioral categories. For each model, every point in the resulting projection represents a single sample, positioned according to its cosine similarity 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), 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.

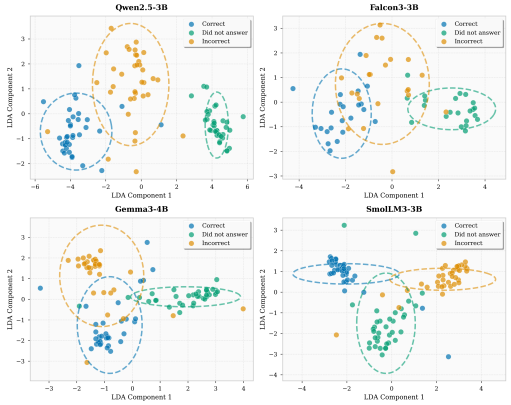
I.1 Full Task (Correct vs. Incorrect vs. Did Not Answer)

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.

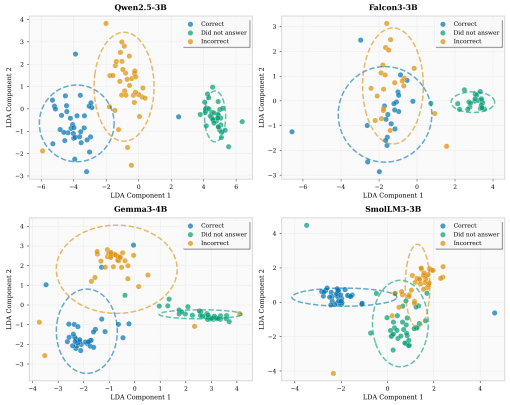


(a) Ref: affirming, Train: affirming

(b) Ref: affirming, Train: rejection



(c) Ref: rejection, Train: affirming



(d) Ref: rejection, Train: rejection

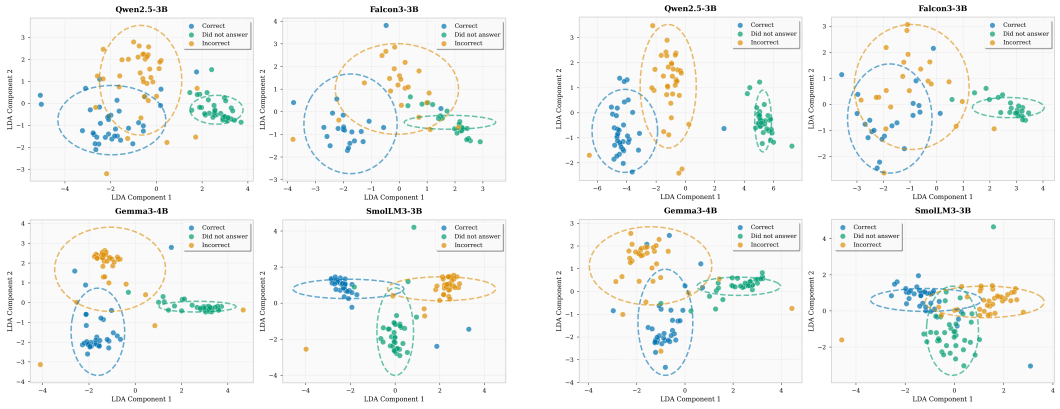
Figure 3: Full task — **Correct** reference gradient. LDA projections of layer-wise cosine similarities for all four combinations of reference and training response types.

I.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.

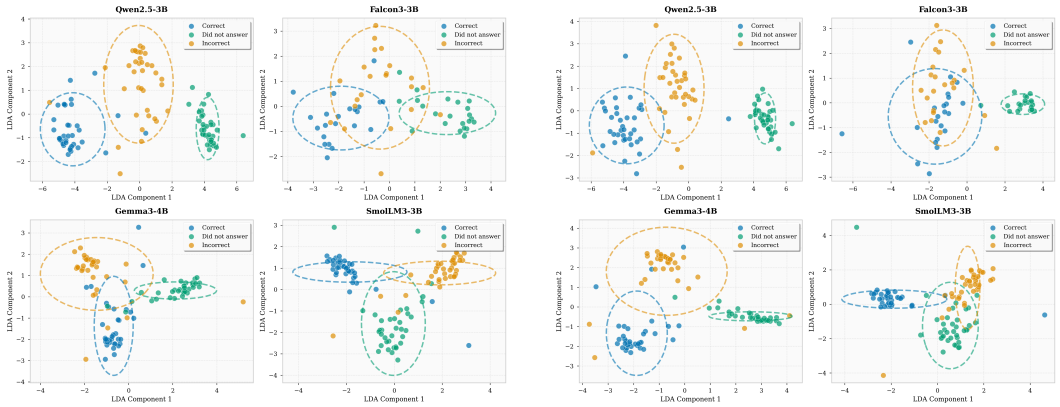
I.3 Response Task (Answered vs. Did Not Answer)

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 occupy largely



(a) Ref: affirming, Train: affirming

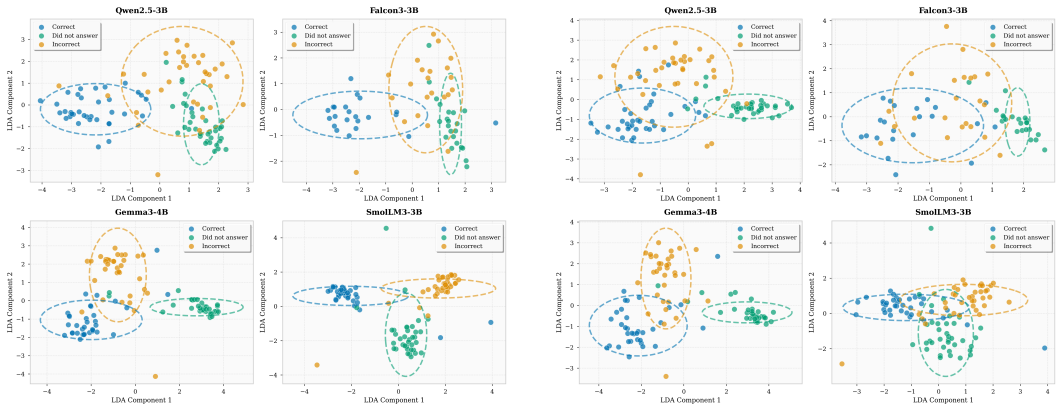
(b) Ref: affirming, Train: rejection



(c) Ref: rejection, Train: affirming

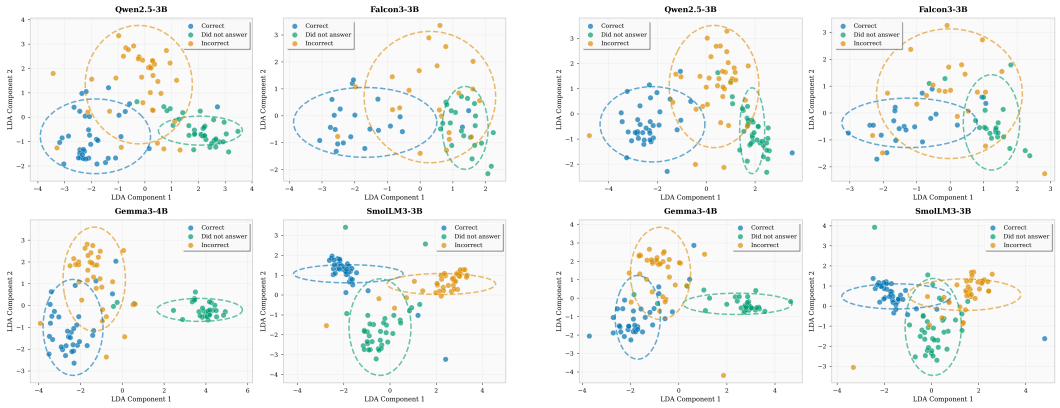
(d) Ref: rejection, Train: rejection

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

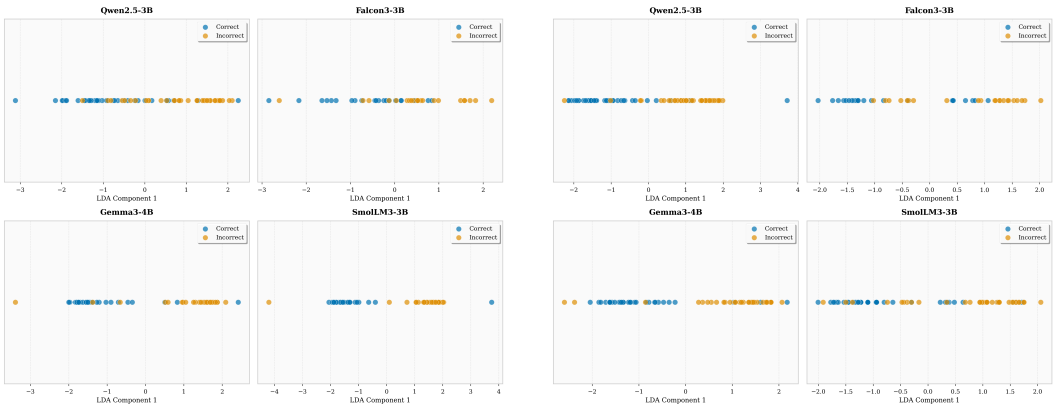
(b) Ref: affirming, Train: rejection



(c) Ref: rejection, Train: affirming

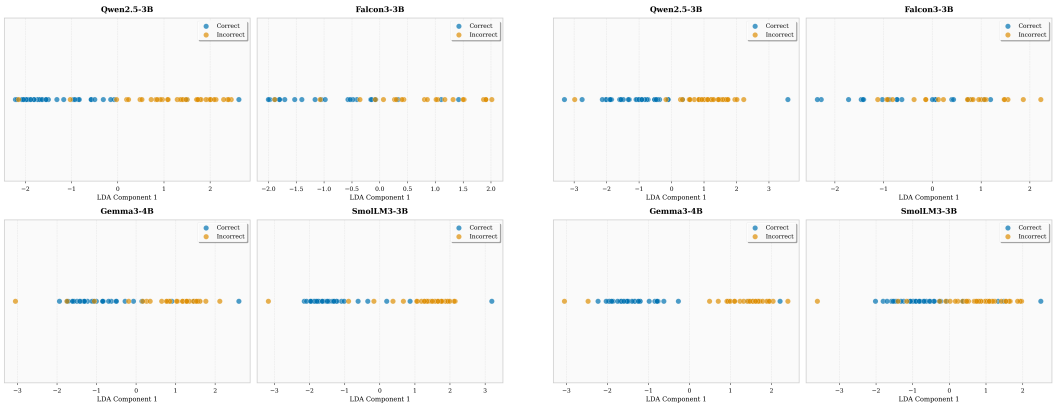
(d) Ref: rejection, Train: rejection

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

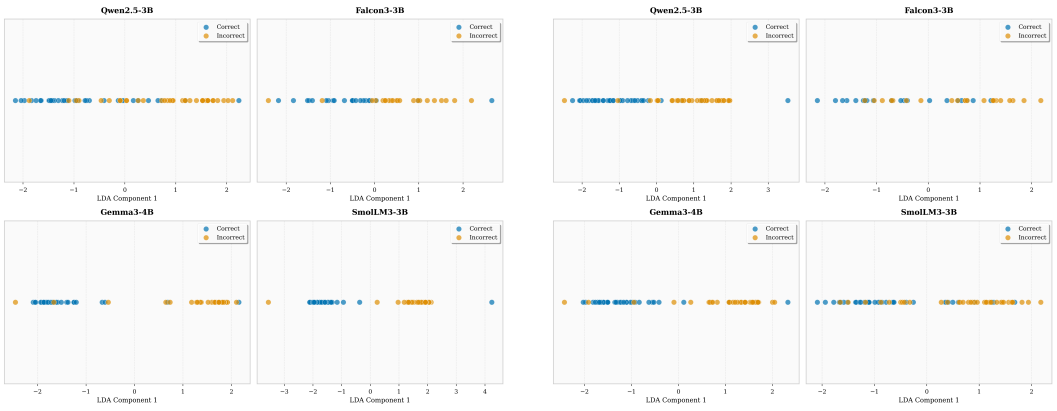
(b) Ref: affirming, Train: rejection



(c) Ref: rejection, Train: affirming

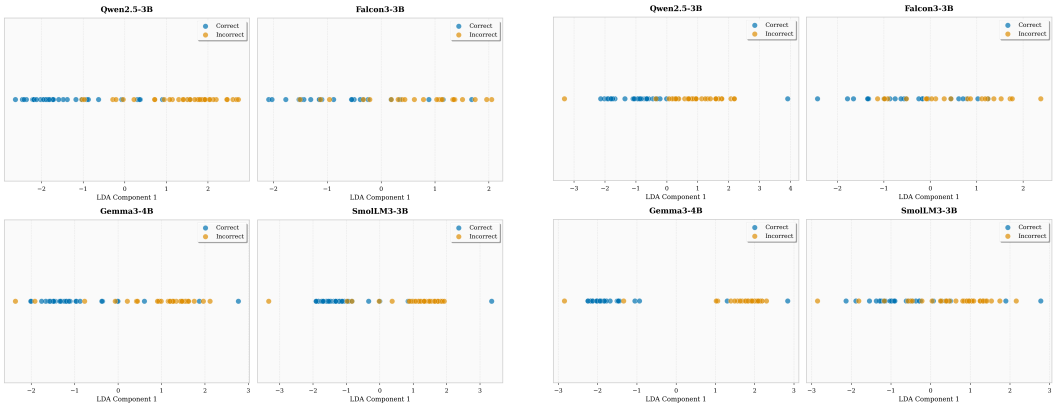
(d) Ref: rejection, Train: rejection

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



(a) Ref: affirming, Train: affirming

(b) Ref: affirming, Train: rejection

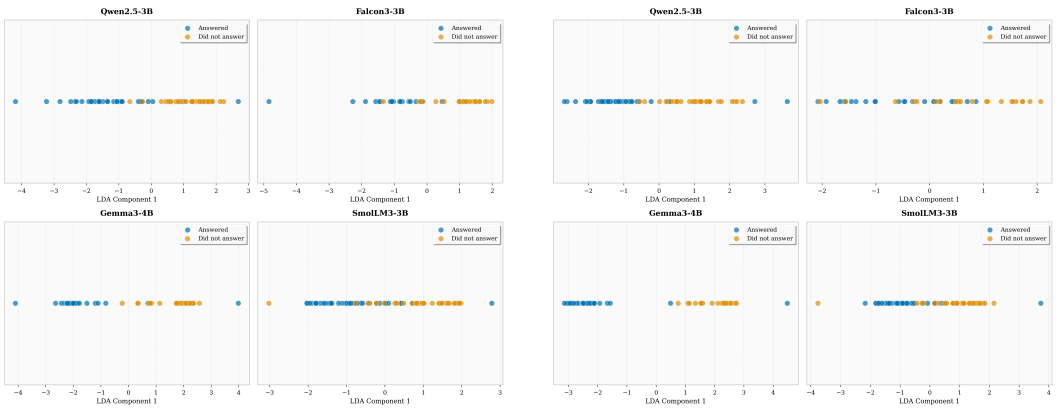


(c) Ref: rejection, Train: affirming

(d) Ref: rejection, Train: rejection

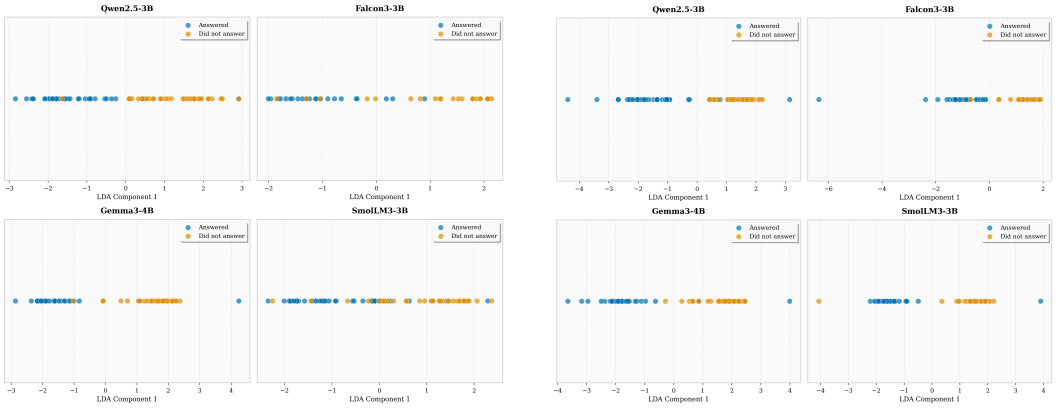
Figure 7: Correctness task — **Incorrect** reference gradient. LDA projections of layer-wise cosine similarities.

distinct regions, although some overlap remains for Qwen2.5-3B, Falcon3-3B, and SmolLM3-3B. The separation is larger than that observed in the Correctness task (Figures 6 and 7), visually confirming that abstention produces a qualitatively distinct gradient signature.



(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. Answered and Did Not Answer samples occupy largely distinct regions.

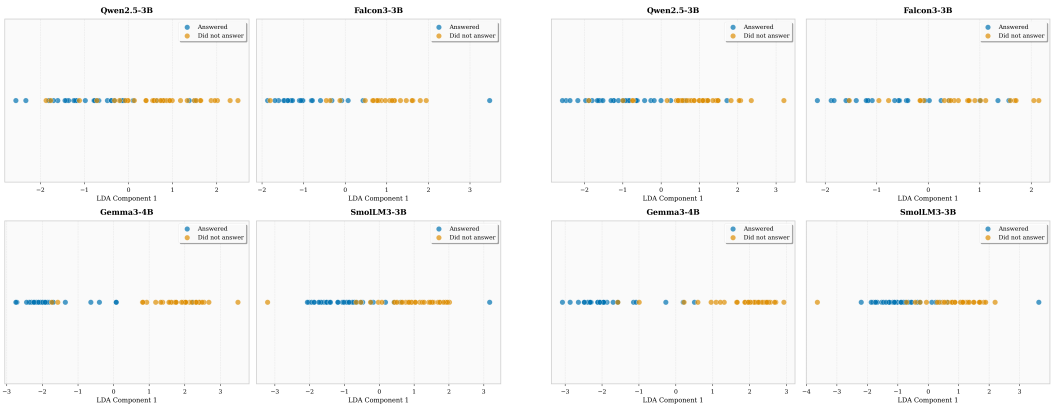
I.4 Summary of Visualization Observations

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 largely distinct region of the LDA projection, with some overlap for a few models. This visual separation corresponds to the 94–99% accuracy observed on the Response task (Table 2).

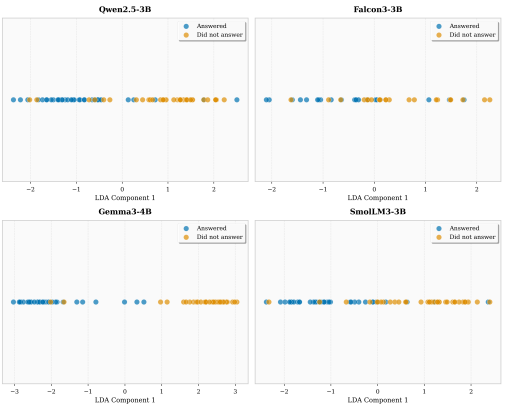
Correct vs. Incorrect separation is tighter. The Correct and Incorrect clusters overlap more than either does with the DNA cluster, consistent with our finding that the Correctness distinction is the primary bottleneck in three-way classification. The overlap is greatest for the rejection response type, 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

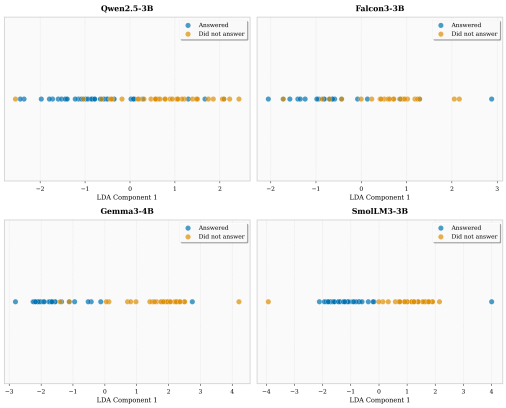


(a) Ref: affirming, Train: affirming

(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. Abstention remains more separable than correctness under this reference.

probes) is sufficient for deployment, as predictors trained on each of the twelve configurations achieve accuracy within a 2-point range.

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), selected on the validation split, is sufficient at inference time.