From Predictive Separability to Causal Validity: Quantization Robustness of Correctness Probes in LLMs

Anonymous ACL submission

Abstract

Linear probing has been widely used to extract high-level concepts such as truthfulness and factuality from the internal representations of large language models (LLMs), yet probe validity is typically evaluated using predictive metrics such as AUROC. In this work, we show that high predictive separability does not necessarily indicate a stable correctness-related separability structure. In preliminary analysis, high-AUROC probes exhibited unstable behavior under activation steering, suggesting that predictive performance alone may be insufficient to establish probe validity. To address this issue, we propose a multi-level validation framework consisting of predictive separability, confound and structure ablation, and causal validation. Correctness-related probes constructed from Contrast-Consistent Search (CCS)-style minimal-variation pairs consistently passed the proposed validation criteria across multiple models and datasets, while remaining relatively stable under five quantization settings with only limited degradation. Furthermore, we observe correlated quantization error structures and probe-dependent geometric alignment patterns during the robustness analysis, and find that degradation magnitude is closely associated with sample-level score-gap shrinkage. Our findings highlight the limitations of relying solely on predictive metrics and demonstrate the importance of rigorous probe validation prior to representation-level robustness analysis.

1 Introduction

Large language models (LLMs) have demonstrated remarkable performance across a wide range of natural language processing tasks and are rapidly being deployed in real-world applications (Vaswani et al., 2017; Brown et al., 2020; Touvron et al., 2023). However, reliability issues such as hallucination and overconfident generation remain significant limitations (Lin et al., 2022; Manakul et al., 2023). Accordingly, there has been growing interest in analyzing and controlling the internal representations of LLMs, particularly as internal representations are increasingly utilized in controllable generation settings such as downstream intervention, activation steering, and representation editing (Turner et al., 2024; Zou et al., 2023; Meng et al., 2022a,b). Within this context, verifying whether such representations genuinely capture stable semantic signals has become an increasingly important problem.

Recently, a growing body of work has employed linear probing to extract high-level concepts such as truthfulness and factuality from LLM hidden states (Burns et al., 2022; Azaria and Mitchell, 2023). Linear probing is widely used to extract semantic concepts from hidden representations and to support downstream interventions such as activation steering and behavior control (Alain and Bengio, 2016; Belinkov, 2022; Burns et al., 2022). However, most prior studies evaluate probe validity primarily through predictive metrics such as AUROC or classification accuracy, despite longstanding concerns that probing methods may exploit superficial statistical regularities or dataset artifacts rather than capturing the intended semantic representation (Hewitt and Liang, 2019; Pimentel et al., 2020). As a result, interpreting probing results solely through predictive metrics risks leading to distorted representation-level conclusions.

At the same time, low-bit quantization techniques such as GPTQ, AWQ, and SmoothQuant have become widely adopted for efficient LLM deployment (Frantar et al., 2022; Lin et al., 2024; Xiao et al., 2023; Lin et al., 2025). Accordingly, increasing attention has been directed toward understanding how stably internal representations are preserved under quantization perturbations. In particular, the robustness of correctness-related representations is important not only for inference accuracy, but also for representation-level interpretabil-

ity and the reliability of downstream interventions.

In preliminary analysis, we observe that probes exhibiting high predictive separability can nevertheless display unstable behavior in actual intervention settings. This observation suggests that predictive metrics alone may be insufficient to guarantee the semantic validity or causal consistency of a probe, and highlights the need for systematic probe validation prior to conducting representation-level robustness analysis.

To address this issue, we propose a multi-level validation framework consisting of predictive separability (Level 1), confound and structure ablation (Level 2), and causal validation (Level 3). Beyond simple predictive correlation, the proposed framework is designed to progressively verify whether a probe relies on confounding structures and whether it maintains consistent semantic behavior under intervention. Using CCS-style (Contrast-Consistent Search) correctness probes that successfully pass this validation framework (Burns et al., 2022), we further analyze representation robustness across diverse quantization settings and investigate the geometric structures associated with robustness.

The main contributions of this work are as follows:

- We show that high predictive separability alone does not guarantee a stable correctness-related representation, and propose a multi-level validation framework for progressively verifying probe validity.
- We demonstrate that validated correctness probes remain relatively stable across diverse weight-only and weight-activation quantization settings.
- We analyze the correlated error structures and probe-dependent geometric alignment patterns associated with quantization robustness, and show that sample-level gap shrinkage is closely related to degradation magnitude.

2 Related Work

LLM Probing and Internal Representations. Recent probing studies have demonstrated that high-level concepts can be extracted from LLM hidden states through linear representations (Burns et al., 2022; Zou et al., 2023). These studies show that truth-related representations can be discovered using Contrast-Consistent Search (CCS) and that diverse latent concepts can be extracted and controlled through Representation Engineering. However, prior work has primarily evaluated probe validity using predictive metrics such as AUROC or accuracy, and whether high predictive separability truly reflects the intended latent representation has not been sufficiently validated (Hewitt and Liang, 2019; Pimentel et al., 2020). In particular, recent work has suggested that although LLMs may internally encode reliable meta-cognitive signals, extracting such signals critically depends on the quality of the probing lens itself (Ma et al., 2025).

Activation Steering and Causal Validation. Activation steering has been actively studied as a method for analyzing the functional role of internal representations and controlling generation behavior (Li et al., 2023; Turner et al., 2024; Zou et al., 2023). In particular, prior work has demonstrated the possibility of truthfulness steering through inference-time intervention and behavior control via activation manipulation. In contrast to prior work, we employ activation steering not as a controllability tool, but as a causal validation tool for verifying probe validity.

Quantization and Model Interpretability. Recent quantization methods, including AWQ (Activation-aware Weight Quantization) (Lin et al., 2024), GPTQ (Generative Pre-trained Transformer Quantization) (Frantar et al., 2022), and SmoothQuant (Xiao et al., 2023), have primarily evaluated robustness using output-level metrics such as perplexity and task accuracy. Recent work has begun investigating quantization effects at the hidden-representation level, including representation sensitivity and interpretability under quantization perturbations (LeVi et al., 2025; Wang et al., 2025). In comparison, the effects of quantization perturbations on the semantic geometry of hidden representations and on the concept-level structures captured by linear probes remain relatively underexplored. In this work, we analyze quantization robustness at the representation level using validated correctness probes, while further investigating the geometric error structures and probe-dependent alignment patterns that emerge under quantization.

3 Methodology

3.1 Probe Construction

Pair Construction. Correctness probes can be constructed in various ways depending on the pair

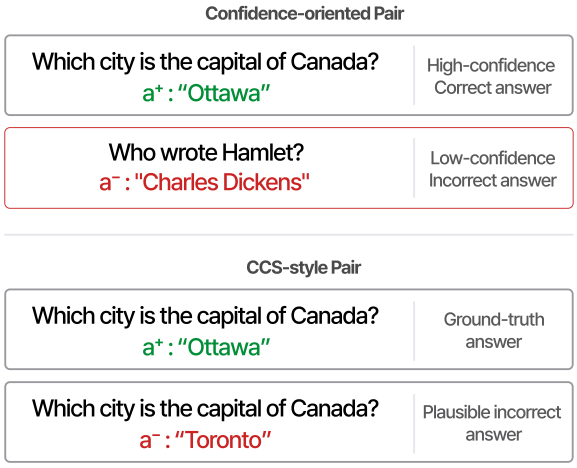

Figure 1: Comparison of confidence-oriented and CCS-style pair construction. Confidence-oriented pairs may include confounding factors beyond correctness, whereas CCS-style pairs isolate correctness by fixing the question and varying only the answer.

construction strategy. In confidence-oriented construction, correct and incorrect generations collected from different questions and generation contexts are used as positive and negative examples. In contrast, CCS-style construction employs minimal-variation contrastive pairs in which the question is held fixed while only the correctness of the proposed answer is varied (Burns et al., 2022). Figure 1 illustrates this distinction. While confidence-oriented pairs may simultaneously contain differences arising from questions and generation contexts, CCS-style pairs isolate correctness under an identical question condition.

Mean-Difference Probe. The goal of this work is to verify whether a model distinguishes the correctness of a proposed answer at the hidden-representation level for a given question q and proposed answer a. To this end, we construct contrastive pairs $(q, a^+), (q, a^-)$, where a^+ and a^- denote correct and incorrect answers to the same question, respectively. Unless otherwise specified, hidden states are extracted from the final token position of the proposed answer. To estimate the correctness direction from these representations, we define a mean-difference probe at layer l as

$$v_l = \mathbb{E}[h_l^+] - \mathbb{E}[h_l^-], \tag{1}$$

where h_l^+ and h_l^- denote hidden representations from the correct and incorrect answer conditions, respectively. The final probe is obtained by normalizing v_l to a unit vector. The mean-difference probe directly captures the contrastive direction between correct and incorrect conditions, and is naturally suited for steering-based causal validation.

Probe-Family Extension. Additionally, to examine whether our findings depend on a specific probe construction, we also train logistic regression (LR) probes. The LR probe uses the normal direction of a linear decision boundary that separates h_l^+ and h_l^-, and is subsequently used to analyze the probe-family generality of representation robustness and geometric behavior (implementation details are provided in Appendix C.3).

3.2 Multi-Level Validation Framework

High predictive performance alone does not guarantee that a probe captures the intended latent representation. A probe may instead rely on superficial confounds unrelated to correctness or on statistical patterns in the training distribution (Hewitt and Liang, 2019; Pimentel et al., 2020). To address this issue, we propose a multi-level validation framework that progressively verifies whether a probe (1) can separate representations, (2) bases this separability on a stable semantic structure, and (3) exerts a causal influence in actual intervention settings.

Level 1: Predictive Separability In the first stage, we evaluate whether the probe can distinguish between the h^+ and h^- representations. AUROC is measured based on the projection scores onto the layer-wise probe direction v_l, allowing us to assess whether the probe captures representational differences between the two conditions.

Level 2: Structural Robustness In the second stage, we examine whether predictive separability depends on prompt artifacts or superficial confounds. To this end, we conduct two types of ablations. First, we evaluate whether AUROC remains stable across different prompt templates in order to analyze template dependency (template stability). Second, we measure the effect of answer-length differences between correct and incorrect conditions on probe performance to assess the presence of length-related confounds (length independence).

Level 3: Causal Validation In the final stage, we evaluate the causal influence of the probe direction through activation steering. For a probe direction at layer l, denoted as v_l, we apply the following intervention to the hidden state at layer l:

$$h_l' = h_l \pm \alpha \cdot v_l. \tag{2}$$

Positive steering $(+)$ is expected to shift model outputs toward the correct-answer direction, while negative steering $(-)$ is expected to shift outputs toward the incorrect-answer direction. When both bidirectional steering interventions exhibit consistent directional effects, this suggests that the probe direction is associated with a correctness-related representation.

3.3 Quantization Robustness Analysis

To evaluate whether probes that pass the validation framework in Section 3.2 remain preserved after quantization, we introduce a representation-level robustness analysis and geometric characterization.

Evaluation Principle. Probes are trained on hidden states extracted from the FP16 model and evaluated on hidden states from quantized models. This setting directly evaluates whether the same representation is preserved after quantization. Representation degradation is defined as follows:

$$\Delta_l = \mathrm{AUROC}_l^{\mathrm{FP16}} - \mathrm{AUROC}_l^{\mathrm{Q}}, \tag{3}$$

where a smaller value of Δ_l indicates that the correctness signal at layer l is more robust to quantization.

Geometric Characterization. To analyze the geometric structure underlying robustness, we examine the alignment between quantization error and the probe direction. Let the probe score for a hidden state h be defined as $s = h \cdot v$, where v denotes the probe direction. If the quantization error vector is denoted by e, the probe score after quantization can be expressed as

$$s_{\mathrm{Q}} = s_{\mathrm{FP16}} + e \cdot v. \tag{4}$$

Thus, changes in probe scores are determined not by the overall error magnitude, but by how strongly the error projects onto the probe direction.

To analyze this effect, we define the mean error vectors for the correct and incorrect conditions as

$$\bar{e}^{+} = \mathbb{E}[h_{\mathrm{Q}}^{+} - h_{\mathrm{FP16}}^{+}], \quad \bar{e}^{-} = \mathbb{E}[h_{\mathrm{Q}}^{-} - h_{\mathrm{FP16}}^{-}], \tag{5}$$

respectively.

We introduce the following metrics to characterize the geometric structure of quantization error. First, to measure the directional similarity of quantization errors arising under the correct and incorrect conditions, we use Error Cosine Similarity (`ErrorCos`), defined as

$$\texttt{ErrorCos} = \cos(\bar{e}^{+}, \bar{e}^{-}). \tag{6}$$

Second, to measure the alignment between the probe direction and the error vectors, we compute $\cos(v, \bar{e}^{+})$ and $\cos(v, \bar{e}^{-})$.

Finally, to analyze changes in sample-level correctness separation, we introduce a relative gap shrinkage metric (`RelShrink`). For each sample i, the sample-level gap shrinkage is defined as the reduction in score separation caused by quantization:

$$\mathrm{shrink}_i = \mathrm{gap}_i^{\mathrm{FP16}} - \mathrm{gap}_i^{Q}, \tag{7}$$

where $\mathrm{gap}_i^{(\cdot)}$ denotes the probe score gap for sample i under the FP16 and quantized models, respectively:

$$v^{\top} h_{i,(\cdot)}^{+} - v^{\top} h_{i,(\cdot)}^{-}.$$

We then normalize this quantity by the FP16 gap scale to compute the final `RelShrink` metric:

$$\texttt{RelShrink} = \frac{\mathbb{E}_i[\mathrm{shrink}_i]}{\mathbb{E}_i[\mathrm{gap}_i^{\mathrm{FP16}}]}, \tag{8}$$

where v denotes the probe direction learned from FP16 hidden states, with its orientation aligned such that the mean FP16 gap is positive. Positive shrinkage indicates that the probe score separation between correct and incorrect conditions decreases after quantization. These metrics are used in Section 6 to analyze the geometric structures associated with quantization robustness.

4 Experimental Setup

Models. We evaluate three instruction-tuned LLMs: Llama 3.1 8B Instruct (Touvron et al., 2023), Qwen2.5 7B Instruct (Qwen et al., 2025), and Mistral 7B Instruct v0.3 (Jiang et al., 2023). These models cover different model families and pretraining distributions to evaluate generality.

Datasets. CCS-style contrastive pairs are constructed from TriviaQA(Joshi et al., 2017), ARC-Easy (Clark et al., 2018), and CommonsenseQA (Talmor et al., 2019). TriviaQA is an open-ended factual QA dataset, where plausible incorrect answers are generated using GPT-4.1 ($N = 4{,}999$). ARC-Easy ($N = 5{,}197$) and CommonsenseQA (CSQA; $N = 10{,}962$) are multiple-choice datasets, where non-ground-truth options are used as incorrect answers. All datasets are split 50/50 into train and test sets. In addition, a preliminary confidence-oriented probing experiment is conducted on TriviaQA to motivate the proposed validation framework.

Quantization. FP16 is used as the baseline model. We evaluate three weight-only INT4 quantization methods: AWQ (Lin et al., 2024), GPTQ (Frantar et al., 2022), and Round-to-Nearest (RTN). We also evaluate two weight-activation settings: W8A8 via SmoothQuant (Xiao et al., 2023) and W4A8. For all quantization conditions, we use the same CCS pairs, prompt templates, extraction layers, and answer-final token positions as in the FP16 baseline.

Probes. Mean-difference probes are used as the primary probe family. To evaluate probe-family generality, logistic regression (LR) probes are additionally trained. All probes are trained using FP16 hidden states from the training split.

Steering. Causal steering was performed across all model–dataset combinations using 200 evaluation samples and multiple steering strengths ranging from $\alpha = 0$ to $\alpha = 10$. For the random baseline, we sampled $N = 10$ random unit vectors of identical dimensionality from $\mathcal{N}(0, I)$ and applied the same steering procedure. The difference in effect size between the probe direction and the random-direction baseline was measured using the following z-score:

$$z = \frac{|\Delta P_{\text{probe}}| - \mu_{|\text{random}|}}{\sigma_{|\text{random}|}}. \tag{9}$$

A probe was considered to provide evidence for Level 3 causal validity when it exhibited consistent bidirectional steering behavior and substantially outperformed random-direction baselines. We report z-scores as an effect-size reference rather than a strict binary criterion. Detailed steering protocols and random-baseline construction are provided in Appendix D.

5 Results

5.1 Probe Validity: From Predictive Separability to Causal Validation

Preliminary Confidence-Probe Observation. As shown in Figure 2(a), the confidence-oriented probe failed causal validation despite achieving high AUROC values (Llama 0.921, Qwen 0.893). In Llama, increasing steering strength led to bidirectional model collapse, while in Qwen, almost no steering response was observed under intervention ($\max \Delta P < 0.01$). In contrast, the CCS-style probe in Figure 2(b) exhibited substantially

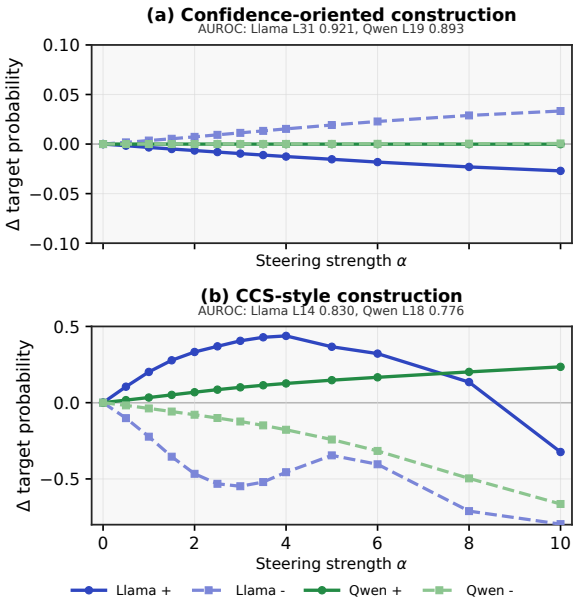

Figure 2: Activation steering behavior of confidence-oriented and CCS-style pair constructions across steering strengths. Panel (a) uses a different y-axis scale to visualize the comparatively smaller steering responses of confidence-oriented constructions.

stronger bidirectional causal effects despite having relatively lower AUROC values. These results demonstrate that predictive metrics alone may be insufficient to establish probe validity.

CCS Probe – Predictive Separability. Accordingly, we construct probes using the CCS-style minimal-variation pairs defined in Section 3.1. As shown in the Level 1 results of Table 1, all conditions achieved AUROC ≥ 0.776, suggesting the existence of correctness-related separability at the hidden-representation level. Although the AUROC values of CCS probes were lower than those of confidence-oriented probes, CCS pairs provide stricter semantic control by employing minimal-variation contrasts that vary only the correctness condition while keeping the question fixed. In contrast, confidence-oriented probes may simultaneously reflect confidence-related factors unrelated to correctness itself.

Confound & Structure Ablation. As shown in the Level 2 results of Table 1, CCS probes consistently exhibited low sensitivity to both prompt-template variation and answer-length differences across all model–dataset conditions. AUROC variation across prompt templates remained below 0.022 in all conditions, while the AUROC gap across answer-length subsets remained below 0.036

Dataset	Llama-3.1-8B-Instruct			Qwen2.5-7B-Instruct			Mistral-7B-Instruct-v0.3		
	L1 AUROC	L2 Temp./Len.	L3 ΔP	L1 AUROC	L2 Temp./Len.	L3 ΔP	L1 AUROC	L2 Temp./Len.	L3 ΔP
TriviaQA	.830	.022/.036	+.44/-.80	.776	.020/.010	+.24/-.66	.824	.014/.013	+.42/-.93
ARC-Easy	.931	.022/.016	+.87/-.73	.964	.014/.011	+.11/-.63	.923	.013/.010	+.85/-.86
CSQA	.838	.010/.000	+.63/-.64	.865	.022/.007	+.19/-.56	.856	.012/.017	+.65/-.77

Table 1: Multi-level validation results for CCS-style probes evaluated at the selected FP16 layer. L1, L2, and L3 denote the three stages of the proposed validation framework. Temp./Len. denotes prompt-template variation and answer-length sensitivity, respectively. The two values in ΔP correspond to $\Delta P(+)$ and $\Delta P(-)$ observed over the steering-strength sweep. Across all model–dataset conditions, CCS-style probes exhibit high predictive separability, low structural sensitivity, and correct-sign steering behavior.

(full condition-level results are provided in Appendix B.6). These results suggest that the observed separability is not primarily driven by superficial lexical artifacts or formatting biases.

Causal Validation. For bidirectional causal effects to hold, positive steering must shift model outputs toward the correct-answer direction, while negative steering must shift outputs toward the incorrect-answer direction. Accordingly, Steer $\Delta P(+)$ should be positive, whereas Steer $\Delta P(-)$ should be negative. As shown in Level 3 of Table 1, CCS probes consistently exhibited correct-sign steering behavior under all conditions, and in most cases substantially outperformed random directions with identical dimensionality (Appendix F.3, Table 21). These results suggest that the probe direction is associated with a correctness-related representation rather than merely reflecting a generic positivity bias.

5.2 Quantization Robustness

In this section, following the protocol described in Section 3.3, we evaluate whether validated correctness signals remain preserved after quantization across five quantization settings.

Overall Robustness. Figure 3(a–c) shows that the layer-wise AUROC profiles are largely preserved after quantization across all models. Quantized models closely follow the characteristic layer trajectories observed in FP16, with only limited degradation around the peak correctness layers. Figure 3(d) summarizes the degradation at the FP16-selected best layer. Across all conditions, only limited degradation was observed, with $\Delta < 0.04$ in every case. The mean degradation values were 0.015 for Llama, 0.009 for Qwen, and 0.025 for Mistral. These results suggest that validated correctness signals remain stable under quantization.

Robustness Across Quantization and Probe Families. Among the quantization methods, W8A8 exhibited the strongest robustness (mean $\Delta = 0.003$), while W4A8 and low-bit weight-only methods (AWQ, GPTQ, and RTN) showed relatively larger degradation. This robustness was not limited to the mean-difference probe. LR probes also exhibited limited degradation across all conditions with $\Delta < 0.06$, but consistently showed larger mean Δ values compared to mean-difference probes (Llama 0.025, Qwen 0.031, Mistral 0.031). These results suggest that the magnitude of quantization-induced degradation can vary depending on the probe family. The geometric characteristics associated with this difference are further analyzed in Section 6.

6 Geometric Analysis of Quantization Robustness

In Section 5.2, we showed that correctness-related representations remain relatively stable after quantization. In this section, we analyze the quantization error structures associated with this robustness using the geometric metrics defined in Section 3.3.

6.1 Correlated Error Structure

Shared Quantization Error Across CCS Pairs. Hidden-state perturbations induced by quantization primarily arise from shared weight rounding patterns. Since CCS pairs share the same question q, the forward passes for the correct condition (q, a^+) and the incorrect condition (q, a^-) largely share the same input context. As a result, quantization error is likely dominated more by shared weight perturbation structures than by condition differences.

As shown in Table 2, `ErrorCos` consistently exhibits high values across all models (Llama 0.968, Qwen 0.955, Mistral 0.941), suggesting that the mean quantization errors arising from the

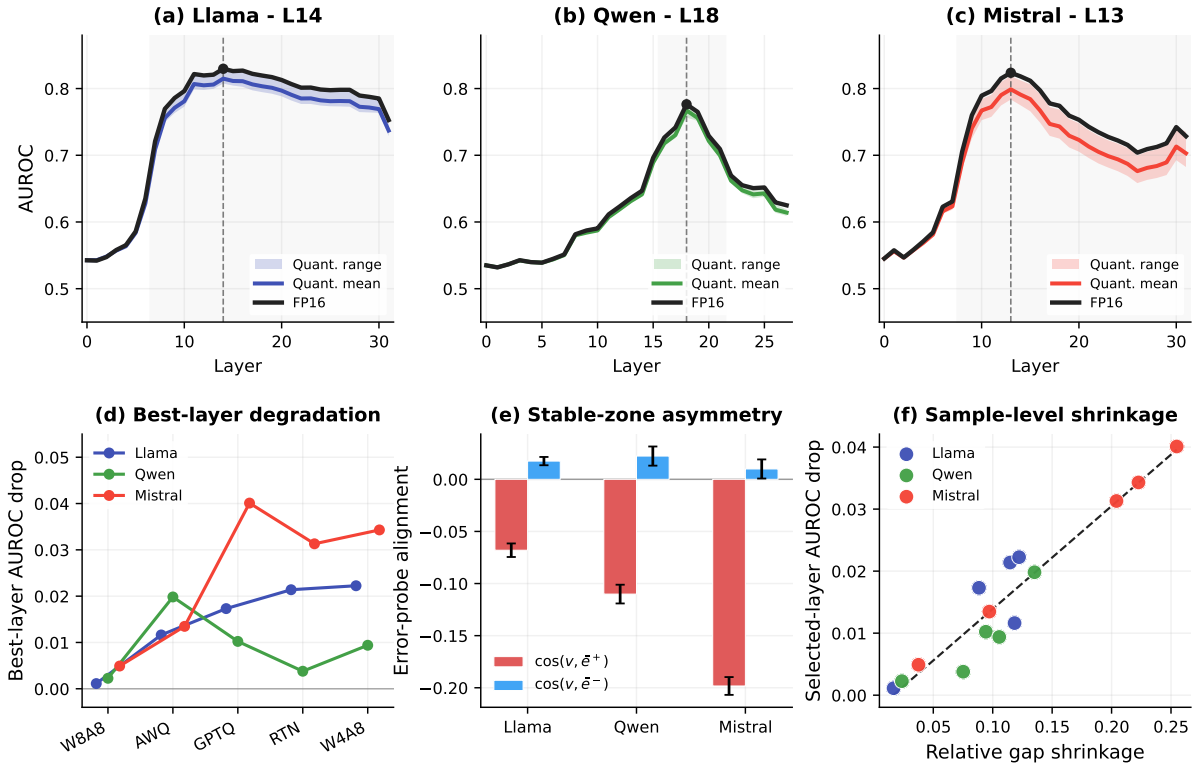

Figure 3: Layer-wise quantization robustness and geometric structure of correctness-signal degradation. (a–c) Layer-wise AUROC profiles for FP16 and quantized models. Colored curves indicate mean AUROC across five quantization settings, with shaded bands showing min–max ranges. Dashed lines denote the selected FP16-best layer. (d) AUROC degradation at the selected layer. (e) Alignment between quantization errors and the CCS probe direction. $\bar{e}^+$ shows consistent anti-alignment, whereas $\bar{e}^-$ remains near zero. (f) Relationship between relative sample-wise gap shrinkage and AUROC degradation. Larger reductions in correctness-score separation are associated with larger AUROC drops (Spearman $\rho = 0.918$, Pearson $r = 0.954$).

correct and incorrect CCS conditions are formed along largely shared directions. Since the mean-difference probe is defined as $v = \mathbb{E}[h^+] - \mathbb{E}[h^-]$, the signal after quantization can be decomposed as

$$v_{\mathrm{Q}} = v + (\bar{e}^+ - \bar{e}^-). \tag{10}$$

A high `ErrorCos` indicates that $\bar{e}^+$ and $\bar{e}^-$ are perturbed in similar directions, partially suppressing the perturbation difference term $(\bar{e}^+ - \bar{e}^-)$. As a result, the probe direction can remain relatively stable even after quantization. However, the subsequent alignment analysis reveals clear asymmetry in the projection structure onto the probe direction.

Asymmetric Probe–Error Alignment. Consistent asymmetric alignment structures were observed between the probe direction v and quantization error. As shown in Table 2, $\cos(v, \bar{e}^+)$ for the mean-difference probe was consistently negative across all models and quantization conditions (range: $-0.035 \sim -0.317$), whereas $\cos(v, \bar{e}^-)$ was distributed mostly near zero. This asymmetry remained statistically significant under all conditions after Benjamini–Hochberg FDR correction (Benjamini and Hochberg, 1995) (paired test, $p < 0.05$; mean $|\text{Cohen's } d|$: Llama 2.12, Qwen 3.66, Mistral 3.89). These results suggest that quantization systematically shifts the h^+ representations in the direction opposite to the probe direction, thereby reducing the projection separation between correct and incorrect conditions. However, the magnitude of the mean-level alignment itself did not directly explain degradation magnitude ($|\cos(v, \bar{e}^+)|$ vs. Δ: Spearman $\rho = 0.13$, $p = 0.64$, $N = 15$).

Sample-Level Gap Shrinkage. Accordingly, we further analyzed sample-level score-gap shrinkage. `RelShrink` exhibited a strong correlation with selected-layer AUROC degradation (Figure 3(f); Spearman $\rho = 0.918$), indicating that quantization systematically acts to compress sample-level correctness separation. In this view, mean error geometry explains the directional tendency of degradation, while sample-level separation compression more

Model	Probe	Degradation		Shared Error	Alignment		Gap Shrinkage		$\cos(\mathrm{MD}, \mathrm{LR})$
		Mean Δ	Max Δ	`ErrorCos`	$\cos(v, \bar{e}^+)$	$\cos(v, \bar{e}^-)$	`RelShrink`	$P(\mathrm{shrink}_i > 0)$	
Llama	MD	0.015	0.022	0.968	-0.121	+0.012	0.092	0.618	0.125
	LR	0.025	0.040		-0.008	+0.005	0.096	0.587	
Qwen	MD	0.009	0.020	0.955	-0.149	0.000	0.087	0.593	0.087
	LR	0.031	0.045		-0.009	+0.004	0.110	0.568	
Mistral	MD	0.025	0.040	0.941	-0.120	+0.113	0.163	0.684	0.139
	LR	0.031	0.056		+0.002	+0.023	0.137	0.612	

Table 2: Probe-family geometric robustness under quantization at the selected FP16 TriviaQA layer, aggregated over five quantization settings. Mean Δ and Max Δ denote the average and maximum AUROC degradation relative to FP16. `ErrorCos` measures shared quantization-error similarity across CCS conditions, while `RelShrink` summarizes sample-level correctness-gap shrinkage under quantization. $\cos(\mathrm{MD}, \mathrm{LR})$ reports the cosine similarity between mean-difference and logistic-regression probe directions within each model.

directly reflects degradation magnitude.

6.2 Probe-Family Geometric Alignment

To verify that our findings are not specific to mean-difference probes, we further analyze the geometric structure of LR probes.

Distinct Alignment Patterns Across Probe Families. As shown in Table 2, mean-difference probes and LR probes exhibited distinct alignment patterns with respect to quantization error. For mean-difference probes, $\cos(v, \bar{e}^+)$ was consistently negative across all models, whereas for LR probes, both $\cos(v, \bar{e}^+)$ and $\cos(v, \bar{e}^-)$ were distributed near zero ($|\cos| < 0.025$). In other words, mean-difference probes exhibited a correlated error pattern characterized by anti-alignment, while LR probes achieved robustness through near-orthogonal mean error distributions relative to the probe direction.

Robustness Differences Across Probe Families. These geometric differences were also consistent with the degradation patterns observed across probe families (Table 2). Mean-difference probes exhibited relatively small degradation across all models, whereas LR probes consistently showed larger degradation. This difference may be related to how the two probes form decision boundaries: mean-difference probes rely on centroid-level contrasts, while LR probes are more sensitive to individual sample perturbations. Consistent with this interpretation, LR probes exhibited positive sample-level shrinkage tendencies across all models, with an average $P(\mathrm{shrink}_i > 0)$ of 0.589.

Geometric Diversity of Robustness Patterns. Notably, the directional similarity between the two probe families was very low, with $\cos(\mathrm{MD}, \mathrm{LR}) < 0.14$ (Table 2). In other words, although the two probes capture different correctness-related directions in hidden-state space, both exhibited robustness accompanied by distinct geometric structures—anti-alignment in mean-difference probes and near-orthogonality in LR probes. These findings suggest that the geometric framework proposed in Section 3.3 can be used to analyze quantization robustness structures beyond a specific probe family.

7 Conclusion

This work shows that high predictive performance in LLM probing does not necessarily imply a stable correctness-related representation, and introduces a multi-level validation framework for rigorous probe verification. While a preliminary confidence-oriented probe exhibited unstable causal steering behavior despite high AUROC, the CCS-style probe consistently passed predictive, structural, and causal validation across all evaluated models and datasets. Using the validated probes, we further analyzed quantization robustness and found that correctness-related representations remain relatively stable across diverse quantization settings. Robustness was consistently accompanied by correlated quantization error structures and probe-dependent geometric alignment patterns, despite distinct geometric behaviors across probe families. Overall, our findings highlight that probe validation should precede robustness analysis, and that quantization robustness is closely tied to the geometry of hidden representations.

Limitations

This work has several limitations. First, our evaluation was conducted on only three models in the 7–8B parameter range. Whether the same validation results and robustness patterns hold for larger models (e.g., 13B or 70B) remains unverified, and understanding how correctness-related representations evolve with model scale requires further investigation. Second, our main correctness analysis focuses on factual QA settings constructed from CCS-style minimal-variation pairs. As a result, the generalizability of our findings to other settings, such as free-form generation, open-ended reasoning, or alternative concepts including sentiment and toxicity, is beyond the scope of this work. Third, our geometric characterization combines mean error vector analysis with sample-level probe-gap shrinkage analysis. Through this, we find that quantization systematically reduces probe score separation for individual paired samples, and that relative gap shrinkage is strongly associated with AUROC degradation. However, this analysis is limited to one-dimensional score dynamics projected onto the probe direction, and therefore does not fully explain higher-order variation in the full hidden-representation space or the detailed causes of degradation observed in LR probes. Investigating these finer-grained dynamics remains an important direction for future work. Finally, although we employ causal steering as a key validation mechanism for probe validity, successful steering alone does not guarantee that a probe fully captures the underlying latent representation, as probing consistency itself remains a broader challenge in LLM representation analysis (Zhao et al., 2025). In this work, steering is used as an empirical validation signal to verify that the probe direction exerts a causal influence on model outputs.

Ethical Considerations

This study does not involve human subjects, personally identifiable information, or sensitive private data. We reviewed the dataset documentation and data fields to check for personally identifying information or offensive content. The datasets used in this study do not contain human-subject data, names, unique personal identifiers, or offensive content; therefore, no additional anonymization was required. We use publicly available research datasets and external artifacts, including datasets, pretrained models, and code, in accordance with their respective licenses and terms of use. Our use of these artifacts is consistent with their intended research purposes and original access conditions. We cite the original creators of these artifacts and do not redistribute any third-party artifacts beyond what is permitted by their licenses.

References

Guillaume Alain and Yoshua Bengio. 2016. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*.

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

Yonatan Belinkov. 2022. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219.

Yoav Benjamini and Yosef Hochberg. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. 2022. Discovering latent knowledge in language models without supervision. *arXiv preprint arXiv:2212.03827*.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

Jacob Cohen. 2013. *Statistical power analysis for the behavioral sciences*. routledge.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. 2022. Gptq: Accurate post-training quantization for generative pre-trained transformers. *arXiv preprint arXiv:2210.17323*.

John Hewitt and Percy Liang. 2019. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (emnlp-ijcnlp)*, pages 2733–2743.

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

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, and 1 others. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

Amit LeVi, Raz Lapid, Rom Himelstein, Chaim Baskin, Ravid Shwartz Ziv, and Avi Mendelson. 2025. You had one job: Per-task quantization using llms' hidden representations. *arXiv preprint arXiv:2511.06516*.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023. Inference-time intervention: Eliciting truthful answers from a language model. *Advances in Neural Information Processing Systems*, 36:41451–41530.

Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Wei-Ming Chen, Wei-Chen Wang, Guangxuan Xiao, Xingyu Dang, Chuang Gan, and Song Han. 2024. Awq: Activation-aware weight quantization for on-device llm compression and acceleration. *Proceedings of machine learning and systems*, 6:87–100.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Truthfulqa: Measuring how models mimic human falsehoods. In *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: long papers)*, pages 3214–3252.

Yujun Lin, Haotian Tang, Shang Yang, Zhekai Zhang, Guangxuan Xiao, Chuang Gan, and Song Han. 2025. Qserve: W4a8kv4 quantization and system co-design for efficient llm serving. *Proceedings of Machine Learning and Systems*, 7.

Ziyang Ma, Qingyue Yuan, Zhenglin Wang, and Deyu Zhou. 2025. Large language models have intrinsic meta-cognition, but need a good lens. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 3460–3477.

Potsawee Manakul, Adian Liusie, and Mark Gales. 2023. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pages 9004–9017.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022a. Locating and editing factual associations in gpt. *Advances in neural information processing systems*, 35:17359–17372.

Kevin Meng, Arnab Sen Sharma, Alex Andonian, Yonatan Belinkov, and David Bau. 2022b. Mass-editing memory in a transformer. *arXiv preprint arXiv:2210.07229*.

Tiago Pimentel, Josef Valvoda, Rowan Hall Maudslay, Ran Zmigrod, Adina Williams, and Ryan Cotterell. 2020. Information-theoretic probing for linguistic structure. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4609–4622.

Qwen, 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, and 24 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. Commonsenseqa: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Ulisse Mini, and Monte MacDiarmid. 2024. Activation addition: Steering language models without optimization.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Qianli Wang, Mingyang Wang, Nils Feldhus, Simon Ostermann, Yuan Cao, Hinrich Schütze, Sebastian Möller, and Vera Schmitt. 2025. Through a compressed lens: Investigating the impact of quantization on factual knowledge recall. *arXiv preprint arXiv:2505.13963*.

Jason Wei, Nguyen Karina, Hyung Won Chung, Yunxin Joy Jiao, Spencer Papay, Amelia Glaese, John Schulman, and William Fedus. 2024. Measuring short-form factuality in large language models. *arXiv preprint arXiv:2411.04368*.

Guangxuan Xiao, Ji Lin, Mickael Seznec, Hao Wu, Julien Demouth, and Song Han. 2023. Smoothquant: Accurate and efficient post-training quantization for large language models. In *International conference on machine learning*, pages 38087–38099. PMLR.

Raoyuan Zhao, Abdullatif Köksal, Ali Modarressi, Michael A Hedderich, and Hinrich Schütze. 2025. Do we know what llms don't know? a study of consistency in knowledge probing. *arXiv preprint arXiv:2505.21701*.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, and 1 others. 2023. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*.

A Preliminary Confidence-Oriented Probe Details

This appendix provides details on the construction procedure and failure modes of the confidence-oriented probe discussed in Sections 1 and 5.1.

A.1 Pair Construction

The confidence-oriented probe constructs h^+/h^- pairs by exploiting the correlation between model confidence and factual correctness. For TriviaQA (N = 5,000), we performed free-form generation with the FP16 versions of Llama 3.1 8B Instruct and Qwen2.5 7B Instruct, and extracted the softmax probability of the first generated token for each generation.

Correctness was judged by case-insensitive substring matching against the ground-truth answer and its aliases. We then constructed pairs according to the first-token probability as follows:

- h^+ (positive class): a high-confidence correct generation, where the model generates the correct answer and the first-token probability lies in the upper part of the distribution.
- h^- (negative class): a low-confidence incorrect generation, where the model generates an incorrect answer and the first-token probability lies in the lower part of the distribution.

As shown in Table 1, correct generations have substantially higher average probability than incorrect generations for both models, satisfying the prerequisite for probability-based h^+/h^- separation.

Model	Accuracy	Avg Prob (Correct)	Avg Prob (Incorrect)	Gap
Llama 3.1 8B	72.6%	0.80	0.50	0.30
Qwen2.5 7B	58.4%	0.80	0.55	0.25

Table 1: Basic statistics for constructing confidence-oriented pairs (TriviaQA, FP16).

The key difference from CCS-style pairs is as follows. In confidence-oriented pairs, h^+ and h^- are extracted from generations for different questions: h^+ comes from questions answered correctly by the model, whereas h^- comes from questions answered incorrectly. Consequently, representation differences between h^+ and h^- may include not only correctness information but also semantic variation such as topic, difficulty, and input length. In contrast, CCS-style pairs replace only the true/false answer for the same question, structurally controlling these confounding factors.

The hidden-state extraction position also differs. In the confidence-oriented probe, we extract from the last prompt token, immediately before generation starts. In the CCS probe, we extract from the last token of the proposed answer, after the model has read the true or false answer.

A.2 Predictive Performance

Both models show high predictive separability with the confidence-oriented probe (Table 2). The best-layer AUROC is 0.921 for Llama (Layer 31) and 0.893 for Qwen (Layer 19), higher than the CCS probe in the main text (Llama 0.830, Qwen 0.776). This suggests that question-level feature differences, introduced because h^+ and h^- come from different questions, may contribute additional separability.

Model	Best AUROC	Best Layer
Llama 3.1 8B	0.921	L31
Qwen2.5 7B	0.893	L19

Table 2: Confidence-oriented probe AUROC (FP16).

A.3 Causal Steering Failure

The confidence-oriented probe passes Level 1 (predictive separability) and does not show clear problems in the Level 2 surface-level confound checks (template stability and length independence), but fails at Level 3 (causal steering). This suggests that Level 2 ablations may miss deeper confounds such as question-level semantic variation. The failure modes are qualitatively distinct across the two models.

Llama — Bidirectional Model Collapse. When activation steering is performed with the probe direction v, both positive steering ($+\alpha v$) and negative steering ($-\alpha v$) decrease the target probability P(Yes). As α increases, P(Yes) decreases monotonically and generation collapses at high α (accuracy: $84\% \rightarrow 0\%$). This suggests that the probe direction may capture a direction strongly coupled to stable generation behavior rather than a correctness axis. Perturbing in either direction pushes the model out of distribution.

The steering results at the best AUROC layer (L31) are as follows:

Condition	Change in P as α increases	Full-gen Accuracy
$+\alpha v$ (positive steering)	P decreases ↓	0% (collapse)
$-\alpha v$ (negative steering)	P increases ↑ (partial)	0% (collapse)

Table 3: Llama confidence-oriented steering failure summary.

Only at L31 did negative steering partially increase P, but the effect was very small (max $\Delta P \approx +3.8\%$ at $\alpha = 16$). Full-generation analysis confirmed that the model produced degenerate near-zero-accuracy outputs (accuracy: 84% → 0%) without hedging behavior.

Qwen — Steering Effect Absence. Steering with the probe direction produced no meaningful change in P(Yes) (max $|\Delta P| < 0.01$). Increasing α from 0 to 16 left the model output essentially unchanged, suggesting that the probe direction is causally irrelevant to the model's generation behavior.

Condition	Max $\|\Delta P\|$	Change in Full-gen Accuracy
$+\alpha v$ / $-\alpha v$	< 0.01	82% → 82% (no change)

Table 4: Qwen confidence-oriented steering failure summary.

A.4 Root Cause Analysis

The two models fail in different ways, but the common structural factor is that h^+ and h^- in the confidence-oriented pairs are extracted from different questions. Therefore, the mean-difference vector $v = E[h^+] - E[h^-]$ may capture question-level semantic variation (topic, difficulty, input length, etc.) in addition to correctness-related information. Such a vector can separate h^+/h^- well under projection (yielding high AUROC), but if the direction is dominated by question-level features, activation steering may not induce behavioral change along a correctness axis.

This observation motivates the main paper: high predictive separability (AUROC) does not guarantee causal validity of a probe(Hewitt and Liang, 2019; Pimentel et al., 2020), and multi-level validation including causal steering is needed to characterize what the representation captures.

A.5 Comparison with CCS Probe

Comparing the confidence-oriented probe and the CCS probe on the same Llama model shows a clear difference:

Probe	AUROC	$+\alpha v$ Steering	$-\alpha v$ Steering	Verdict
Confidence	0.921	P decreases (collapse)	P decreases (collapse)	×
CCS	0.830	P(Yes) increases ↑	P(Yes) decreases ↓	✓

Table 5: Comparison of confidence-oriented and CCS-probe steering (Llama, TriviaQA).

In the CCS probe, positive steering consistently moves the model toward the "true-answer" direction, whereas negative steering moves it toward the "false-answer" direction; the two directions act oppositely. This contrasts with the bidirectional collapse observed for the confidence-oriented probe and shows that the minimal-variation structure of CCS pairs is important for stable causal validation.

B Dataset Construction & Statistics

This section describes CCS pair construction, dataset selection, and quantitative statistics of the final datasets.

B.1 CCS Pair Construction

A CCS-style pair combines a true answer a^+ and a false answer a^- for the same question q, forming a minimal-variation contrast in which most of the input is shared and only correctness changes. The false-answer construction differs by dataset format.

TriviaQA (Open-ended). TriviaQA is an open-ended factual QA dataset and does not provide false answers. We therefore used GPT-4.1 to generate a plausible wrong answer for each question. The generation prompt was:

> *Given a question and its correct answer, generate ONE plausible but WRONG answer that: (1) Is the same type/category as the correct answer, (2) Is believable — someone who doesn't know the answer might think it's correct, (3) Is clearly factually wrong (not an alternative correct answer or alias), (4) Is similar in length to the correct answer.*

We used temperature = 0.7 and max_tokens = 100. If the generated wrong answer matched the ground-truth answer or one of its aliases, we regenerated it up to three times. Among 5,000 questions, 4,999 valid CCS pairs were retained.

ARC-Easy (4-choice MC). ARC-Easy is a four-choice science commonsense dataset with three provided distractors for each correct answer. We randomly selected one distractor as the false answer a^- (seed = 42). We merged the train, test, and validation splits, yielding 5,197 CCS pairs.

CommonsenseQA (5-choice MC). CommonsenseQA is a five-choice everyday commonsense dataset. We selected one false answer randomly using the same procedure as ARC-Easy. Merging the train and validation splits yielded 10,962 CCS pairs.

For each dataset, CCS pairs were split 50/50 into train and test sets (seed = 42), and three prompt templates were assigned uniformly.

B.2 Dataset Selection Rationale

Our probe extraction assumes that the model has sufficient parametric knowledge for the dataset questions. To check this prerequisite, we screened six candidate datasets using the FP16 accuracy and best-layer AUROC of Llama 3.1 8B Instruct.

Dataset	Format	Model Accuracy	Best AUROC	Decision
TriviaQA	Open-ended	67.6%	0.830	✓ selected
ARC-Easy	4-choice MC	87%	0.931	✓ selected
CommonsenseQA	5-choice MC	75%	0.838	✓ selected
SimpleQA	Open-ended	4.2%	0.595	× excluded
NQ-Open	Open-ended	35.2%	0.642	× excluded
PopQA	Open-ended	27.0%	—	× excluded

Table 6: Candidate dataset screening results.

SimpleQA (Wei et al., 2024) and NQ-Open (Kwiatkowski et al., 2019) have low model accuracies (4.2% and 35.2%, respectively), and their AUROC remains much lower than that of the selected datasets (0.595 and 0.642). This suggests that when the model lacks sufficient parametric knowledge, the probe may fail to extract a reliable correctness-related signal.

B.3 False Answer Quality

A potential concern is that false answers may contain trivial artifacts, allowing the probe to learn surface properties rather than correctness. We therefore applied the following quality controls.

Type consistency. The GPT-4.1 prompt explicitly required the false answer to have the same type/category as the true answer. For example, if the true answer is a person name, the false answer is also a person name; if the true answer is a year, the false answer is also a year. In the multiple-choice datasets, the choices are naturally drawn from the same category.

Length similarity. As shown in Table 7, the mean token-length ratio between true and false answers lies within 0.8–1.2 for all datasets, reducing the risk that answer length becomes the dominant discriminative cue. This is also consistent with the length-independence test in the main text (AUROC gap < 0.036).

Alias filtering. In TriviaQA, false answers that matched a ground-truth alias were automatically regenerated. In the final pairs, there were zero cases where true_answer == false_answer.

B.4 Dataset Statistics

Statistic	TriviaQA	ARC-Easy	CSQA
Total pairs	4,999	5,197	10,962
Train / Test	2,499/2,500	2,598/2,599	5,481/5,481
False source	GPT-4.1	MC distractor	MC distractor
Question len.	13.9	19.4	13.2
True ans. len.	1.7	3.8	1.5
False ans. len.	1.7	3.7	1.5
False/true len.	0.99	0.99	1.02
Templates	3	3	3

Table 7: Quantitative statistics of the final CCS-pair datasets.

B.5 Prompt Templates

We used three system-prompt templates when extracting hidden states. Each CCS pair was randomly assigned a template_id (0, 1, or 2), and probe training and evaluation used hidden states extracted under the assigned template.

Template ID	System Prompt
0	"You are a helpful assistant."
1	"You are a helpful assistant. Answer the user's question directly."
2	"You are a knowledgeable AI assistant that helps users with various tasks."

Table 8: System prompt templates used for hidden-state extraction.

The user message used the same format for all templates: `"Q: question A: answer"`. Since Mistral does not support a system prompt in the same way, we used only the user turn for Mistral.

Template-level AUROC variation was at most 0.022 across all conditions, indicating that the probe does not rely on template-specific artifacts.

B.6 Level 2 Structural Robustness

Table 9 reports the Level 2 structural robustness checks used in the main text. Template range is the AUROC range across the three prompt templates at the selected FP16 layer. Length gap is the AUROC difference between examples with small and large answer-length differences. Across all evaluated model–dataset conditions, template range is at most 0.022 and length gap is at most 0.036, supporting that the probe separability is not driven by

template-specific artifacts or answer-length differences.

Dataset	Model	Layer	Template Range	Length Gap
TriviaQA	Llama	14	0.022	0.036
TriviaQA	Qwen	18	0.020	0.010
TriviaQA	Mistral	13	0.014	0.013
ARC-Easy	Llama	14	0.022	0.016
ARC-Easy	Qwen	18	0.014	0.011
ARC-Easy	Mistral	15	0.013	0.010
CSQA	Llama	14	0.010	0.000
CSQA	Qwen	18	0.022	0.007
CSQA	Mistral	17	0.012	0.017

Table 9: Level 2 structural robustness results at the selected FP16 layer. Template Range denotes the AUROC range across prompt templates, and Length Gap denotes the AUROC gap between small and large answer-length-difference subsets.

B.7 CCS Pair Examples

The following are randomly sampled CCS-pair examples from each dataset (seed = 42).

TriviaQA:

Question	True Answer (a^+)	False Answer (a^-)
What type of animal can be found on a tin of 'Lyle's Golden Syrup'?	A LION	A TIGER
What homonymic word refers both to a pit and a hunted animal?	Quarry	Mole

Table 10: TriviaQA CCS pair examples.

Question	True Answer (a^+)	False Answer (a^-)
Many scientists theorize that global warming will melt the polar ice caps. What would most likely happen if the polar ice caps melted?	Ocean levels would increase.	Global precipitation would decrease.
What is the role of decomposers in a food chain?	They break down dead organic matter.	They consume other organisms.

Table 11: ARC-Easy CCS pair examples.

CommonsenseQA:

Question	True Answer (a^+)	False Answer (a^-)
The program kept getting errors, the amateur end user began to what?	get frustrated	write code
He fantasied about getting a what while driving to work and the pros and cons of the extra responsibilities and benefits?	promotion	new car

Table 12: CommonsenseQA CCS pair examples.

C Experimental Setup

This section describes the full experimental setup, including hidden-state extraction, probe training, causal steering, and quantization implementation.

C.1 Models

We used three instruction-tuned models in the 7–8B parameter range. The full HuggingFace identifiers are listed in Table 13.

Model	HuggingFace ID	Precision
Llama 3.1 8B	`meta-llama/Llama-3.1-8B-Instruct`	FP16
Qwen2.5 7B	`Qwen/Qwen2.5-7B-Instruct`	FP16
Mistral 7B v0.3	`mistralai/Mistral-7B-Instruct-v0.3`	FP16

Table 13: Models used in the experiments.

All models were loaded with `torch_dtype=torch.float16` and `device_map="cuda"`.

C.2 Hidden State Extraction

Extraction layer. We extracted from the residual stream corresponding to the output of each transformer block. Concretely, we captured the hidden state after the FFN sub-layer residual addition, immediately before it is passed to the next transformer block. We extracted from all layers of Llama (32 layers), Qwen (28 layers), and Mistral (32 layers).

Extraction token. For the CCS-pair input `"Q: question A: answer"`, we extracted the hidden state at the last token of the proposed answer. The answer-final token position was determined by separately tokenizing the answer text and matching it backwards within the full input_ids sequence (reverse token alignment). When repeated substrings were present, we used the last matching occurrence; when exact matching failed, we attempted partial token matching for robustness.

Input format. All templates used the following format:

```
[System: {template}]
[User: "Q: {question} A: {answer}"]
```

For Mistral, which does not support a system prompt in the same way, we used only the user turn.

Batching. We used batch_size = 4 and processed each sample without padding through an individual forward pass.

C.3 Probe Training

Mean-difference probe (primary). Using hidden states from the train split (50%), we define:

$$v_l = \mathbb{E}[h_l^+] - \mathbb{E}[h_l^-], \quad \hat{v}_l = v_l / |v_l|$$

Each layer was trained independently, and the train/test split was fixed with seed = 42.

Logistic Regression probe (secondary). We trained an L2-regularized logistic regression model (C = 1.0) after standardization with StandardScaler (solver = lbfgs, max_iter = 1000, random_state = 42). To obtain a probe direction, we transformed the learned weight vector back into the original hidden-state space by dividing by the StandardScaler scale, and then normalized it.

PCA probe (secondary). We used the first principal component of pairwise differences $(h^+ - h^-)$ as the probe direction (n_components = 1, random_state = 42). If its dot product with the mean-difference direction was negative, we flipped its sign.

C.4 Decoding Parameters

For P(Yes)-based causal steering, we used greedy decoding (temperature = 0). For free-form generation analyses (free-form α-sweep and negative steering test), we used max_new_tokens = 100 and temperature = 0. All evaluations used N = 200 samples. The steering vector was added to the residual stream at the same location described in Appendix C.2; the detailed steering protocol is described in Appendix D.

C.5 Quantization Implementation

We evaluated five quantization conditions in addition to the FP16 baseline. Table 14 summarizes the implementation setup.

Condition	Method/Backend	Implementation
AWQ	INT4 AWQ	Pre-quantized HuggingFace checkpoints
GPTQ	INT4 GPTQ	Pre-quantized HuggingFace checkpoints
RTN	INT4 round-to-nearest	Manual simulated weight quantization
W8A8	INT8 weights + INT8 activations	Manual simulated weight and activation quantization
W4A8	INT4 weights + INT8 activations	Manual simulated weight and activation quantization

Table 14: Quantization implementation summary.

For AWQ/GPTQ, we used community checkpoints from HuggingFace Hub with the same bit-width (INT4): Llama uses the `hugging-quants` AWQ/GPTQ checkpoints, Qwen uses the official Qwen AWQ/GPTQ checkpoints, and Mistral uses the `solidrust` AWQ and `thesven` GPTQ checkpoints. For RTN, W8A8, and W4A8, we started from the FP16 base models and applied manual simulated quantization as described below.

RTN (Round-to-Nearest). After loading the FP16 model, we manually applied per-group symmetric INT4 quantization to all `nn.Linear` weights (group_size = 128). We immediately dequantized the quantized weights back to FP16, yielding a simulate-quantize setup:

$$\begin{aligned} q &= \mathrm{round}(w/s), \\ w_{\text{quant}} &= q \cdot s, \\ s &= \max(|w_g|)/7. \end{aligned}$$

Here, w_g denotes a weight group of size 128, and q is clamped to the range [-7, 7].

W8A8. We combined RTN-style INT8 weight quantization with per-token dynamic INT8 activation quantization. Activation quantization was simulated by registering a forward hook on each transformer layer output and applying per-token symmetric INT8 quantize-dequantize:

$$\begin{aligned} q &= \mathrm{round}(a/s_t), \\ a_{\text{quant}} &= q \cdot s_t, \\ s_t &= \max(|a_t|)/127. \end{aligned}$$

Here, s_t is computed independently at each token position, and q is clamped to [-128, 127]. This implementation does not include the smoothing factor used by SmoothQuant and may therefore be somewhat more aggressive than a production W8A8 implementation.

W4A8. We combined RTN-style INT4 weight quantization with the same per-token dynamic INT8 activation quantization used for W8A8.

C.6 Hardware & Software

Component	Specification
GPU	NVIDIA GeForce RTX 3090 (24 GB VRAM)
PyTorch	torch, dtype=float16
transformers	4.46.0
autoawq	0.2.7.post3
auto-gptq	0.7.1
optimum	1.23.3
scikit-learn	Used for LR/PCA probe training
Random seed	42 (train/test split, template assignment, LR probe)

Table 15: Hardware and software environment.

For AWQ model loading, we set `fuse_layers=False` so that hidden-state hooks would function correctly.

D Causal Steering Implementation Details

This section details the implementation and evaluation protocol for Level 3 validation (causal steering).

D.1 Activation Addition

We performed causal steering with activation addition (Turner et al., 2024). At a target layer l, we add the normalized probe direction $\hat{v}_l$ with scaling factor α to the residual stream:

$$h'_l = h_l + \alpha \cdot \hat{v}_l$$

Here, $\hat{v}_l$ is the normalized mean-difference probe direction for the corresponding layer. A forward hook adds the steering vector to the target layer output, and subsequent layers proceed normally using the modified hidden state. The steering vector is applied identically to all token positions (broadcast addition across sequence positions). For P(Yes) measurement, it is applied once during the prefill forward pass; for free-form generation, it is applied at every decoding step.

D.2 Steering Protocol

We performed bidirectional steering under two conditions:

False + Positive steering ($+\alpha v$). We input the false answer (a^-) from a CCS pair and steer in the positive probe direction (+v). The expected effect is to make the model more likely to endorse the statement, increasing P(Yes).

True + Negative steering ($-\alpha v$). We input the true answer (a^+) from a CCS pair and steer in the opposite direction ($-v$). The expected effect is to make the model more likely to reject the statement, decreasing P(Yes).

D.3 P(Yes) Measurement

For each CCS pair, we input `"Q: question A: answer"` and extracted the next-token logits for the "Yes" and "No" tokens to compute P(Yes):

$$P(\text{Yes}) = \frac{\exp(\text{logit}_{\text{Yes}})}{\exp(\text{logit}_{\text{Yes}}) + \exp(\text{logit}_{\text{No}})}$$

For each tokenizer, we verified the token IDs corresponding to single-token "Yes"/"No" forms. We checked both leading-whitespace forms (e.g., " Yes", " No") and non-whitespace forms, and selected the form encoded as a single token in the model vocabulary. Steering effect was quantified as $\Delta P = P(\text{Yes})_{\text{steered}} - P(\text{Yes})_{\text{baseline}}$ relative to the baseline ($\alpha = 0$).

D.4 Alpha Selection

We used different α ranges for different evaluation goals. For main P(Yes) steering and random-baseline comparisons, we used $\alpha \in \{0, 0.5, 1.0, \ldots, 4.0\}$ to avoid saturation. For free-form behavioral sweeps, we used a wider range $\alpha \in \{2, 4, 6, 8, 10\}$ to ensure observable generation flips (Mistral saturated at $\alpha = 8$, so we report up to $\alpha = 8$). Both settings were chosen after a pilot sweep (N=20, $\alpha \in \{0.5, 1, 2, 4, 8, 16\}$) confirmed the absence of abnormal collapse in the selected range.

D.5 Random Baseline

To test whether the probe direction produces stronger steering than random directions, we constructed a random baseline. We sampled 10 random vectors from a standard normal distribution $\mathcal{N}(0, I)$, normalized them to unit length (seed = 123), and performed the same P(Yes)-based steering procedure with $\alpha \in \{0, 0.5, 1.0, \ldots, 4.0\}$.

We computed the z-score as follows:

$$z = \frac{|\Delta P_{\text{probe}}| - \text{mean}(|\Delta P_{\text{random}}|)}{\text{std}(|\Delta P_{\text{random}}|)}.$$

In the main validation analysis, z-scores are used as an effect-size reference together with random-direction comparisons, providing a conservative control for Level 3 steering validation.

D.6 Free-form Generation Evaluation

Because P(Yes) is based on a single next-token probability, it may diverge from actual generation behavior. We therefore added a free-form generation evaluation.

Protocol. For N = 200 test samples, we generated free-form responses with greedy decoding (temperature = 0, max_new_tokens = 100) and classified each response into one of three categories:

- Affirm: the model judges the proposed answer to be correct (e.g., "Yes", "correct", "that's right").
- Deny: the model judges the proposed answer to be incorrect (e.g., "No", "incorrect", "wrong").

- Ambiguous: the response does not fall into either category (e.g., hedging or irrelevant output).

We used a rule-based classifier. If the first generated token was "Yes" or "No", the response was classified immediately. Otherwise, we compared the frequency of affirm/deny keywords within the first 100 characters. Appendix E.7 discusses classifier implementation and validity.

Flip definition. We measured the causal effect of steering through the following behavioral flips:

- False + positive steering: a response that was deny at baseline becomes affirm after steering.
- True + negative steering: a response that was affirm at baseline becomes deny after steering.

D.7 Steering Layer Selection

The AUROC, steering, quantization robustness, and geometric analyses in the main paper were all performed at the FP16 layer with the highest AUROC. This criterion was not adjusted after observing the analysis results. To check for possible layer cherry-picking, Appendix G reports a full-layer sweep. The selected layers used in the main paper are:

Model	TriviaQA	ARC-Easy	CommonsenseQA
Llama	L14	L14	L14
Qwen	L18	L18	L18
Mistral	L13	L15	L17

Table 16: Selected layer for each model–dataset pair.

In the full-layer behavioral sweep, the layer with the strongest steering effect may differ from Table 16. This sweep is an additional validation that steering effects are observed across a broad neighborhood around the selected layer; detailed results are given in Appendix G.3.

D.8 Quantized Steering Preservation

To evaluate whether steering is preserved under quantization, we applied the probe direction learned from FP16 directly to the corresponding quantized model and ran the same steering protocol. Concretely:

1. Apply $\hat{v}_l$ learned at the selected FP16 layer to the same layer of the quantized model.
2. Measure ΔP on the same test samples and with the same α.
3. Quantify the FP16-relative steering preservation rate as $\Delta P_{\text{quant}}/\Delta P_{\text{fp16}}$ when ΔP_{fp16} is sufficiently large for the ratio to be stable.

For quantized models, we evaluated both monotonicity preservation in P(Yes) and preservation of the absolute magnitude of ΔP. Preservation rates were computed only when $|\Delta P_{\text{fp16}}| \geq 0.05$; otherwise, we compared absolute ΔP values directly.

E Statistical Testing Methodology

This section summarizes the statistical testing procedures used in the main paper and appendices.

E.1 AUROC Computation

For each layer, we computed the AUROC of the mean-difference probe as follows. We projected test-split h^+ and h^- hidden states onto the probe direction v to obtain scalar scores, labeled h^+ as the positive class (label = 1) and h^- as the negative class (label = 0), and used sklearn's `roc_auc_score`. The sign of the probe direction was canonicalized on the train split so that $\text{mean}(h^+) \cdot v > \text{mean}(h^-) \cdot v$. Thus, test AUROC is always at least 0.5. LR and PCA probes use the same projection-based AUROC computation.

E.2 Bootstrap Confidence Intervals

To evaluate AUROC stability, we performed bootstrap resampling (Appendix F.1). We resampled the train split with replacement, retrained the probe for each trial, and measured AUROC on the fixed test split. The 95% confidence interval was computed from the 2.5th and 97.5th percentiles of the bootstrap distribution (100 trials, seed = 42). We chose the number of trials after confirming CI convergence, and the bootstrap standard deviation was below 0.01 in all conditions.

E.3 Random Baseline Z-Score

To compare causal-steering effects against random directions, we computed a z-score. We sampled 10 unit vectors from $\mathcal{N}(0, I)$ (seed = 123), steered with the same α, and constructed an empirical distribution of $|\Delta P|$:

$$z = \frac{|\Delta P_{\text{probe}}| - \mu_{\text{random}}}{\sigma_{\text{random}}}$$

This z-score is reported as an effect-size summary that contextualizes probe steering relative to the random-direction distribution, rather than

as a strict binary validity criterion. In the high-dimensional hidden space ($d \approx 4096$), 10 random directions have expected pairwise cosine approximately $1/\sqrt{d} \approx 0.016$, so they function as approximately orthogonal control directions.

E.4 Asymmetry Statistical Testing (BH FDR)

To test the asymmetric structure of quantization error discussed in the main text, we performed statistical tests over layers within the stable probe zone.

Stable Probe Zone. We analyze only layers satisfying $\cos(v, \text{signal_dir}) \geq 0.90$, where signal_dir denotes the normalized vector $\text{mean}(h^+) - \text{mean}(h^-)$ on the test split. We additionally verified that varying either the signal-alignment threshold or the FP16 AUROC threshold preserved the same qualitative asymmetry pattern. Full threshold sensitivity results are provided in Appendix I.

Because stable-zone layer values are used as observations, we interpret results with possible correlations between adjacent layers in mind.

We used three tests: (1) a one-sample t-test for $\cos(v, \bar{e}^+)$, (2) a one-sample t-test for $\cos(v, \bar{e}^-)$, and (3) a paired t-test comparing $\cos(v, \bar{e}^+)$ and $\cos(v, \bar{e}^-)$ within the same layer.

Each test type was run independently for the 3 models $\times$ 5 quantization settings, yielding 15 model–quantization conditions. Benjamini–Hochberg (BH) false discovery rate (FDR) correction (Benjamini and Hochberg, 1995) was applied separately for each test family (family size = 15), with adjusted $p < 0.05$ as the significance criterion.

E.5 Effect Size: Cohen's d

To quantify the size of the asymmetry, we computed Cohen's d for the paired differences between cos(v, $\bar{e}^+$) and cos(v, $\bar{e}^-$) over stable-zone layers:

$$d = \frac{\bar{x}_{\text{diff}}}{s_{\text{diff}}}$$

Here, $\bar{x}_{\text{diff}}$ is the mean paired difference and s_{diff} is its standard deviation. Under the conventional interpretation (Cohen, 2013), $|d| > 0.8$ is a large effect. The mean $|d|$ observed in this study is 2.12–3.89, well above this threshold in all conditions.

E.6 Spearman Rank Correlation

We evaluated the cross-condition association between $|\cos(v, \bar{e}^+)|$ and AUROC degradation Δ at the selected layer using Spearman rank correlation (3 models $\times$ 5 quantization settings; $N = 15$). The correlation reported in the main text is not significant ($\rho = 0.13$, $p = 0.64$), indicating that mean-level alignment magnitude alone does not explain degradation magnitude. We therefore complement this analysis with the sample-level gap-shrinkage metric shown in Figure 2(f) of the main paper.

E.7 Free-form Generation Classification

The behavioral effect of causal steering was quantified with a rule-based classifier (Appendix D.6). If the first token of the generated response was "Yes" or "No", we immediately labeled it as affirm/deny. Otherwise, we compared affirm/deny keyword counts in the first 100 characters. The same classifier was applied to both probe and random-baseline generations.

F Additional Validation Results

This section provides detailed values for the causal-validation results summarized in the main text.

F.1 Bootstrap AUROC Stability

To verify that the AUROC of the confidence-oriented probe is not an artifact of a particular set of train examples, we performed a bootstrap stability test. We resampled the train split with replacement, retrained the probe at each trial, and measured AUROC on the fixed test split (100 trials, seed = 42).

Model	Layer	Original AUROC	Bootstrap Mean	Std	95% CI
Llama	L29	0.9082	0.9022	0.0114	[0.8777, 0.9164]
Llama	L30	0.9204	0.9171	0.0077	[0.9016, 0.9284]
Llama	L31	0.9207	0.9159	0.0048	[0.9052, 0.9230]
Qwen	L18	0.8891	0.8842	0.0093	[0.8593, 0.8968]
Qwen	L19	0.8926	0.8901	0.0064	[0.8771, 0.9011]
Qwen	L20	0.8901	0.8878	0.0047	[0.8771, 0.8959]

Table 17: AUROC bootstrap stability (Llama and Qwen, TriviaQA, confidence-oriented probe). We report this test for the confidence-oriented probe because its question-disjoint pair structure makes it more susceptible than CCS pairs to artifacts from the particular h^+/h^- pairing. The table reports the best layer and adjacent layers for each model.

Across all conditions, bootstrap standard deviations are small and the original AUROC lies within the 95% CI. Thus, the predictive separability of the confidence-oriented probe itself is stable, but as shown in Appendix A.3, stable AUROC does not guarantee causal validity.

F.2 Freeform α-Sweep

To check agreement between P(Yes) and actual generation behavior, we measured behavioral flip counts as α increases, using layers where the full-layer sweep showed strong behavioral effects. For each test sample (N=200), we aggregated flips under false + positive steering and true + negative steering.

α	Llama (L16)	Qwen (L17)	Mistral (L19)
2	64	56	31
4	137	122	110
6	252	193	149
8	277	208	177
10	281	226	—

Table 18: Bidirectional flip counts for the probe direction, summed over valid-baseline subsets. Mistral is evaluated up to $\alpha_{\max} = 8$.

α	Llama	Qwen	Mistral
2	3.4	0.5	0.7
4	7.2	1.7	1.3
6	11.0	1.9	2.1
8	17.9	2.6	3.5
10	27.4	3.2	—

Table 19: Random-baseline flip counts, averaged over 10 random directions. Mistral is evaluated up to $\alpha_{\max} = 8$.

Flip counts sum the two steering conditions (false + positive steering and true + negative steering) and are normalized over the valid-baseline subset where the baseline behavior is in the expected direction.

Model	Flips	Rate	Rand.	Ratio
Llama (α=10)	281	100.0%	27.4	10×
Qwen (α=10)	226	77.9%	3.2	71×
Mistral (α=8)	177	63.9%	3.5	51×

Table 20: Summary at the largest α (Llama/Qwen: $\alpha = 10$, Mistral: $\alpha = 8$). Valid-baseline counts are 281, 290, and 277 for Llama, Qwen, and Mistral, respectively; Rand. denotes the random-baseline average.

For all models, the probe direction yields consistently larger behavioral flip rates than random directions.

F.3 Random Baseline Z-Score

To evaluate the strength of P(Yes)-based steering, we computed z-scores ($\alpha = 4.0$, 10 random directions; see Appendix D.5). The z-score contextualizes the probe steering effect against the empirical random-direction distribution.

Dataset	Model	Probe ΔP	Random mean ΔP	Z-score
TriviaQA	Llama	+0.447	−0.011	5.99
TriviaQA	Qwen	+0.120	−0.002	15.55
TriviaQA	Mistral	+0.495	+0.146	1.79†
ARC-Easy	Mistral	+0.854	+0.047	9.11
CSQA	Mistral	+0.243	−0.007	7.39

Table 21: Random baseline z-score at the selected FP16 layer (α=4.0, 10 random directions). †Lower than the remaining conditions, but still stronger than all sampled random directions. Mistral satisfies larger z-scores on both remaining datasets.

For all models, the steering effect of the probe direction is consistently larger than the random-direction distribution.

F.4 Negative Steering Test

To test whether the probe direction is correctness-aligned rather than merely reflecting a positivity bias, we evaluated behavioral changes under the full 2×2 set of conditions (true/false × positive/negative steering).

Condition	Llama	Qwen	Mistral
true+(+v): affirm retain	98.7%	100.0%	100.0%
false+(−v): deny retain	100.0%	82.9%	100.0%
false+(−v): affirm→deny	—	—	23/93 (24.7%)
true+(+v): deny→affirm	—	—	9/21 (42.9%)

Table 22: Negative steering test results (TriviaQA, $\alpha = 4.0$).

Interpretation: the high retention rates under aligned steering conditions (true+(+v), false+(−v)) indicate that the probe direction is not merely pushing all outputs toward one polarity. Instead, it preserves behavior in a direction consistent with the original judgment state (affirm/deny). For Mistral, steering also corrects some misclassified samples.

F.5 h^- Randomization Test

To test whether the probe depends on the CCS paired structure, we performed two randomization tests.

Test A: Random Label Swap. We randomly swapped the h^+/h^- labels for 50% of train-split samples and retrained the probe (100 trials).

Test B: $h^- \rightarrow$ Random h^+ Replacement. We replaced h^- with h^+ from another sample, destroying

the true/false contrast, and retrained the probe (100 trials).

Test	Metric	Llama	Qwen	Mistral
Base	Original AUROC	0.830	0.776	0.824
A	Swapped AUROC	0.617 ±0.069	0.573 ±0.058	0.616 ±0.078
A	AUROC Drop	0.212	0.204	0.208
B	Replaced AUROC	0.533 ±0.024	0.530 ±0.023	0.528 ±0.026
B	AUROC Drop	0.297	0.247	0.296

Table 23: h^- randomization test results. All models show structure-dependent behavior.

In Test B, all three models converge to near-chance performance (AUROC ≈ 0.53). This shows that the probe strongly depends on the CCS paired structure and is unlikely to be explained by a simple one-sided feature bias.

G Layer Selection & Full-Layer Analysis

This section presents full-layer AUROC distributions and steering results to show that the best-layer results reported in the main paper are not due to cherry-picking.

G.1 Layer Selection Rationale

The main results are reported at the selected layer with the highest FP16 AUROC. The same criterion is used for AUROC, steering, quantization, and geometric analyses.

- AUROC reporting: select the FP16 layer with the highest AUROC after a full-layer sweep.
- Steering evaluation: evaluate at the same selected layer.
- Quantization robustness: compute Δ at the same selected layer.
- Geometric analysis: use the same selected layer.

The full-layer behavioral sweep is an additional check that this choice is not a layer-local artifact.

G.2 Full-Layer AUROC Profiles

Table 24 summarizes full-layer AUROC profiles for 3 models $\times$ 3 datasets. Many layers beyond the selected layer also show high AUROC, indicating that correctness separability is not confined to a single layer.

Model	Dataset	Selected Layer	Best AUROC	Top-5 Layers	Mean AUROC
Llama	TriviaQA	L14	0.830	L14, L16, L15, L17, L11	0.749
Llama	ARC-Easy	L14	0.931	L14, L13, L15, L12, L16	0.813
Llama	CommonsenseQA	L14	0.838	L14, L11, L13, L12, L15	0.759
Qwen	TriviaQA	L18	0.776	L18, L19, L17, L20, L16	0.629
Qwen	ARC-Easy	L18	0.964	L18, L19, L20, L21, L17	0.761
Qwen	CommonsenseQA	L18	0.865	L18, L17, L19, L15, L16	0.683
Mistral	TriviaQA	L13	0.824	L13, L14, L12, L15, L16	0.711
Mistral	ARC-Easy	L15	0.923	L15, L14, L16, L13, L17	0.804
Mistral	CommonsenseQA	L17	0.856	L17, L18, L12, L16, L13	0.780

Table 24: Full-layer AUROC summary (FP16).

For all models, the top-5 layers are contiguous or near-contiguous, and the mean AUROC ranges from 0.629 to 0.813, clearly above chance-level separability (0.5). This suggests that correctness separability is distributed across a broad layer range.

G.3 Full-Layer Steering Sweep

To test whether the causal steering effect is confined to the AUROC-selected layer or distributed across neighboring layers, we performed steering sweeps across all transformer layers. For each layer, we evaluated steering with $\alpha \in \{0, 0.5, \ldots, 4, 5, 6, 8, 10\}$ using 5 seeds and 200 samples per seed.

Across all models, steering effects emerge across contiguous mid-to-late layer regions rather than at a single isolated layer. In particular, neighboring layers around the AUROC-selected layer often exhibit similar monotonic steering behavior, suggesting that the selected layer reflects a broader causal region rather than a narrow layer-local artifact.

PASS layers satisfy both monotonic response with increasing α and max $|\Delta P| > 0.01$. Table 25 summarizes the corresponding PASS ranges and maximum steering magnitudes.

Model	Report Layer	fp ($+\alpha v$) PASS Range	tn ($-\alpha v$) PASS Range	Max fp ΔP	Max tn ΔP
Llama	L16	L16–L29, L31	L1–L8, L12–L29, L31	+0.604	−0.805
Qwen	L17	L8–L23	L0–L1, L9–L23	+0.451	−0.801
Mistral	L19	L16–L31	L0, L3, L5–L8, L15–L31	+0.477	−0.889

Table 25: Layer ranges with consistent steering effects on TriviaQA in FP16. The report layer denotes the layer with the strongest behavioral effect in the free-form sweep. PASS indicates layers satisfying both monotonic response across α and max $|\Delta P| > 0.01$.

Figure 1 provides compact summaries of steering curves for the top-3 AUROC layers for all models. Figure 2 shows the full-layer steering sweep across all transformer layers on TriviaQA dataset.

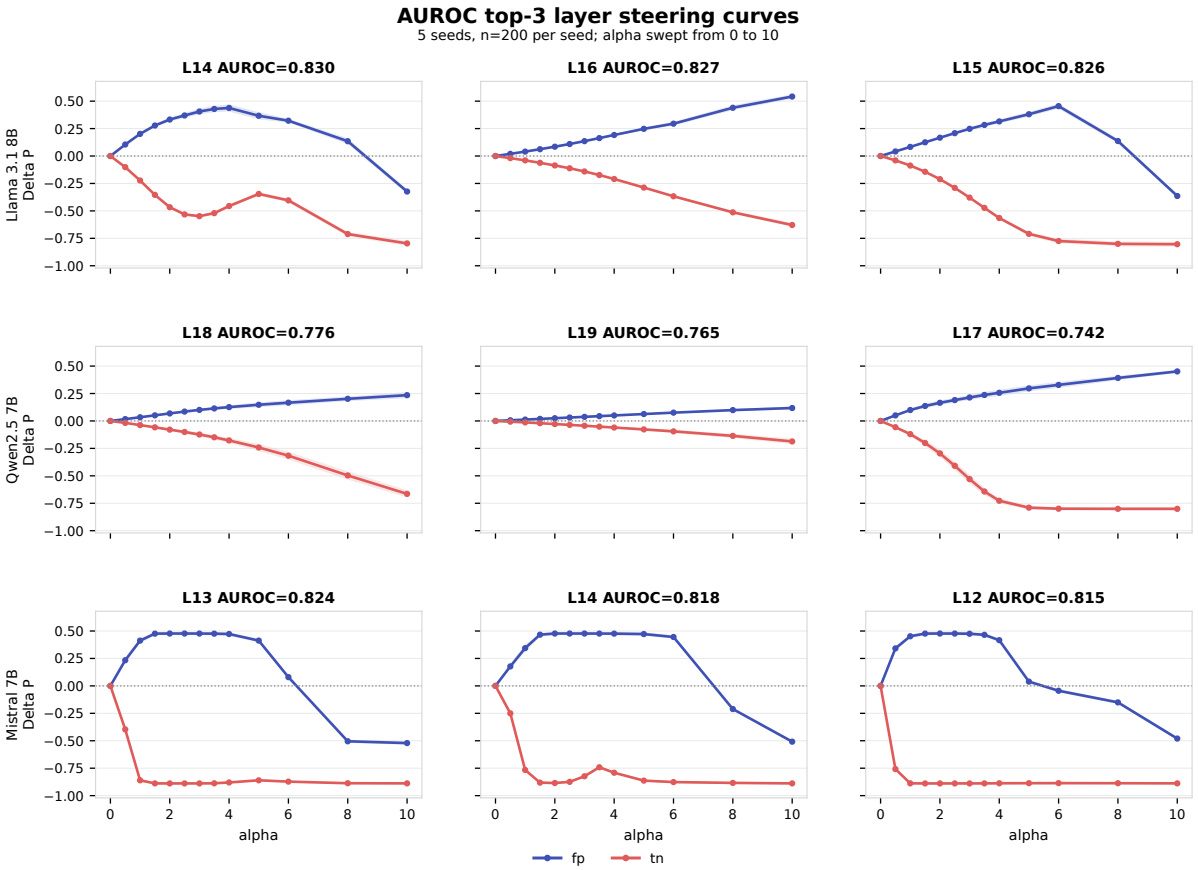

Figure 1: Compact summaries of steering behavior at the top-3 AUROC layers for Llama 3.1 8B, Qwen2.5 7B, and Mistral 7B.

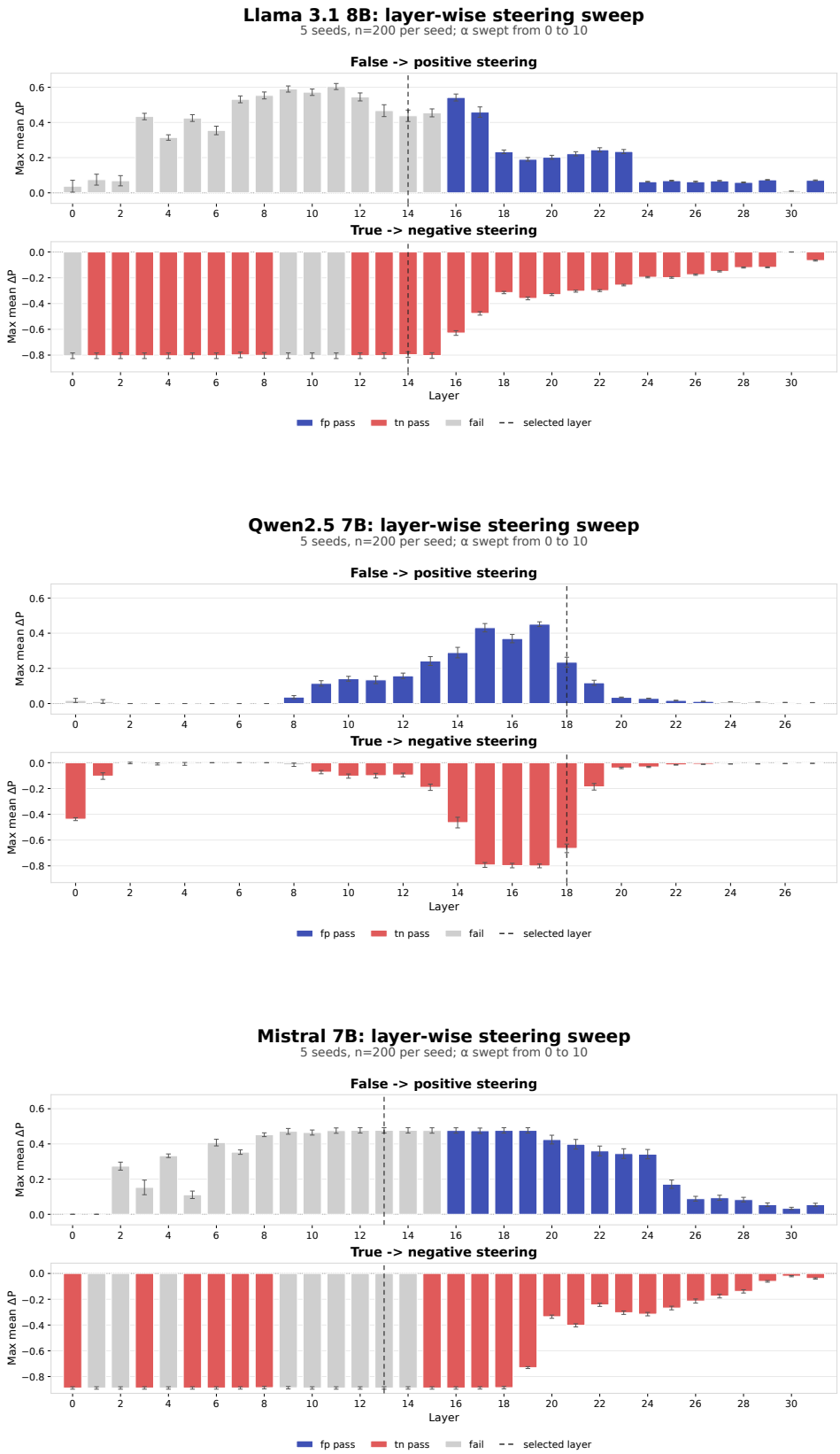

Figure 2: Full-layer steering sweep across all transformer layers. Bars indicate the maximum mean ΔP observed at each layer, and error bars denote standard deviation across five seeds. Colored bars indicate condition-specific PASS layers, while gray bars denote layers failing either the monotonicity criterion or the minimum effect-size threshold. The dashed vertical line marks the AUROC-selected layer used in the main analysis.

G.4 Quantization AUROC Degradation: Layer-Averaged vs Selected-Layer

To verify that quantization-robustness results are not an artifact of selected-layer cherry-picking, we compare selected-layer Δ with the all-layer mean Δ.

Condition	Llama Selected	Llama Mean	Qwen Selected	Qwen Mean	Mistral Selected	Mistral Mean
W8A8	0.001	0.002	0.002	0.002	0.005	0.003
AWQ	0.012	0.010	0.020	0.006	0.013	0.010
GPTQ	0.017	0.014	0.010	0.007	0.040	0.034
RTN	0.021	0.018	0.004	0.004	0.031	0.028
W4A8	0.022	0.019	0.009	0.007	0.034	0.029

Table 26: Comparison between selected-layer Δ and all-layer mean Δ (TriviaQA). The selected layer is the FP16 layer with the highest AUROC for each model.

Although selected-layer and all-layer mean Δ values differ slightly by condition, the main ranking and interpretation are preserved. W8A8 has the smallest degradation for all models, and Mistral GPTQ/RTN/W4A8 shows relatively large Δ both at the selected layer and on average. Thus, the quantization-robustness conclusion does not depend on a single selected layer.

G.5 Layer-wise Pattern of Geometric Structure

To check that the asymmetric anti-alignment pattern observed in the Phase 3 geometric analysis (ErrorCos, cos(v, $\bar{e}^+$), cos(v, $\bar{e}^-$)) is not confined to a single layer, we analyzed the layer-wise pattern. The asymmetric pattern becomes consistently observable after a transition region around L6 for Llama and around L15 for Qwen, and this transition is similar across all five quantization conditions. This suggests that the observed geometric pattern is model-consistent rather than specific to a particular quantization method.

H Per-Condition Quantization Results

This section expands the quantization-robustness results by condition. All values are computed at the FP16 selected layer of the TriviaQA CCS probe. We define $\Delta = \text{AUROC}_{\text{FP16}} - \text{AUROC}_{\text{quant}}$; positive values indicate AUROC degradation after quantization.

H.1 Selected-Layer Quantization Results

Model	Metric	FP16	W8A8	AWQ	GPTQ	RTN	W4A8	Max Δ
Llama L14	AUROC	0.830	0.829	0.818	0.812	0.808	0.807	–
Llama L14	Δ	–	0.001	0.012	0.017	0.021	0.022	0.022
Qwen L18	AUROC	0.776	0.774	0.757	0.766	0.773	0.767	–
Qwen L18	Δ	–	0.002	0.020	0.010	0.004	0.009	0.020
Mistral L13	AUROC	0.824	0.819	0.810	0.784	0.792	0.789	–
Mistral L13	Δ	–	0.005	0.013	0.040	0.031	0.034	0.040

Table 27: Selected-layer AUROC and degradation Δ by quantization condition (TriviaQA).

Table 27 reports both the selected-layer AUROC and its FP16-relative degradation. AUROC decreases relative to FP16 in all conditions, but the drop is limited. W8A8 is generally the most robust condition, and the largest degradation is observed for Mistral GPTQ. These values provide the detailed selected-layer counterpart to the main-paper quantization summary.

H.2 Geometric Error Summary

`ErrorCos` summarizes the directional similarity between the mean quantization errors of the positive and negative CCS conditions, $\cos(\bar{e}^+, \bar{e}^-)$. Table 28 reports condition-level averages over stable probe-zone layers. Across all model–quantization conditions, `ErrorCos` remains high (0.939–0.987), indicating that quantization errors for h^+ and h^- tend to occur in highly correlated directions.

Model	W8A8	AWQ	GPTQ	RTN	W4A8
Llama	0.987	0.969	0.979	0.982	0.983
Qwen	0.976	0.966	0.939	0.968	0.968
Mistral	0.979	0.964	0.954	0.944	0.943

Table 28: Condition-level mean `ErrorCos` over stable probe-zone layers (TriviaQA).

To examine the directionality of quantization error, we compare $\cos(v, \bar{e}^+)$ and $\cos(v, \bar{e}^-)$ for each condition. Under the stable probe-zone criterion, $\cos(v, \bar{e}^+)$ is generally negative (range: -0.150 to 0.022; median: -0.091), whereas $\cos(v, \bar{e}^-)$ lies near zero (range: -0.029 to 0.051; median: 0.017).

This result shows that quantization error has a consistent asymmetric orientation with respect to the correctness probe direction. Applying paired t-tests with BH FDR correction in the stable probe zone, we find that the difference between $\cos(v, \bar{e}^+)$ and $\cos(v, \bar{e}^-)$ is significant for all model–quantization conditions (paired t-test p <

1.34e-02 across conditions; mean |Cohen's d| = 2.89). On average, 95% of stable-zone layers satisfy $\cos(v, \bar{e}^+) < \cos(v, \bar{e}^-)$.

Degradation magnitude is not explained by mean-level alignment alone; the sample-level gap-shrinkage analysis in Figure 3(f) of the main paper complements this view. The sample-level gap shrinkage measures how much quantization reduces probe-score separation for each paired sample. For sample i, define $\text{gap}_i = v^\top h_i^+ - v^\top h_i^-$. Let $\text{gap}_i^{\text{FP16}}$ and gap_i^Q denote the gaps under FP16 and quantized model Q, respectively. We compute:

$$\text{shrink}_i = \text{gap}_i^{\text{FP16}} - \text{gap}_i^Q,$$
$$\texttt{RelShrink} = \frac{\mathbb{E}_i[\text{shrink}_i]}{\mathbb{E}_i[\text{gap}_i^{\text{FP16}}]}.$$

Equivalently, when $e_i^+ = h_{i,Q}^+ - h_{i,\text{FP16}}^+$ and $e_i^- = h_{i,Q}^- - h_{i,\text{FP16}}^-$, $\text{shrink}_i = v^\top e_i^- - v^\top e_i^+$. Thus, positive shrinkage means that quantization reduces the h^+/h^- probe-score separation for that sample.

For both the mean-difference and LR probes, the mean shrinkage is positive in all conditions (Table 29), and condition-wise one-sided Wilcoxon tests all have $p < 0.05$.

Probe	Positive shrinkage	Relative shrinkage	$P(\text{shrink}_i > 0)$	AUROC Δ
MD	15/15	0.1140	0.632	0.016
LR	15/15	0.1145	0.589	0.029

Table 29: Sample-level gap shrinkage summary (TriviaQA, 3 models $\times$ 5 quantization conditions).

I Sensitivity Analysis

This section checks that the stable probe-zone definition in Appendix E.4 is not overly sensitive to a particular threshold choice. We use two threshold families.

- Signal-alignment threshold: cos(v, signal_dir) $\geq$ 0.70, 0.75, 0.80, 0.85, 0.90, 0.95
- Predictive-separability threshold: FP16 AUROC $\geq$ 0.60, 0.65, 0.70, 0.75, 0.80

I.1 Threshold Robustness

Criterion	Threshold	Conditions	Layers	Mean cos $(v, \bar{e}^+)$	Mean cos $(v, \bar{e}^-)$	Asym. rate	BH sig.
cos(v, signal)	0.70	10	220	-0.086	0.009	0.958	10/10
cos(v, signal)	0.80	10	210	-0.087	0.013	0.958	10/10
cos(v, signal)	0.90	10	185	-0.089	0.020	0.954	10/10
cos(v, signal)	0.95	10	150	-0.094	0.025	0.952	10/10
FP16 AUROC	0.60	10	215	-0.087	0.011	0.958	10/10
FP16 AUROC	0.70	10	155	-0.098	0.022	0.956	10/10
FP16 AUROC	0.80	5	65	-0.086	0.018	0.954	5/5

Table 30: Stable-zone threshold sensitivity.

Even under more conservative thresholds, $\cos(v, \bar{e}^+)$ remains consistently negative, while $\cos(v, \bar{e}^-)$ remains near zero or positive. After BH FDR correction, the asymmetry test remains significant for all included conditions. Therefore, the asymmetric error-alignment conclusion is robust to the stable-zone threshold.

I.2 Interpretation of Reduced Condition Counts

Under FP16 AUROC $\geq$ 0.75 or 0.80, the number of included conditions decreases to five because some model–quantization conditions do not contain enough layers satisfying the threshold. However, the sign pattern and asymmetry rate remain in the high-separability subset. Thus, increasing the threshold does not remove the conclusion; the same directional result persists in a more conservative subset.

This analysis is therefore used only to check sensitivity to the stable-zone definition, not to select a favorable threshold. The main geometric conclusion is preserved under both signal-alignment and predictive-separability criteria.

The reduced-condition rows should be interpreted as stricter diagnostic subsets rather than separate experimental settings. They retain only layers where the probe direction is already highly stable or highly separable, so they test whether the observed asymmetry survives after excluding weaker layers. The persistence of the same sign pattern in these subsets supports the claim that the effect is not driven by marginal layers near the inclusion boundary.

J LR / PCA Probe Comparison

The main analysis uses the mean-difference CCS probe. This section compares logistic regression (LR) and PCA directions on the same paired representations to test whether the results depend on a single probe-construction method.

J.1 Predictive AUROC Comparison

Model	Dataset	Mean-diff	LR	PCA
Llama	TriviaQA	0.830	0.813	0.776
Llama	ARC-Easy	0.931	0.926	0.907
Llama	CommonsenseQA	0.838	0.820	0.727
Qwen	TriviaQA	0.776	0.772	0.531
Qwen	ARC-Easy	0.964	0.956	0.550
Qwen	CommonsenseQA	0.865	0.849	0.603
Mistral	TriviaQA	0.824	0.810	0.773
Mistral	ARC-Easy	0.923	0.918	0.907
Mistral	CommonsenseQA	0.856	0.815	0.788

Table 31: Best FP16 AUROC across layers by probe type.

Mean-difference and LR achieve similar AUROC in most conditions. PCA captures useful separability in some conditions, but its performance drops sharply for Qwen, making it less stable as a probe direction.

J.2 Quantization Degradation by Probe Type

Table 32 compares quantization-induced AUROC degradation for mean-difference and LR probes on TriviaQA. Values are computed at the selected FP16 layer for each model and averaged across the five quantization settings. Both probe types show limited degradation, but LR exhibits consistently larger mean degradation than the mean-difference probe.

Model	MD mean Δ	LR mean Δ	MD max Δ	LR max Δ
Llama	0.015	0.025	0.022	0.040
Qwen	0.009	0.031	0.020	0.045
Mistral	0.025	0.031	0.040	0.056

Table 32: Quantization-induced AUROC degradation by probe type at the selected TriviaQA layer. Mean and maximum Δ are computed across five quantization settings.

J.3 Steering Comparison at the Selected Layer

Model	Dataset	Probe	AUROC	max fp ΔP	min tn ΔP
Llama	TriviaQA	LR	0.813	+0.073	-0.068
Llama	TriviaQA	PCA	0.754	+0.565	-0.776
Qwen	TriviaQA	LR	0.771	+0.000	+0.000
Qwen	TriviaQA	PCA	0.517	+0.035	-0.048
Mistral	TriviaQA	LR	0.803	+0.425	-0.850
Mistral	TriviaQA	PCA	0.753	+0.427	-0.857
Llama	ARC-Easy	LR	0.923	+0.320	-0.501
Llama	ARC-Easy	PCA	0.906	+0.861	-0.789
Qwen	ARC-Easy	LR	0.956	+0.006	-0.054
Qwen	ARC-Easy	PCA	0.508	+0.000	+0.000
Mistral	ARC-Easy	LR	0.903	+0.823	-0.855
Mistral	ARC-Easy	PCA	0.901	+0.823	-0.855

Table 33: LR/PCA steering effect at selected layers. Values are max ΔP for false-positive correction (+v) and min ΔP for true-negative suppression (-v).

LR can have high predictive AUROC while showing weak steering in some Qwen conditions. PCA provides strong steering directions for Llama and Mistral, but both AUROC and steering are weak for Qwen. Therefore, the main paper uses the mean-difference CCS probe because it gives the most stable balance between predictive performance and causal manipulability.

K Multi-Dataset Steering Results

This section expands the multi-dataset causal-steering results from the main text. All experiments use the FP16 selected layer for each model–dataset pair, and steering strength is summarized by the maximum positive correction and maximum negative suppression observed over the α sweep.

K.1 Main Multi-Dataset Results

Dataset	Model	Layer	AUROC	max fp ΔP	min tn ΔP
TriviaQA	Llama	L14	0.830	+0.44	-0.80
TriviaQA	Qwen	L18	0.776	+0.24	-0.66
TriviaQA	Mistral	L13	0.824	+0.42	-0.93
ARC-Easy	Llama	L14	0.931	+0.87	-0.73
ARC-Easy	Qwen	L18	0.964	+0.11	-0.63
ARC-Easy	Mistral	L15	0.923	+0.85	-0.86
CommonsenseQA	Llama	L14	0.838	+0.63	-0.64
CommonsenseQA	Qwen	L18	0.865	+0.19	-0.56
CommonsenseQA	Mistral	L17	0.856	+0.65	-0.77

Table 34: Multi-dataset CCS probe validation.

Across all three datasets, the CCS probe shows both high AUROC and bidirectional steering effects. ARC-Easy has the highest AUROC, and TriviaQA and CommonsenseQA also show consistent behavioral change as α increases. Qwen's positive-steering effect is smaller than that of the other models, but its negative-steering effect is clear on all three datasets.

K.2 Excluded Dataset Candidates

SimpleQA and Natural Questions were excluded after preliminary screening. Under Level 1/2 criteria, these datasets did not reliably satisfy the prerequisites for CCS-style true/false pair construction because model parametric knowledge or answer-generation stability was insufficient. We therefore use TriviaQA, ARC-Easy, and CommonsenseQA as the main validation datasets, and report excluded candidates only in the dataset-screening results in Appendix B.

L Use of AI Assistants

AI assistants were used only for limited English proofreading and grammar correction. All research design, experiments, analysis, and conclusions were conducted and verified by the authors.