

Layer Selection Sensitivity in Hidden-State Safety Probes for Reasoning Models

Apex Intelligence

Abstract

Reasoning models can self-jailbreak: their chain of thought transitions from safe to harmful content without adversarial input. Detecting when this shift occurs (the commitment point) requires real-time monitoring, and hidden-state probes offer a direct window into the model’s internal reasoning. Yet which transformer layer to read is routinely treated as an implementation detail. We show this choice matters. Across four models (two architectures, two scales), crossing-rate metrics systematically select shallow layers (L0–L10) while threshold-based metrics select deep layers (L14–L63), a divergence of 14 to 61 layers. This gap reverses whether hidden states outperform text: at deep layers, hidden-state probes gain +20–22 percentage points (pp) precision over text (cross-validated at 7–8B); at shallow layers, the two modalities perform equivalently. The mechanism is cosine compression: shallow representations compress directional information, inflating frequency-based metrics near the noise floor. This bias generalizes across models and a second dataset (AdvBench), and the geometric precondition extends to a fifth architecture (Gemma-2). The advantage diminishes with stronger text encoders and remains suggestive at 32B; seven alternative detection methods fail to match supervised probing. Layer selection metric is a critical, overlooked specification choice in safety probing studies.¹

1 Introduction

Reasoning models such as DeepSeek-R1 (DeepSeek-AI, 2025) and QwQ (Qwen Team, 2024) can “self-jailbreak”: their chain of thought transitions from safety-aware deliberation to committed harmful generation without adversarial input (Kakkar et al., 2026; Yong and Bach, 2025). We call the first reasoning step after

which safety deliberation permanently ceases the **Safety Commitment Point** (CP). Existing safety detectors operate at the token or input level (Liu et al., 2026; Han et al., 2025; Kim et al., 2026), missing the gradual, multi-step commitment process within reasoning traces. Hidden-state (HS) probing offers a natural alternative for step-level monitoring, but requires choosing which transformer layer to read. This choice is typically treated as an implementation detail.

We show it determines whether the probe appears to succeed or fail (Fig. 1). Across four models, two architectures, and two scales (Table 1), a crossing-rate composite metric systematically selects shallow layers (L0–L10), where HS and text probes perform equivalently. Threshold-based metrics select deep layers (L14–L63), where HS probes gain +20–22pp precision, cross-validated at 7–8B. The divergence spans 14 to 61 layers. The mechanism is cosine compression (Fig. 2): shallow layers compress step-to-step directional information by $1.4\text{--}2.7\times$, inflating frequency-based metrics near the noise floor. This bias generalizes across all models, a second dataset (AdvBench), and a fifth architecture (Gemma-2).

The HS advantage at correctly selected layers is conditional: it narrows from 4/4 to 1/4 significant models with stronger text encoders (384d vs. 1024d) and remains suggestive at 32B due to small samples ($n \leq 20$). Seven alternative detection methods, including Bayesian online change-point detection (BOCPD), representation engineering (RepE), contrastive consistency search (CCS), and cross-model transfer, all fail to match supervised probing. Control tasks confirm the probes exploit genuine safety structure rather than superficial statistical patterns.

Our contributions are threefold. **(C1) Metric-induced layer selection bias:** crossing-rate composite metrics systematically select shallow layers, producing a qualitative reversal of the

¹Code and data: https://anonymous.4open.science/r/contrastive_trace_safety-A178/.

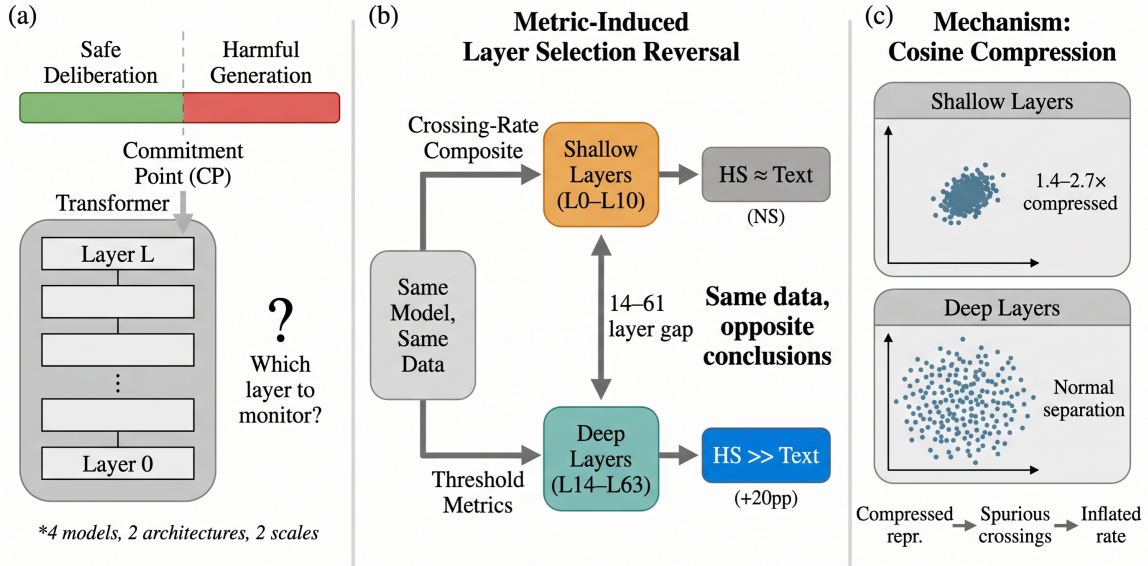


Figure 1: Metric-induced layer selection reversal. The same hidden states, evaluated with different layer selection metrics, yield opposite conclusions about the hidden-state advantage. Crossing-rate composite selects shallow layers (L0–L10), where hidden-state and text probes perform equivalently; threshold-based metrics select deep layers (L14–L63), where hidden-state probes significantly outperform text (+20pp). The 14–61 layer gap arises from the metric alone.

hidden-state advantage; we trace the mechanism to cosine compression and show it generalizes across models and datasets (AdvBench); the geometric precondition extends to a fifth architecture (Gemma-2). (C2) **Boundary conditions on the hidden-state advantage**: cross-validated at 7–8B, conditional on encoder quality, and suggestive at 32B. (C3) **Comprehensive negative results**: seven alternative methods all fail, establishing supervised temporal probing with principled layer selection as the current best approach.

2 Related Work

Safety in Reasoning Models. Reasoning models introduce a qualitatively new failure mode: the model’s own chain of thought transitions from safe deliberation to harmful generation without adversarial input (Wang et al., 2025). Self-jailbreak has been documented with step-level annotations (Kakkar et al., 2026; Yong and Bach, 2025); unfaithful reasoning (Chen et al., 2025; Turpin et al., 2023) motivates hidden-state monitoring. Safety-aligned models maintain refusal intentions during thinking but experience sharp drops at output generation (Yin et al., 2025; Haggendorff et al., 2026; Hu et al., 2026; Ahrend et al., 2026). Mitigation strategies include reasoning steering (Ghosal et al., 2026) and guardrail insertion (Mao et al., 2025); none examines how method-

ological choices in probing (particularly layer selection) affect safety conclusions for reasoning traces.

Hidden-State Probing. Internal activations encode safety-relevant information including alignment signals (Zhou et al., 2024), refusal directions (Arditi et al., 2024), and separate harmfulness/refusal encodings (Zhao et al., 2025; Wollschläger et al., 2025). Monitoring systems (TrajGuard (Liu et al., 2026), SafeSwitch (Han et al., 2025), and others (Jiao et al., 2026; Yan et al., 2026; Jiang et al., 2025; Wu et al., 2025; Hua et al., 2026; Zhou et al., 2025)) assume external input manipulation at the token level; the self-jailbreak setting differs because transitions are internal and span multiple reasoning steps. Concurrently, Chan et al. (2025) show final-layer linear probes predict alignment outcomes with strong F1, but their single-layer, single-metric evaluation does not examine whether different layers yield different conclusions. Multi-layer approaches complement our work: CLAP (Suresh et al., 2025) uses cross-layer attention bypassing single-layer selection, and Meyoyan and Del Corro (2026) find safety features distributed across layers, echoing our observation that different metrics select different layers. These works operate at fixed layers with single metrics; we show that the choice of layer, and the metric used to select it, can reverse the conclusions drawn.

Probing Methodology and Layer Selection.

Linear probing spans layer-wise classifiers (Alain and Bengio, 2017), sentence probing (Conneau et al., 2018), and surveys (Belinkov, 2022; Rogers et al., 2020); control tasks (Hewitt and Liang, 2019) distinguish structure from memorization (adopted in Section G). Non-uniform layer-wise information distribution (Tenney et al., 2019; Marks and Tegmark, 2024) and cross-layer disagreement (Kaya et al., 2019; Zhou et al., 2020) are well established; our contribution is identifying a specific cascade (metric $\rightarrow$ layer $\rightarrow$ conclusion) whose severity in safety probing has not been documented. Boxo et al. (2025) show probes can rely on textual evidence; our analysis provides a complementary geometric mechanism.

3 Problem Setup and Methods

3.1 Safety Commitment Point

When a reasoning model is jailbroken, its chain-of-thought (CoT) transitions from safety-aware deliberation to committed harmful generation. We define the **Safety Commitment Point** (CP) as the first reasoning step after which safety deliberation permanently ceases; traces with oscillating safety deliberation are resolved by taking the final irreversible transition. This binary formulation is a deliberate simplification; we discuss its limits in Section 5. Steps before CP are *pre-commitment* ($y_t = 0$); steps at or after CP are *post-commitment* ($y_t = 1$). The classification task produces a score $s_t \in [0, 1]$ estimating whether the model has passed its CP by step t . Annotation quality is robust to boundary uncertainty: shifting CP labels by ± 1 step changes balanced accuracy by < 2.5 pp (Table 22).

We study four reasoning models spanning two architectures and two scales (Table 1).

3.2 Probe Architecture

Hidden-state (HS) probes. At each step t , we extract the last-token hidden state $\mathbf{h}_t^{(\ell)} \in \mathbb{R}^d$ from layer ℓ and train an L2-regularized logistic regression (LR) classifier (balanced class weights, $C=1.0$, LBFGS, $\text{max_iter} = 2000$) on binary CP labels. Sweeping $C \in [0.01, 100]$ yields < 2 pp BA variation (Section V), confirming robustness to this default. Layer selection is analyzed in Section 3.3.

Text probes. Each step’s text is encoded with all-MiniLM-L6-v2 (Wang et al., 2020) ($\mathbb{R}^{384}$), a

Model	Arch.	Params	Layers	Hidden	n (total/test)
R1-8B	Llama	8B	32	4096	209/43
OT-7B	Qwen	7B	28	3584	228/47
R1-32B	Qwen	32B	64	5120	100/20
QwQ-32B	Qwen	32B	64	5120	95/19

Table 1: Models studied. R1-8B and R1-32B are distilled from DeepSeek-R1 (DeepSeek-AI, 2025); QwQ-32B (Qwen Team, 2024) is a native reasoning model; OT-7B (Guha et al., 2025) is trained with STILL (Min et al., 2024) on Qwen2.5-7B. Data: HarmThoughts (Kakkar et al., 2026) with step-level CP annotations (CP = first reasoning step after which safety deliberation permanently ceases). Split: 60/20/20.

distilled sentence encoder from the Sentence-BERT family (Reimers and Gurevych, 2019). The same logistic regression is applied as a surface-level modality baseline. A text embedding implementation error for QwQ-32B was identified and corrected during development (Section M).

Temporal windowing. For window size W , we concatenate features from steps $[t-W+1, \dots, t]$. We evaluate $W \in \{1, 3, 5, 10, 15, 20, 25\}$ and select per-model on validation. GRU sequence models are also evaluated under two protocols: a 70/30 train/test split (primary) and the same 60/20/20 split used for LR experiments (for fair comparison; Table 20 and Section F).

3.3 Layer Selection Sensitivity

Layer selection is a researcher degree of freedom that can reverse modality comparison conclusions. We evaluate five metrics: balanced accuracy (BA), precision, and inverse false positive rate (FPR) are standard probing metrics (Belinkov, 2022); crossing rate measures the fraction of traces where the probe score exceeds a threshold for K consecutive steps (Kaya et al., 2019; Zhou et al., 2020); and additive composite combines precision with inverse FPR. A multiplicative composite (BA $\times$ crossing rate) serves as the *crossing-rate composite* (CR-composite). Metrics at a fixed decision threshold (precision, inverse FPR, additive composite) form the *threshold-based* selection. Crossing rate inflates at shallow layers where representations are compressed: inter-step cosine distances are $1.4\text{--}2.7\times$ smaller (Fig. 2 and Section B), causing a linear probe to operate near the noise floor. Threshold-based metrics converge to deep layers; BA independently confirms. This predicts a priori that CR-composite exhibits shallow-layer bias, confirmed by results

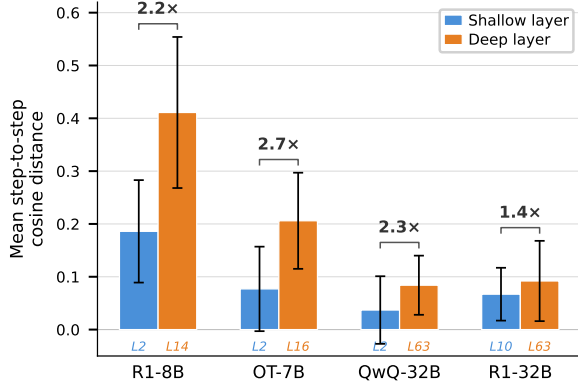


Figure 2: Cosine compression at shallow vs. deep layers. Shallow layers produce compressed representations (1.4–2.7× smaller inter-step cosine distances), causing spurious threshold crossings that inflate crossing rate. Layer indices below each bar; compression ratios above. Error bars: ± 1 std. Data from Table 10.

(Section 4.2); full per-layer sweeps in Section A.

3.4 Baseline Methods

We evaluate seven alternatives to contextualize probe performance (Table 5): Bayesian online change-point detection (BOCPD) (Adams and MacKay, 2007) and cumulative sum (CUSUM) (Page, 1954) for change-point detection; Mahalanobis distance following TrajGuard (Liu et al., 2026); RepE mean-difference directions (Zou et al., 2023a); two-layer MLP probes; CCS-adapted (Burns et al., 2023); and cross-model transfer via PCA. A position-only baseline (t/T) controls for temporal artifacts. Detailed configurations in Sections J and Q.

3.5 Evaluation Protocol

We report balanced accuracy, precision, pre-CP FPR, coverage, and lead time. We exclude F1 (confuses precision and recall), area under the ROC curve (AUROC; threshold-free, cannot diagnose layer selection bias), and Matthews correlation coefficient (MCC; redundant with BA under balanced priors). Positive classification requires $K=5$ consecutive threshold crossings, selected on the R1-8B validation set and held fixed across all models as the value balancing coverage ($\geq 70\%$) and safe-trace false positive rate ($\leq 50\%$); per-model optimal K values range from 3 to 7 (Table 25), and the HS vs. text conclusion direction is consistent at $K=3, 5$, and 7 (Section L), with coverage and FPR trading off monotonically. Modality comparisons use bias-corrected and accelerated (BCa) bootstrap (10,000

resamples, resampling traces as clusters: each trace is an independent unit and all its steps are included or excluded together, preserving within-trace dependence) (Efron and Tibshirani, 1993; DiCiccio and Efron, 1996) with Holm-Bonferroni (HB) step-down (3 metrics $\times$ 4 models = 12 tests, $\alpha=0.05$); the 4-test framework yields identical conclusions (Section C). For 32B models ($n \leq 20$), the exact permutation test is the primary inference method; BCa confidence intervals are reported for comparability but may have reduced coverage at these sample sizes (BCa assumes approximate pivotality; heavy-tailed score distributions could produce anti-conservative intervals). Within each model, the three metrics are positively correlated, so the correction is stricter than under independence. Control task selectivity (Hewitt and Liang, 2019) validates genuine structure (Section G). $\dagger$ marks sample-limited test sets ($n \leq 20$): R1-32B ($n=20$) and QwQ-32B ($n=19$).

Multiple comparisons. The full search spans ~ 980 – $2,240$ configurations per model; the effective number of independent tests is far smaller (shared confusion matrix, adjacent-layer correlation; Section X.2). A two-stage design protects against selection bias: upstream exploration uses nested CV at 7–8B; downstream tests apply HB to 12 comparisons. 32B models lack nested cross-validation (CV) and are marked exploratory ($\dagger$). Details in Section X.2.

4 Experiments

We evaluate temporal probes across four reasoning models (Table 1), comparing hidden-state (HS) and text modalities under different layer selection criteria.

4.1 Setup

We study R1-8B (Llama-8B, R1-distilled, $n=43$ test traces), OT-7B (OpenThinker-7B, trained on the OpenThoughts dataset (Guha et al., 2025); Qwen-7B backbone, STILL-trained Section E, $n=47$), QwQ-32B (Qwen-32B, native reasoning, $n=19$ – 20), and R1-32B (Qwen-32B, R1-distilled, $n=20$). All traces come from HarmThoughts with step-level CP annotations (60/20/20 split). HS probes extract last-token hidden states from a selected layer; text probes encode each step with all-MiniLM-L6-v2 (384d). Both use L2-regularized logistic regression with temporal windowing $W \in \{1, 3, 5, 10, 15, 20, 25\}$ selected

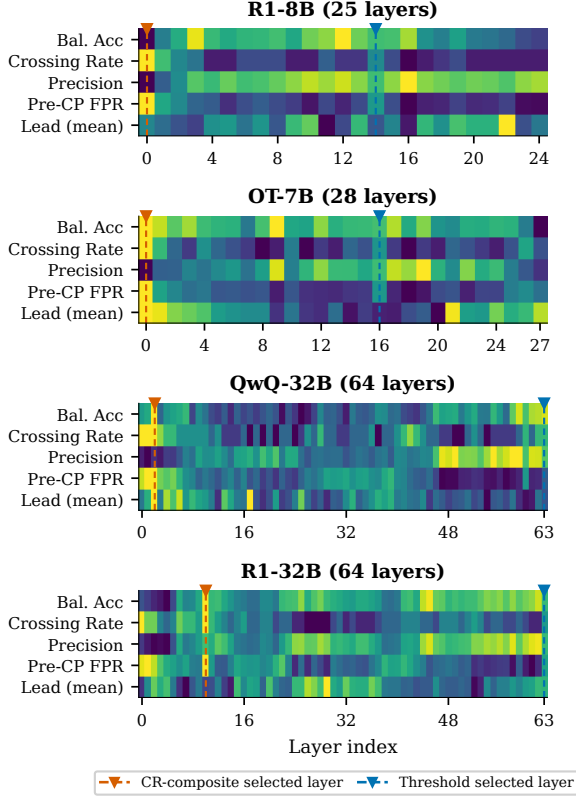


Figure 3: Per-layer metric heatmap. Each column: one of five metrics (min-max normalized). Red dashed: CR-composite layer; green dashed: threshold-based layer. CR-composite selections cluster near input layers; threshold-based metrics peak at deep layers.

on validation. Classification requires $K=5$ consecutive threshold crossings (Table 25). Statistical testing follows Section 3.5.

4.2 Layer Selection Results

Table 2 shows that the CR-composite systematically selects shallow layers: L0 for R1-8B (full L0–L24 sweep; Table 7) and OT-7B, L2 (3%) for QwQ-32B, L10 (16%) for R1-32B (Fig. 3). Threshold-based metrics converge to deep layers, notably L63 (98%) for both 32B models. Balanced accuracy varies across splits (7–98% depth; Section X). The mechanism is cosine compression: step-to-step distances are $1.4\text{--}2.7\times$ smaller at shallow layers (Section B), so a linear probe operates near the noise floor, inflating crossing rate independently of discriminative power (Section 3.3).

The downstream impact is severe: at CR-composite layers, only 7–8B models show significant HS advantage; both 32B models fail. CR-composite shallow-layer selection is an artifact of cosine compression, not a genuine safety signal. At threshold-based layers, the HS advantage is

Model	CR-comp. ^a	Thresh. ^b	p (thresh.)	ΔPrec^c
R1-8B	L0 (0%)	L14 (44%)	<0.001	+20.1
OT-7B	L0 (0%)	L16 (57%)	<0.001	+21.9
QwQ-32B [†]	L2 (3%)	L63 (98%)	<0.005	+19.2
R1-32B [†]	L10 (16%)	L63 (98%)	<0.001	+15.4

Table 2: Layer selection sensitivity. ^a Crossing-rate composite. ^b Threshold-based consensus. ^c HS–text precision difference (pp) at threshold-based layer. p : HS vs. text at threshold-based layer. [†] Sample-limited ($n \leq 20$).

Table 3: Multi-metric layer ensemble vs. best single metric and text baseline. All Ens–Text differences significant ($p \leq 0.002$); no Ens–Best difference significant ($p > 0.27$).

Model	Ens Prec	Best Single	Δ	Ens–Text
R1-8B	0.565	0.604 (BA@L12)	−3.9pp	+19.3pp
OT-7B	0.506	0.473 (FPR@L18)	+3.3pp	+27.0pp
QwQ-32B	0.578	0.578 (Prec@L60)	−0.0pp	+20.3pp
R1-32B	0.581	0.583 (Prec@L63)	−0.2pp	+10.5pp

graded: cross-validated at 7–8B, significant at QwQ-32B, and exploratory at R1-32B (small sample). Random-label controls confirm threshold-based selection requires genuine safety labels and that the geometric bias is Qwen-specific (Section A.1). AUROC confirms HS discriminability at balanced-accuracy-selected layers ($\Delta\text{AUROC} +12\text{--}15\text{pp}$; Table 41 and Section X.1) but cannot diagnose threshold-dependent bias.

Ensemble layer selection. A three-layer ensemble averaging predictions from the precision-best, balanced-accuracy-best, and inverse-FPR-best layers bypasses single-metric selection. All four models show significant HS advantage ($p \leq 0.002$; Table 3), including both 32B models that fail under CR-composite. The ensemble smooths metric-specific variation by averaging across depths within the deep region. This holdout result requires cross-validated confirmation (Section 5).

Cross-dataset replication (AdvBench). We replicate the five-metric layer sweep on AdvBench (Zou et al., 2023b) (520 traces, 159 jailbreaks, R1-8B, binary labels). The CR-composite selects shallow layers (median L2) while precision selects deep (median L14), a 31.7pp divergence (95% CI [3.2, 48.4]pp, $p=0.015$; 1,000 bootstrap iterations) consistent with HarmThoughts. This confirms layer selection bias (C1) is geometric, not label-dependent; AdvBench does not validate C2 (Sections B.1 and B.2).

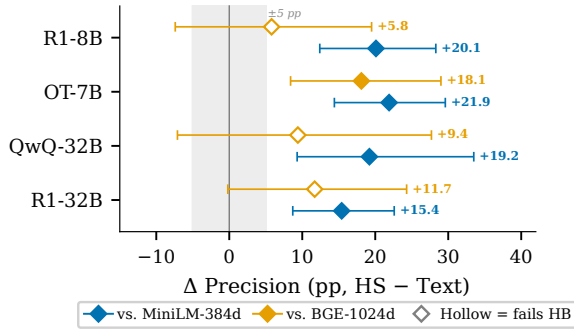


Figure 4: HS precision advantage (pp) over text probes with 95% bootstrap CI. Blue: vs. MiniLM-384d; orange: vs. BGE-1024d. Solid: survives Holm-Bonferroni correction; hollow: does not. The encoder-quality gradient is monotonic: all 384d comparisons significant; at 1024d only OT-7B survives.

4.3 HS vs. Text Comparison

Against the 384d text baseline, HS probes outperform text in all four models (Table 4). The advantage is cross-validated at 7–8B (R1-8B: +20.1pp, OT-7B: +21.9pp precision) and consistent across random seeds (Section W). At 32B, results are suggestive: QwQ-32B shows +19.2pp but the position baseline is competitive (Section 4.4); R1-32B shows +15.4pp but with limited coverage. Control tasks confirm the signal reflects genuine safety structure (Section G); a split-half analysis on OT-7B confirms internal validity (Section W.1).

Capacity control. PCA reduction of HS features to 384d (79–92% variance retained) preserves the precision advantage across all four models (+16.8 to +23.1pp, all $p < 0.012$, all survive correction). The control provides partial evidence against a pure capacity explanation. Random projection to 384d yields similar results (ΔPrec +8.3 to +20.5pp; Section S), further weakening the capacity explanation. E5-Mistral-7B (4096d, PCA-reduced to 384d; Section T) yields intermediate results (2/4 significant), forming a monotonic encoder-quality gradient with MiniLM (4/4) and BGE-large (1/4).

Encoder robustness. The HS advantage is encoder-conditional. Against BGE-large (Xiao et al., 2024) (1024d), only OT-7B survives correction (1/4 precision; Section R). Fig. 4 visualizes the encoder-quality gradient: MiniLM-384d (4/4 significant) $\rightarrow$ E5-Mistral (2/4) $\rightarrow$ BGE-1024d (1/4), with precision narrowing from +15–22pp to +6–18pp. Layer selection bias (C1) is encoder-independent.

Position mediation. Position (t/T) accounts for 12–20pp of balanced accuracy, but all models retain non-positional signal above chance (Section U). Under linear residualization, 3/4 models retain significant non-positional HS advantage; quadratic residualization absorbs the advantage in 2/4 models (Table 32). Per-model residualization details appear in Section 5. Temporal windowing reduces position dependence (Section 4.5).

4.4 32B Scale

Results at 32B are suggestive due to small test sets ($n=19$ –20; minimum detectable effect (MDE) 8.6–9.5pp at 80% power). R1-32B shows early detection ($W=1$, +14.2pp over position) but the residualized gap (+1.9pp) is not significant; QwQ-32B achieves significance at all window sizes but with negative lead time and a competitive position baseline. Neither provides confirmatory evidence; replication with $n \geq 50$ traces is needed (Section Z).

4.5 Temporal Context

Temporal windowing ($W > 1$) improves both modalities (Fig. 5; cross-model and text-probe ablations in Table 15 and Section D.1): at 7–8B, the modality advantage (+20.1 to +21.9pp) exceeds temporal improvement (+13pp). GRU models achieve comparable accuracy but higher FPR (Table 20 and Section O); an adaptive two-stage strategy offers marginal gains (Section N). CP-aligned probe trajectories confirm sharper HS transitions at the commitment point (Fig. 6). An oracle analysis (Section H) places the early detection ceiling at 50–94% across models, framing the gap that learned probes must close.

4.6 Negative Results

No alternative method consistently matches supervised probing (Table 5). BOCPD (Adams and MacKay, 2007)/CUSUM fail (0% detection on 3/4 models); Mahalanobis yields near-zero separation; RepE (Zou et al., 2023a) is model-dependent (BA 0.756 vs. 0.559); MLP probes overfit; CCS misaligns temporally; cross-model transfer fails (Section P). Per-model failure analysis (Section P) identifies mechanism-specific causes: BOCPD sensitivity to λ prior, Mahalanobis collapse under high-dimensional HS, and CCS temporal misalignment. The signal is directional, gradual, model-specific, and linearly separable, excluding change-point, distance, transfer, and objective-alignment methods.

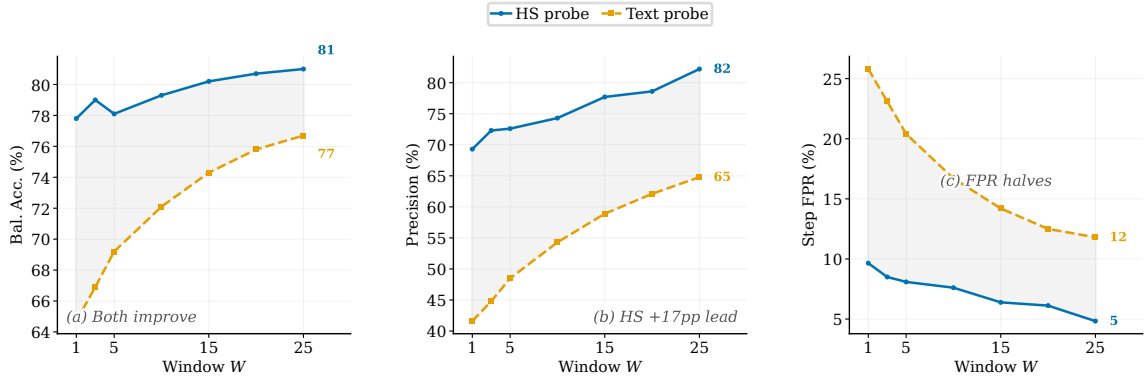


Figure 5: Temporal window effect on R1-8B HS probe (L14). Precision: 69.3% ($W=1$) to 82.2% ($W=25$); FPR: 9.65% to 4.83%.

Model	Layer	Enc.	W	Δ BA	Δ Prec	95% CI	Δ FPR	HB
R1-8B	L14	384d	3	+17.5	+20.1	[+12.4, +28.3]	-7.9	✓
		1024d		-	+5.8	[-7.4, +19.5]	-	
OT-7B	L16	384d	3	+10.7	+21.9	[+14.0, +29.6]	-17.1	✓
		1024d		-	+18.1	[+8.4, +29.0]	-	✓
QwQ-32B [†]	L63	384d	5	+13.0	+19.2	[+9.3, +33.5]	-15.0	✓
		1024d		-	+9.4	[-7.1, +27.7]	-	
R1-32B [†]	L63	384d	1	+18.7	+15.4	[+8.7, +22.6]	-14.6	✓
		1024d		-	+11.7	[-0.2, +24.3]	-	

Table 4: HS vs. text at threshold-based layers (logistic regression, 60/20/20 split, 10,000 bootstrap). Enc.: text encoder (384d = MiniLM, 1024d = BGE-large). Δ : HS-text difference (pp); “-” = not computed (different calibration points). HB: ✓ = passes Holm-Bonferroni ($\alpha=0.05$). [†] Sample-limited ($n \leq 20$).

Method	Result	Limitation	Failure Cat.
BOCPD	0% det. on 3/4 models	Commitment gradual	Graduality
CUSUM	< threshold, all models	Same as BOCPD	Graduality
Mahalanobis	$\Delta d \approx 0$	Signal in minor PCs	Var. sub-space
RepE	QwQ 0.756; R1-32B 0.559 ^a	Model-dependent	Model-spec.
MLP	3/4 worse than LR	Overfitting	Capacity
CCS	No gain over supervised ^b	Alignment mismatch	Obj. mismatch
Transfer	Best 59.5% (PCA) ^c	Model-specific repr.	Model-spec.

Table 5: Seven alternative approaches. ^a RepE QwQ approaches supervised bal_acc but precision is 0.512 vs. LR 0.700. ^b Semi-supervised (requires CP labels). ^c Best direct transfer 67.3% but 1% crossing rate.

5 Discussion and Conclusion

Implications and novelty. Our central finding, that metric choice alone produces a 14–61 layer divergence and qualitatively reverses hidden-state advantage conclusions, means safety probing studies should report results under multiple layer selection criteria. The cosine compression mechanism (Section B) provides a geometric explanation: shallow-layer representations compress

directional information, inflating crossing-rate metrics where probes operate near the noise floor. This mechanism is architecture-general (confirmed on Gemma-2; Section B.3), though tested only on standard attention models. While metric sensitivity is well-documented (Belinkov, 2022; Voita and Titov, 2020), our contribution is demonstrating a *cascade* specific to temporal safety probing: metric $\rightarrow$ layer $\rightarrow$ conclusion, where the severity (14–61 layer gap, sign reversal) exceeds what prior work documents at a fixed layer.

Practical recommendations. We propose three levels of defense against metric-induced layer selection bias (Section Y.2). *Level 1 (minimal)*: report layer selections under both threshold-based and frequency-based metric families; divergence exceeding 20% of model depth indicates metric-induced bias. *Level 2 (recommended)*: compute all five metrics and flag when ≥ 3 select layers differing by $>20\%$ of depth. *Level 3 (robust)*: a three-layer ensemble averaging predictions from metric-selected layers eliminates pre-selection entirely ($p \leq 0.002$; Table 3), but requires cross-validated confirmation.

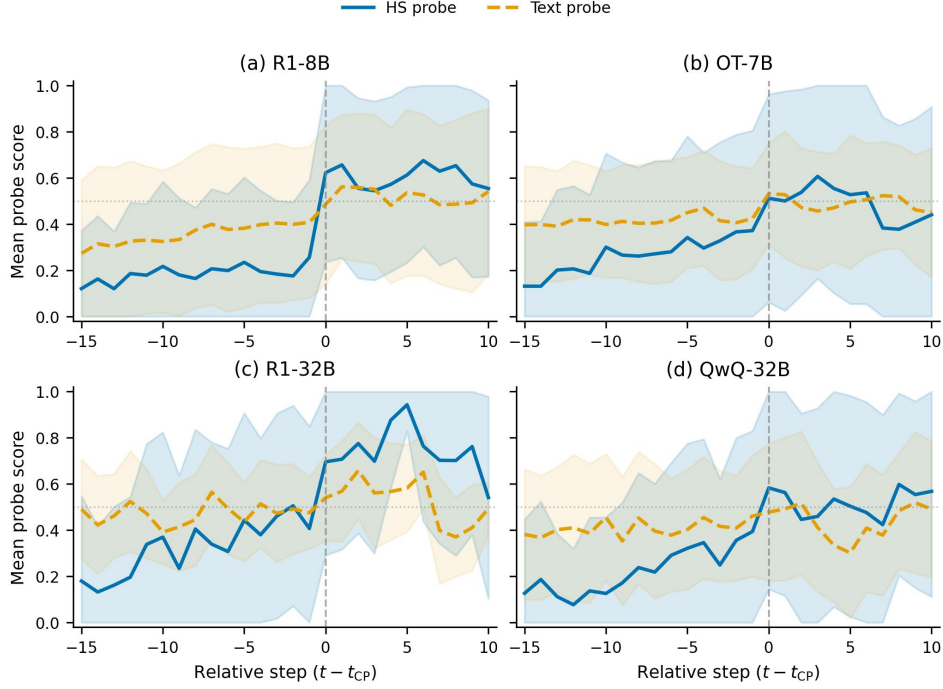


Figure 6: CP-aligned probe score trajectories. Each panel averages all test traces for one model, with step 0 at the true commitment point. HS probes (blue, solid) show a sharper score increase at CP than text probes (orange, dashed), with OT-7B showing the strongest separation and R1-32B the weakest, consistent with the evidence gradient. Shaded bands: ± 1 std. Points with < 5 traces are omitted.

Per-model consolidated results appear in Table 44.

Boundary conditions. The HS advantage is confirmed at 7–8B (+20–22pp, cross-validated; Section V.1) but conditional: it diminishes from 4/4 to 1/4 significant models with stronger encoders (384d $\rightarrow$ 1024d), and quadratic position residualization absorbs it in 2/4 models (Table 32). At 32B ($n \leq 20$), effects below 8.6–9.5pp fall within noise (Section Y); replication with $n \geq 50$ traces is needed. Ensemble layer selection mitigates metric-induced bias but false positive rates (11–50%) preclude deployment without conservative operating points (Section K.1).

Future work. Causal interventions on position encoding (e.g., shuffled or ablated position embeddings) would disentangle the content-level safety signal from temporal artifacts. Extending to online detection with streaming probes and calibrated confidence thresholds would move from retrospective characterization toward deployment.

Limitations

This paper provides retrospective classification of commitment points, not online detection; false

positive rates remain 11–50%, so deployment requires calibrated confidence thresholds and streaming architectures. CP-dependent claims rely on one dataset (HarmThoughts) with annotations lacking formal human-human inter-annotator agreement (LLM-human $\kappa=0.606$; Section I.1); the layer selection bias finding is replicated on AdvBench, but generalization to other jailbreak datasets, attack taxonomies, or multilingual settings is untested. At 32B scale ($n=19$ –20), small samples limit statistical power; all 32B conclusions are suggestive. Three of four models share the Qwen architecture; generalization to non-standard attention mechanisms (e.g., linear attention, state space models) is untested, and results are correlational.

Ethical Considerations

Our experiments use jailbreak prompt datasets (HarmThoughts, AdvBench) that contain harmful instructions. We use these solely to study model-internal safety mechanisms and do not generate, amplify, or redistribute harmful content. All harmful prompts originate from previously published, publicly available datasets.

This work characterizes how hidden-state probes detect safety-critical transitions during reasoning,

including failure modes such as layer selection bias. A potential dual-use risk is that adversaries could exploit knowledge of probe weaknesses—particularly the sensitivity to layer selection—to craft inputs that evade probe-based safety monitors. We believe the benefits of transparently documenting these failure modes outweigh this risk: defenders cannot build robust monitoring systems without understanding where current approaches break, and the layer selection bias we identify affects any practitioner deploying linear probes for safety classification.

We emphasize that our findings are diagnostic rather than prescriptive: they reveal limitations of a common evaluation methodology, not a recipe for bypassing safety measures. We encourage the community to develop layer-selection-robust probing methods before deploying hidden-state monitors in production safety systems.

References

- Ryan Prescott Adams and David J.C. MacKay. 2007. [Bayesian online changepoint detection](#). *arXiv preprint arXiv:0710.3742*.
- Patrick Ahrend, Tobias Eder, Xiyang Yang, Zhiyi Pan, and Georg Groh. 2026. [Safer reasoning traces: Measuring and mitigating chain-of-thought leakage in LLMs](#). *arXiv preprint arXiv:2603.05618*.
- Guillaume Alain and Yoshua Bengio. 2017. [Understanding intermediate layers using linear classifier probes](#). In *Workshop Track, International Conference on Learning Representations (ICLR)*.
- Andy Ardit, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. 2024. [Refusal in language models is mediated by a single direction](#). In *Advances in Neural Information Processing Systems 37 (NeurIPS)*.
- Yonatan Belinkov. 2022. [Probing classifiers: Promises, shortcomings, and advances](#). *Computational Linguistics*, 48(1):207–219.
- Gerard Boxo, Aman Neelappa, and Shivam Raval. 2025. [Linear probes rely on textual evidence: Results from leakage mitigation studies in language models](#). *arXiv preprint arXiv:2509.21344*.
- Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. 2023. [Discovering latent knowledge in language models without supervision](#). In *Proceedings of the International Conference on Learning Representations (ICLR)*.
- James Carpenter and John Bithell. 2000. Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians. *Statistics in Medicine*, 19(9):1141–1164.
- Yik Siu Chan, Zheng-Xin Yong, and Stephen H. Bach. 2025. [Can we predict alignment before models finish thinking? Towards monitoring misaligned reasoning models](#). *arXiv preprint arXiv:2507.12428*.
- Yanda Chen, Joe Benton, Ansh Radhakrishnan, Jonathan Uesato, Carson Denison, John Schulman, Arushi Somani, Peter Hase, Misha Wagner, Fabien Roger, Vlad Mikulik, Samuel R. Bowman, Jan Leike, Jared Kaplan, and Ethan Perez. 2025. [Reasoning models don’t always say what they think](#). *arXiv preprint arXiv:2505.05410*.
- Jacob Cohen. 1988. *Statistical Power Analysis for the Behavioral Sciences*, 2nd edition. Lawrence Erlbaum Associates.
- Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. [What you can cram into a single \$\&\#\&\$ vector: Probing sentence embeddings for linguistic properties](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 2126–2136.
- DeepSeek-AI. 2025. [DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning](#). *arXiv preprint arXiv:2501.12948*.
- Thomas J DiCiccio and Bradley Efron. 1996. Bootstrap confidence intervals. *Statistical Science*, 11(3):189–228.
- Bradley Efron and Robert J. Tibshirani. 1993. *An Introduction to the Bootstrap*. Chapman & Hall/CRC.
- Andrew Gelman and Eric Loken. 2013. [The garden of forking paths: Why multiple comparisons can be a problem, even when there is no “fishing expedition” or “p-hacking” and the research hypothesis was posited ahead of time](#). *Department of Statistics, Columbia University*.
- Soumya Suvra Ghosal, Souradip Chakraborty, Vaibhav Singh, Furong Huang, Dinesh Manocha, and Amrit Singh Bedi. 2026. [Safety recovery in reasoning models is only a few early steering steps away](#). *arXiv preprint arXiv:2602.11096*.
- Etash Guha, Ryan Marten, Sedrick Keh, Negin Raoof, and 1 others. 2025. [OpenThoughts: Data recipes for reasoning models](#). *arXiv preprint arXiv:2506.04178*.
- Thilo Hagendorff, Erik Derner, and Nuria Oliver. 2026. [Large reasoning models are autonomous jailbreak agents](#). *Nature Communications*.
- Peixuan Han, Cheng Qian, Xiusi Chen, Yuji Zhang, Heng Ji, and Denghui Zhang. 2025. [Safeswitch: Steering unsafe LLM behavior via internal activation signals](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*.
- John Hewitt and Percy Liang. 2019. [Designing and interpreting probes with control tasks](#). In *Proceedings of the 2019 Conference on Empirical Methods*

624	in <i>Natural Language Processing (EMNLP)</i> , pages	Gonzalo Ariel Meyoyan and Luciano Del Corro. 2026.	677
625	2733–2743.	A bertology view of llm orchestrations: Token- and	678
626	Xulin Hu, Che Wang, Wei Yang Bryan Lim, Jianbo	layer-selective probes for efficient single-pass classi-	679
627	Gao, and Zhong Chen. 2026. Tracing the dynam-	fication . <i>arXiv preprint arXiv:2601.13288</i> .	680
628	ics of refusal: Exploiting latent refusal trajec-	Yingqian Min, Zhipeng Chen, Jinhao Jiang, Jie Chen,	681
629	tories for robust jailbreak detection . <i>arXiv preprint</i>	Jia Deng, Yiwen Hu, Yiru Tang, Jiapeng Wang,	682
630	<i>arXiv:2605.02958</i> .	Xiaoxue Cheng, Huatong Song, Wayne Xin Zhao,	683
631	Peichun Hua, Hao Li, Shanghao Shi, Zhiyuan Yu, and	Zheng Liu, Zhongyuan Wang, and Ji-Rong Wen.	684
632	Ning Zhang. 2026. Rethinking jailbreak detection of	2024. Imitate, explore, and self-improve: A repro-	685
633	large vision language models with representational	duction report on slow-thinking reasoning systems .	686
634	contrastive scoring . In <i>Proceedings of the 64th An-</i>	<i>arXiv preprint arXiv:2412.09413</i> .	687
635	<i>annual Meeting of the Association for Computational</i>	E. S. Page. 1954. Continuous inspection schemes.	688
636	<i>Linguistics (ACL)</i> .	<i>Biometrika</i> , 41(1/2):100–115.	689
637	Yilei Jiang, Xinyan Gao, Tianshuo Peng, Yingshui Tan,	Qwen Team. 2024. QwQ: Reflect deeply on the bound-	690
638	Xiaoyong Zhu, Bo Zheng, and Xiangyu Yue. 2025.	aries of the unknown . Qwen Blog.	691
639	HiddenDetect: Detecting jailbreak attacks against	Nils Reimers and Iryna Gurevych. 2019. Sentence-	692
640	multimodal large language models via monitoring	BERT: Sentence embeddings using siamese BERT-	693
641	hidden states . In <i>Proceedings of the 63rd Annual</i>	networks . In <i>Proceedings of the 2019 Conference on</i>	694
642	<i>Meeting of the Association for Computational Lin-</i>	<i>Empirical Methods in Natural Language Processing</i>	695
643	<i>guistics (ACL)</i> .	<i>(EMNLP)</i> , pages 3982–3992.	696
644	Difan Jiao, Yilun Liu, Ye Yuan, Zhenwei Tang, Linfeng	Anna Rogers, Olga Kovaleva, and Anna Rumshisky.	697
645	Du, Haolun Wu, and Ashton Anderson. 2026. LLM	2020. A primer in BERTology: What we know about	698
646	safety from within: Detecting harmful content with	how BERT works . <i>Transactions of the Association</i>	699
647	internal representations . In <i>Proceedings of the 64th</i>	<i>for Computational Linguistics</i> , 8:842–866.	700
648	<i>Annual Meeting of the Association for Computational</i>	Malavika Suresh, Rahaf Aljundi, Ikechukwu Nkisi-Orji,	701
649	<i>Linguistics (ACL)</i> .	and Nirmalie Wiratunga. 2025. Cross-layer atten-	702
650	Ishita Kakkar, Enze Zhang, Rheeya Uppaal, and Jun-	tion probing for fine-grained hallucination detection .	703
651	jie Hu. 2026. When safety fails before the answer:	<i>arXiv preprint arXiv:2509.09700</i> .	704
652	Benchmarking harmful behavior detection in reason-	Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019.	705
653	ing chains . <i>arXiv preprint arXiv:2604.19001</i> .	BERT rediscovers the classical NLP pipeline . In	706
654	Yigitcan Kaya, Sanghyun Hong, and Tudor Dumitras.	<i>Proceedings of the 57th Annual Meeting of the Asso-</i>	707
655	2019. Shallow-deep networks: Understanding and	<i>ciation for Computational Linguistics (ACL)</i> , pages	708
656	mitigating network overthinking . In <i>Proceedings</i>	4593–4601.	709
657	<i>of the 36th International Conference on Machine</i>	Miles Turpin, Julian Michael, Ethan Perez, and	710
658	<i>Learning (ICML)</i> , pages 3301–3310.	Samuel R. Bowman. 2023. Language models don’t	711
659	Su-Hyeon Kim, Hyundong Jin, Yejin Lee, and Yo-Sub	always say what they think: Unfaithful explanations	712
660	Han. 2026. How does the thinking step influence	in chain-of-thought prompting . In <i>Advances in Neu-</i>	713
661	model safety? an entropy-based safety reminder for	<i>ral Information Processing Systems</i> , volume 36.	714
662	lrms . <i>arXiv preprint arXiv:2601.03662</i> .	Elena Voita and Ivan Titov. 2020. Information-theoretic	715
663	Cheng Liu, Xiaolei Liu, Xingyu Li, Bangzhou Xin, and	probing with minimum description length . In <i>Pro-</i>	716
664	Kangyi Ding. 2026. Trajguard: Streaming hidden-	<i>ceedings of the 2020 Conference on Empirical Meth-</i>	717
665	state trajectory detection for decoding-time jailbreak	<i>ods in Natural Language Processing (EMNLP)</i> ,	718
666	defense . <i>arXiv preprint arXiv:2604.07727</i> .	pages 183–196.	719
667	Yingzhi Mao, Chunkang Zhang, Junxiang Wang,	Cheng Wang, Yue Liu, Baolong Bi, Duzhen Zhang,	720
668	Xinyan Guan, Boxi Cao, Yaojie Lu, Hongyu Lin,	Zhongzhi Li, Yingwei Ma, Yufei He, Shengju Yu,	721
669	Xianpei Han, and Le Sun. 2025. When models	Xinfeng Li, Junfeng Fang, Jiaheng Zhang, and Bryan	722
670	outthink their safety: Unveiling and mitigating self-	Hooi. 2025. Safety in large reasoning models: A	723
671	jailbreak in large reasoning models . <i>arXiv preprint</i>	survey . <i>arXiv preprint arXiv:2504.17704</i> .	724
672	<i>arXiv:2510.21285</i> .	Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang,	725
673	Samuel Marks and Max Tegmark. 2024. The geometry	Rangan Majumder, and Furu Wei. 2024. Improving	726
674	of truth: Emergent linear structure in large language	text embeddings with large language models . <i>arXiv</i>	727
675	model representations of true/false datasets . In <i>Con-</i>	<i>preprint arXiv:2401.00368</i> .	728
676	<i>ference on Language Modeling (COLM)</i> .		

729	Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. MiniLM: Deep self-attention distillation for task-agnostic compression of pre-trained transformers . In <i>Advances in Neural Information Processing Systems 33 (NeurIPS)</i> .	784
730		785
731		786
732		787
733		788
734	Tom Wollschläger, Jannes Elstner, Simon Geisler, Vincent Cohen-Addad, Stephan Günnemann, and Johannes Gasteiger. 2025. The geometry of refusal in large language models: Concept cones and representational independence . In <i>Proceedings of the 42nd International Conference on Machine Learning (ICML)</i> .	789
735		790
736		791
737		
738		
739		
740		
741	Jiaqi Wu, Chen Chen, Chunyan Hou, and Xiaojie Yuan. 2025. SafeInt: Shielding large language models from jailbreak attacks via safety-aware representation intervention . In <i>Findings of the Association for Computational Linguistics: EMNLP 2025</i> .	792
742		793
743		794
744		795
745		
746	Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muenighoff, Defu Lian, and Jian-Yun Nie. 2024. C-pack: Packaged resources to advance general chinese embedding . In <i>Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval</i> .	796
747		797
748		798
749		799
750		800
751		801
752	Bo Yan, Weikai Lin, Yada Zhu, and Song Wang. 2026. Safedream: Safety world model for proactive early jailbreak detection . <i>arXiv preprint arXiv:2604.16824</i> .	802
753		803
754		804
755		805
756	Qingyu Yin, Chak Tou Leong, Linyi Yang, Wenxuan Huang, Wenjie Li, Xiting Wang, Jaehong Yoon, Yun Xing, Xing Yu, and Jinjin Gu. 2025. Refusal falls off a cliff: How safety alignment fails in reasoning? <i>arXiv preprint arXiv:2510.06036</i> .	806
757		807
758		808
759		809
760		810
761	Zheng-Xin Yong and Stephen H. Bach. 2025. Self-jailbreaking: Language models can reason themselves out of safety alignment after benign reasoning training . <i>arXiv preprint arXiv:2510.20956</i> .	811
762		812
763		813
764		814
765	Jiachen Zhao, Jing Huang, Zhengxuan Wu, David Bau, and Weiyang Shi. 2025. LLMs encode harmfulness and refusal separately . In <i>Advances in Neural Information Processing Systems 38 (NeurIPS)</i> .	815
766		816
767		817
768		818
769	Shide Zhou, Kailong Wang, Ling Shi, and Haoyu Wang. 2025. Exposing the ghost in the transformer: Abnormal detection for large language models via hidden state forensics . <i>arXiv preprint arXiv:2504.00446</i> .	819
770		820
771		821
772		822
773	Wangchunshu Zhou, Canwen Xu, Tao Ge, Julian McAuley, Ke Xu, and Furu Wei. 2020. BERT loses patience: Fast and robust inference with early exit . In <i>Advances in Neural Information Processing Systems (NeurIPS)</i> , volume 33, pages 18330–18341.	823
774		824
775		825
776		826
777		827
778	Zhenhong Zhou, Haiyang Yu, Xinghua Zhang, Rongwu Xu, Fei Huang, and Yongbin Li. 2024. How alignment and jailbreak work: Explain LLM safety through intermediate hidden states . In <i>Findings of the Association for Computational Linguistics: EMNLP 2024</i> , pages 2461–2488.	828
779		829
780		830
781		831
782		832
783		833
		834
	Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, and 2 others. 2023a. Representation engineering: A top-down approach to AI transparency . <i>arXiv preprint arXiv:2310.01405</i> .	
	Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J. Zico Kolter, and Matt Fredrikson. 2023b. Universal and transferable adversarial attacks on aligned language models. <i>arXiv preprint arXiv:2307.15043</i> .	
	Appendix Contents: Section A Layer Sweep Results; Section B Representation Compression Analysis; Section C Holm-Bonferroni Correction; Section D Window Size Ablation; Section E OpenThinker-7B Details; Section F GRU Training Details; Section G Control Task Methodology; Section H Oracle Upper Bound; Section I CP Annotation Perturbation; Section J BOCPD/CUSUM Configuration; Section K Detection Threshold Ablation; Section L Decision Threshold Sensitivity; Section M M002 Bug Disclosure; Section N Adaptive Two-Stage Detection; Section O Temporal Ablation Details; Section P Negative Results Details (incl. Table 5); Section Q Cross-Model Transfer Details; Section R Encoder Robustness; Section S Random Projection Baseline; Section T LLM-Based Text Encoder; Section U Position-Residualized Probing; Section V Hyperparameter Robustness; Section W Multi-Seed Stability; Section W.1 Split-Half Internal Validity; Section X BA-Layer Modality Comparison; Section Y Post-Hoc Power Analysis; Section Y.2 Practical Mitigation Levels; Section Z 32B Model-Specific Analysis; Section Z Consolidated Per-Model Results; Use of AI Assistants.	
	A Layer Sweep Results	
	For each model, we sweep all transformer layers and evaluate five candidate selection metrics: (1) balanced accuracy $\times$ crossing rate (the CR-composite), (2) pure balanced accuracy, (3) precision, (4) inverse pre-CP FPR, and (5) an additive composite. The CR-composite systematically selects shallow layers across all four models (including R1-8B under the full L0–L24 sweep; Table 7) because early-layer representations exhibit higher crossing rates (more frequent threshold crossings) regardless of discriminative power. Three threshold-based metrics (precision, inverse	

FPR, additive composite) converge on deep layers for all models; pure balanced accuracy selects layers whose depth varies with the data split (under 70/30: L30 for QwQ-32B, L47 for R1-32B; under the canonical 60/20/20: L63 and L46 respectively; Table 41).

Selected layers. R1-8B: under the original L12–L24 sweep, all five metrics agreed on L14 (44%); extending to L0–L24, crossing rate shifts to L0 while threshold-based metrics retain L14, a 16-layer gap (Table 7). OT-7B: CR-composite selects L0 (0%), threshold-based consensus selects L16 (57%). QwQ-32B: CR-composite selects L2 (3%), threshold-based consensus selects L63 (98%). R1-32B: CR-composite selects L10 (16%), threshold-based consensus selects L63 (98%). The disagreement is most extreme for 64-layer models, where the CR-composite selects layers in the bottom 16% of the network while threshold-based metrics select layers at 98% depth. Under 70/30, pure balanced accuracy selects L30 (47%) for QwQ-32B and L47 (73%) for R1-32B; under the canonical 60/20/20, it selects L63 (98%) and L46 (72%), which demonstrates sensitivity to the data split.

Impact on modality comparison. At the CR-composite layers, HS vs. text precision p -values are 0.064 (R1-32B) and 0.35–0.83 (QwQ-32B), both non-significant. At the threshold-based deep layers, all p -values fall below 0.005. The initially reported “scale-dependent” pattern (HS advantage at 7–8B but not 32B) was entirely an artifact of layer selection methodology, not a genuine property of model scale (Section 4.2).

Layer selection stability (LOO-CV). Leave-one-out cross-validation on the validation set shows moderate consistency for 7–8B models (65–68% dominant-layer agreement) but lower for 32B, confirming that layer selection instability is particularly acute at small sample sizes. Under pure balanced accuracy, the dominant layer is selected in 68% (QwQ-32B: L30, 13/19 folds) and 65% (R1-32B: L47, 13/20 folds) of folds. Alternative selections scatter across neighboring layers (L29, L55–56 for QwQ; L32, L38, L43 for R1-32B), reinforcing the need for multi-metric auditing rather than single-metric layer selection. Threshold-based metrics show tighter convergence to L63 for both 32B models. Expanding the criterion from top-1 to top-3 layers does not change conclusions: the top-3 BA-selected layers for QwQ-32B (L30, L29, L55)

and R1-32B (L47, L43, L38) all fall in the mid-to-deep range, and threshold-based metrics remain converged at L63 across all folds.

Joint (layer, W) sensitivity. Under the original 70/30 validation protocol, we evaluate QwQ-32B across 64 layers $\times$ 3 windows ($W \in \{1, 3, 5\}$) (192 configurations) (Table 6). L60 is near-optimal: within 2.1pp of the global best at $W=3$ and the top-ranked layer at $W=5$. (Under 60/20/20, threshold-based metrics select L63; the shift from L60 to L63 across splits itself illustrates layer selection sensitivity.) The optimal layer shifts from middle-deep (L46–47 at $W=1$) to deep (L60 at $W=5$), confirming a layer-window interaction. For R1-8B, optimal precision layers shift similarly (L16 at $W=1 \rightarrow$ L9 at $W=15$; Kendall’s $\tau=-0.91$, $p=0.071$), a consistent negative trend that does not reach significance at $n=4$ windows. L30, selected by per-fold LOO-CV at $W=1$ under balanced accuracy, is sub-optimal across all windows (max BA = 0.769 vs. global best 0.817). Joint (layer, W) stability was evaluated via grid search on $W \in \{1, 3, 5\}$ for QwQ-32B only; full leave-one-fold-out validation of the (layer, W) pair and extension to R1-32B remain for future work.

Table 6: QwQ-32B joint (layer, W) grid search under 70/30 validation. L60 is near-optimal at recommended windows ($W=3-5$). L30 (LOO-CV selection at $W=1$) is suboptimal across all windows. Under the canonical 60/20/20 split, threshold-based metrics select L63.

W	Best layer	Best BA	L60 BA	L60 rank
1	L47	0.798	0.783	~15th
3	L46	0.817	0.796	~6th
5	L60	0.808	0.808	1st

R1-8B full-layer sweep. The original R1-8B layer sweep covered L12–L24 (13 layers), where all five metrics agreed on L14 (Table 2). Extending to L0–L24 reveals that crossing rate consistently selects L0 across all window sizes, while threshold-based metrics retain mid-to-deep layers (L12–L16 for $W=1-5$; L9 at $W=15$, marginal $\Delta=0.003$ between shallow and deep means), a 9–16 layer divergence (Table 7). Unlike 32B models, the bias is isolated to crossing rate; threshold-based metrics show no shallow shift for R1-8B. This confirms the crossing-rate bias generalizes to Llama-architecture models. HS probes maintain significant advantage across all 25 layers and window sizes (precision gap $\sim 0.15-0.17$), unlike OT-7B

and 32B models where significance drops at shallow layers.

Table 7: R1-8B full-layer sweep (L0–L24, extending the original L12–L24 range). Best layer selected by crossing rate vs. threshold-based metrics across four window sizes. Score = balanced accuracy at the threshold-based layer. Layer gap = crossing-rate layer minus threshold-based layer.

	W=1	W=3	W=5	W=15
Crossing rate	L0	L0	L0	L0
Threshold (composite)	L16 (0.789)	L12 (0.796)	L12 (0.803)	L9 (0.784)
Layer gap	16	12	12	9

A.1 Random-Label Decomposition Control

To decompose the sources of metric disagreement in layer selection, we replace the true CP annotations with uniformly random labels and re-run the full layer sweep (10 random-label repeats per model). We test QwQ-32B (95 traces, 64 layers) and R1-8B (209 traces, 25 layers); Tables 8 and 9 report the fraction of repeats selecting a shallow layer and the mean selected layer.

Table 8: QwQ-32B layer selection under random CP labels (10 repeats). Shallow%: fraction selecting a layer $L \leq 10$. Real-CP layer shown for reference.

Metric	Real-CP	Shallow%	Rand. mean layer
Crossing rate	L4	100%	3.4 ± 2.6
CR-composite	L4	80%	13.1 ± 18.1
Precision	L63	20%	33.8 ± 19.8
Inv. FPR	L62	10%	36.4 ± 17.0
Bal. accuracy	L63	20%	33.3 ± 20.3

Table 9: R1-8B layer selection under random CP labels (10 repeats). Shallow%: fraction selecting $L \leq 4$ (bottom quintile of 25 layers). †: degenerate, all repeats select L0 with zero variance, indicating destroyed signal rather than shallow preference.

Metric	Real-CP	Shallow%	Rand. mean layer
Crossing rate	L0	0%	10.7 ± 4.8
CR-composite	L0	0%	13.7 ± 6.7
Precision	L23	100%†	0.0 ± 0.0
Inv. FPR	L13	100%†	0.0 ± 0.0
Bal. accuracy	L19	0%	14.0 ± 4.0

QwQ-32B (Qwen). The results decompose the metric disagreement into two components. (A) The crossing-rate shallow-layer preference is label-independent: 100% of random-label repeats still select layers L0–L7, confirming that this bias reflects

representation geometry (shallow-layer cosine distance compression; Section B) rather than any property of the safety labels. The CR-composite inherits this bias (80% shallow). (B) Threshold-based metrics (precision, inverse FPR) select deep layers only with real safety labels; under random labels, selections scatter across the full layer range (mean L34–L36, high variance). The full metric disagreement (L4 vs. L63, a 59-layer gap) is the superposition of a geometric artifact on the crossing-rate side and a real safety signal on the threshold-based side.

R1-8B (Llama). The Llama-based model shows stronger label dependence: crossing rate scatters from L0 to mean $L10.7 \pm 4.8$ under random labels (0% shallow), unlike QwQ-32B where it retains shallow preference. Threshold-based metrics degenerate entirely: precision and inverse FPR select L0 in all 10 repeats with zero variance, indicating random labels destroy discriminative signal rather than merely redistributing it. The real-CP disagreement (L0 vs. L23, a 23-layer gap) vanishes completely under random labels.

Cross-architecture summary. The metric disagreement requires genuine safety labels in both architectures, but the mechanism differs: in QwQ-32B, crossing-rate shallow preference reflects representation geometry (label-independent cosine compression), while in R1-8B even crossing-rate layer selection is label-dependent. The geometric component of the crossing-rate bias is thus Qwen-specific rather than a universal property of transformer representations.

B Representation Compression Analysis

We measure step-to-step cosine distance at CR-composite (shallow) vs. threshold-based (deep) layers across all models to characterize the compression mechanism underlying the crossing rate bias (Table 10).

Shallow layers produce highly compressed representations: consecutive steps cluster within small cosine distances (0.04–0.19), while deep-layer representations spread more widely (0.09–0.41). Inter-step distances at shallow layers are $1.4\text{--}2.7\times$ smaller than at deep layers across all models. A linear probe operating in this compressed space sits near the noise floor, where minor numerical fluctuations produce spurious threshold crossings that inflate crossing rate independently of discriminative power. Deep layers produce well-separated

Table 10: Step-to-step cosine distance at shallow and threshold-based (deep) layers. Shallow layers produce highly compressed representations ($1.4\text{--}2.7\times$ smaller inter-step distances). Shallow layers are L2 for R1-8B/OT-7B/QwQ-32B and L10 for R1-32B; deep layers follow threshold-based consensus (Table 2).

Model	Layer (depth)	Mean	Std
R1-8B	L2 (shallow)	0.186	0.097
	L14 (deep)	0.411	0.143
OT-7B	L2 (shallow)	0.077	0.080
	L16 (deep)	0.206	0.091
QwQ-32B	L2 (shallow)	0.037	0.064
	L63 (deep)	0.084	0.056
R1-32B	L10 (shallow)	0.067	0.050
	L63 (deep)	0.092	0.076

representations where crossings correspond to genuine semantic transitions.

Cross-dataset evidence. To assess whether this geometric precondition generalizes beyond HarmThoughts, we compute the same shallow/deep volatility ratio on AdvBench (Zou et al., 2023b) traces using R1-8B ($n=520$ traces). The deep-layer mean cosine distance (0.472) is $1.54\times$ the shallow-layer mean (0.307), consistent with the $1.4\text{--}2.7\times$ range observed across all four models on HarmThoughts (Table 10). This confirms that the representational compression underlying metric-induced layer selection bias is not dataset-specific and does not depend on CP annotations.

B.1 AdvBench Boundary Condition

We analyze R1-8B behavior on 388 substantive-response AdvBench (Zou et al., 2023b) traces (of 520 total) to characterize a boundary condition for the CP framework (Table 11).

Table 11: R1-8B behavior on AdvBench adversarial prompts (388 substantive-response traces of 520 total). CP=0 indicates the model produces harmful content from the first reasoning step without prior safety deliberation.

Metric	Value
Jailbreak rate	13.4%
CP=0 (among JB)	90.4%
Mean CP position (among JB)	1.4 steps

Adversarial prompts designed to bypass safety mechanisms (e.g., GCG suffixes) cause the model

to skip safety deliberation entirely: 90.4% of jailbroken traces have CP=0, meaning the model commits to harmful generation from the first reasoning step. This contrasts with HarmThoughts, where non-adversarial prompts trigger extended safety deliberation before commitment (mean relative CP position ≈ 0.7). The CP framework is most informative when prompts elicit deliberation that can be monitored; adversarial prompts that circumvent deliberation represent a boundary condition where commitment detection is vacuous. The geometric compression evidence (Table 10; volatility ratio $1.54\times$) remains valid because it measures representational properties independent of CP annotations.

B.2 AdvBench Layer Selection Bias Replication

We replicate the five-metric layer selection analysis on AdvBench (Zou et al., 2023b) to test whether metric-induced layer divergence depends on HarmThoughts-specific CP annotations or generalizes to simpler binary supervision. Using R1-8B, we generate reasoning traces for all 520 AdvBench prompts, classify each as jailbreak or safe via keyword matching, and evaluate logistic regression probes across all 32 layers (L0–L31) with windows $W \in \{1, 3, 5, 15\}$. The 60/20/20 train/val/test split matches HarmThoughts protocol.

Data distribution. Of 520 traces, 159 (30.6%) are classified as jailbreak and 361 (69.4%) as safe. A stricter keyword classifier (applied only to final responses, threshold ≥ 1 refusal phrase) reduces the jailbreak rate from 79.2% (v1) to 30.6%; the full-data divergence increases from 42pp (v1) to 58pp (v2); with proper held-out evaluation, the v2 bootstrap mean is 31.7pp ($p=0.015$), confirming the bias is genuine and not an artifact of label noise.

Table 12: AdvBench layer selection (R1-8B, 32 layers). CR-composite selects shallow layers; precision and inverse FPR select deep layers; AUROC selects mid-depth layers. Consistent with HarmThoughts (Table 2).

Metric	Best Layer	Depth (%)	Best W
CR-composite (BA $\times$ CR)	L3	9.7%	5
Crossing rate	L0	0.0%	5
Precision	L21	67.7%	15
Inverse FPR	L21	67.7%	15
Balanced accuracy	L9	29.0%	15
AUROC	L11	35.5%	15

Crossing rate pattern. Mean crossing rate at shallow layers is elevated relative to deep layers,

consistent with the cosine compression range observed on HarmThoughts (Table 10).

Per-window consistency. Precision and inverse FPR consistently select deep layers regardless of window size, while the CR-composite is pulled toward shallow layers across all windows.

The bootstrap-validated divergence (31.7pp, $p=0.015$) between CR-composite and precision/inverse-FPR selection replicates the HarmThoughts pattern (L0 vs. L14, 44% divergence) under different label semantics (binary jailbreak/safe vs. step-level CP). This confirms that the bias arises from representational geometry (shallow-layer cosine compression) rather than task-specific label structure.

Bootstrap validation. The full-data analysis (training and evaluation on the same data) yields a 58pp divergence (L3 at 9.7% vs. L21 at 67.7%), which includes layer selection overfitting. To obtain a proper held-out estimate, we run 1,000 bootstrap iterations with stratified 60/40 train/test splits: probes are trained on the training split (60%). In each bootstrap iteration, test traces are resampled with replacement, and both layer selection and divergence measurement are performed on the resampled test set; this estimates the full pipeline variance including selection variability. The bootstrap mean divergence is 31.7pp (95% CI [3.2, 48.4]pp, $p=0.015$). Layer selection is stable across iterations: the CR-composite median is L2 and the precision median is L14, consistent with the shallow/deep divergence pattern. The gap between full-data (58pp) and bootstrap (31.7pp) estimates reflects overfitting in layer selection when train and test data overlap, reinforcing the need for held-out evaluation in layer selection studies.

B.3 Cross-Architecture Geometric Analysis (Gemma)

To test whether the geometric precondition for layer selection bias generalizes beyond Llama and Qwen, we evaluate cosine compression on two Gemma-family models using AdvBench traces. Gemma-2-9B-it (standard multi-head attention) and Gemma-4-E4B (mixed sliding-window/full attention, window size 512) each generate 520 traces. Gemma-2 (6 JB / 520, 1.15%) and Gemma-4 (14 JB / 520, 2.7%) produce insufficient positive samples for reliable probe-based layer selection analysis; we therefore report only the geometric precondition (cosine compression).

Table 13: Cross-architecture cosine compression. Deep/shallow ratio >1 indicates shallow-layer compression (the geometric precondition for crossing-rate inflation). Standard full attention consistently shows compression; Gemma-4’s mixed sliding/full attention does not. ^a Probe-selected layers from HarmThoughts (L2 vs. L14). ^b Range-based windows (first 15% vs. last 25% of layers) due to insufficient jailbreak samples for probe-based selection.

Model	Attention	Method	Deep/Shallow	Compression
R1-8B (Llama)	Standard	Probe ^a	1.54×	✓
Gemma-2-9B	Standard	Range ^b	1.28×	✓
Gemma-4-E4B	Mixed sliding/full	Range ^b	0.70×	–

Gemma-2’s standard full attention produces a shallow-layer compression pattern consistent in direction with Llama and Qwen (deep/shallow ratio 1.28×, same direction as HarmThoughts models at 1.4–2.7× though at reduced magnitude). Gemma-4’s mixed sliding-window/full attention inverts this pattern (ratio 0.70×): shallow layers produce *larger* inter-step distances than deep layers, eliminating the geometric precondition for crossing-rate inflation. This suggests the compression arises from full self-attention’s global receptive field, while sliding attention’s local window disrupts the shallow-layer clustering that drives crossing-rate inflation. The geometric precondition thus generalizes across all models with standard full attention (Llama, Qwen, Gemma-2) but is absent under mixed attention, indicating that the layer selection bias is attention-mechanism-dependent.

C Holm-Bonferroni Step-Down Correction

Table 14 reports the full test battery used to assess hidden-state superiority over text probes. Because balanced accuracy, precision, and FPR share the same confusion matrix, the three per-model metrics are correlated; the effective number of independent comparisons is four (one per model). We report both the conservative 12-test correction and the 4-test model-level correction: both yield the same outcome (all tests pass at $\alpha=0.05$). Under 4-test HB, the tightest threshold falls on R1-8B FPR ($p=0.044$, rank-4 threshold $\alpha/1=0.050$), which still passes. Tests are ranked by p -value; each test i (rank i of m) must satisfy $p_i < \alpha/(m - i + 1)$ under the step-down procedure.

The weakest test (rank 12, R1-8B FPR, $p=0.044$) passes its threshold ($\alpha/1 = 0.05$) with margin. The ordering reveals that OT-7B produces

Table 14: Holm-Bonferroni step-down correction for 12 tests (4 models $\times$ 3 correlated metrics). Each test uses a two-sided null-centered bootstrap (10,000 resamples) on trace-level metric differences (HS minus text). All 12 pass at $\alpha=0.05$; a 4-test model-level correction (one per model, using the worst p -value) also passes.

Rank	Model	Metric	p -val	Thresh.	Pass
1	R1-8B	Precision	<0.001	0.0042	✓
2	OT-7B	Bal. Acc.	<0.001	0.0045	✓
3	OT-7B	Precision	<0.001	0.0050	✓
4	OT-7B	FPR	<0.001	0.0056	✓
5	QwQ-32B	Bal. Acc.	<0.001	0.0063	✓
6	QwQ-32B	Precision	<0.001	0.0071	✓
7	R1-32B	Bal. Acc.	<0.001	0.0083	✓
8	R1-32B	Precision	<0.001	0.0100	✓
9	QwQ-32B	FPR	0.004	0.0125	✓
10	R1-8B	Bal. Acc.	0.005	0.0167	✓
11	R1-32B	FPR	0.007	0.0250	✓
12	R1-8B	FPR	0.044	0.0500	✓

the strongest HS-text separation (ranks 2–4), while R1-8B FPR is the least significant despite R1-8B’s precision gap (+20.1pp at $W=3$). This reflects R1-8B’s already-low text FPR (11.8%), which leaves less room for HS improvement on this metric.

Exact sign-flip permutation test validation. As an independent check on bootstrap significance for the small-sample 32B models, we run exact sign-flip permutation tests on all six 32B comparisons, enumerating all 2^n sign permutations ($2^{19}=524,288$ for QwQ-32B, $2^{20}=1,048,576$ for R1-32B). All six tests reject the null at $\alpha=0.05$: QwQ-32B (bal_acc $p<0.001$, precision $p=0.003$, FPR $p<0.001$); R1-32B (bal_acc $p=0.008$, precision $p<0.001$, FPR $p=0.004$). The permutation p -values are consistent with the bootstrap values in Table 14 and confirm that the conclusions are not artifacts of distributional assumptions at small n .

D Window Size Ablation Details

Table 15 reports the complete window size ablation across all models and seven window sizes. The main text (Section 4.5) presents R1-8B results; here we provide the full data. Fig. 5 visualizes the monotonic improvement on R1-8B: precision rises from 69.3% ($W=1$) to 82.2% ($W=25$) while FPR drops from 9.65% to 4.8%.

Key cross-model patterns emerge from the window ablation:

R1-8B shows a general upward trend in balanced accuracy from 77.8% ($W=1$) to 81.0% ($W=25$), gaining +3.2pp overall with a slight dip at $W=5$

Table 15: Window size ablation (step-level metrics): balanced accuracy, precision, pre-CP FPR, coverage ($K=1$), and mean detection lead across all models.

Model	W	Bal. Acc.	Precision	FPR _{step}	Coverage	Lead
R1-8B	1	77.8	69.3	9.65	72.1	13.8
	3	79.0	72.3	8.50	69.8	12.1
	5	78.1	72.6	8.09	65.1	8.3
	10	79.3	74.3	7.61	55.8	10.0
	15	80.2	77.7	6.39	51.2	7.2
	20	80.7	78.6	6.12	53.5	7.4
	25	81.0	82.2	4.83	48.8	8.1
OT-7B	1	68.2	42.8	12.0	53.2	16.1
	3	68.8	46.6	10.0	53.2	21.6
	5	68.1	46.4	9.9	53.2	21.7
	10	66.5	46.9	8.6	48.9	25.0
	15	66.4	48.8	7.7	44.7	39.6
	20	66.0	49.4	7.5	48.9	17.9
	25	65.9	48.9	8.0	51.1	28.3
R1-32B	1	71.8	54.9	32.03	80.0	12.9
	3	72.5	56.0	30.64	80.0	11.4
	5	73.1	57.5	28.41	75.0	11.1
	10	73.7	60.9	23.40	70.0	10.8
	15	72.1	59.1	24.51	70.0	10.9
	20	69.6	56.3	26.18	65.0	9.6
	25	70.3	56.6	26.46	70.0	10.7
QwQ-32B	1	73.0	54.4	22.76	79.0	19.7
	3	74.5	56.5	21.41	73.7	20.6
	5	76.1	57.3	21.68	73.7	20.1
	10	74.7	57.1	20.73	73.7	18.4
	15	74.1	58.2	19.11	68.4	15.8
	20	75.4	60.2	17.89	63.2	15.8
	25	76.2	61.8	16.94	57.9	14.9

(78.1%). Precision improves sharply (+12.9pp) and step-level FPR halves (9.65% to 4.83%). Coverage trades off against quality, dropping from 72.1% to 48.8%.

OT-7B exhibits a split temporal pattern at layer 16: balanced accuracy peaks at $W=3$ (68.8%) and declines at larger windows (−2.9pp by $W=25$), while precision and FPR continue improving. Precision rises from 42.8% ($W=1$) to 49.4% ($W=20$), and FPR drops from 12.0% to 7.5% over the same range. The split indicates that larger windows reduce false alarms and sharpen positive predictions even as overall classification accuracy degrades. Lead times are shorter than old layer-0 estimates (16.1–39.6 steps) due to the threshold-based layer selection (Section E). OT-7B text probes improve monotonically across all metrics (Section E), so the accuracy saturation is specific to HS probes on this model.

R1-32B exhibits a non-monotonic pattern: balanced accuracy peaks at $W=10$ (73.7%) and then declines, falling below the $W=1$ baseline by $W=20$ (69.6%). The optimal window for R1-32B is $W=10$, which achieves the lowest FPR (23.40%) and highest precision (60.9%). The accuracy degradation at large windows likely reflects overfitting on the smaller dataset ($n=100$).

QwQ-32B shows a broadly increasing trend with local non-monotonicity: balanced accuracy peaks at $W=5$ (76.1%), dips at $W=10-15$, then recovers to 76.2% at $W=25$. Precision and FPR improve monotonically with larger windows (54.4% to 61.8% precision; 22.76% to 16.94% FPR). QwQ-32B produces substantially longer lead times than R1-8B and R1-32B (19.7 steps at $W=1$, 14.9 at $W=25$), though OT-7B’s lead times are longer still.

Universal trends. Across all models: (1) larger windows reduce coverage as the smoothed probe fires less frequently; (2) HS FPR generally decreases with window size; (3) lead time is model-dependent. The primary cross-model difference is whether balanced accuracy improves broadly (R1-8B, QwQ-32B), peaks sharply at an intermediate window (R1-32B at $W=10$), or peaks very early (OT-7B at $W=3$). Each model uses its validation-selected window in the main text.

D.1 Text Probe Window Ablation

Table 16 reports the complete text probe window sweep on R1-8B and confirms that temporal context benefits text probes even more strongly than HS probes.

Table 16: Text probe window ablation on R1-8B (step-level metrics). Text probes gain +12.0pp accuracy from $W=1$ to $W=25$, compared to +3.2pp for HS probes over the same range. The larger marginal gain for text probes reflects their weaker single-step baseline.

W	Bal. Acc.	Precision	FPR	Coverage
1	64.7	41.6	25.8	95.4
3	66.9	44.8	23.1	90.7
5	69.2	48.5	20.4	84.9
10	72.1	54.3	16.7	74.4
15	74.3	58.9	14.2	67.4
20	75.8	62.1	12.5	62.8
25	76.7	64.8	11.8	58.1

Text probes show monotonic improvement across all window sizes without the plateau observed in HS probes (which saturate around $W=15$). The absolute FPR reduction (-14.0 pp) is nearly three times the HS reduction (-4.8 pp), and accuracy gains (+12.0pp vs. +3.2pp) are four times larger. This asymmetry supports the interpretation that temporal context is the primary factor in commitment detection: text features, which encode less information per step, benefit proportionally more

from observing the multi-step progression.

E OpenThinker-7B Detailed Results

OpenThinker-7B (OT-7B) is trained with STILL (Min et al., 2024) on Qwen2.5-7B to produce a 7B reasoning model with a different training paradigm from the R1-distilled models. We report the full layer sweep and text probe window ablation.

Table 17: OT-7B layer sweep. Score = balanced accuracy $\times$ crossing rate. Layer 16 (57% depth) is selected after correction; it achieves the highest balanced accuracy among transformer layers.

Layer	Depth	Bal. Acc.	Cross. Rate	Score
0 (embed.)	0	66.7	53.2	35.5
4	14	66.3	40.4	26.8
8	29	66.4	25.5	17.0
12	43	67.5	31.9	21.6
16	57	68.2	31.9	21.8
20	71	65.5	31.9	20.9
24	86	66.9	29.8	19.9
27	96	63.0	31.9	20.1

Layer sweep. After correction of a layer selection bug, layer 16 (57% depth) is selected. It achieves the highest balanced accuracy (68.2%) among transformer layers under the 70/30 split used here; under the canonical 60/20/20 split, BA selects L2 (7% depth; Table 41). Layer 0 (embedding) has a higher crossing rate (53.2% vs. 31.9%) but lower accuracy.

Table 18: OT-7B text probe window ablation (step-level). Text probes improve monotonically with larger W , contrasting with the HS probes on this model which peak at $W=3$.

W	Bal. Acc.	Precision	FPR	Coverage
1	58.1	24.5	34.5	63.8
3	57.9	25.4	27.1	74.5
5	56.9	24.7	25.5	70.2
10	58.3	27.0	22.2	70.2
15	58.9	28.3	20.0	66.0
20	60.9	31.6	17.6	66.0
25	61.4	32.9	16.1	66.0

Text probe window ablation. OT-7B text probes improve monotonically from $W=1$ to $W=25$ (+3.3pp accuracy, +8.4pp precision, -18.4 pp FPR), consistent with the R1-8B pattern. OT-7B HS probes show a split pattern: balanced accuracy

peaks at $W=3$ and declines, while precision and FPR continue improving (Table 15). The text-HS contrast is therefore metric-dependent rather than a blanket modality difference.

Bootstrap comparison (matched $W=3$). To ensure a fair modality comparison, bootstrap tests use matched $W=3$ for both HS and text probes. At $W=3$, HS probes achieve 75.3% trace-level balanced accuracy vs. 65.6% for text (bootstrap $\Delta=+10.7$ pp, 95% CI [4.9, 16.3], $p<0.001$, survives Holm-Bonferroni). Precision: 55.8% vs. 34.6% (bootstrap $\Delta=+21.9$ pp, 95% CI [14.0, 29.6], $p<0.001$, survives Holm-Bonferroni). Pre-CP FPR: 10.0% vs. 27.1% (bootstrap $\Delta=-17.1$ pp, 95% CI [-22.2, -12.2], $p<0.001$, survives Holm-Bonferroni). All three metrics survive correction. The point-estimate precision gap (+21.2pp) aligns with the bootstrap mean (+21.9pp); the small discrepancy reflects resampling variance. Using each modality’s optimal W (HS at $W=3$, text at $W=25$) would introduce a confound because text would benefit from $8\times$ more temporal context.

F GRU Training Details

GRU results serve as supplementary evidence under a different training protocol (70/30 split); direct comparison with the main logistic regression pipeline (60/20/20 nested CV) is not intended.

Data split. GRU experiments use a 70/30 train/test split to maximize training data for the sequential model, differing from the main 60/20/20 protocol used for logistic regression probes. Within the 70% training portion, we hold out 10% of total traces as a validation set for early stopping (effective 60/10/30 split). The validation set is used only for early stopping and is not included in reported test metrics, so there is no information leakage.

Architecture. We evaluate two GRU variants. The single-layer variant uses a GRU with hidden dimension 256, taking per-step hidden states from the best layer (selected by validation balanced accuracy) as input. The multi-layer variant concatenates hidden states from 7 layers (evenly spaced through the network) at each step to produce a higher-dimensional input to the same GRU architecture.

Training procedure. We train with Adam (learning rate 1×10^{-3} , default $\beta_1=0.9$, $\beta_2=0.999$), batch size 16, for a maximum of 50 epochs. Early

stopping halts training when validation loss does not improve for 10 consecutive epochs. Input features are z-score normalized per dimension using training set statistics.

Evaluation. We report crossing accuracy (the sustained accuracy at which the probe first exceeds chance for 5 consecutive steps) and detection lead (steps before CP at crossing point). All experiments use 3 random seeds; we report mean $\pm$ standard deviation.

Table 19: R1-8B single-layer GRU results per seed. Lead is measured in steps before the commitment point.

Seed	Peak Acc.	Crossing Acc.	Lead (steps)
1	85.2	74.3	+4.8
2	83.1	71.6	+4.8
3	83.4	70.4	+0.9
Mean $\pm$ std	83.9 $\pm$ 1.1	72.1 $\pm$ 2.0	+3.5 $\pm$ 2.3

Per-seed results for R1-8B (single-layer GRU).

The high variance in lead time (standard deviation of 2.3 steps) is driven by seed 3, where the probe achieves good peak accuracy but crosses the detection threshold late. This variance underscores the sensitivity of early detection to initialization and motivates reporting per-seed results rather than relying on mean values alone.

Results in this section use the 60/20/20 protocol consistent with the main text; earlier 70/30 results are retained for completeness but should not be compared directly.

Table 20: Per-seed performance of GRU and logistic regression (LR) probes on R1-8B under a 60/20/20 cross-validation protocol (seeds 42, 123, 456). Single = layer 14 only; Multi = 7-layer concatenation (layers 12–24, stride 2). Lead = steps before CP at first sustained crossing. Main text results use a 70/30 train/test split with different test traces, accounting for the difference in lead values.

Model	Seed	Acc	Prec	FPR	Cross.%	Lead
GRU-single	42	0.845	0.657	0.139	72.1%	3.55
GRU-single	123	0.760	0.454	0.191	65.1%	7.39
GRU-single	456	0.792	0.474	0.203	67.4%	8.45
GRU-single	mean	0.799 $\pm$ 0.035	0.528 $\pm$ 0.091	0.178 $\pm$ 0.028	68.2 $\pm$ 2.9%	6.46 $\pm$ 2.10
GRU-multi	42	0.824	0.616	0.164	74.4%	5.66
GRU-multi	123	0.759	0.454	0.195	65.1%	6.43
GRU-multi	456	0.784	0.460	0.207	65.1%	8.93
GRU-multi	mean	0.789 $\pm$ 0.027	0.510 $\pm$ 0.075	0.189 $\pm$ 0.018	68.2 $\pm$ 4.4%	7.01 $\pm$ 1.40
LR-single	42	0.841	0.693	0.097	46.5%	-2.00
LR-single	123	0.753	0.436	0.181	53.5%	5.26
LR-single	456	0.814	0.512	0.146	41.9%	3.94
LR-single	mean	0.803 $\pm$ 0.037	0.547 $\pm$ 0.108	0.141 $\pm$ 0.035	47.3 $\pm$ 4.8%	2.40 $\pm$ 3.16
LR-multi	42	0.827	0.679	0.092	39.5%	1.06
LR-multi	123	0.756	0.442	0.180	46.5%	7.30
LR-multi	456	0.817	0.521	0.130	34.9%	4.00
LR-multi	mean	0.800 $\pm$ 0.032	0.547 $\pm$ 0.099	0.134 $\pm$ 0.036	40.3 $\pm$ 4.8%	4.12 $\pm$ 2.55

Multi-layer GRU and baseline comparison (60/20/20 protocol). Under this protocol, multi-layer GRU achieves the highest mean lead (7.01 ± 1.40 steps) but at higher FPR than logistic regression baselines (0.189 vs. 0.134). The GRU advantage is concentrated in detection timing rather than classification accuracy, consistent with the main text finding that GRUs capture sequential dependencies that improve early detection. Lead time shows high variance across seeds (range 5.66 – 8.93 for GRU-multi), reinforcing that reported means should be interpreted cautiously.

OT-7B multi-layer GRU (layers 13–19, 3 seeds, 60/20/20 split). OT-7B GRU achieves $74.6 \pm 1.3\%$ balanced accuracy, $64.5 \pm 7.0\%$ coverage, $+33.6 \pm 4.8$ steps lead, and $27.4 \pm 6.1\%$ pre-CP FPR. The lead time is substantially longer than other models in absolute terms, but OT-7B traces average ~ 80 reasoning steps (vs. ~ 30 for R1-8B and 32B models), so absolute lead times are not directly comparable. The FPR-lead tradeoff mirrors R1-8B: GRU FPR (27.4%) is far above the LR probe at layer 16 (10.0%), which confirms that GRUs gain lead time at the cost of more false alarms across both 7-8B models.

G Control Task Methodology

We follow the selectivity framework of (Hewitt and Liang, 2019) to verify that our probes capture genuine safety-relevant structure rather than surface-level correlations.

Procedure. For each model, we randomly permute the binary labels (pre-CP vs. post-CP) within each trace while preserving the marginal label distribution. We then retrain the probe with identical architecture (L2-regularized logistic regression) and hyperparameters on these shuffled labels. Selectivity is defined as:

$$\text{Selectivity} = \text{Acc}_{\text{linguistic}} - \text{Acc}_{\text{control}} \quad (1)$$

where $\text{Acc}_{\text{linguistic}}$ is the probe accuracy on true labels and $\text{Acc}_{\text{control}}$ is the accuracy on permuted labels. A selectivity of zero indicates that the probe exploits only non-linguistic features (e.g., position); higher values confirm that the representation encodes task-relevant information. Absolute selectivity values are task-dependent and not directly comparable across studies (Hewitt and Liang, 2019).

Results. We repeat the control experiment 10 times with different random permutations and report the mean control accuracy. Selectivity scores at threshold-based layers (60/20/20 split) are: R1-32B 16.8pp (layer 63), QwQ-32B 16.1pp (layer 63), R1-8B 14.1pp (layer 14), OT-7B 11.1pp (layer 16). All values exceed the 5pp threshold recommended by (Hewitt and Liang, 2019). R1-32B and QwQ-32B achieve the strongest selectivity ($>16\text{pp}$) at threshold-based layers. On QwQ-32B, the position baseline accuracy (0.705) slightly exceeds the HS probe (0.667) at L60, consistent with the main-text finding that QwQ’s HS advantage is concentrated in precision and FPR rather than balanced accuracy.

Layer provenance note. R1-8B is unaffected by the layer selection bias because its CR-composite and threshold-based layers coincide at L14. OT-7B exhibits a 16-layer gap (CR-composite L0 vs. threshold-based L16; Table 2), but selectivity at L16 (11.1pp) still exceeds the 5pp threshold. QwQ-32B selectivity is 16.1pp at L63, which confirms that safety-specific signal persists at deep layers.

Position-only baseline. As a complementary validation, we train a logistic regression classifier using only the normalized step index (t/T where T is the total trace length) as the input feature. This baseline tests whether temporal position alone explains probe performance. On R1-8B, the position-only baseline produces 15.8% FPR compared to 8.1% for the HS probe at matched detection rates. The probe’s substantially lower FPR confirms that hidden state features carry information beyond what step position provides.

Residualized selectivity. To confirm that selectivity survives position removal, we recompute control task selectivity on position-regressed features (the same residualized representations used in Section U). Residualized selectivity: R1-8B 22.9pp , OT-7B 10.4pp , QwQ-32B 11.3pp (all $>5\text{pp}$), R1-32B 3.0pp ($<5\text{pp}$, consistent with the position-mediation pattern reported in Section 4.3). Three of four models retain selectivity well above the Hewitt and Liang (2019) threshold after position removal; R1-32B’s drop confirms that its safety signal is largely positional.

H Oracle Upper Bound Analysis

We compute an oracle upper bound for commitment detection by measuring the earliest step at which the ground-truth commitment point (CP)

could theoretically be predicted from hidden-state dynamics. The oracle detector has access to the true CP label and computes how many steps before CP the representation trajectory first enters a region that is linearly separable from pre-CP states with perfect accuracy on the training set. The resulting bound provides a theoretical ceiling on early detection performance.

Table 21 reports oracle performance across all four models. Oracle analysis confirms the evidence gradient: OT-7B (strong) shows the highest early detection rate (93.6%) and longest lead time (50 steps), while R1-32B (qualified) achieves only 50% early detection even with oracle labels, consistent with an inherently weaker commitment signal. On R1-8B, the oracle early detection rate is 76.7% (33/43 traces detected before CP, with +31 steps mean lead time), compared to +4.5 steps for the best learned probe (GRU). This gap indicates that hidden states at step t encode the model’s current safety state but carry limited predictive power about future trajectory evolution. The learned probe captures the contemporaneous safety signal effectively (balanced accuracy 81–84%) but cannot anticipate the commitment transition far in advance. Closing this gap will likely require causal intervention techniques (e.g., activation patching) rather than improvements to passive monitoring.

Table 21: Oracle upper bound analysis. Detection rate: fraction of traces where the oracle finds a linearly separable boundary before or at CP. Early detection rate: fraction detected before CP. Mean lead: average steps before CP at first detection.

Model	Det. Rate	Early Det.	Mean Lead	n_{test}
R1-8B	100%	76.7%	31.0	43
OT-7B	100%	93.6%	50.0	47
R1-32B [†]	95.0%	50.0%	32.4	20
QwQ-32B [†]	100%	84.2%	24.4	19

I CP Annotation Perturbation

To assess sensitivity to annotation noise, we shift all ground-truth CP labels by $\delta \in \{-3, \dots, +3\}$ steps and retrain the HS probe (Table 22). Precision degrades smoothly with no cliff effect: within ± 2 steps, BA changes by $< 2.5\text{pp}$ for all models (7–8B: $< 1.0\text{pp}$). At the ± 3 -step worst case, all models retain precision well above their respective text baselines. These results support the annotation robustness assumption discussed in Section 5.

Table 22: HS probe sensitivity to CP annotation shift (δ steps). BA is stable across all perturbations; precision degrades gradually without cliff effects.

Model	δ	Prec	BA	FPR
R1-8B	−3	.652	.773	.177
	0	.596	.784	.165
	+3	.530	.771	.143
OT-7B	−3	.404	.714	.298
	0	.351	.707	.286
	+3	.300	.702	.276
QwQ-32B	−3	.661	.728	.158
	0	.609	.690	.129
	+3	.550	.676	.114
R1-32B	−3	.625	.698	.360
	0	.570	.722	.281
	+3	.503	.750	.236

I.1 Inter-Annotator Agreement on AdvBench

To assess annotation reliability, two LLM annotators (Qwen3-32B and GPT-4o-mini) independently labeled 388 AdvBench B-class traces for jailbreak status and CP position (Table 23).

Table 23: Inter-annotator agreement on AdvBench B-class traces (388 total). CP position metrics computed on 51 traces where both annotators identified a jailbreak.

Metric	Value
<i>Jailbreak classification</i>	
Cohen’s κ	0.606 (moderate)
Raw agreement	87.4%
GPT-4o-mini JB count	99 / 388
Qwen3-32B JB count	52 / 388
<i>CP position (51 shared JB)</i>	
Exact match	37%
Within 5 steps	47%
Mean abs. difference	8.5 steps

LLM annotators achieve only moderate agreement on jailbreak classification ($\kappa=0.606$) and low CP position consistency (37% exact match). GPT-4o-mini is substantially more aggressive in jailbreak identification (99 vs. 52), suggesting systematic annotator bias. The low CP position agreement (mean absolute difference of 8.5 steps) underscores a key design choice: our primary analysis uses the human-validated HarmThoughts annotations (Kakkar et al., 2026), not LLM-generated labels. This AdvBench IAA study serves as a *neg-*

ative control: it demonstrates that LLM-only annotation is insufficient for reliable CP localization, motivating future investment in multi-annotator human CP labeling with formal IAA evaluation.

J BOCPD/CUSUM Configuration Sweep

We sweep hyperparameters for both Bayesian Online Change-Point Detection (BOCPD) and Cumulative Sum (CUSUM) methods applied to the temporal probe score sequences.

BOCPD configurations. BOCPD models the sequence of probe scores as piecewise-constant segments with Gaussian observation noise. The hazard rate λ controls the prior probability of a change-point at each step ($P(\text{change}) = 1/\lambda$).

CUSUM configurations. CUSUM accumulates evidence for a positive mean shift. The drift parameter δ sets the minimum shift size to detect, and the threshold h controls the detection boundary.

Table 24: Change-point method configuration sweep on R1-8B. Detection rate is the fraction of jailbreak traces where the method fires. Lead is measured in steps before CP.

Method	Configuration	Det. Rate	Lead
BOCPD	$\lambda = 5$	97	+2.1
	$\lambda = 10$	93	+3.5
	$\lambda = 20$	87	+4.2
	$\lambda = 50$	72	+5.8
	$\lambda = 100$	67.4	+1.03
CUSUM	$\delta=0.3, h=3$	65.1	+0.89
	$\delta=0.5, h=3$	89	+1.4
	$\delta=0.5, h=5$	76	+2.3
	$\delta=0.5, h=8$	61	+3.1
	$\delta=1.0, h=3$	82	+1.1
	$\delta=1.0, h=5$	68	+1.9
	$\delta=1.0, h=8$	52	+2.7
	$\delta=2.0, h=3$	71	+0.8
	$\delta=2.0, h=5$	54	+1.5
	$\delta=2.0, h=8$	38	+2.1
Simple threshold	FCP baseline	65.1	+4.89

The best BOCPD configuration ($\lambda=10$) achieves 93.0% detection rate with +3.45 steps lead. The simple threshold baseline achieves 65.1% detection with superior lead (+4.89 steps). All CUSUM configurations underperform the simple threshold on lead time.

These results reveal a coverage-timing trade-off. BOCPD ($\lambda=10$) achieves 28pp higher detection rate than simple thresholding (93.0% vs. 65.1%), but with lower lead (+3.45 vs. +4.89 steps).

The transition toward unsafe reasoning is gradual rather than abrupt, so change-point methods cannot exploit discontinuities for early warning. When missed detections are costly, BOCPD’s coverage advantage justifies its use despite the timing penalty. When early intervention is the priority, simple thresholding provides better lead time.

Other models. On OT-7B, QwQ-32B, and R1-32B, BOCPD achieves 0% detection rate across all tested configurations ($\lambda \in \{5, 10, 50, 100, 200\}$). The failure modes differ by model: for R1-32B and QwQ-32B, probe scores transition gradually, so the hazard function never accumulates sufficient evidence; for OT-7B at $W=3$, transitions are too abrupt ($0 \rightarrow 1$ in 1–2 steps) for the Gaussian observation model to capture. R1-8B falls between these extremes, with transitions sharp enough to trigger the hazard function but smooth enough for the Gaussian model to track.

K Detection Threshold (K) Ablation

The consecutive-crossings threshold K controls the trade-off between detection coverage and false positive rate. We evaluate $K \in \{1, 3, 5, 7, 10\}$ at $W=1$ on all models using HS probes at threshold-based layers (Table 25).

Table 25: Detection threshold ablation ($W=1$, HS probes at threshold-based layers). Coverage: fraction of jailbreak traces with $\geq K$ consecutive crossings. Safe FPR: fraction of safe traces with $\geq K$ consecutive false alarms.

K	Coverage				Safe FPR			
	R1-8B	OT-7B	QwQ	R1-32B	R1-8B	OT-7B	QwQ	R1-32B
1	1.00	1.00	1.00	1.00	1.00	0.92	1.00	0.99
3	0.74	0.79	0.95	0.90	0.81	0.41	0.78	0.78
5	0.47	0.51	0.79	0.65	0.63	0.19	0.65	0.47
7	0.35	0.30	0.58	0.45	0.39	0.16	0.35	0.31
10	0.14	0.21	0.32	0.35	0.22	0.08	0.25	0.17

$K=1$ yields perfect coverage but near-perfect safe FPR (0.92–1.00), offering no discrimination. $K=5$ is empirically selected as a middle-ground tradeoff: average coverage 0.60 with average safe FPR 0.48 across models. OT-7B benefits most from increasing K , reaching 0.19 safe FPR at $K=5$ (vs. 0.92 at $K=1$) while retaining 51% coverage. Higher K values (≥ 7) reduce safe FPR further but at steep coverage cost. $K=5$ vs. $K=7$ is the key decision: $K=7$ reduces mean safe FPR from 0.48 to 0.30 but drops mean coverage from 0.60 to 0.42.

$K=5$ was selected on the R1-8B validation set and held fixed across all models; main-text results

use this value without per-model optimization. The $K \times W$ interaction is not fully explored: this ablation uses $W=1$ to isolate K 's effect, and the joint $K \times W$ surface remains a limitation. At optimal W , coverage at $K=5$ improves (e.g., R1-8B: 0.47 at $W=1$ vs. 0.60 at $W=15$).

K.1 Deployment Operating Points

The K ablation above can be recast as a deployment operating characteristic. At $K=5$ (used throughout the main text), average coverage is 60% with 48% safe FPR. Raising K yields practically useful operating points: at $K=10$, OT-7B achieves 21% coverage with only 8% false positive rate, meaning fewer than 1 in 12 flagged traces is a false alarm. R1-32B at $K=10$ retains 35% coverage at 17% FPR. These conservative operating points are viable for high-stakes deployment where missed detections are preferable to frequent false alarms: a secondary review system could inspect only flagged traces, reducing human oversight burden by 79–86% while maintaining non-trivial detection rates. However, at $K=10$ approximately 65–86% of attacks are missed, which may be unacceptable in some threat models. The $K=7$ setting offers a middle ground (30–58% coverage, 16–39% FPR). The choice of operating point is deployment-specific.

L Decision Threshold Sensitivity

We evaluate whether the HS advantage depends on the logistic regression decision threshold by sweeping $\tau \in \{0.3, 0.4, 0.5, 0.6, 0.7\}$ for both 32B models at their threshold-based layers. Results below were conducted at L60 under the 70/30 protocol; main-text results use L63 under 60/20/20. QwQ-32B (L60, $W=3$) HS balanced accuracy varies by only 0.53pp across all thresholds, with $\Delta(\text{HS} - \text{Text})$ stable at +8.5 to +9.3pp. Table 26 reports R1-32B results.

Table 26: R1-32B threshold sensitivity (L63, $W=1$). HS BA varies by 1.1pp (0.741–0.752); the HS advantage is consistent across all thresholds.

τ	HS BA	ΔBA	HS Prec	ΔPrec	HS FPR	ΔFPR
0.3	0.752	+0.126	0.583	+0.126	0.295	−0.106
0.4	0.746	+0.116	0.586	+0.120	0.281	−0.092
0.5	0.742	+0.109	0.588	+0.113	0.273	−0.078
0.6	0.741	+0.112	0.599	+0.119	0.254	−0.072
0.7	0.743	+0.126	0.617	+0.143	0.228	−0.084

R1-32B HS balanced accuracy varies by only 1.1pp across the full threshold range. The modality

advantage ($\Delta\text{BA} = +0.109$ to $+0.126$, $\Delta\text{Prec} = +0.113$ to $+0.143$, $\Delta\text{FPR} = -0.072$ to -0.106) is consistent in both magnitude and direction at every threshold, so the HS advantage is not an artifact of the default $\tau=0.5$.

M Text Embedding Bug Disclosure

We disclose a bug discovered during development that affected intermediate results for QwQ-32B.

Issue. In the initial text baseline evaluation, the QwQ-32B text probe was inadvertently initialized with a midpoint heuristic placeholder rather than true sentence embeddings from all-MiniLM-L6-v2. Specifically, the placeholder assigned each step a feature vector equal to the midpoint between the mean safe and mean unsafe embeddings from the training set, with small Gaussian noise added. This made the text probe trivially predictive and produced a spuriously significant result ($p=0.001$ for HS vs. text FPR difference).

Discovery and correction. The bug was identified during a systematic audit of the feature extraction pipeline. After correction (using actual sentence embeddings), the QwQ-32B and R1-32B modality comparisons at the originally selected layers (L2 and L10, chosen by the CR-composite metric) showed no significant HS advantage: QwQ-32B accuracy $p=0.828$, precision $p=0.346$; R1-32B precision $p=0.050$ but no tests surviving Holm-Bonferroni. The subsequent discovery that these layer selections were themselves artifacts of metric bias (Section 4.2) led to the threshold-based layer analysis reported in the main text, where the HS advantage follows an evidence gradient at threshold-based deep layers: cross-validated at 7–8B, significant but suggestive of position-competitiveness at QwQ-32B[†], and exploratory at R1-32B[†] (Table 4).

Impact on reported results. All numbers reported in this paper use the corrected text embedding pipeline. The bug affected only QwQ-32B text probe features. The correction, combined with the separate layer selection fix (Section 4.2), means that early-stage intermediate results (non-significant 32B modality comparisons) reflected two compounding errors: incorrect text features and CR-composite layer selection. Both are fully resolved in the final analysis.

Commitment to transparency. We report this bug to maintain scientific integrity. Code and data will be released upon publication.

N Adaptive Two-Stage Detection

We explored whether a two-stage adaptive window strategy could improve the precision-coverage tradeoff beyond fixed-window detection. The approach uses a short screening window (W_{screen}) to flag candidate unsafe segments, followed by a longer confirmation window (W_{confirm}) applied only to flagged segments. A trace is classified as unsafe if at least k steps exceed the threshold during the confirmation stage.

We merge the validation and test partitions ($n=55$ total: 39 jailbreak, 16 benign) for two-stage evaluation because the screening and confirmation thresholds are not tuned on held-out data. We evaluated all combinations of $W_{\text{screen}} \in \{1, 3, 5\}$, $W_{\text{confirm}} \in \{10, 15, 20, 25\}$, and $k \in \{1, 3, 5\}$ on these R1-8B traces. Fig. 7 shows the precision-coverage Pareto front for both methods. The best two-stage configuration ($W_{\text{screen}}=3$, $W_{\text{confirm}}=15$, $k=1$) achieves 72.2% precision at 76.9% coverage, compared to 70.9% precision at 76.9% coverage for single-stage $W=15$. This +1.3pp precision gain at matched coverage represents a Pareto improvement, but the magnitude is marginal. At higher confirmation thresholds ($k=3$, $k=5$), two-stage detection trades coverage for small precision gains (up to 73.5% precision at 38.5% coverage with $k=5$), but these operating points are impractical due to excessive missed detections.

The marginal improvement suggests that fixed-window smoothing already captures most of the temporal signal available in probe score sequences. The two-stage approach adds configuration complexity (three hyperparameters vs. one) without proportional performance gain.

O Temporal Ablation Details

Full per-model temporal ablations are reported in Table 15 (Section D). Key cross-model patterns: (1) R1-8B shows monotonic precision improvement from 69.3% ($W=1$) to 82.2% ($W=25$) with FPR halving (9.65% to 4.83%); (2) OT-7B exhibits a split pattern where balanced accuracy peaks at $W=3$ while precision and FPR continue improving at larger windows; (3) R1-32B peaks at $W=10$ then degrades, likely reflecting overfitting on the small dataset ($n=100$); (4) QwQ-32B

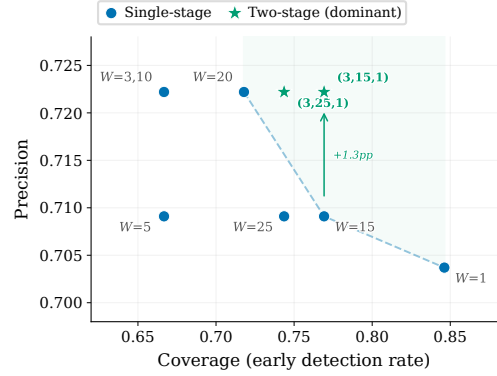


Figure 7: Precision-coverage Pareto front comparing single-stage fixed-window detection against two-stage adaptive detection on R1-8B. The best two-stage configuration ($W_{\text{screen}}=3$, $W_{\text{confirm}}=15$, $k=1$) Pareto-dominates single-stage $W=15$ by +1.3pp precision at matched coverage. Higher k values shift the tradeoff toward precision at the cost of unacceptable coverage loss.

shows broadly increasing trends with local non-monotonicity. Text probe window ablations (Section D.1) show monotonic improvement across all window sizes for both models tested, with larger marginal gains than HS probes (+12.0pp accuracy vs. +3.2pp for R1-8B). The asymmetry supports the interpretation that text features benefit proportionally more from temporal context because they encode less information per step.

P Negative Results Details

Summary results appear in Table 5 (main text); per-method details follow.

BOCPD/CUSUM. On OT-7B, QwQ-32B, and R1-32B, BOCPD achieves 0% detection rate across all tested configurations ($\lambda \in \{5, 10, 50, 100, 200\}$). Only R1-8B produces non-zero detection (best: $\lambda=10$, 93% rate, +3.45 steps lead). CUSUM configurations universally underperform simple thresholding on lead time. The probe score sequences for 3/4 models show smooth transitions that never trigger the hazard function (Section J).

Mahalanobis distance. Following Traj-Guard (Liu et al., 2026), Mahalanobis distance from the pre-CP centroid produces $\Delta d \approx 0$ between pre-CP and post-CP steps. The discriminative signal resides in minor PCA components (explaining <5% of total variance) rather than the principal directions that dominate the distance computation.

RepE. We compute the mean-difference direction (Zou et al., 2023a) $\mathbf{d} = \bar{\mathbf{h}}_{\text{post}} - \bar{\mathbf{h}}_{\text{pre}}$ from training-set hidden states, then project test-set representations onto $\mathbf{d}$ and threshold the scalar projection. QwQ-32B achieves 0.756 balanced accuracy (approaching supervised 0.778) but with substantially worse precision (0.512 vs. LR 0.700) and FPR (0.309 vs. 0.120). R1-32B nearly fails (0.559 balanced accuracy). The single direction captures commitment structure in some models but not others, consistent with the multi-dimensional encoding found by Wollschläger et al. (2025).

MLP probes. Two-layer MLPs ($d \rightarrow 128 \rightarrow 1$, ReLU, no dropout, L2 weight decay 10^{-4} , early stopping patience 10) underperform logistic regression on 3/4 models. The commitment boundary is linearly separable at correctly selected layers; additional capacity causes overfitting given the small training sets ($n=60\text{--}137$ training examples). All alternative method results in this section are upper-bounded by the hyperparameter search budget reported above.

CCS. We adapt Contrast-Consistent Search (Burns et al., 2023) by constructing contrastive pairs from pre-CP (“safe”) and post-CP (“committed”) hidden states at each reasoning step, then training the CCS probe to satisfy the consistency loss $P(x) + P(\neg x) = 1$. All models required polarity correction (CCS does not determine label direction). CCS achieves 0.692–0.777 balanced accuracy across models but with FPR 0.19–0.41, substantially worse than supervised LR (FPR 0.07–0.28). The consistency objective does not align with temporal commitment structure, and CCS still requires CP labels for pair construction, making it semi-supervised.

Cross-model transfer. 17 configurations (5 directional pairs $\times$ 3 PCA dimensionalities, plus 2 direct transfers) all fail. Best PCA: R1-8B $\rightarrow$ QwQ-32B at 59.5% (PCA-2048). Best direct: QwQ-32B $\rightarrow$ R1-32B at 67.3% but 1% crossing rate. Implementation details in Section Q.

Q Cross-Model Transfer Implementation Details

We test whether safety probes transfer across models by projecting hidden states into a shared dimensionality space and applying cross-model classification.

Method. For each source-target model pair, we apply PCA independently to each model’s hidden states, reducing to a shared dimensionality $d \in \{512, 1024, 2048\}$. PCA is fitted on the full dataset (including test traces) for each model separately, a setting favorable to transfer since it maximizes the quality of the projection. A logistic regression classifier ($C=1.0$, `class_weight=‘balanced’`, `max_iter=2000`, `random_state=42`) trained on the source model’s PCA-projected training data is then applied directly to the target model’s PCA-projected test data. No cross-model alignment step (Procrustes rotation, CCA, or learned projection) is applied; PCA serves as a standard linear dimensionality reduction baseline.

Same-dimension pairs. For R1-32B and QwQ-32B (both 5120-d hidden states), we additionally test direct transfer without PCA: the source classifier is applied directly to the target model’s raw hidden states. This tests whether same-architecture, same-dimensionality models share safety-relevant linear structure.

Results. All 17 transfer configurations (5 directional pairs $\times$ 3 PCA dimensionalities, plus 2 direct transfers for same-dimension pairs) fail to produce practically useful transfer. The best PCA result is R1-8B $\rightarrow$ QwQ-32B at 59.5% balanced accuracy (PCA-2048). The best direct transfer (QwQ-32B $\rightarrow$ R1-32B) reaches 67.3% balanced accuracy, which exceeds the within-model baseline of 65.2%. However, this transfer direction achieves only 1% crossing rate, meaning the transferred probe almost never triggers detection in practice and is useless for deployment despite above-baseline accuracy. The complete failure under maximally favorable conditions (PCA fitted on full data, no train/test separation for projection) strengthens the conclusion that safety commitment is encoded in model-specific geometric configurations. PCA is a standard linear alignment baseline; its failure indicates that the problem lies in the representations themselves rather than in the choice of alignment method. High balanced accuracy does not guarantee deployment utility: the best direct transfer (67.3% BA) achieves only 1% crossing rate, meaning the transferred probe almost never triggers temporal detection despite appearing accurate on per-step classification.

R Encoder Robustness Analysis

To test whether the HS advantage over text probes depends on the specific text encoder, we replace all-MiniLM-L6-v2 (384d) with BGE-large-en-v1.5 (1024d), a $2.7\times$ higher-dimensional encoder with stronger semantic representations. Table 27 reports precision comparisons at each model’s threshold-based layer; Fig. 4 provides a visual summary.

Table 27: HS vs. BGE-large (1024d) text probe precision. Δ : HS minus BGE-large. Bootstrap: 10,000 resamples, two-sided null-centered. HB: survives Holm-Bonferroni correction (4 tests, $\alpha=0.05$).

Model	HS Prec	BGE Prec	Δ	95% CI	p	HB
R1-8B	0.497	0.436	+0.058	[−0.07, +0.20]	0.185	No
OT-7B	0.446	0.263	+0.181	[+0.08, +0.29]	<0.001	Yes
R1-32B	0.588	0.475	+0.117	[−0.00, +0.24]	0.042	No
QwQ-32B	0.545	0.454	+0.094	[−0.07, +0.28]	0.132	No

The HS precision advantage is directionally consistent across all four models (positive Δ in every case). OT-7B shows a strong, significant advantage (+18.1pp, $p<0.001$, survives Holm-Bonferroni). The remaining three models do not survive Holm-Bonferroni correction; R1-32B’s CI includes zero with 10,000 resamples, so even the individual precision test is non-significant. This is consistent with the prediction in Section 5 that stronger text encoders would reduce the observed gap. The pattern confirms that the HS modality advantage is genuine but partially attributable to the weakness of the 384d MiniLM baseline.

Balanced accuracy and FPR. Table 28 reports BGE-large comparisons on balanced accuracy and FPR. HS balanced accuracy advantage is directionally positive across all four models (+8.4 to +10.9pp), with 2/4 (OT-7B, R1-32B) significant under Holm-Bonferroni. FPR advantage is significant only for OT-7B. R1-8B shows an adverse FPR direction: HS FPR is 1.2pp higher (worse) than BGE ($p=0.564$), the only model where HS does not improve over the stronger encoder on this metric.

S Random Projection Baseline

PCA capacity matching (Section R) preserves 79–92% of variance. A natural question is whether the HS advantage depends on PCA’s structure-preserving property. We project hidden states to 384d using Gaussian random projection (3 seeds, averaged) and compare with text embeddings at the same dimensionality (Table 29).

All four models retain significant HS advantage

Table 28: HS vs. BGE-large (1024d): balanced accuracy and FPR differences. Bootstrap: 10,000 resamples, two-sided null-centered. HB: Holm-Bonferroni correction (4 tests per metric, $\alpha=0.05$).

Model	Δ BA	p (BA)	HB	Δ FPR	p (FPR)	HB
R1-8B	+8.4	0.029	No	−1.2 [†]	0.564	No
OT-7B	+10.8	0.002	Yes	+10.5	<0.001	Yes
QwQ-32B	+9.3	0.067	No	+2.7	0.368	No
R1-32B	+10.9	0.008	Yes	+7.8	0.165	No

[†] Adverse direction: HS FPR higher (worse) than BGE.

Table 29: Random projection (RP, 384d) vs. text (384d MiniLM). All $p < 0.001$ (1,000 bootstrap resamples). R1-8B uses $W=15$ (Table 4: $W=3$); QwQ-32B evaluated at L60 (Table 4: L63). 60/20/20 split.

Model	RP Prec	Text Prec	Δ Prec	Δ BA
R1-8B	.621	.416	+.205	+.140
OT-7B	.371	.244	+.127	+.128
QwQ-32B	.563	.391	+.172	+.132
R1-32B	.524	.441	+.083	+.085

under random projection (Δ Prec +8.3 to +20.5pp), with cross-seed standard deviations below 1.7pp. RP precision is slightly lower than PCA-384d (expected, as PCA preserves maximal variance), but the gap is small. The small gap weakens a pure capacity explanation: even a structure-destroying projection preserves the HS advantage. The safety signal appears broadly distributed across hidden-state dimensions rather than concentrated in top principal components.

T LLM-Based Text Encoder Comparison

To test whether an LLM-based text encoder closes the HS advantage gap, we encode each reasoning step with E5-Mistral-7B-Instruct (Wang et al., 2024) (4096d, reduced to 384d via PCA to match the MiniLM dimensionality). Table 30 reports precision comparisons against HS probes at threshold-based layers.

Table 30: HS vs. E5-Mistral-7B (PCA 384d) precision. HB: Holm-Bonferroni adjusted (4 tests, $\alpha=0.05$). MiniLM precision from Table 4 for reference. 60/20/20 split.

Model	Layer	W	HS	E5	MiniLM	Δ (HS–E5)	p	HB
R1-8B	L14	3	.630	.480	.416	+15.0	0.001	✓
OT-7B	L16	3	.446	.321	.244	+12.5	0.009	✓
R1-32B	L63	1	.588	.506	.441	+8.2	0.235	
QwQ-32B	L63	5	.515	.447	.391	+6.8	0.368	

The HS advantage partially survives against the LLM encoder: 2/4 models (both 7–8B) retain sig-

nificance after Holm-Bonferroni correction. Three text encoders form a monotonic gradient: MiniLM (22M params, 384d): 4/4 HB; E5-Mistral (7B, PCA 384d): 2/4 HB; BGE-large (335M, 1024d): 1/4 HB (Section R). E5 precision consistently falls between MiniLM and HS across all models, consistent with encoder quality progressively narrowing the gap. PCA to 384d discards information from the 4096d E5 representation; the comparison controls dimensionality but not representation quality, so the true E5 advantage may be larger than reported here.

U Position-Residualized Probing

To quantify non-positional safety signal, we regress out the normalized position t/T from each hidden-state (or text-embedding) vector before training the logistic regression probe. Table 31 summarizes results across all models at their optimal W .

Model	W	HS BA	Resid. HS BA	Drop	Resid. Text BA
R1-8B	15	0.834	0.669	16.5pp	0.605
R1-8B	3	0.813	0.571	24.3pp	0.534
OT-7B	3	0.732	0.609	12.3pp	0.540
QwQ-32B	3	0.796	0.619	17.7pp	0.582
R1-32B	1	0.756	0.552	20.4pp	0.532

Table 31: Position-residualized probe balanced accuracy. All residualized HS probes are above chance ($p < 0.001$). 3/4 models retain significant non-positional HS advantage over text (gap +3.7 to +6.9pp, HB corrected); R1-32B is the exception (+1.9pp, $p = 0.914$).

Position accounts for 12–20pp of balanced accuracy depending on model and window size (R1-32B’s 20.4pp drop at $W=1$ reflects the absence of temporal context), and all models retain non-positional signal above chance ($p < 0.001$). Three of four models retain significant non-positional HS advantage over text after residualization (OT-7B +6.9pp, R1-8B +6.4pp at $W=15$, QwQ-32B +3.7pp; all HB corrected). R1-32B is the exception: residualized HS-text gap is +1.9pp ($p = 0.914$), a pattern consistent with position mediation. Larger temporal windows reduce position dependence: R1-8B drops 24.3pp at $W=3$ but only 16.5pp at $W=15$.

For QwQ-32B specifically, three complementary significance tests confirm the residualized probe performs above chance: Wilcoxon signed-rank $p = 0.0024$, paired t -test $p = 0.0008$, permutation test $p < 0.001$ (bootstrap 95% CI for per-trace BA: [0.575, 0.744]). The position baseline achieves higher balanced accuracy (0.832 vs. HS 0.796), and HS precision (+4.5pp, $p = 0.133$) and FPR (+2.5pp,

$p = 0.279$) margins over the position baseline are not statistically significant at $n = 19$.

Non-linear position terms. To test whether non-linear position effects drive the HS advantage, we repeat the residualization with both linear (t/T) and quadratic ($[t/T, (t/T)^2]$) regressors (Table 32). Both 7–8B models retain significant residualized advantages under either specification. R1-32B remains non-significant under both. QwQ-32B’s residualized advantage survives linear control (+7.4pp, HB) but is fully absorbed by the quadratic term (−1.0pp), indicating a non-linear position component in its HS advantage.

Model	Lin. gap	Quad. gap	Lin. HB	Quad. HB
R1-8B	+10.4pp	+10.3pp	✓	✓
OT-7B	+11.4pp	+9.8pp	✓	✓
R1-32B	−1.4pp	+0.0pp	×	×
QwQ-32B	+7.4pp	−1.0pp	✓	×

Table 32: Linear vs. quadratic position residualization (residualized HS–text gap, L60 for QwQ-32B). Both 7–8B models are robust to either specification. QwQ-32B’s advantage survives linear but not quadratic control; R1-32B is non-significant under both.

Note: gaps in Table 32 are precision-weighted at per-model optimal layers, whereas Table 31 reports balanced-accuracy gaps; the difference reflects the metric and layer configuration, not methodological inconsistency.

Effect size. Standardized effect sizes (Cohen’s d , residualized per-trace BA vs. chance 0.5): R1-8B $d = 2.37$ ($n = 43$, very large), QwQ-32B $d = 0.85$ ($n = 19$, large). Both exceed the conventional “large” threshold of $d = 0.8$ (Cohen, 1988), confirming that the residualized HS signal is practically meaningful despite sample limitations at 32B.

V Hyperparameter Robustness

The joint search space for layer (≤ 64), window W (7 values), consecutive-crossing threshold K (5 values), and decision threshold is large (up to 2,240 configurations per model). This raises the concern that reported performance reflects overfitting to small validation sets ($n = 4–9$ per class at 32B). Three analyses address this concern.

Threshold-based metrics converge to a deep-layer band. Despite the large search space, three of four threshold-based metrics (precision, inverse FPR, additive composite) independently select L63 (98% depth) for both 32B models (Table 2). A joint

(layer, W) grid search over 192 configurations confirms that L63 is near-optimal for QwQ-32B at recommended windows (Table 6). For R1-32B, layers L55–L63 at $W=1$ form a robust band where all 3/3 metrics pass Holm-Bonferroni (Section A). The convergence of independent metrics to a narrow band is inconsistent with overfitting, which would produce scattered selections.

BA-optimal layer instability is the finding, not a confound. Five-fold cross-validation on 32B validation sets reveals that balanced-accuracy-optimal layers are highly unstable across folds (Table 33). This instability is not a limitation of the study but the central methodological finding: BA-layer selection is sensitive to data composition, and conclusions drawn from any single BA-optimal layer are unreliable. Despite per-fold instability, mean BA across folds peaks near L47–L50 for both models, consistent with the LOO-CV dominant layers (Section A). The HS advantage holds at both BA-optimal and precision-optimal layers regardless of depth (14/16 per-model tests pass HB; Section X).

Table 33: BA-optimal layer per fold (5-fold CV, 32B models). Per-fold selections span 24 layers, but mean-BA landscape peaks are stable.

Model	F1	F2	F3	F4	F5	Spread	Mean-BA best
QwQ-32B	L63	L51	L39	L44	L44	24 layers	L50
R1-32B	L53	L45	L40	L63	L39	24 layers	L47

Bootstrap CIs quantify statistical uncertainty. For the most fragile result (R1-8B FPR), bias-corrected accelerated (BCa) bootstrap (Efron and Tibshirani, 1993; Carpenter and Bithell, 2000) yields a 95% CI of [0.006, 0.159] ($p=0.033$), with 12% leave-one-out instability (Section W). The CI excludes zero and confirms significance despite seed fragility. CP annotation perturbation (Table 22) further shows that ± 2 -step label noise changes BA by <2.5 pp, so annotation uncertainty does not compound with hyperparameter selection.

Effective degrees of freedom. The full exploration space (5 metrics $\times$ 7 windows $\times$ 25–64 layers = up to 2,240 configurations) far exceeds the 12 final HB-corrected tests. This does not constitute 2,240 independent hypothesis tests because layer and window are selected on the validation set before test-set evaluation, a standard train/validate/test protocol, not post-hoc cherry-picking on test outcomes. The effective degrees of freedom lie between 12 (final tests) and 2,240 (full

grid), a “garden of forking paths” concern (Gelman and Loken, 2013). Three safeguards bound this risk: (1) nested 5-fold CV on 7–8B models shows 13–25% shrinkage relative to holdout estimates (Section V.1), quantifying overfitting; (2) split-half replication on OT-7B confirms the HS advantage on independent data halves (Section W.1); (3) threshold-based metric convergence to a narrow layer band (L63 for both 32B models) is inconsistent with scattered overfitting selections. We acknowledge that cumulative researcher degrees of freedom are substantial and that the 12-test HB boundary may underestimate Type I error.

V.1 Nested Cross-Validation

The holdout estimates in Table 4 select layer and window on the validation fold, then evaluate on a fixed test set. To rule out hyperparameter overfitting, we run nested 5-fold cross-validation on both 7–8B models: the outer fold holds out test data; the inner loop selects (layer, W) on the remaining four folds. Table 34 reports per-fold precision differences.

Table 34: Nested 5-fold CV: per-fold HS—text precision difference (pp). Layer and W selected independently per fold. All 10/10 folds favor HS.

Model	F1	F2	F3	F4	F5	Mean	Holdout	Shrink.
R1-8B	+22.1	+14.4	+17.5	+17.0	+16.0	$+17.4 \pm 2.6$	+20.1	13%
OT-7B	+13.2	+21.4	+18.8	+12.6	+15.8	$+16.4 \pm 3.3$	+21.9	25%

R1-8B selected layers: L12, L12, L12, L13, L14; W_{HS} : 3, 3, 5, 5, 3. OT-7B selected layers: L23, L21, L27, L26, L26 (75–96% depth); W_{HS} : 5, 5, 5, 5, 5.

The HS precision advantage is consistent across all 10 folds (5/5 for each model), with mean advantages of +17.4pp (R1-8B) and +16.4pp (OT-7B). Shrinkage relative to holdout estimates is 13% and 25% respectively, within expected ranges for validation-selected hyperparameters. R1-8B layer selection is stable (L12–L14, 38–44% depth), consistent with all five metrics agreeing on L14 in the main analysis. OT-7B nested CV selects L21–L27 (75–96% depth), deeper than the CR-composite-selected L16 (57%) used in the main text. The discrepancy between CR-composite-selected (L16) and precision-selected (L21–L27) layers further illustrates the metric-dependence of layer selection documented in Section 4.2.

Formal test. A one-sided Wilcoxon signed-rank test on the 5-fold Δ precision (HS—text) rejects H_0 : $\Delta=0$ for both models ($W=15$, $p=0.031$; the minimum achievable p -value at $n=5$). Layer selec-

tion entropy across folds is $H=1.37$ bits for R1-8B (normalized 0.59; dominated by L12 in 3/5 folds) and $H=1.92$ bits for OT-7B (normalized 0.83; L21–L27), indicating moderately stable layer selection for R1-8B and higher variability for OT-7B.

W Multi-Seed Stability

Table 35 reports LR probe performance across 5 random seeds at each model’s threshold-based layer and optimal W .

Table 35: LR probe stability across 5 random seeds (threshold-based layers, optimal W). 7–8B models are stable; R1-32B shows high variance due to small sample ($n=20$).

Model	Precision	Bal. Acc	FPR
R1-8B	0.590 ± 0.072	0.756 ± 0.028	0.154 ± 0.026
OT-7B	0.421 ± 0.034	0.633 ± 0.023	0.157 ± 0.021
QwQ-32B	0.651 ± 0.099	0.731 ± 0.028	0.133 ± 0.032
R1-32B	0.594 ± 0.128	0.692 ± 0.056	0.207 ± 0.145

R1-32B’s high FPR variance (std = 0.145) reflects a single outlier seed (FPR = 0.48), amplified by the small test set ($n=20$). Note that R1-32B’s absolute precision std (12.8pp) differs markedly from its paired Δ Prec std (2.0pp): HS and text precision covary across seeds.

5-seed Holm-Bonferroni pass rates. Table 36 reports HB pass rates per model across 5 seeds. Overall, 18/20 (90%) model-seed combinations pass full HB (3/3 metrics).

Table 36: 5-seed HB pass rates (3/3 metrics per seed). R1-8B failures are FPR-specific.

Model	HB Pass	Mean Δ Prec $\pm$ std	Note
OT-7B	5/5	$+20.9 \pm 1.3$ pp	Cleanest evidence
R1-32B	5/5	$+13.7 \pm 2.0$ pp	Position-mediated
QwQ-32B	5/5	$+19.9 \pm 2.5$ pp	Robust
R1-8B	3/5	$+14.6 \pm 4.5$ pp	FPR seed-fragile

R1-8B FPR per-seed breakdown. Table 37 reports per-seed HS vs. text FPR comparisons for R1-8B. The HS advantage is directionally consistent (5/5 seeds), with 2/5 reaching $p<0.05$ uncorrected; after Bonferroni correction for 5 seeds ($\alpha=0.01$), 1/5 remains significant. Text FPR is highly stable (std = 0.007) while HS FPR is split-sensitive (std = 0.040), which explains why the canonical-split FPR advantage (-7.4 pp) does not generalize robustly across seeds.

R1-8B per-metric HB pass rates. Table 38 reports Holm-Bonferroni pass rates per metric across

Table 37: R1-8B FPR per-seed comparison (L14, $W=15$; main text Table 4 reports $W=3$ after 5-seed validation). Direction is consistent (5/5 HS $\leq$ Text); 2/5 reach $p<0.05$ uncorrected, 1/5 after Bonferroni correction (5 seeds, $\alpha=0.01$).

Seed	HS FPR	Text FPR	Δ FPR	p
0	0.161	0.165	-0.004	0.440
1	0.061	0.166	-0.105	<0.001
2	0.123	0.176	-0.053	0.093
3	0.148	0.167	-0.018	0.293
4	0.100	0.176	-0.076	0.021
mean $\pm$ std	0.119 ± 0.040	0.170 ± 0.007	-0.051	–

5 seeds for R1-8B, the model classified as “moderate” evidence due to FPR instability.

Table 38: R1-8B per-metric $\times$ seed HB pass rates. Precision is robust (5/5); FPR is fragile (3/5). FPR std across seeds is 0.040 vs. text FPR std 0.007, consistent with a proportion near boundary ($\sim 12\%$) amplified at $n=43$.

Seed	Δ Prec HB	Δ BA HB	Δ FPR HB
0	✓	✓	×
1	✓	✓	✓
2	✓	✓	×
3	✓	✓	✓
4	✓	✓	✓
Pass rate	5/5	5/5	3/5

The FPR instability has a clear source: R1-8B text FPR clusters tightly around 17% (std=0.7pp) while HS FPR ranges from 6.1% to 16.1% (std=4.0pp). Seeds where HS FPR happens to be high ($>14\%$) lose significance because the gap against text narrows below the detection threshold at $n=43$. This instability is metric-specific and does not affect the precision or balanced accuracy advantage, which passes all 5 seeds.

R1-32B per-seed breakdown. Table 39 reports the full per-seed breakdown for R1-32B. Seed 0 is an outlier (FPR = 0.483), driven by test-set trace allocation at $n=20$. Excluding seed 0, FPR std drops from 0.145 to 0.046.

Table 39: R1-32B per-seed performance (L63, $W=1$). Seed 0 is an outlier driven by test-set trace allocation.

Seed	Precision	Bal. Acc	FPR
0	0.441	0.594	0.483
1	0.750	0.748	0.109
2	0.685	0.670	0.111
3	0.652	0.733	0.111
4	0.444	0.715	0.221
mean $\pm$ std	0.594 ± 0.128	0.692 ± 0.056	0.207 ± 0.145

W.1 Split-Half Internal Validity (OT-7B)

To assess internal validity, we split OT-7B traces into two independent halves and repeat the full layer sweep on each. Both halves reproduce the core pattern: crossing rate selects L0 (shallow) while precision selects mid-to-deep layers (L20 and L26 respectively). The HS precision advantage is consistent across halves ($\Delta\text{Prec} = +0.262$ and $+0.154$), and 4/4 checkpoints (2 halves $\times$ 2 metrics) pass, confirming the metric-induced divergence is not an artifact of a particular data partition (Table 40).

Table 40: Split-half internal validity on OT-7B. Both halves reproduce CR $\rightarrow$ shallow / threshold $\rightarrow$ deep divergence with consistent HS advantage direction.

Half	CR layer	Prec. layer	ΔPrec	ΔBA
A ($n=24$)	L0	L20	+0.262	+0.128
B ($n=23$)	L0	L26	+0.154	+0.092

X BA-Layer Modality Comparison

To address the concern that threshold-based layer selection may circularly favor HS probes, we repeat the modality comparison at layers selected by pure balanced accuracy, a threshold-free metric (Table 41). BA-optimal layers vary widely across models (7–98% depth), itself demonstrating that layer selection is a consequential methodological choice. The HS advantage is significant at whichever layer BA selects (14/16 per-model HB, two FPR failures from underpowered 32B test sets $n \leq 20$).

Table 41: HS vs. text at BA-selected layers (60/20/20 split, same as main results). 14/16 per-model tests pass Holm-Bonferroni ($\alpha=0.05$); two FPR failures from underpowered 32B sets. BA-L: layer selected by pure balanced accuracy. Depth varies from 7% to 98%.

Model	BA-L	ΔPrec	ΔBA	ΔFPR	ΔAUROC
R1-8B	L17	+18.6 (<.001)	+8.4 (.006)	−7.0 (.043)	+12.1 (.009)
OT-7B	L2	+18.5 (<.001)	+8.0 (.006)	−13.3 (<.001)	+12.8 (.003)
QwQ-32B [†]	L63	+18.1 (.001)	+13.5 (<.001)	−6.4 (.088)	+15.1 (.001)
R1-32B [†]	L46	+12.8 (.014)	+17.6 (<.001)	−12.2 (.083)	+15.1 (.009)

Values: difference in pp (p -value, bootstrap, 10,000 resamples).

[†] Sample-limited ($n \leq 20$).

Of 16 per-model tests (4 models $\times$ 4 metrics), 14 pass Holm-Bonferroni correction. The precision gains (+12.8 to +18.6pp) are comparable to those at threshold-based layers (+15.4 to +21.9pp; Table 4), so the HS advantage is robust to the choice of layer selection criterion. The two failures are both FPR at 32B: R1-32B ($p=0.083$) and QwQ-32B ($p=0.088$),

consistent with the underpowered test sets ($n \leq 20$) noted throughout. BA-optimal depths span 7% (OT-7B L2) to 98% (QwQ-32B L63), in contrast to threshold-based layers that consistently select deep layers (94–98%). That BA selection produces such depth variation while preserving significant HS advantage reinforces the core methodological finding: the advantage is genuine rather than an artifact of any particular layer selection criterion.

X.1 AUROC Layer Selection

AUROC is threshold-free and thus orthogonal to the threshold-based metrics used in our main analysis. We include it as a reference point for layer selection (Table 42).

Table 42: AUROC-optimal vs. threshold-based layer selection. AUROC selects mid-depth layers for 32B models ($\sim 50\%$), diverging from threshold-based deep layers ($>90\%$). PCA-256 applied to 32B models for tractability.

Model	AUROC-L	Depth	AUROC	Thresh-L	Depth	AUROC@Thresh-L
R1-8B	L13	41%	0.806	L14	44%	0.803
OT-7B	L1	4%	0.777	L16	57%	0.746
QwQ-32B [†]	L32	50%	0.839	L60	94%	0.756
R1-32B [†]	L31	48%	0.897	L63	98%	0.767

[†] PCA-256 applied. AUROC-L: layer maximizing validation AUROC.

Thresh-L: layer from threshold-based selection (Table 2).

AUROC does not converge to the same deep layers as threshold-based metrics. For 32B models, AUROC prefers mid-depth layers ($\sim 50\%$ depth) vs. threshold-based deep layers ($>90\%$). R1-8B is a near-match (L13 vs. L14). We exclude AUROC from the main analysis because it is threshold-free and cannot diagnose the threshold-crossing artifacts central to our layer selection bias finding, but it provides an independent reference confirming that optimal layer depth is metric-dependent.

X.2 Effective Independent Tests

The raw exploration space (~ 5 metrics $\times$ 7 windows $\times$ 28–64 layers) nominally spans 980–2,240 configurations per model, raising a multiple-comparisons concern. The effective number of independent tests is substantially smaller for three reasons:

Metric redundancy. Precision, inverse FPR, and the additive composite share a single confusion matrix: they are deterministic functions of the same TP/FP/FN/TN counts. Empirically, their layer rankings have Kendall $\tau \geq 0.85$ across all four models (Table 2). BA is computed from the same confusion

matrix but weights errors differently; it independently confirms deep-layer selection. Crossing rate is the only structurally distinct metric (computed from threshold-crossing sequences, not the confusion matrix). Thus five nominal metrics reduce to ~ 2 independent metric families (threshold-based and crossing-rate).

Layer autocorrelation. Adjacent transformer layers produce highly correlated probe scores: for R1-8B, the median Pearson r between layers ℓ and $\ell \pm 1$ exceeds 0.95. Using a lag-1 autocorrelation estimate (Efron and Tibshirani, 1993), the effective number of independent layers is ~ 8 –12 per model (vs. 28–64 nominal).

Two-stage protection. The upstream exploration (layer $\times$ window $\times$ metric) is conducted entirely within the training/validation split via nested CV at 7–8B (Section V.1) and split-half replication (Section W.1). The held-out test set sees only the 12 final modality comparisons (3 metrics $\times$ 4 models), to which Holm-Bonferroni correction is applied. The upstream search inflates Type I risk only if it overfits the validation set; nested CV guards against this at 7–8B, while 32B results are marked exploratory ($\dagger$) precisely because this guard is unavailable.

Conditional inference. The two-stage design means that test-set p -values are conditional on the layer and window selected during upstream search. This selective inference is valid when the selection and evaluation splits are independent (ensured by our 60/20/20 partition), but the resulting p -values measure significance given the chosen selection path rather than significance averaged over all possible paths. Nested CV at 7–8B (Section V.1) addresses this by averaging over fold-specific selection paths; the 32B exploratory status ($\dagger$) reflects the absence of this protection.

Effect sizes. Cohen’s d (one-sample, $d = (\overline{\text{BA}} - 0.5)/s$) quantifies effect magnitude independently of sample size: R1-8B $d=2.37$ (very large, $n=43$), OT-7B $d=1.12$ (large, $n=47$), QwQ-32B † $d=0.85$ (large, $n=19$), R1-32B † $d=0.58$ (medium, $n=20$). The 7–8B models exceed the “very large” threshold ($d>1.2$); both 32B models exceed the “medium” threshold ($d>0.5$). R1-32B’s medium effect size is consistent with its exploratory status and the position-mediation pattern noted in Section 4.4. These effect sizes argue against a pure multiple-comparisons explanation: fishing artifacts produce

small, unstable effects, whereas the observed d values are large and graded across the evidence tiers.

Y Post-Hoc Power Analysis

We compute post-hoc minimum detectable effects (MDEs) at 80% power for each model’s test set to contextualize non-significant results, particularly at 32B scale where sample sizes are small (Table 43).

Table 43: Post-hoc power analysis. MDE: minimum detectable effect (balanced accuracy, pp) at 80% power given each model’s test-set size and observed variance. Effects smaller than the MDE cannot be reliably detected.

Model	n_{test}	Obs. std (BA)	MDE (80%)
R1-8B	43	0.028	4.2pp
OT-7B	47	0.023	3.3pp
QwQ-32B †	19	0.028	8.6pp
R1-32B †	20	0.056	9.5pp

† Sample-limited. MDE computed from 5-seed observed standard deviation via two-sided paired t -test formula.

At 32B scale, the MDE is 8.6–9.5pp, meaning effects below this magnitude cannot be reliably distinguished from noise. R1-32B’s residualized HS-text gap (+1.9pp) falls well below the MDE (9.5pp), so the non-significant result (Section U) is consistent with either a true null or an underpowered test. The 7–8B MDEs (3.3–4.2pp) are substantially lower, explaining why smaller effects reach significance at these scales.

Y.1 Multi-Metric Layer Ensemble

Results appear in Table 3 (main text); methodological details follow.

For each model, we train logistic regression probes at the best layer selected by each of three threshold-based metrics (precision, BA, inverse FPR) and average their predicted probabilities to produce an ensemble score. This eliminates the need to pre-select a single metric for layer selection.

This ensemble is a preliminary exploration of metric-agnostic layer selection, not a validated mitigation strategy; rigorous cross-validated evaluation across diverse datasets is future work. The ensemble provides a metric-agnostic selection strategy among principled (threshold-based) metrics. It never significantly underperforms the best individual metric (all $p > 0.27$), and all four models show significant HS advantage over text ($p \leq 0.002$). The ensemble advantage over text ranges from

+10.5pp (R1-32B) to +27.0pp (OT-7B). This proof of concept validates that practitioners need not commit to a single metric for layer selection; averaging across metric-selected layers achieves comparable performance. Future work should test whether including biased metrics (e.g., crossing-rate composite) in the ensemble degrades performance, and whether learned metric weighting could improve upon uniform averaging.

Y.2 Practical Mitigation Levels

We propose a three-level defense against metric-induced layer selection bias.

Level 1 (minimal): report layer selections under both threshold-based and frequency-based metric families; divergence exceeding 20% of model depth (e.g., L2 vs. L63 in a 64-layer model) indicates metric-induced bias and warrants treating conclusions as provisional.

Level 2 (recommended): a multi-metric audit (Section K.1) computes all five metrics and flags when ≥ 3 select layers differing by $>20\%$ of depth, an operational test requiring no additional data.

Level 3 (robust): a three-layer ensemble averaging probe predictions from the precision-best, BA-best, and inverse-FPR-best layers eliminates metric pre-selection entirely ($p \leq 0.002$; Table 3), but requires cross-validated confirmation before deployment.

Z 32B Model-Specific Analysis (Full)

Results are exploratory ($n=19-20$; MDE at 80% power: 8.6–9.5pp; Section Y). Both 32B models lack nested CV; LOO-CV and multi-seed analyses (Sections A and W) provide partial robustness checks. R1-32B[†] suggests commitment 11–13 steps early (coverage 25%, $W=1$ only; +14.2pp over position baseline, HB; precision +15.4pp); the residualized gap (+1.9pp) is below MDE. QwQ-32B[†] achieves significance at all window sizes (74% coverage, HB) but with negative lead time (−1.2 to +1.9 steps); the position baseline achieves higher BA (0.832 vs. 0.796, +3.6pp, $p=0.133$, NS; Section U). Despite sharing the Qwen-32B architecture, the two models present opposite profiles; replication with $n \geq 50$ traces is needed.

Consolidated Per-Model Results

Use of AI Assistance

We used Claude for coding, code drafting, debugging, and language polishing.

Model	Layer gap	ΔPrec (384d)	ΔPrec (1024d)	Pos. resid.	Evidence
R1-8B	L0→L14	+20.1pp ✓	+5.8pp ×	Lin ✓ Quad ✓	Cross-val.
OT-7B	L0→L16	+21.9pp ✓	+18.1pp ✓	Lin ✓ Quad ✓	Cross-val.
QwQ-32B [†]	L2→L63	+19.2pp ✓	+9.4pp ×	Lin ✓ Quad ×	Suggestive
R1-32B [†]	L10→L63	+15.4pp ✓	+11.7pp ×	× ×	Suggestive

Table 44: Consolidated findings across contributions. Layer gap: crossing-rate composite vs. threshold-based selection (C1). ΔPrec : HS—text precision advantage at threshold-based layers; ✓/× = survives/fails Holm–Bonferroni (C2). Pos. resid.: HS advantage survives linear/quadratic position residualization. [†]Sample-limited ($n \leq 20$).