Pre-CoT Signals Encode Difficulty Awareness, Not Answer Commitment: Evidence from Activation Steering Across Training Regimes

Anonymous ACL submission

Abstract

A growing line of work shows that the final answer of a chain-of-thought (CoT) is already *decodable* from a model's hidden state *before* any CoT token is emitted—suggesting the model has pre-committed to an answer. But does the pre-CoT signal reflect a committed answer, or merely the model's assessment of how hard the problem is? We distinguish these readings by probing and then steering the pre-CoT residual stream across a four-stage post-training ladder (base, instruction-tuned, RLVR-tuned, and reasoning-distilled) of QWEN2.5-7B and a cross-family QWEN3-4B ladder. Pre-CoT correctness probes reach AUC 0.83–0.94 on trained variants, confirming the model can predict its own correctness before reasoning. Steering along this direction does not change final-answer accuracy in any cell—but it does shift the *length* of the CoT, and the direction of the shift depends on training regime: on MATH-500, instruction-tuned models write *more* when steered toward the hard centroid (slope -7.7 words/α) while reasoning-distilled models write *less* ($+60.7$ words/α), with SimpleRL in between ($+2.1$). The same pre-CoT signal thus produces opposite behavioral responses—hedging versus compression—depending on how the model was trained. We conclude that the pre-CoT signal encodes *difficulty awareness*: the model knows how hard the problem is before it reasons, and this knowledge modulates how it reasons, not whether it reasons correctly.

1 Introduction

Whether chain-of-thought (CoT) reasoning is causally responsible for a model's final answer, or merely a post-hoc narrative attached to an answer that was already decided, is a central question for both the safety and the science of large reasoning models. A recent wave of probing studies shows that a linear classifier on the residual stream immediately before any CoT token is generated can predict the final answer with high accuracy (Cox et al., 2026; Zhang et al., 2025; Boppana et al., 2026; David, 2025). This decodability is often interpreted as evidence that the model has *pre-committed* to an answer before reasoning. But what does the pre-CoT state actually encode? It could reflect a committed answer—or it could reflect the model's assessment of how hard the problem is, which correlates with correctness without determining it.

Two gaps in this literature motivate our work. First, the bulk of evidence is collected on *instruction-tuned* models, with reasoning-trained models—models that have undergone reinforcement learning with verifiable rewards (RLVR) or large-scale SFT on R1-style reasoning traces—left under-examined. Cox et al. (2026) flag this explicitly as a limitation: in reasoning models, the CoT is optimized as a latent that contributes to task reward, which could plausibly change the faithfulness-usefulness trade-off. Second, the relationship between *probe AUC* (a correlational quantity) and *what the pre-CoT state represents* is often left implicit. A probe can succeed because the model has a private commitment to a specific answer, or simply because the pre-CoT state encodes "how likely this model is to solve this problem at all." These two readings make opposite predictions under intervention: if the signal is commitment, steering should change the answer; if it is difficulty awareness, steering should change behavior (e.g., effort allocation) without changing the answer.

This paper distinguishes the two readings by steering along the pre-CoT correctness direction across a four-stage post-training ladder built on a shared base—QWEN2.5-7B in base, instruction-tuned, RLVR-trained (SIMPLERL-ZOO), and reasoning-distilled (R1-DISTILL) variants—replicated on a second base family (QWEN3-4B). On GSM8K and MATH-500, we train pre-CoT correctness probes and then steer the residual stream along the probe direction, measur-

ing both final-answer accuracy and CoT length.

Findings.

1. **The correlational ladder is mostly a base-vs-trained gap, not a method-by-method gap.** On MATH-500, base models reach AUC 0.63 ± 0.05 at best layer while all three trained variants land in 0.81–0.85 (5-seed mean). Single-seed apparent gaps among trained variants—e.g., R1-Distill 0.94 vs. instruct 0.85—shrink to within-std bands when seeds are varied, because the minority class is small at high accuracy.

2. **Steering the pre-CoT signal changes CoT length, not correctness—and the direction depends on training.** In all 8 (model, benchmark) cells, steering the pre-CoT residual toward the wrong-answer centroid produces no detectable accuracy drop (every recovery-rate CI includes 1.0). But the intervention is not inert: on MATH-500, it shifts CoT length, and the slope increases monotonically with reasoning training—instruct -7.7 words/α (steering toward hard *lengthens*), SimpleRL $+2.1$, R1-Distill $+60.7$ (steering toward hard *shortens*). The same difficulty signal produces opposite behavioral responses—hedging versus compression—across training regimes.

The takeaway: models encode how hard a problem is for them *before* generating any CoT tokens, and this difficulty awareness modulates reasoning behavior (longer or shorter CoT) without modulating reasoning outcome (correctness). The behavioral signature of that modulation—hedging versus compression—depends on how the model was trained.

2 Related Work

CoT faithfulness. Lanham et al. (2023) and Turpin et al. (2023) establish that CoT does not always reflect the actual computation: under biasing prompts, models generate CoT explanations that rationalize biased answers without acknowledging the bias (Turpin et al., 2023), and final answers can survive systematic corruption of intermediate steps (Lanham et al., 2023). Arcuschin et al. (2025) extend this to natural in-the-wild CoT, finding implicit post-hoc rationalization at rates of 0.04–2.17% across frontier thinking models (and substantially higher on smaller non-thinking models). Recent RL-specific analyses argue that the ends-justify-thoughts pattern can be *induced* by RL training (Howe and Carroll, 2026; Mirtaheri and Belkin, 2026).

Pre-CoT and intra-CoT probing. Wang et al. (2025) probe confidence step-by-step during CoT generation, finding that models often have their answer well before CoT completes ("early answering") and that high answer accuracy is not a reliable indicator of correct intermediate reasoning. Zhang et al. (2025) probe intermediate states of reasoning models for self-verification, achieving high AUC and using this signal for early-exit. Cox et al. (2026) sharpen the focus to the position immediately before any CoT token: a linear probe at that location predicts the final answer with >0.9 AUC, and steering along the probe direction flips $>50\%$ of answers. Boppana et al. (2026) report final-answer probes at $\sim 88\%$ accuracy averaged across token positions during CoT, and use them to characterize *performative* reasoning—cases where the model has internally committed before its CoT reflects that commitment. Yuan et al. (2026) train probes during CoT and obtain 0.95 AUROC for trace correctness from the full trace and 0.79 from the very first reasoning step. David (2025) push the prediction window earlier in the token sequence (AUC ~ 0.84 at $t = 4$ reasoning tokens). Among these, Yuan et al. (2026) most directly attempt to separate *probe accuracy* from *causal commitment*, reporting that four interventions on the diagnostic signal fail—in their setting the probe is diagnostic but not causal. Our work re-examines this question in a different setting: rather than intervening on a trace-correctness probe during CoT, we run the steering experiment on a pre-CoT answer-commitment probe (Cox et al., 2026) as an independent measurement, decoupling the predictive and causal sides of the same direction.

Activation steering on reasoning models. Venhoff et al. (2025) extract steering vectors from reasoning models and show that small linear directions modulate behaviors such as backtracking, uncertainty expression, and example generation for hypothesis validation. Sinii et al. (2026) provide a mechanistic analysis of trained steering vectors on reasoning models, finding that last-layer vectors act as token-substitution biases concentrated on the first generated token while penultimate-layer vectors reshape MLP and unembedding contribu-

tions; complementary work shows that base models already contain analogous directions that only induce backtracking after reasoning finetuning (Ward et al., 2025). Arditi et al. (2024) demonstrate the same single-direction phenomenon in the refusal setting. We use the same difference-of-means construction (Arditi et al., 2024) for the steering vector, but apply it to the probe direction itself rather than to behavioral attributes, and read out final-answer correctness rather than fine-grained CoT properties.

RLVR analysis. The question of whether RLVR teaches new reasoning or merely sharpens distributions already present in the base is contested. Yue et al. (2025) argue the latter; Guo et al. (2025); Zeng et al. (2025) are read by the community as supporting the former. We do not adjudicate this directly, but our ladder—base, general SFT, RLVR, reasoning-distill—spans the relevant comparisons.

3 Method

3.1 Pre-CoT correctness probe

For a model M and a prompt p, let $h_l(p) \in \mathbb{R}^d$ be the residual stream activation at decoder layer l at the *last token of the templated prompt*, i.e., the position immediately before M emits its first generation token. We collect $(h_l(p), y(p))$ pairs where $y(p) \in \{0, 1\}$ indicates whether M's final answer to p (extracted from the full generation produced by M under sampling temperature 0.6) is correct. For each layer l, we train a logistic regression with balanced class weights on an 80/20 split and report the test AUC. The *best probe layer* for a (model, benchmark) pair is $\arg\max_l \mathrm{AUC}_l$.

This is the correctness-probe variant of the Cox-style protocol (Cox et al., 2026); it asks whether the pre-CoT residual stream linearly separates problems the model will solve from those it will fail.

3.2 Activation steering

At the best probe layer l^*, we compute centroid vectors $\mu_1 = \mathrm{mean}_{p:y(p)=1} h_{l^*}(p)$ and μ_0 analogously, and define the steering direction $v = \mu_1 - \mu_0$ (the wrong−correct direction at layer l^*). For inference, we add a hook to layer l^* that modifies the residual stream as

$$h'_{l^*} = h_{l^*} + \alpha \, \|v\| \, \hat{v}, \quad \hat{v} = \frac{v}{\|v\|} \quad (= h_{l^*} + \alpha v) \tag{1}$$

i.e., we sweep $\alpha \in \{-2, -1, 0, +1, +2\}$ in units of $\|v\|$.[1] $\alpha < 0$ pushes the pre-CoT state toward the wrong centroid; $\alpha > 0$ pushes it toward correct. The hook stays active throughout generation, adding αv to layer l^*'s output at every token position rather than only at the pre-CoT position; under it, the model produces the full CoT and final answer with the same temperature and decoding parameters as the baseline run.

Recovery rate. For each (model, benchmark) we report the *recovery rate*

$$\mathrm{Rec} = \frac{\mathrm{acc}(\alpha = -1)}{\mathrm{acc}(\alpha = 0)}, \tag{2}$$

the fraction of baseline accuracy retained when the pre-CoT state is pushed one $\|v\|$ toward the wrong centroid. Rec $= 1$ means the CoT fully overrides the pre-CoT nudge; Rec $\ll 1$ means the model's final answer is driven by the pre-CoT residual, which the CoT does not correct. Rec > 1 is sampling noise around full override.

The recovery rate has the property that the *similar* probe direction whose AUC was reported in the correlational protocol is the direction whose causal role is now being tested. This couples the two experiments: a high probe AUC combined with a low recovery rate would be a *positive* commitment finding; a high probe AUC combined with a high recovery rate is the configuration our results fall into.

Confidence intervals and splits. Because recovery is a ratio of two accuracies estimated on the same n_{test} problems, we report a paired bootstrap: resampling problems with replacement (10^4 iterations, the problem as the resampling unit), we obtain a 95% CI for $\mathrm{acc}(\alpha)$ at each α and for the recovery rate jointly. Steering has a *detectable effect* when the recovery CI excludes 1.0. We run the protocol under two splits: *fresh*—a random held-out set of $n_{\text{test}} = 200$ problems, independent of the probe, so the best-layer choice never saw the steered problems—and *probe*, which reuses the exact held-out fraction `train_probe.py` reports AUC on. The fresh split is our primary condition; the probe split is a robustness check (App. C).

[1] In our setup $\|v\|$ is small relative to the residual stream norm at layer l^*, so $\alpha = \pm 1$ is a perturbation of roughly 10% of the residual norm. We confirmed in pilot runs that larger relative magnitudes start to degrade text fluency.

3.3 Implementation

We generate the CoT in vLLM (Kwon et al., 2023) for throughput, then *replay* the prompt in Hugging Face Transformers to obtain hidden states—vLLM does not currently expose residual streams. For steering, we run inference in Transformers with a forward hook on layer l^* implementing Eq. 1, using the same sampling configuration as generation.

4 Experimental Setup

Models. We construct two cross-family ladders. The primary ladder is built on QWEN2.5-7B: (i) the pretrained *base*; (ii) the general-purpose *instruction-tuned* variant (QWEN2.5-7B-INSTRUCT); (iii) an *RLVR* variant trained with GRPO on math reward (SIMPLERL-ZOO, Zeng et al., 2025); and (iv) a *reasoning-distilled* variant trained by SFT on DeepSeek-R1 traces (R1-DISTILL-QWEN-7B, Guo et al., 2025). This contrast isolates "RL on math reward" vs. "SFT on R1 reasoning traces" on a shared base—the exact comparison that prior pre-CoT work has not done (Cox et al., 2026).

The secondary ladder is built on QWEN3-4B (Yang et al., 2025): *base*, *Instruct-2507* (an instruction-tuned variant), and *Thinking-2507* (an explicit reasoning variant). This family uses a different architecture depth (36 decoder layers vs. 28) and serves as a cross-family replication.

Benchmarks. GSM8K (Cobbe et al., 2021) test set ($n = 1319$) for grade-school math, and the MATH-500 subset (Lightman et al., 2024; Hendrycks et al., 2021) ($n = 500$) for competition-level math with difficulty levels L_1 through L_5.

Probe training. For each (model, benchmark) we run vLLM generation once, extract hidden states for all n prompts in a single Transformers forward pass, and train per-layer probes on a 80/20 split. The best layer is selected on a single seed (seed 0; the value plotted in Figure 1); at that layer we re-estimate AUC over 5 seeds and report the mean $\pm$ std in Table 1.

Steering. For text generation under steering we use $n_{\text{test}} = 200$ held-out problems per cell, with `max_new_tokens` $= 8192$ for MATH-500 and 3072 for GSM8K to accommodate long CoTs in reasoning models. All 8 cells of the 4×2 main ladder are run under the *fresh* split and re-run under the *probe* split (Sec. 3.2); a four-layer sweep (L5, L14, L20, L23) is run for R1-Distill on MATH-500.

Model	Acc	Best L	AUC ($\pm$ std)
Math-500			
qwen2.5-7b-base	0.232	14	0.631 ± 0.049
qwen2.5-7b-instruct	0.748	16	0.854 ± 0.026
qwen2.5-7b-simplerl	0.750	13	0.809 ± 0.057
r1-distill-qwen-7b	0.906	20	0.826 ± 0.080
qwen3-4b-base	0.532	1	0.637 ± 0.060
qwen3-4b-instruct	0.946	14	0.882 ± 0.077
qwen3-4b-thinking	0.904	21	0.851 ± 0.035
Gsm8K			
qwen2.5-7b-base	0.700	21	0.700 ± 0.010
qwen2.5-7b-instruct	0.902	17	0.672 ± 0.042
qwen2.5-7b-simplerl	0.912	28	0.744 ± 0.070
r1-distill-qwen-7b	0.907	9	0.653 ± 0.025
qwen3-4b-base	0.631	23	0.611 ± 0.030
qwen3-4b-instruct	0.939	36	0.586 ± 0.055
qwen3-4b-thinking	0.909	21	0.794 ± 0.033

Table 1: Pre-CoT correctness probe results. Best-layer AUC with 5-seed mean $\pm$ std. Acc is the model's baseline final-answer accuracy ($n = 500$ for MATH-500, $n = 1319$ for GSM8K).

5 Study 1: Pre-CoT Correctness Is Decodable

Pre-CoT correctness is decodable from trained models, but not from base. Figure 1 and Table 1 show that on MATH-500, the post-training step is the dominant source of probe AUC. Qwen2.5-7B-Base reaches only 0.63 ± 0.05 at its best layer (L14); all three trained variants exceed 0.80. The same pattern repeats on QWEN3-4B: base 0.72 vs. trained 0.89–0.94. The pattern is qualitatively consistent with prior reports (Cox et al., 2026; Zhang et al., 2025).

Apparent method-by-method ladders do not survive seed variation. A single-seed reading of Figure 1 would suggest a clear ordering on MATH-500: base < SimpleRL < instruct < R1-Distill. With 5 seeds, however, the AUC of the three trained variants overlaps within one standard deviation: 0.85 ± 0.03, 0.81 ± 0.06, 0.83 ± 0.08. The R1-Distill single-seed 0.94 is a high-variance draw; the multi-seed mean falls back to 0.83 with std 0.08, driven by the small negative class ($n_{\text{neg}} = 47$ out of 500) which makes probe AUC sensitive to which negatives are in the validation split. **On MATH-500, the only conclusion that survives seed re-runs is binary—base vs. trained**, not which training method produces a stronger signal. On GSM8K, even this binary collapses: all AUCs sit in 0.65–0.74, and the base model (0.70 ± 0.01) is

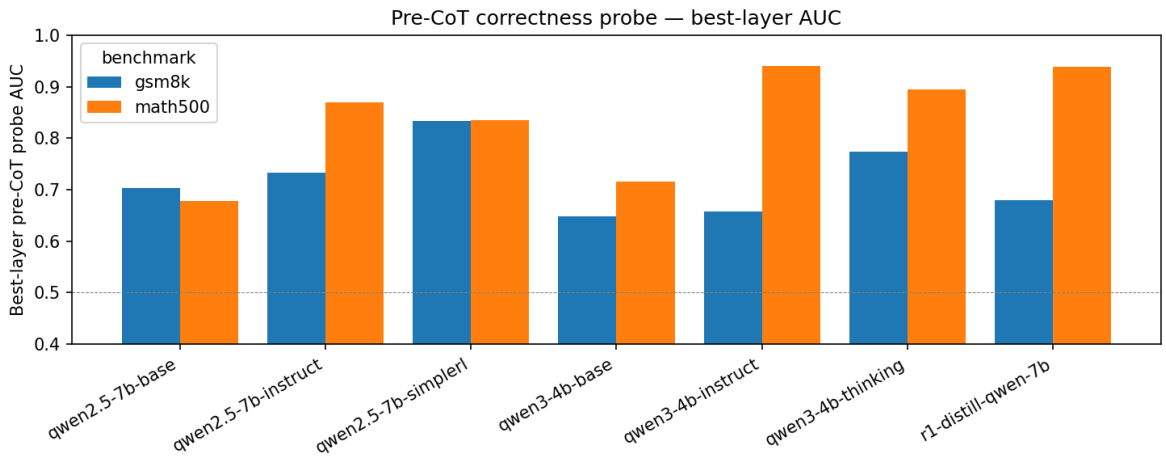

Figure 1: **Pre-CoT correctness probe AUC at the best layer**, by model and benchmark. *Bars show single-seed (seed-0) values*. Because the layer is selected on this same seed, these readings are optimistic relative to the seed-averaged numbers; Table 1 reports the authoritative 5-seed mean $\pm$ std. On MATH-500, base models sit at 0.68–0.72 and trained variants at 0.83–0.94; on GSM8K, the gap is smaller and the ordering noisier.

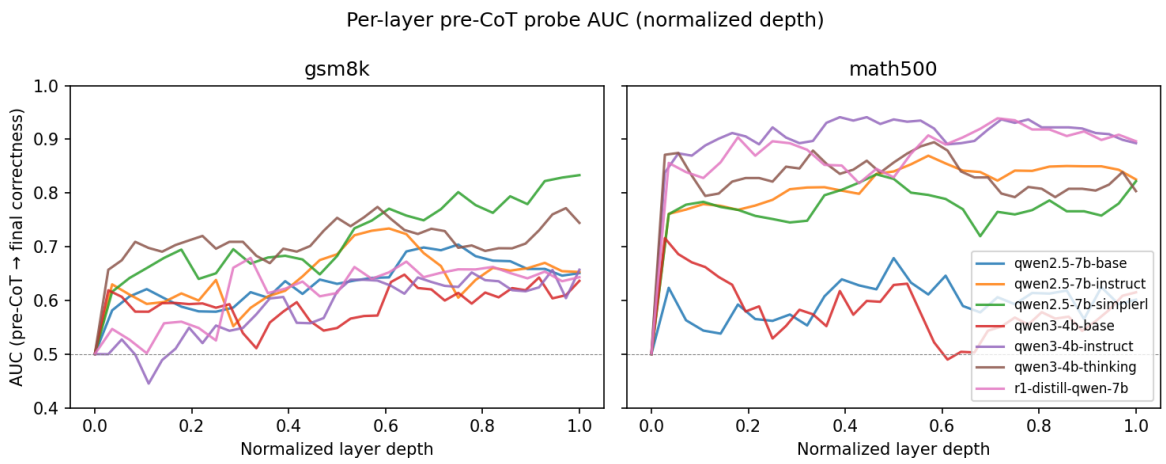

Figure 2: **Per-layer pre-CoT probe AUC.** On MATH-500, reasoning-trained variants (R1-Distill pink, Qwen3-Instruct purple, Qwen3-Thinking brown) reach near-peak AUC by the first decoder layer and stay there. Base models climb slowly. X-axis is normalized layer depth so the two model families can be compared on a shared axis.

statistically tied with—or above—the trained variants, so no clean separation of any kind survives.

Per-layer dynamics: depth of the decodable signal differs across training. Figure 2 plots per-layer AUC on a normalized depth axis. Reasoning-trained variants reach near-peak AUC by the first decoder layer and stay there: pre-CoT correctness in those models is encoded broadly across the residual stream stack, not localized to a deep layer. Base models climb gradually and only peak in the middle of the stack. SimpleRL on MATH-500 shows a non-monotonic curve, peaking around layer 13. *The depth at which the signal emerges changes more than its magnitude across training regimes.* This will become important when interpreting the steering results: we now ask whether *any* layer of these high-AUC residual streams is causally driving the final answer.

6 Study 2: Steering the Pre-CoT Signal

Steering result, headline. Table 2 and Figures 3–4 report the 8-cell steering ladder at $n_{\text{test}} = 200$. Seven of the eight recovery-rate CIs include 1.0: pushing the pre-CoT residual one $\|v\|$ toward the wrong-answer centroid produces no detectable drop in accuracy in any cell—not for base models, not for the instruction-tuned model, and not for either reasoning-trained variant. The eighth CI, R1-Distill on GSM8K (Rec 1.08, $[1.02, 1.15]$), excludes 1.0 from *above*: negative steering there

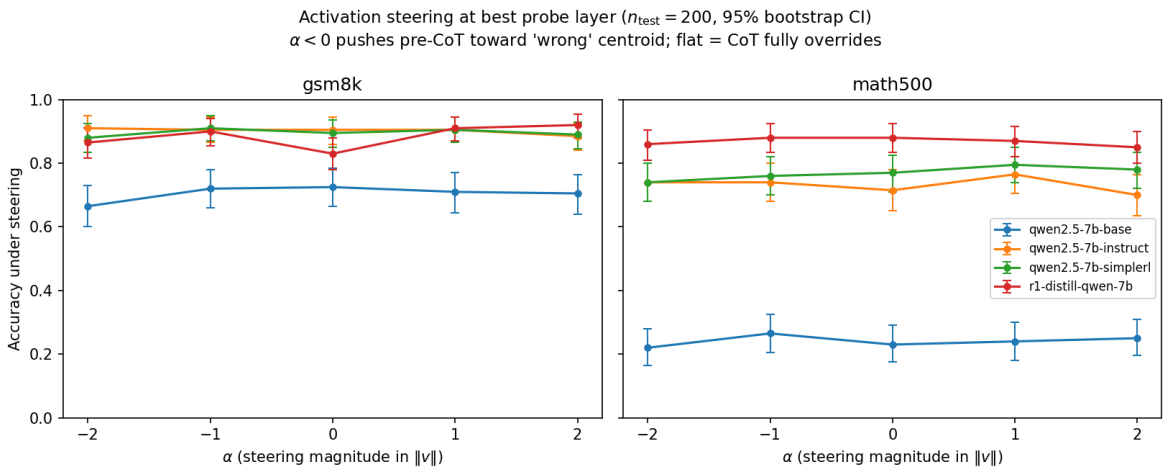

Figure 3: **Accuracy vs. steering magnitude** at the best probe layer for each (model, benchmark), $n_{\text{test}} = 200$. $\alpha < 0$ nudges the pre-CoT residual toward the "wrong" centroid. Flat lines mean the CoT fully overrides the nudge. Every curve is flat within paired-bootstrap CIs: the nudge has no detectable effect on accuracy in any cell.

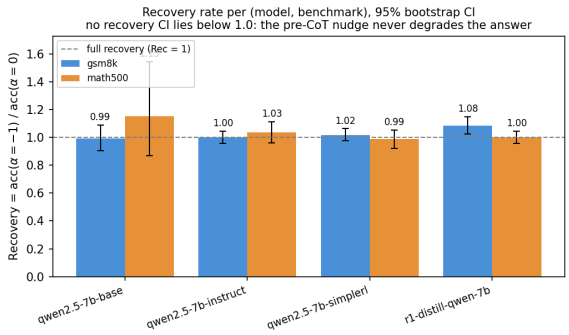

Figure 4: **Recovery rate** = **acc**(α=−1)/**acc**(α=0), with 95% paired-bootstrap CIs. Values near 1.0 indicate that the CoT fully overrides a wrong-direction pre-CoT nudge. Every CI includes 1.0: the nudge has no detectable effect on accuracy in any of the 8 cells.

Model	Bench	L	Probe	α=0	Rec [95% CI]
qwen2.5-7b-base	math500	14	0.68	0.230	1.15 [0.87, 1.54]
qwen2.5-7b-instruct	math500	16	0.87	0.715	1.04 [0.96, 1.11]
qwen2.5-7b-simplerl	math500	13	0.84	0.770	0.99 [0.92, 1.05]
r1-distill-qwen-7b	math500	20	0.94	0.880	1.00 [0.96, 1.05]
qwen2.5-7b-base	gsm8k	21	0.70	0.725	0.99 [0.90, 1.09]
qwen2.5-7b-instruct	gsm8k	17	0.73	0.905	1.00 [0.96, 1.05]
qwen2.5-7b-simplerl	gsm8k	28	0.83	0.895	1.02 [0.98, 1.06]
r1-distill-qwen-7b	gsm8k	9	0.68	0.830	1.08 [1.02, 1.15]

Table 2: Steering results at the best probe layer, fresh split, $n_{\text{test}} = 200$. **L** is the steering layer, **Probe** is single-seed probe AUC at that layer, α=0 is baseline accuracy under the hooked-but-unsteered forward pass, **Rec** is recovery rate (Eq. 2) with a 95% paired-bootstrap CI. All eight CIs include 1.0 except R1-Distill/GSM8K, whose CI lies *above* 1.0—negative steering there slightly *raises* accuracy, the opposite of a commitment effect (see text).

slightly *raises* accuracy, the opposite of what a pre-CoT correctness commitment predicts. Its α=0 baseline (0.83) is the minimum of its own α-curve, and the same cell recovers 0.95 and 0.97 under the two robustness conditions (App. C); we read it as baseline noise. The picture is uniform: the pre-CoT direction the probe identifies is not what causally drives the final answer.

Steering does not change the answer. For R1-DISTILL-QWEN-7B on MATH-500, the probe AUC at layer 20 is 0.94—one of the strongest pre-CoT decoding results we observe. Yet steering at that same layer leaves accuracy flat across the entire sweep (0.86/0.88/0.88/0.87/0.85 for $\alpha \in \{-2, -1, 0, +1, +2\}$; Rec 1.00, CI $[0.96, 1.05]$). The same holds for SIMPLERL on MATH-500 (Rec 0.99) and for the instruction-tuned model (Rec 1.04). The high probe AUC in these models reads correctness off the residual stream, but the residual stream's contribution to the final answer is not along the direction the probe identifies.

6.1 Steering changes form, not function

A uniformly null accuracy result invites the objection that the intervention simply failed to take effect—that $\alpha\|v\|$ is too small to do anything. It is not. While final-answer accuracy is flat, the *length* of the CoT responds to steering. We quantify this with a within-problem regression of CoT length (whitespace-split word count) on α and report the per-problem slope averaged across the held-out set, with significance assessed by a one-sample t-test against zero.

Model	AUC	Words/α	Sweep $\Delta\%$
instruct	0.87	-7.7^*	-9.5%
simplerl	0.83	$+2.1$	$+1.7\%$
r1-distill	0.94	$+60.7^*$	$+14.0\%$

Table 3: CoT length slope on MATH-500 (words per unit α, paired one-sample t-test; * denotes $|t| \geq 1.96$). *AUC*: probe AUC at the steered layer. *Sweep* $\Delta\%$: percentage change from $\alpha{=}-2$ to $\alpha{=}+2$ relative to the $\alpha{=}0$ baseline. The slope increases monotonically across the training ladder: instruct (-7.7) $\rightarrow$ SimpleRL ($+2.1$) $\rightarrow$ R1-Distill ($+60.7$).

Layer	$\alpha{=}-1$	$\alpha{=}0$	$\alpha{=}+1$	Rec
L5 (early)	0.860	0.880	0.910	0.977
L14 (middle)	0.875	0.875	0.855	1.000
L20 (best probe)	0.880	0.880	0.870	1.000
L23 (late)	0.885	0.880	0.875	1.006

Table 4: Layer sweep for R1-DISTILL-QWEN-7B on MATH-500 ($n = 200$). Recovery rate is statistically indistinguishable from 1.0 at every layer tested.

Length slopes on MATH-500. Table 3 reports the length slope for each trained variant on MATH-500. The three models fall on a continuum ordered by degree of reasoning training: the instruction-tuned model has a *negative* slope (-7.7 words/α, $t{=}-4.4$), meaning that steering toward the correct ("easy") centroid shortens the CoT while steering toward the wrong ("hard") centroid lengthens it; the reasoning-distilled model has a *positive* slope ($+60.7$ words/α, $t{=}+5.0$), with the opposite sign; and SimpleRL, the lightest reasoning training, falls in between with a small positive slope ($+2.1$ words/α, significant in 2 of 3 splits). In percentage terms relative to the $\alpha{=}0$ baseline, the instruction-tuned model's CoT shrinks by 8.4% at $\alpha{=}+2$, while R1-Distill's grows by 11.8% at $\alpha{=}+2$—a 14.0% swing across the full sweep. The sign reversal is not an artifact of non-converging generation: restricting to problems answered correctly at $\alpha{=}0$, R1-Distill's slope is $+68$ words/α ($t{=}+5.4$, $n{=}176$), larger than the all-problem estimate.

7 Robustness Checks

Layer choice is not the trick: a layer sweep on R1-Distill. A natural objection to the negative steering result is that we may have picked the wrong layer. We test this on R1-Distill (the highest single-seed probe AUC in the ladder) on MATH-500 by repeating the $\alpha \in \{-1, 0, +1\}$ steering at three additional layers: L5 (early), L14 (middle), L23 (late). Table 4 shows that at every layer tested, the recovery rate is within sampling noise of 1.0. The pre-CoT residual stream of R1-Distill is broadly correlation-rich but causal-poor across the entire stack, not just at the layer we initially chose.

Difficulty stratification. Figure 5 (appendix) repeats the probe training within each MATH-500 difficulty level. On L5 problems (hardest, $n = 134$, well-separated by accuracy), the ladder is preserved: base 0.63, SimpleRL 0.55, instruct 0.82, R1-Distill 0.94. On L1–L3 problems, R1-Distill is degenerate (it solves nearly every one—no negatives to learn from), and SimpleRL drops below 0.6 at most levels.[2]

Cross-family replication. The Qwen3-4B family confirms the base-vs-trained pattern on MATH-500: base 0.72 $\rightarrow$ instruct 0.94 / thinking 0.89. Per-layer curves (Figure 2) show the same "early decoder layer saturation" for trained variants in Qwen3-4B as for the Qwen2.5-7B reasoning models, despite the different architecture depth (36 vs. 28 decoder layers).

Statistical power. At $n_{\text{test}} = 200$ the paired-bootstrap CI on each recovery rate has a half-width of roughly 0.04–0.10 for the high-accuracy cells; base-model cells, whose low accuracy inflates the ratio's variance, are wider. All eight CIs are reported in Table 2, and the null result is that every one of them includes 1.0. We re-ran the full 8-cell ladder at $n_{\text{test}} = 100$ and under the probe-aligned split; the conclusion—no cell with a recovery CI below 1.0—is unchanged in both (App. C).

8 Discussion

Pre-CoT representations encode difficulty, not commitment. A model that reaches 0.94 probe AUC at a pre-CoT position is *predicting* whether it will solve the problem, with high reliability. Our steering results show that this prediction does not causally drive the answer—nudging the residual stream in the probe direction has no detectable effect on correctness in any of the 8 cells, and no effect at any R1-Distill layer we tested—but it does change *how much the model writes*, and the direction of that change depends on training regime

[2] A subject-level cut (Algebra/Geometry/etc.) yields broadly similar picture; see Appendix B.

(Sec. 6.1). The pre-CoT signal is therefore better read as *difficulty awareness*—a representation of "how hard this problem is for me"—than as a private commitment to a specific final answer. The signal modulates reasoning *strategy* (hedging vs. compression) without modulating reasoning *outcome* (correctness). This is consistent with the fact that the strongest probe AUC cells correspond to the highest-accuracy models: the pre-CoT state can be highly correlated with correctness simply because the model can tell which problems it will solve.

Difficulty awareness as metacognition. Prior work on calibration and self-assessment shows that models can verbalize uncertainty when prompted; our results show that a difficulty signal is present *before* any token is generated and *without* any explicit prompt to self-assess, and that this signal is functionally active—it causally influences CoT length even though it does not causally influence the answer. The pre-CoT difficulty representation is thus not a passive correlate of eventual correctness but an active input to the model's generation process, albeit one that modulates effort rather than outcome. This distinction—between representations that predict behavior and representations that determine behavior—matters for interpretability and safety: a model that "knows" it will fail but proceeds to the same answer anyway is a different kind of system from one that decides the answer in advance, and the appropriate interventions differ accordingly. For instance, if difficulty awareness is the right reading, then intervening on the pre-CoT state to improve reliability should target effort allocation (e.g., prompting more careful reasoning) rather than trying to override a committed answer.

Relation to answer-identity steering. Cox et al. (2026) report that steering along an answer-*identity* direction (which specific option the model will choose) flips answers in multiple-choice settings. Our experiment targets a different direction—correctness (right vs. wrong) rather than answer identity (A vs. B)—and a different task format (free-form math rather than multiple choice). The null accuracy result here does not contradict their finding: a model can have a committed answer identity without the correctness direction being causal. What our result does bound is the scope of the "rationalization" reading: the pre-CoT correctness-predictive direction is not a lever on the answer in free-form math reasoning. More broadly, the contrast highlights that "what the probe decodes" and "what the representation controls" are different questions—a direction that is highly decodable (AUC 0.94) need not be causally efficacious for the outcome the probe predicts.

Implications for RLVR analysis. The RLVR-vs-base debate (Yue et al., 2025; Guo et al., 2025; Zeng et al., 2025) centers on whether RL produces new reasoning capability or merely amplifies existing distributions. Our length-steering results add a new angle: whatever RLVR does, it changes *how the model uses its difficulty signal*. The instruction-tuned model hedges when it perceives difficulty (negative length slope); the reasoning-distilled model compresses (positive slope); SimpleRL sits in between. This ordering suggests that reasoning training progressively shifts the model from a hedging response to a compression response when it perceives a problem as hard—perhaps because instruction-tuned models treat perceived difficulty as a signal to invest more effort, while reasoning-distilled models, trained on traces that compress routine steps and elaborate only on obstacles, treat the same signal as a cue to adjust the *allocation* of reasoning effort rather than its *quantity*.

9 Conclusion

We replicate the pre-CoT decoding finding of prior probing work on a four-stage post-training ladder and a cross-family extension, confirming that pre-CoT probe AUC discriminates base from trained models. Steering along the same probe direction does not change final-answer accuracy in any cell—but it does change CoT length, and the direction of that change depends on training regime. On MATH-500, the length slope orders monotonically with reasoning training: instruction-tuned models lengthen when steered toward "hard" (-7.7 words/α), while reasoning-distilled models shorten ($+60.7$ words/α), with SimpleRL in between. The pre-CoT direction encodes difficulty awareness: the model knows how hard the problem is before it reasons, and this knowledge modulates *how* it reasons, not *whether* it reasons correctly. The same internal signal thus produces opposite behavioral signatures across training regimes—hedging in instruction-tuned models, compression in reasoning-distilled ones—revealing that what a probe decodes and what a representation controls are distinct questions.

Limitations

Scope of models. Our ladder includes one base family (Qwen2.5-7B) plus a smaller cross-family check (Qwen3-4B). We do not include Llama-3.1-8B or larger ($\geq$ 32B) models, and we cover one open RLVR variant (SIMPLERL-ZOO) rather than the full set of community RL'd checkpoints. The "RLVR" axis is thinner than the SFT axis.

Scope of benchmarks. We use GSM8K and MATH-500. We do not include AIME, multi-domain MCQ (e.g., MMLU-Pro), logic benchmarks (KnowLogic), or multimodal reasoning. The uniformly null steering result might not generalize to non-math reasoning, where pre-CoT difficulty signals could play out differently.

Probe class. We use linear probes (logistic regression), following standard practice in this line of work. Non-linear probes might detect difficulty or correctness signals that the linear probe misses. We see the linear restriction as a feature for the correlational protocol (it makes the steering direction in Eq. 1 well-defined), but we acknowledge that "not linearly steering-causal" is weaker than "not causal at all."

Statistical power. We steer $n_{\text{test}} = 200$ problems per cell and report paired-bootstrap CIs. The CIs are tight enough to exclude recovery effects larger than $\sim$0.10 in the high-accuracy cells, but not arbitrarily small ones; base-model cells, with low accuracy, have wider CIs. A null result here is evidence of *no large causal effect* of the steering direction, not proof of exactly zero.

Single direction. We model the pre-CoT correctness signal as a single direction via difference-of-means, following Arditi et al. (2024). The Cox-style protocol assumes that the relevant signal is along the mean-difference direction. Recent work using sparse autoencoders suggests CoT-relevant computation may be distributed across multiple monosemantic features at scale (Chen et al., 2026), raising the question of whether richer geometric structure better captures pre-CoT information. Reasoning models may carry difficulty information in multi-dimensional subspaces or in attention-output channels we do not steer. We view our result as showing that the natural Cox-style direction modulates effort but not correctness, not as proof that no direction could be causal for the answer.

References

Iván Arcuschin, Jett Janiak, Robert Krzyzanowski, Senthooran Rajamanoharan, Neel Nanda, and Arthur Conmy. 2025. Chain-of-Thought reasoning in the wild is not always faithful. *Preprint*, arXiv:2503.08679.

Andy Arditi, 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*.

Siddharth Boppana, Annabel Ma, Max Loeffler, Raphael Sarfati, Eric Bigelow, Atticus Geiger, Owen Lewis, and Jack Merullo. 2026. Reasoning theater: Disentangling model beliefs from Chain-of-Thought. *Preprint*, arXiv:2603.05488.

Xi Chen, Aske Plaat, and Niki van Stein. 2026. How does Chain of Thought think? Mechanistic interpretability of Chain-of-Thought reasoning with sparse autoencoding. In *AAAI Conference on Artificial Intelligence*, pages 30297–30305.

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

Kyle Cox, Darius Kianersi, and Adrià Garriga-Alonso. 2026. Decoding answers before Chain-of-Thought: Evidence from pre-CoT probes and activation steering. *Preprint*, arXiv:2603.01437.

Joey David. 2025. Temporal predictors of outcome in reasoning language models. *Preprint*, arXiv:2511.14773.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, and 175 others. 2025. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645(8081):633–638.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the MATH dataset. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*.

Nikolaus Howe and Micah Carroll. 2026. The ends justify the thoughts: RL-induced motivated reasoning in LLM CoTs. *Preprint*, arXiv:2510.17057.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient

memory management for large language model serving with PagedAttention. In *Proceedings of the Symposium on Operating Systems Principles*, pages 611–626.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, Kamilė Lukošiūtė, Karina Nguyen, Newton Cheng, Nicholas Joseph, Nicholas Schiefer, Oliver Rausch, Robin Larson, Sam McCandlish, Sandipan Kundu, and 11 others. 2023. Measuring faithfulness in Chain-of-Thought reasoning. *Preprint*, arXiv:2307.13702.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2024. Let's verify step by step. In *International Conference on Learning Representations*.

Parsa Mirtaheri and Mikhail Belkin. 2026. Catching rationalization in the act: Detecting motivated reasoning before and after CoT via activation probing. *Preprint*, arXiv:2603.17199.

Viacheslav Sinii, Nikita Balagansky, Gleb Gerasimov, Daniil Laptev, Yaroslav Aksenov, Vadim Kurochkin, Alexey Gorbatovski, Boris Shaposhnikov, and Daniil Gavrilov. 2026. Small vectors, big effects: A mechanistic study of RL-induced reasoning via steering vectors. *Preprint*, arXiv:2509.06608.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. 2023. Language models don't always say what they think: Unfaithful explanations in Chain-of-Thought prompting. In *Advances in Neural Information Processing Systems*.

Constantin Venhoff, Iván Arcuschin, Philip Torr, Arthur Conmy, and Neel Nanda. 2025. Understanding reasoning in thinking language models via steering vectors. *Preprint*, arXiv:2506.18167.

Zezhong Wang, Xingshan Zeng, Weiwen Liu, Yufei Wang, Liangyou Li, Yasheng Wang, Lifeng Shang, Xin Jiang, Qun Liu, and Kam-Fai Wong. 2025. Chain-of-Probe: Examining the necessity and accuracy of CoT step-by-step. In *Findings of the Association for Computational Linguistics: NAACL*, pages 2586–2606.

Jake Ward, Chuqiao Lin, Constantin Venhoff, and Neel Nanda. 2025. Reasoning-finetuning repurposes latent representations in base models. *Preprint*, arXiv:2507.12638.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Aojie Yuan, Zhiyuan Julian Su, Haiyue Zhang, Yi Nian, and Yue Zhao. 2026. Hidden error awareness in Chain-of-Thought reasoning: The signal is diagnostic, not causal. *Preprint*, arXiv:2605.09502.

Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Yang Yue, Shiji Song, and Gao Huang. 2025. Does reinforcement learning really incentivize reasoning capacity in LLMs beyond the base model? *Preprint*, arXiv:2504.13837.

Weihao Zeng, Yuzhen Huang, Qian Liu, Wei Liu, Keqing He, Zejun Ma, and Junxian He. 2025. SimpleRL-Zoo: Investigating and taming zero reinforcement learning for open base models in the wild. *Preprint*, arXiv:2503.18892.

Anqi Zhang, Yulin Chen, Jane Pan, Chen Zhao, Aurojit Panda, Jinyang Li, and He He. 2025. Reasoning models know when they're right: Probing hidden states for self-verification. *Preprint*, arXiv:2504.05419.

A Difficulty-stratified probe AUC

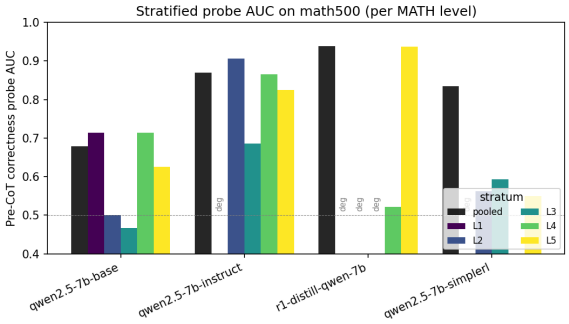

Figure 5: Probe AUC on MATH-500 stratified by difficulty level L_1–L_5. Cells marked "deg" have fewer than 5 negatives at that level and are excluded.

Figure 5 re-trains the correctness probe *within* each MATH-500 difficulty level, at the model's pooled best layer, to test whether the pooled AUC merely tracks problem difficulty. The level-stratified estimates are only suggestive: a cell is degenerate wherever the model is accurate enough to leave fewer than 5 negatives in that level (all of L_1–L_3 for R1-Distill, and L_1 for instruct and SimpleRL), and the surviving cells are scored on a 20% test split of a single level and are correspondingly noisy. Subject to those caveats, the within-level AUC is generally lower and markedly less ordered than the pooled value, indicating that the pooled probe signal is partly—but not wholly—a difficulty-correlated artifact.

B Subject-level probe AUC on MATH-500

The pooled probe is partially confounded with difficulty. We re-train the probe within each MATH sub-

ject (Algebra, Geometry, Number Theory, Counting & Probability, Intermediate Algebra, Prealgebra, Precalculus) for the four Qwen2.5-7B variants. Subject-internal AUCs are noisier than pooled (cells have $n_{\text{neg}} \in [5, 95]$) but the qualitative ladder—base low, trained higher—holds in 5 of 7 subjects (the exceptions are Geometry and Intermediate Algebra). For R1-Distill, three of seven subject cells (Algebra, Number Theory, Precalculus) are degenerate due to high accuracy.

C Steering robustness: split and sample size

Table 5 reports the recovery rate and its 95% paired-bootstrap CI for all 8 cells under three conditions: the primary *fresh* split at $n_{\text{test}} = 200$, the same split at $n_{\text{test}} = 100$, and the *probe* split (the exact held-out fraction the probe AUC is reported on; GSM8K $n = 264$, MATH-500 $n = 100$). The conclusion is identical in all three: no cell has a recovery CI lying below 1.0. The single CI that excludes 1.0 anywhere in the table—R1-Distill/GSM8K at n=200—excludes it from above, and is not reproduced at n=100 or under the probe split.

Model	Bench	fresh n=200	fresh n=100	probe
base	math500	1.15 [.87,1.54]	1.14 [.76,1.76]	1.30 [.89,2.00]
instruct	math500	1.04 [.96,1.11]	0.95 [.86,1.04]	1.01 [.93,1.11]
simplerl	math500	0.99 [.92,1.05]	1.03 [.95,1.11]	0.94 [.84,1.04]
r1-dist.	math500	1.00 [.96,1.05]	0.97 [.91,1.02]	1.05 [.99,1.11]
base	gsm8k	0.99 [.90,1.09]	0.96 [.84,1.09]	0.93 [.85,1.01]
instruct	gsm8k	1.00 [.96,1.05]	1.03 [.97,1.10]	0.99 [.96,1.03]
simplerl	gsm8k	1.02 [.98,1.06]	1.02 [.97,1.08]	1.02 [.99,1.05]
r1-dist.	gsm8k	1.08 [1.02,1.15]	0.95 [.88,1.01]	0.97 [.94,1.00]

Table 5: Recovery rate [95% paired-bootstrap CI] across split and sample size. Every CI includes 1.0 except R1-Distill/GSM8K at n=200, which lies *above* 1.0 (negative steering raises accuracy—not a commitment effect).

D Hyperparameters

Generation. vLLM 0.14, temperature 0.6, top-p 0.95, `max_new_tokens` as in the main text. Each (model, benchmark) is generated once.

Probe. `sklearn.linear_model.LogisticRegression(C=1.0, max_iter=1000, class_weight='balanced')`, 80/20 split, seed $\in \{0, 1, 2, 3, 4\}$ for the 5-seed runs.

Steering. HF Transformers with a forward pre-hook on the chosen decoder layer, $v = \mu_1 - \mu_0$ computed from the probe-training split. $\alpha \in \{-2, -1, 0, +1, +2\}$ in units of $\|v\|$. Same temperature/top-p as generation.

E Prompts

Base Model Prompt (Qwen2.5-7B base)

```
Question: {question}
Answer: ⟨generation⟩
```

Instruct Model Prompt (Qwen2.5-7B-Instruct)

SYSTEM
```
You are Qwen, created by Alibaba Cloud. You
are a helpful assistant.
```
USER
```
{question}
Please reason step by step, and put your
final answer within \boxed{}.
```
ASSISTANT
```
⟨generation⟩
```

Instruct Model Prompt (Qwen2.5-7B-SimpleRL)

SYSTEM
```
You are a helpful assistant.
```
USER
```
{question}
Please reason step by step, and put your
final answer within \boxed{}.
```
ASSISTANT
```
⟨generation⟩
```

Reasoning Model Prompt (DeepSeek-R1-Distill-Qwen)

USER
```
{question}
Please reason step by step, and put your
final answer within \boxed{}.
```
ASSISTANT
```
<think>
⟨generation⟩
```

F Computational Resources

Experiments were run on NVIDIA H100 80GB GPUs (single-GPU array jobs, up to 4 concurrent on the partition; some early pilot runs additionally used tensor-parallel across 2 H100s for the 7B family). Generation passes (vLLM, the 8 main model–benchmark cells plus the 6 Qwen3-4B cross-family cells) took approximately 8 GPU-hours total; hidden-state extraction (HuggingFace Transformers replay over the saved generations, 14 cells at ∼25 ex/s) took approximately 1 GPU-

hour; steering runs across the α sweep and robustness conditions (main n=200 fresh split, the n=100 and probe-aligned robustness splits, and the n=200 layer sweep at L5/L14/L23 on R1-Distill) took approximately 70 GPU-hours. Full project compute, including pilot runs (the n=40 / 3-α ladder, the initial n=20 layer sweep, the R1-1.5B exploratory pilot) and discarded configurations (contamination re-runs, smoke tests, and cancelled jobs), is estimated at 110 GPU-hours.