SHEAR: HIDDEN-STATE EDITING FOR MID-THOUGHT REASONING CORRECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

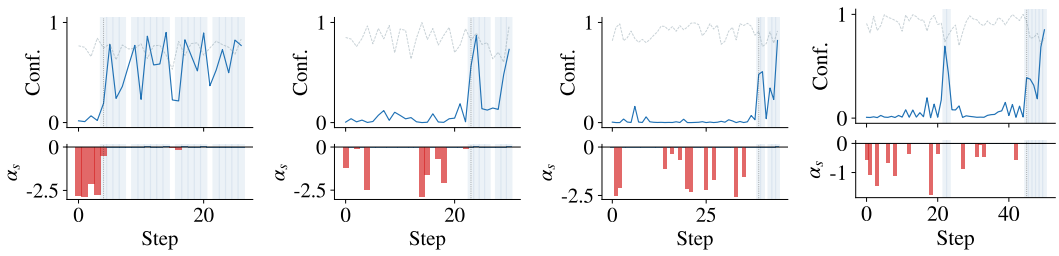
Large reasoning models (LRMs) generate extended chains of thought, but *do they internally know when they have already found the correct answer?* We investigate this question by probing hidden states during reasoning. A lightweight probe on hidden states identifies per-step reasoning correctness with high accuracy and shows a sharp jump at the moment of answer stabilization, providing strong affirmative evidence. This information is substantially more accessible through hidden states than through token-level signals: in controlled equal-dimension comparisons using the same classifier, hidden-state probes markedly outperform token-level probes. We propose SHEAR (**S**emantic **H**idden-layer **E**diting with **A**nchored **R**easoning), which directly exploits this hidden knowledge through a semantic confidence probe and a stability-anchored editing direction that jointly determine when and how to intervene. SHEAR achieves the highest or tied-highest results on six mathematical reasoning benchmarks across three LRM scales, with statistically significant improvements against the strongest baseline in most configurations. Reversing the editing direction causes catastrophic collapse, indicating that it operates on genuine reasoning structure. Hidden states are therefore not merely a readout of reasoning quality but a control surface for correcting it mid-thought. Code is available at this URL.

1 INTRODUCTION

When a large language model (LLM) reasons through a mathematical problem, does it internally "know" the moment it arrives at the correct answer? As a subclass of LLMs, large reasoning models (LRMs) generate extended chains of thought to solve complex problems. Consider the examples in Figure 1: a semantic confidence probe applied to the model's hidden states tracks reasoning correctness in real time, jumping sharply when the answer stabilizes, while token-level confidence shows no systematic response. In Figure 1a, the DEEPSEEK-R1-1.5B model finds the correct answer at step 4 yet continues reasoning for 23 more steps; the probe detects the answer immediately, but the model lacks a mechanism to stop. In Figures 1b–1d, three models at different scales struggle through extended incorrect reasoning before reaching the answer; the probe remains low for most of the trajectory and rises sharply at the moment of stabilization. If the model's hidden states indeed encode such information, can we directly exploit it to improve the reasoning process?

Existing inference-time reasoning control methods, including prompt-based approaches (Xu et al., 2025; Ma et al., 2025) and confidence-based early exit (Yang et al., 2025), are limited to token-level signals as proxies for reasoning quality. Throughout, *token-level* refers to statistics of the output distribution, such as token max-probability, its entropy, or their aggregates over a sliding window. However, such signals pass through the language modeling head, a projection optimized for next-token prediction—an objective under which the correctness of the reasoning so far need not be preserved, even when the residual stream encodes it. In §3, we show that hidden-state features substantially outperform engineered token features under controlled equal-dimension comparisons that separate this projection effect from dimensionality, and a full-dimensional MLP probe reaches near-perfect discrimination. The signal fires in real time: at the step where the model's answer first stabilizes, the probe output jumps sharply, while token confidence barely changes. Several concurrent methods apply activation steering (Chen et al., 2025; Li et al., 2026; Huang et al., 2025), but steer along precomputed directions targeting thought-type patterns (Chen et al., 2025), confidence-based prototypes (Li et al., 2026), or overthinking tendencies (Huang et al., 2025), without per-step

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



(a) 1.5B, s^* at step 4 **(b)** 1.5B, s^* at step 23 **(c)** 7B, s^* at step 39 **(d)** 32B, s^* at step 45

Figure 1: Reasoning trajectories across three model scales. Top: semantic probe score (blue) and token confidence (gray dashed). Bottom: editing magnitude α_s that SHEAR would apply. Blue shading marks correct steps; dotted lines mark s^* (defined in §3.2). In (a), DEEPSEEK-R1-1.5B finds the answer at step 4 but continues for 23 more steps; the probe detects the transition while token confidence shows no systematic change at s^*. In (b)–(d), models at DEEPSEEK-R1-1.5B, DEEPSEEK-R1-7B, and QWQ-32B scales struggle through extended incorrect reasoning before stabilizing; the probe remains low for most of the trajectory, with occasional spikes at transiently correct steps, then rises sharply at s^* and editing attenuates.

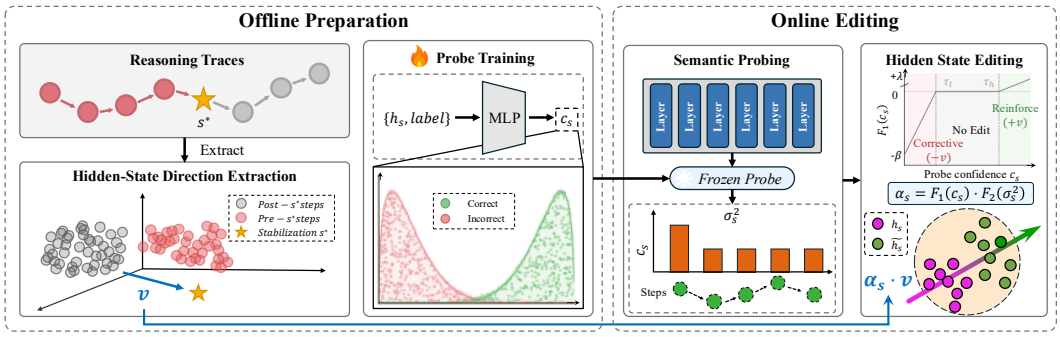


Figure 2: Overview of SHEAR. **Offline preparation**: (1) Reasoning traces are generated and the stability index s^* is identified. (2) The editing direction v is extracted from pre- and post-s^* hidden states. (3) An MLP probe is trained to output a semantic confidence score c_s. **Online editing**: (4) The frozen probe computes c_s and local variance σ_s^2 at each reasoning step. (5) The editing magnitude $\alpha_s = F_1(c_s) \cdot F_2(\sigma_s^2)$ steers hidden states toward the answer-stabilized region via $\tilde{h}_s = h_s + \alpha_s \cdot v$.

monitoring of reasoning correctness. We term this disparity the *signal accessibility gap* and show that it can be closed by directly reading and editing hidden states.

Based on these findings, as illustrated in Figure 2, we propose SHEAR (**S**emantic **H**idden-layer **E**diting with **A**nchored **R**easoning), which reads and edits hidden states during inference as detailed in §4. A semantic confidence probe assesses each step's reasoning quality, and a stability-anchored direction defines the intervention. Reversing the editing direction on DEEPSEEK-R1-7B collapses performance from 95.5% to 0.7%, confirming that the direction encodes reasoning-relevant structure.

Our contributions are:

- We show that hidden-state probes identify per-step reasoning correctness substantially more accurately than any token-level signal, across equal-dimension comparisons. This reveals a *signal accessibility gap*: reasoning quality information readily accessible in hidden states is largely invisible to token-level statistics.
- We propose SHEAR, which edits hidden states using a semantic confidence probe and a stability-anchored editing direction. Reverse-direction editing causes catastrophic collapse, establishing directional specificity.
- Evaluation on six math and two cross-domain benchmarks shows statistically significant improvements against the strongest baseline in the majority of configurations, with cross-domain generalization to graduate-level science reasoning.

Table 1: Signal comparison for predicting step-level correctness on DEEPSEEK-R1-7B with ground-truth labels. All classifiers use logistic regression except the last row, which uses an MLP. At equal dimension of 10D, hidden-state PCA outperforms token features by 13.7 percentage points under the same classifier.

Signal Source	Dim	AUC
Token max-prob (raw)	1D	0.563
Step position	1D	0.577
Token features (max-prob + window, LR)	10D	0.716
Hidden state (PCA, LR)	5D	0.766
	10D	0.853
	50D	0.935
Hidden state (LR)	3584D	0.945
Hidden state (MLP)	3584D	**0.983**

2 RELATED WORK

Inference-time reasoning control and process reward models. Recent methods control reasoning at inference time through prompt modification (Xu et al., 2025; Ma et al., 2025), early exit (Yang et al., 2025), or activation steering (Chen et al., 2025; Li et al., 2026; Huang et al., 2025). Prompt-based (Xu et al., 2025; Ma et al., 2025) and early-exit methods (Yang et al., 2025) rely on token-level signals, whereas §3 shows that hidden-state features provide substantially stronger prediction of per-step correctness. Dynasor-CoT (Fu et al., 2025) probes internal LRM states for certainty-based termination, but does not perform corrective editing during reasoning. Activation steering methods (Chen et al., 2025; Li et al., 2026; Huang et al., 2025) operate in the hidden-state space but without per-step semantic monitoring of reasoning correctness. A complementary approach trains process reward models (PRMs) to judge step-level correctness for ranking or search (Lightman et al., 2023; Wang et al., 2024; Luo et al., 2024). SHEAR's probe differs from PRMs in that it operates on intermediate-layer hidden states rather than output tokens, enables real-time editing rather than post-hoc ranking, and requires minimal training data.

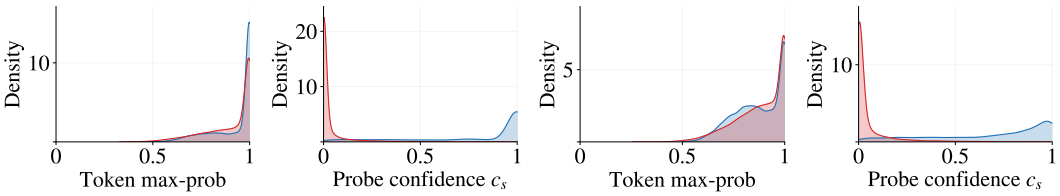
Probing and editing internal representations. Linear probes have decoded diverse knowledge from hidden states, including linguistic structure (Belinkov et al., 2017; Hewitt & Manning, 2019), factual knowledge (Burns et al., 2023; Li et al., 2023), and truthfulness (Azaria & Mitchell, 2023; Marks & Tegmark, 2023). Hidden-state signals have also been used to verify intermediate answers for confidence-based early exit (Zhang et al., 2025), to score and prune unpromising traces under parallel sampling (Liang et al., 2026), and for output-free self-evaluation at the response level (Wang et al., 2025b). On the editing side, methods such as ROME (Meng et al., 2022) and MEMIT (Meng et al., 2023) modify model parameters, while activation steering (Turner et al., 2024; Li et al., 2023) edits inference-time representations to shift LLM behavior. These works *read* the signal to stop, prune, or score a trace, leaving the underlying computation untouched; SHEAR instead closes the loop, using the probed signal for *dynamic* hidden-state editing with adaptive magnitude control that alters the computation within a single trace.

3 DO HIDDEN STATES ENCODE REASONING PROGRESS?

If the account above is right, per-step reasoning correctness should be decodable from the hidden states of LRMs yet largely absent from the token-level signals that existing methods rely on; we test both halves of this prediction. The probe in this section serves purely as a *diagnostic tool* for measuring information content; its use as an online control signal in SHEAR is introduced separately in §4.

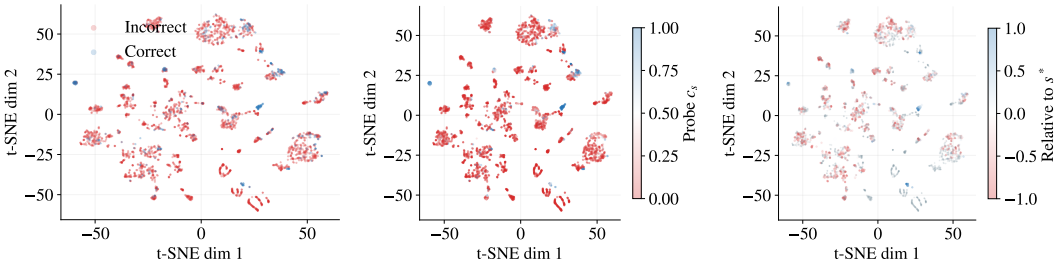
3.1 REPRESENTATIONS RELIABLY PREDICT STEP-LEVEL CORRECTNESS

Setup. We run DEEPSEEK-R1-7B on 500 MATH training problems with greedy decoding and record intermediate hidden states; a 300-problem subset is cached for visualization, as described in Appendix C. Reasoning steps are segmented by "`\n\n`" delimiters. Step correctness is determined by extracting the running answer at each step and comparing it to the gold answer via symbolic equivalence (`math_equal`). This answer-based label is an approximation that may occasionally



(a) 1.5B token **(b)** 1.5B probe **(c)** 7B token **(d)** 7B probe

Figure 3: Distributions of token max-probability and semantic probe scores, separated by step correctness under ground-truth labels. Blue: correct steps; red: incorrect steps. Token distributions overlap for both model scales (a,c), while probe distributions are well separated (b,d), consistent with the AUC gap in Table 1.



(a) By correctness (GT) **(b)** By probe score **(c)** By reasoning progress

Figure 4: t-SNE visualization of hidden states from DEEPSEEK-R1-7B at layer 22. (a) Correct and incorrect steps exhibit class-conditional clustering. (b) Probe scores align with ground-truth structure. (c) Steps transition smoothly in embedding space as reasoning progresses from incorrect to correct.

diverge from true reasoning quality; we discuss this limitation in Appendix D and note that the reverse-direction experiment in §5.5 confirms the resulting signal captures genuine reasoning structure. The hidden state $h_s \in \mathbb{R}^D$ is the activation at the first token of step s from a single intermediate layer ℓ, corresponding to layer 22 for DEEPSEEK-R1-7B. We use the first token because it aggregates information from the preceding step via causal attention; layer selection via Ridge regression R^2 is described in Appendix C.

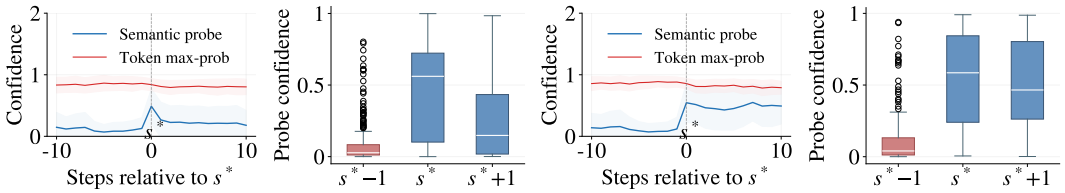
Signal comparison. Table 1 compares signal types from raw token statistics to hidden-state probes. All classifiers use problem-level splits, and to enable a controlled equal-dimension comparison, we reduce hidden states via PCA and extract matched-dimension token features, evaluating both under the same logistic regression classifier.

As Table 1 shows, raw token max-probability is only marginally above chance, and at equal dimension under the same logistic regression classifier, hidden-state PCA features outperform token features by a wide margin. A full-dimensional MLP probe achieves near-perfect discrimination. Figure 3 visualizes this gap through kernel density estimates of each signal, separated by ground-truth step correctness. For token max-probability, the correct and incorrect distributions overlap at both model scales, whereas the probe score distributions are well separated: incorrect steps concentrate near zero while correct steps cluster near one. This separation holds at DEEPSEEK-R1-1.5B scale as well, corroborating that the signal accessibility gap is not scale-specific.

Figure 4 provides a geometric view of this gap through t-SNE projections of layer-22 hidden states. Steps colored by ground-truth correctness exhibit class-conditional clustering, confirming that the representation space organizes around reasoning quality. The continuous probe score colormap in Figure 4b aligns with this cluster structure, validating that the probe faithfully captures the geometric separation rather than exploiting a spurious shortcut. Figure 4c colors each point by its temporal distance to s^*: pre-stabilization steps cluster with incorrect representations while post-stabilization steps cluster with correct ones, and the transition between regions is gradual.

Table 2: Signal dynamics at the stability index s^*, measured on DEEPSEEK-R1-7B across 264 problems with valid s^*. The probe jumps by $+0.451$ at the moment of answer stabilization; token confidence barely changes. The $22.6\times$ ratio compares absolute magnitudes; note that the two changes have opposite signs.

Position	Token Max-Prob	Semantic Probe
$s^* - 1$ (before)	0.875	0.113
s^* (stability index)	0.855	0.564
Change at s^*	-0.020	$+0.451$ ($22.6\times$)



(a) 1.5B trajectories **(b)** 1.5B at s^* **(c)** 7B trajectories **(d)** 7B at s^*

Figure 5: Signal dynamics around the stability index s^* at two model scales. (a,b) At DEEPSEEK-R1-1.5B, the semantic probe rises sharply at s^* while token max-probability shows no systematic change (a), though probe confidence partially reverts after s^* (b). (c,d) At DEEPSEEK-R1-7B, the same sharp rise holds (c), and probe confidence remains elevated after s^* (d).

3.2 THE SIGNAL TRACKS REASONING PROGRESS IN REAL TIME

Beyond encoding correctness as a static property, the signal is *dynamic*: it fires at the precise moment the model's answer stabilizes.

We define the *stability index* s^* as the earliest step after which the model's running answer remains correct for at least 50% of subsequent steps. With $\text{correct}(s) \in \{0, 1\}$ as defined above:

$$s^* = \min\left\{ s \;:\; \frac{|\{s' > s : \text{correct}(s') = 1\}|}{|\{s' > s\}|} \geq \theta \right\}, \quad \theta = 0.5, \tag{1}$$

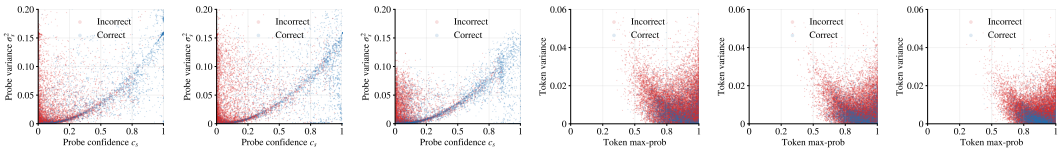
where θ is the stabilization threshold (not model parameters). Of 500 training problems, 444 produce at least one correct step, and 264 of those yield a valid s^*, covering 76.5% of the 345 problems with at least 3 reasoning steps and at least 1 correct step used for direction extraction. The rest oscillate without stabilization. The direction is robust to θ, as shown in Appendix E.

Table 2 quantifies the change at s^*, and Figure 5 visualizes the temporal dynamics. In Figures 5c and 5a, the probe trajectory averaged over all problems with valid s^* rises sharply at offset zero while token max-probability shows no systematic change; the shaded bands indicate that this is consistent across problems, not driven by outliers. Figures 5d and 5b confirm via boxplots that the probe transitions from low to high within a single step at s^*, with the median jumping from near zero to above 0.5 at both scales. This real-time responsiveness makes hidden-state signals suitable for *online* inference-time editing. Figure 1 illustrates this across three model scales: when the model reasons incorrectly, the probe stays low and SHEAR applies corrective editing; once the answer stabilizes at s^*, the probe rises and editing attenuates, while token confidence shows no systematic change.

4 SHEAR: EDITING HIDDEN STATES FOR BETTER REASONING

Section 3 showed that reasoning correctness information is far more accessible in hidden states than in token-level statistics. Activation steering research (Chen et al., 2025; Turner et al., 2024; Li et al., 2023) demonstrates that additive perturbations along meaningful directions in intermediate-layer representations can bias subsequent computation. As illustrated in Figure 2, we therefore propose SHEAR, which directly reads and edits hidden states during inference. The core operation is additive editing at each reasoning step boundary:

$$\tilde{h}_s = h_s + \alpha_s \cdot v, \tag{2}$$



(a) 1.5B probe **(b)** 7B probe **(c)** 32B probe **(d)** 1.5B token **(e)** 7B token **(f)** 32B token

Figure 6: Confidence-variance space across three model scales with ground-truth labels. In the probe's semantic space (a–c), correct and incorrect steps are clearly separated at all scales; in the token space (d–f), the two classes overlap completely.

where $h_s \in \mathbb{R}^D$ is the layer-ℓ hidden state at the first token of step s, $v \in \mathbb{R}^D$ is a stability-anchored editing direction, defined formally in Eq. 3 of §4.1, and $\alpha_s \in \mathbb{R}$ is an adaptive editing magnitude determined by a semantic confidence probe and a control function described in §4.2. The full procedure is given in Algorithm 1; the probe and the edit together add negligible arithmetic overhead per step, as quantified in Appendix C.

4.1 Stability-Anchored Editing Direction

The editing direction v should capture the representational shift that occurs when the model's answer stabilizes. We operationalize this using the stability index s^*, defined in Eq. 1, which marks the step where the model's answer first predominantly stabilizes to the correct value. Both s^* and the probe labels require gold answers and are computed offline; at inference SHEAR reads only hidden states, generating a single trace without answer extraction or resampling. For each training problem i with a valid s_i^*, let $\mathcal{S}_i^> = \{s : s > s_i^*\}$ denote the post-s^* steps. We extract a difference vector between the stability index and the subsequent redundant reasoning:

$$\delta_i = h_{s_i^*} - \frac{1}{|\mathcal{S}_i^>|} \sum_{s \in \mathcal{S}_i^>} h_s, \qquad v = \frac{\mathbb{E}_i[\delta_i]}{\|\mathbb{E}_i[\delta_i]\|}. \tag{3}$$

Each δ_i captures the direction from redundant elaboration toward the answer-stabilized state; averaging and normalizing yields a global editing direction. The sign of α_s determines whether the edit reinforces the answer-stabilized state ($+v$) or pushes away from premature convergence ($-v$), as detailed in §4.2. We validate directional specificity in §5.5 through reverse-direction editing.

Direction consistency. The individual δ_i vectors achieve mean $\cos(\delta_i, v) = 0.399$, far above the near-zero expectation for random vectors in high-dimensional space, with 87.6% pointing in the same hemisphere as v. The stability-anchored direction also outperforms both random directions and alternative extraction strategies, as detailed in Appendix E and F.

4.2 Adaptive Editing Control

The editing magnitude α_s should be strong when the model reasons poorly and gentle when on track. SHEAR determines α_s from a semantic confidence probe and a variance-aware control function.

Semantic confidence probe. A lightweight MLP $f : \mathbb{R}^D \to [0, 1]$ maps the hidden state to a semantic confidence score $c_s = f(h_s)$, the predicted probability that step s is correct. The probe is trained with binary cross-entropy on 500 MATH problems in under one minute; architecture details are in Appendix C, and the layer-wise AUC profile and scale-dependent statistics in Appendix J.

Control function. Let $\bar{c}_s = \frac{1}{k} \sum_{j=s-k+1}^{s} c_j$ denote the windowed mean over k=5 steps, and let $\sigma_s^2 = \frac{1}{k} \sum_{j=s-k+1}^{s} (c_j - \bar{c}_s)^2$ denote the corresponding local variance. Writing $[\cdot]_+ = \max(0, \cdot)$ for

the positive part and $\mathbf{1}[\cdot]$ for the indicator function, the control function is:

$$\alpha_s = F_1(c_s) \cdot F_2(\sigma_s^2), \tag{4}$$

$$F_1(c_s) = \frac{\lambda\,[c_s - \tau_h]_+}{1 - \tau_h} \;-\; \frac{\beta\,[\tau_l - c_s]_+}{\tau_l}, \tag{5}$$

$$F_2(\sigma_s^2) = 1 + \mathbf{1}[\sigma_s^2 > \gamma]\;\min\Big(\frac{\sigma_s^2}{\sigma_0},\; 2\Big), \tag{6}$$

The remaining hyperparameters are: β=3.0 (maximum editing magnitude), λ=0.01 (high-confidence editing strength), τ_l and τ_h (25th and 75th percentiles of the training-set probe confidence distribution), σ_0=0.01 (reference variance scale), and γ=0.005 (variance activation threshold).

When c_s is low, F_1 produces a negative α_s, applying a corrective perturbation along $-v$. When c_s is high, F_1 produces a small positive α_s that gently reinforces the answer-stabilized state along $+v$. No editing occurs in the middle regime. The variance term F_2 amplifies editing when the probe output oscillates, signaling uncertainty. Appendix K motivates the functional form of F_1 and F_2.

The probe remains calibrated under editing because it evaluates the model's own next-step representation rather than the edited hidden state directly; we verify this through orthogonality analysis and empirical case studies in Appendix L.

Figure 6 visualizes the signal accessibility gap in the confidence-variance space: correct and incorrect steps form distinct clusters in the probe's semantic space at all three scales, but overlap completely in the token space.

5 Experiments

5.1 Setup

Models. We evaluate on three model scales: DeepSeek-R1-1.5B (Guo et al., 2025) at layer 19, DeepSeek-R1-7B (Guo et al., 2025) at layer 22, and QwQ-32B (Team, 2025) at layer 58. The editing layer is selected automatically via Ridge regression R^2 on the training set, as described in Appendix C. All main results use greedy decoding (T=0); the variance analysis in Appendix G uses T=0.7 sampling.

Benchmarks. Six mathematical reasoning benchmarks: MATH-500 (Lightman et al., 2023), AIME24 (AI-MO, 2024a), AIME25 (OpenCompass, 2025), GSM8K (Cobbe et al., 2021), AMC23 (AI-MO, 2024b), and Olympiad Bench (He et al., 2024). For cross-domain evaluation: GPQA-Diamond (Rein et al., 2024), a graduate-level science reasoning benchmark, and LiveCodeBench (Jain et al., 2024), a code generation benchmark. The probe and editing direction are extracted from 500 MATH (Hendrycks et al., 2021) training problems. All evaluation benchmarks are fully out-of-distribution except MATH-500, which draws exclusively from the held-out test split.

Baselines. We compare against ten inference-time methods spanning prompt-based (CoD (Xu et al., 2025), NoThinking (Ma et al., 2025)), token-suppression (NoWait (Wang et al., 2025a)), early-exit (DEER (Yang et al., 2025), Dynasor-CoT (Fu et al., 2025)), steering-based (SEAL (Chen et al., 2025), Manifold Steering (Huang et al., 2025), ReBalance (Li et al., 2026)), and verifier-based (FlashThink (Jiang et al., 2025), TrimR (Lin et al., 2025)) approaches. Supervision details and data efficiency analysis are provided in Appendix H and I.

5.2 Main Results

As Table 3 shows, SHEAR achieves the highest or tied-highest Pass@1 on all 6 benchmarks at all three scales. On DeepSeek-R1-1.5B and DeepSeek-R1-7B, gains over Vanilla are largest on competition benchmarks, reaching up to 71.7% on AIME24 and 66.5% on AIME25, with GSM8K and Olympiad also showing clear improvements. On QwQ-32B, gains are smaller but consistent across all six benchmarks, reaching 25.1% on AIME25 and 9.9% on AIME24. Over 10 independent trials with T=0.7 sampling, 13 of the 18 model-dataset configurations reach statistical significance at $p < 0.05$ against the strongest baseline via Welch's t-test, as detailed in Appendix G.

Table 3: Main results on six mathematical reasoning benchmarks (Pass@1 %). **Bold**: best; underline: second best. "–": results not available from prior work. Δ: relative change vs. Vanilla.

Method	MATH-500	AIME24	AIME25	GSM8K	AMC23	Olympiad
			DEEPSEEK-R1-1.5B			
Vanilla	79.6	23.3	20.0	76.0	62.5	41.2
w/ CoD	80.2$_{\blacktriangle 0.8}$	33.3$_{\blacktriangle 42.9}$	13.3$_{\blacktriangledown 33.5}$	69.5$_{\blacktriangledown 8.6}$	62.5$_{\blacktriangle 0.0}$	35.2$_{\blacktriangledown 14.6}$
w/ DEER	67.0$_{\blacktriangledown 15.8}$	20.0$_{\blacktriangledown 14.2}$	23.3$_{\blacktriangle 16.5}$	69.2$_{\blacktriangledown 8.9}$	57.5$_{\blacktriangledown 8.0}$	35.4$_{\blacktriangledown 14.1}$
w/ NoThinking	75.0$_{\blacktriangledown 5.8}$	6.7$_{\blacktriangledown 71.2}$	16.7$_{\blacktriangledown 16.5}$	61.6$_{\blacktriangledown 18.9}$	60.0$_{\blacktriangledown 4.0}$	37.3$_{\blacktriangledown 9.5}$
w/ NoWait	78.0$_{\blacktriangledown 2.0}$	30.0$_{\blacktriangle 28.8}$	16.7$_{\blacktriangledown 16.5}$	75.1$_{\blacktriangledown 1.2}$	60.0$_{\blacktriangledown 4.0}$	40.6$_{\blacktriangledown 1.5}$
w/ Dynasor-CoT	77.2$_{\blacktriangledown 3.0}$	26.7$_{\blacktriangle 14.6}$	26.7$_{\blacktriangle 33.5}$	77.1$_{\blacktriangle 1.4}$	72.5$_{\blacktriangle 16.0}$	42.6$_{\blacktriangle 3.4}$
w/ SEAL	78.6$_{\blacktriangledown 1.3}$	23.3$_{\blacktriangle 0.0}$	26.7$_{\blacktriangle 33.5}$	76.4$_{\blacktriangle 0.5}$	75.0$_{\blacktriangle 20.0}$	32.7$_{\blacktriangledown 20.6}$
w/ Manifold Steering	78.6$_{\blacktriangledown 1.3}$	30.0$_{\blacktriangle 28.8}$	–	77.2$_{\blacktriangle 1.6}$	72.5$_{\blacktriangle 16.0}$	–
w/ ReBalance	83.0$_{\blacktriangle 4.3}$	33.3$_{\blacktriangle 42.9}$	26.7$_{\blacktriangle 33.5}$	78.3$_{\blacktriangle 3.0}$	72.5$_{\blacktriangle 16.0}$	43.9$_{\blacktriangle 6.6}$
w/ **SHEAR (ours)**	**84.6**$_{\blacktriangle 6.3}$	**40.0**$_{\blacktriangle 71.7}$	**33.3**$_{\blacktriangle 66.5}$	**79.5**$_{\blacktriangle 4.6}$	**77.5**$_{\blacktriangle 24.0}$	**45.5**$_{\blacktriangle 10.4}$
			DEEPSEEK-R1-7B			
Vanilla	89.8	46.7	30.0	89.2	85.0	56.1
w/ CoD	90.0$_{\blacktriangle 0.2}$	46.7$_{\blacktriangle 0.0}$	36.7$_{\blacktriangle 22.3}$	84.5$_{\blacktriangledown 5.3}$	85.0$_{\blacktriangle 0.0}$	47.5$_{\blacktriangledown 15.3}$
w/ DEER	87.8$_{\blacktriangledown 2.2}$	50.0$_{\blacktriangle 7.1}$	40.0$_{\blacktriangle 33.3}$	90.4$_{\blacktriangle 1.3}$	80.0$_{\blacktriangledown 5.9}$	53.9$_{\blacktriangledown 3.9}$
w/ NoThinking	80.6$_{\blacktriangledown 10.2}$	26.7$_{\blacktriangledown 42.8}$	20.0$_{\blacktriangledown 33.3}$	87.1$_{\blacktriangledown 2.4}$	65.0$_{\blacktriangledown 23.5}$	45.3$_{\blacktriangledown 19.3}$
w/ NoWait	86.8$_{\blacktriangledown 3.3}$	50.0$_{\blacktriangle 7.1}$	26.7$_{\blacktriangledown 11.0}$	90.2$_{\blacktriangle 1.1}$	85.0$_{\blacktriangle 0.0}$	52.1$_{\blacktriangledown 7.1}$
w/ Dynasor-CoT	88.2$_{\blacktriangledown 1.8}$	46.7$_{\blacktriangle 0.0}$	33.3$_{\blacktriangle 11.0}$	87.6$_{\blacktriangledown 1.8}$	85.0$_{\blacktriangle 0.0}$	55.4$_{\blacktriangledown 1.2}$
w/ SEAL	90.6$_{\blacktriangle 0.9}$	43.3$_{\blacktriangledown 7.3}$	26.7$_{\blacktriangledown 11.0}$	88.4$_{\blacktriangledown 0.9}$	77.5$_{\blacktriangledown 8.8}$	53.9$_{\blacktriangledown 3.9}$
w/ Manifold Steering	88.4$_{\blacktriangledown 1.6}$	53.3$_{\blacktriangle 14.1}$	–	87.6$_{\blacktriangledown 1.8}$	87.5$_{\blacktriangle 2.9}$	–
w/ ReBalance	92.6$_{\blacktriangle 3.1}$	53.3$_{\blacktriangle 14.1}$	40.0$_{\blacktriangle 33.3}$	91.6$_{\blacktriangle 2.7}$	92.5$_{\blacktriangle 8.8}$	57.0$_{\blacktriangle 1.6}$
w/ **SHEAR (ours)**	**94.2**$_{\blacktriangle 4.9}$	**60.0**$_{\blacktriangle 28.5}$	**43.3**$_{\blacktriangle 44.3}$	**95.5**$_{\blacktriangle 7.1}$	**95.0**$_{\blacktriangle 11.8}$	**59.4**$_{\blacktriangle 5.9}$
			QWQ-32B			
Vanilla	94.8	66.7	53.3	96.3	87.5	66.7
w/ CoD	93.8$_{\blacktriangledown 1.1}$	63.3$_{\blacktriangledown 5.1}$	46.7$_{\blacktriangledown 12.4}$	96.2$_{\blacktriangledown 0.1}$	92.5$_{\blacktriangle 5.7}$	67.7$_{\blacktriangle 1.5}$
w/ DEER	94.4$_{\blacktriangle 0.4}$	70.0$_{\blacktriangle 4.9}$	46.7$_{\blacktriangledown 12.4}$	96.2$_{\blacktriangledown 0.1}$	**95.0**$_{\blacktriangle 8.6}$	64.3$_{\blacktriangledown 3.6}$
w/ NoThinking	94.8$_{\blacktriangle 0.0}$	66.7$_{\blacktriangle 0.0}$	**66.7**$_{\blacktriangle 25.1}$	96.5$_{\blacktriangle 0.2}$	90.0$_{\blacktriangle 2.9}$	66.1$_{\blacktriangledown 0.9}$
w/ NoWait	93.8$_{\blacktriangledown 1.1}$	66.7$_{\blacktriangle 0.0}$	63.3$_{\blacktriangle 18.8}$	96.3$_{\blacktriangle 0.0}$	92.5$_{\blacktriangle 5.7}$	62.6$_{\blacktriangledown 6.1}$
w/ Dynasor-CoT	94.2$_{\blacktriangledown 0.6}$	63.3$_{\blacktriangledown 5.1}$	–	95.2$_{\blacktriangledown 1.1}$	90.0$_{\blacktriangle 2.9}$	–
w/ SEAL	92.6$_{\blacktriangledown 2.3}$	63.3$_{\blacktriangledown 5.1}$	56.7$_{\blacktriangle 6.4}$	96.2$_{\blacktriangledown 0.1}$	**95.0**$_{\blacktriangle 8.6}$	67.5$_{\blacktriangle 1.2}$
w/ FlashThink	93.2$_{\blacktriangledown 1.7}$	60.0$_{\blacktriangledown 10.0}$	40.0$_{\blacktriangledown 25.0}$	96.5$_{\blacktriangle 0.2}$	92.5$_{\blacktriangle 5.7}$	–
w/ TrimR	93.8$_{\blacktriangledown 1.1}$	56.7$_{\blacktriangledown 15.0}$	43.3$_{\blacktriangledown 18.8}$	93.7$_{\blacktriangledown 2.7}$	90.0$_{\blacktriangle 2.9}$	–
w/ ReBalance	95.2$_{\blacktriangle 0.4}$	70.0$_{\blacktriangle 4.9}$	63.3$_{\blacktriangle 18.8}$	96.8$_{\blacktriangle 0.5}$	**95.0**$_{\blacktriangle 8.6}$	68.6$_{\blacktriangle 2.8}$
w/ **SHEAR (ours)**	**95.9**$_{\blacktriangle 1.2}$	**73.3**$_{\blacktriangle 9.9}$	**66.7**$_{\blacktriangle 25.1}$	**97.0**$_{\blacktriangle 0.7}$	**95.0**$_{\blacktriangle 8.6}$	**69.0**$_{\blacktriangle 3.4}$

Table 4: Cross-domain generalization (Pass@1 %). The same probe and editing direction trained on MATH transfer to graduate-level science reasoning and code generation with no additional tuning. Δ: relative change vs. Vanilla. **Bold**: best; underline: second best.

	GPQA-Diamond			LiveCodeBench		
Method	DEEPSEEK-1.5B	DEEPSEEK-7B	QWQ-32B	DEEPSEEK-1.5B	DEEPSEEK-7B	QWQ-32B
Vanilla	17.1	33.8	63.1	19.5	44.0	87.5
w/ ReBalance	21.7$_{\blacktriangle 26.9}$	39.4$_{\blacktriangle 16.6}$	67.2$_{\blacktriangle 6.5}$	22.5$_{\blacktriangle 15.4}$	46.5$_{\blacktriangle 5.7}$	88.3$_{\blacktriangle 0.9}$
w/ **SHEAR (ours)**	**29.3**$_{\blacktriangle 71.3}$	**51.8**$_{\blacktriangle 53.3}$	**68.8**$_{\blacktriangle 9.0}$	**26.0**$_{\blacktriangle 33.3}$	**50.2**$_{\blacktriangle 14.1}$	**89.4**$_{\blacktriangle 2.2}$

5.3 CROSS-DOMAIN GENERALIZATION

Using the same probe and direction extracted from MATH, with zero additional training, SHEAR improves over both vanilla and the strongest baseline on GPQA-Diamond and LiveCodeBench across all three model scales, as Table 4 shows. On GPQA-Diamond with DEEPSEEK-R1-7B, SHEAR improves over vanilla by 53.3%. On DEEPSEEK-R1-7B, the 53.3% gain on GPQA exceeds all in-domain gains, consistent with its lower baseline leaving more room for corrective editing. This transfer demonstrates that the editing direction captures a general property of reasoning quality, rather than mathematics-specific features. The running answer that defines s^* is required only when building the offline training set: LiveCodeBench admits no comparable intermediate answer, yet the MATH-derived probe and direction apply to it unchanged. Appendix M further examines architectural generality across the three models' distinct training regimes.

Table 5: Ablation study. "w/o Direction" replaces the stability-anchored direction v with a random unit vector; "w/o Probe" replaces the adaptive α_s with a fixed $\alpha_s{=}\beta$ for all s. Δ: relative change vs. Vanilla. **Bold**: best.

Model	Component	AMC23	GSM8K	MATH-500	Olympiad
DEEPSEEK-R1-7B	**SHEAR**	**95.0**$_{\blacktriangle 11.8}$	**95.5**$_{\blacktriangle 7.1}$	**94.2**$_{\blacktriangle 4.9}$	**59.4**$_{\blacktriangle 5.9}$
	w/o Direction	87.5$_{\blacktriangle 2.9}$	91.2$_{\blacktriangle 2.2}$	92.4$_{\blacktriangle 2.9}$	57.9$_{\blacktriangle 3.2}$
	w/o Probe	87.5$_{\blacktriangle 2.9}$	90.5$_{\blacktriangle 1.5}$	92.2$_{\blacktriangle 2.7}$	57.0$_{\blacktriangle 1.6}$
	Vanilla	85.0	89.2	89.8	56.1
DEEPSEEK-R1-1.5B	**SHEAR**	**77.5**$_{\blacktriangle 24.0}$	**79.5**$_{\blacktriangle 4.6}$	**84.6**$_{\blacktriangle 6.3}$	**45.5**$_{\blacktriangle 10.4}$
	w/o Direction	60.0$_{\blacktriangledown 4.0}$	78.8$_{\blacktriangle 3.7}$	83.0$_{\blacktriangle 4.3}$	42.8$_{\blacktriangle 3.9}$
	w/o Probe	67.5$_{\blacktriangle 8.0}$	78.2$_{\blacktriangle 2.9}$	72.4$_{\blacktriangledown 9.0}$	39.9$_{\blacktriangledown 3.2}$
	Vanilla	62.5	76.0	79.6	41.2

Table 6: Directional specificity of the editing vector (Pass@1 %). Δ: relative change from Vanilla. Flipping the sign of v collapses accuracy to near zero.

Dataset	Model	Vanilla	SHEAR (+)		Reverse (−)	
			Acc	Δ	Acc	Δ
AMC23	DEEPSEEK-R1-1.5B	62.5	**77.5**	+24.0	10.3	−83.5
	DEEPSEEK-R1-7B	85.0	**95.0**	+11.8	15.0	−82.4
GSM8K	DEEPSEEK-R1-1.5B	76.0	**79.5**	+4.6	0.8	−98.9
	DEEPSEEK-R1-7B	89.2	**95.5**	+7.1	0.7	−99.2

5.4 ABLATION STUDY

Table 5 shows that removing either component degrades performance at both scales. On DEEPSEEK-R1-7B, the full system's gain on GSM8K of 7.1% exceeds the sum of individual component gains of 1.5% and 2.2%, suggesting complementarity. The scale-dependent pattern is notable: on DEEPSEEK-R1-1.5B, the w/o Probe variant causes MATH-500 to drop from 84.6% to 72.4%, *below* the 79.6% Vanilla. Similarly, the w/o Direction variant drops AMC23 from 77.5% to 60.0%, below the 62.5% Vanilla. The w/o Probe variant applies editing at fixed positive magnitude ($\alpha_s{=}\beta{>}0$) on every step, restricting all edits to the $+v$ direction; on DEEPSEEK-R1-7B this still improves over Vanilla, but on DEEPSEEK-R1-1.5B the indiscriminate editing is harmful. These results reinforce that *both* components are necessary: the direction provides semantic content for how to edit, and the probe provides per-step modulation that prevents over-editing.

5.5 REVERSE-DIRECTION EDITING

If the improvements were merely a side effect of noise-driven exploration, reversing v should produce comparable results. We test this by flipping the sign of v while keeping the control function identical.

Table 6 shows an extreme asymmetry: forward editing yields improvements of 4.6% to 24.0%, while reverse editing causes collapse to near-zero accuracy, with drops of 82.4% to 99.2%. The asymmetry holds on both DEEPSEEK-R1-1.5B and DEEPSEEK-R1-7B and on both benchmarks. We note that flipping v also inverts the sign semantics of the control function, so the collapse reflects both directional content and control interaction. The random-direction baseline in Appendix F isolates the directional contribution more cleanly: under the same magnitude profile, a random direction achieves only 87.5%, confirming that v encodes reasoning-relevant structure.

6 CONCLUSION

We asked whether a reasoning model's own assessment of its progress can be fed back to correct it mid-thought. Controlled probing experiments show that reasoning correctness is considerably more accessible in hidden states than in token-level statistics, and SHEAR turns this signal into per-step edits that yield statistically significant improvements against the strongest baseline in the majority of configurations across three model scales. More broadly, the gap between what LLMs encode internally and what their outputs reveal suggests that hidden states are an underutilized resource for improving reasoning; limitations and broader discussion are provided in Appendix N.

AI USE STATEMENT

We used generative AI for two subtasks. First, we generated the probe training data with the models under study, labeling each reasoning step automatically (§3, Appendix C). Second, we used generative AI tools to check and correct the grammar of our writing.

ETHICS STATEMENT

We identify no ethical concerns in this work. We run all experiments on publicly released models (DeepSeek-R1-Distill-Qwen-1.5B/7B and QwQ-32B) and publicly available reasoning benchmarks, without human subjects or private and personally identifiable data. We release no new models or datasets, and we have read and adhere to the ICLR Code of Ethics.

REPRODUCIBILITY STATEMENT

We have taken the following steps to make our results reproducible. Appendix C documents the complete offline pipeline, the probe architecture and training procedure, the layer-selection criterion, and every control-function hyperparameter. Appendix D gives the precise definition of step-level correctness used to construct probe labels, and Appendix A collects all notation. Appendix B states the full inference-time editing algorithm in pseudocode. All models and benchmarks used in the paper are publicly available. Appendix G reports mean and standard deviation over ten independent sampling runs together with Welch's t-test significance levels, so that the reported gains can be assessed against run-to-run variance. A complete anonymized implementation is available at https://anonymous.4open.science/r/SHEAR-C9A4/.

REFERENCES

AI-MO. Aime 2024, July 2024a. URL https://huggingface.co/datasets/AI-MO/aimo-validation-aime.

AI-MO. Amc 2023, July 2024b. URL https://huggingface.co/datasets/AI-MO/aimo-validation-amc.

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

Yonatan Belinkov, Nadir Durrani, Fahim Dalvi, Hassan Sajjad, and James Glass. What do neural machine translation models learn about morphology? In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 861–872, 2017.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering latent knowledge in language models without supervision. In *The Eleventh International Conference on Learning Representations*, 2023.

Runjin Chen, Zhenyu Zhang, Junyuan Hong, Souvik Kundu, and Zhangyang Wang. Seal: Steerable reasoning calibration of large language models for free. *arXiv preprint arXiv:2504.07986*, 2025.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Yichao Fu, Junda Chen, Yonghao Zhuang, Zheyu Fu, Ion Stoica, and Hao Zhang. Reasoning without self-doubt: More efficient chain-of-thought through certainty probing. In *ICLR 2025 Workshop on Foundation Models in the Wild*, 2025.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, et al. Olympiadbench: A challenging benchmark for promoting agi with olympiad-level bilingual multimodal scientific problems. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3828–3850, 2024.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*, 2021.

John Hewitt and Christopher D Manning. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4129–4138, 2019.

Yao Huang, Huanran Chen, Shouwei Ruan, Yichi Zhang, Xingxing Wei, and Yinpeng Dong. Mitigating overthinking in large reasoning models via manifold steering. *arXiv preprint arXiv:2505.22411*, 2025.

Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. Livecodebench: Holistic and contamination free evaluation of large language models for code. *arXiv preprint arXiv:2403.07974*, 2024.

Guochao Jiang, Guofeng Quan, Zepeng Ding, Ziqin Luo, Dixuan Wang, and Zheng Hu. Flashthink: An early exit method for efficient reasoning. *arXiv preprint arXiv:2505.13949*, 2025.

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

Yulin Li, Tengyao Tu, Li Ding, Junjie Wang, Huiling Zhen, Yixin Chen, Yong Li, and Zhuotao Tian. Efficient reasoning with balanced thinking. In *Proceedings of the 14th International Conference on Learning Representations*, 2026.

Zhixiang Liang, Beichen Huang, Zheng Wang, and Minjia Zhang. Hidden states as early signals: Step-level trace evaluation and pruning for efficient test-time scaling. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 26800–26813, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-395-1. doi: 10.18653/v1/2026.findings-acl.1336. URL https://aclanthology.org/2026.findings-acl.1336/.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. *arXiv preprint arXiv:2305.20050*, 2023.

Weizhe Lin, Xing Li, Zhiyuan Yang, Xiaojin Fu, Hui-Ling Zhen, Yaoyuan Wang, Xianzhi Yu, Wulong Liu, Xiaosong Li, and Mingxuan Yuan. Trimr: Verifier-based training-free thinking compression for efficient test-time scaling. *arXiv preprint arXiv:2505.17155*, 2025.

Liangchen Luo, Yinxiao Liu, Rosanne Liu, Samrat Phatale, Meiqi Guo, Harsh Lara, Yunxuan Li, Lei Shu, Yun Zhu, Lei Meng, et al. Improve mathematical reasoning in language models by automated process supervision. *arXiv preprint arXiv:2406.06592*, 2024.

Wenjie Ma, Jingxuan He, Charlie Snell, Tyler Griggs, Sewon Min, and Matei Zaharia. Reasoning models can be effective without thinking. *arXiv preprint arXiv:2504.09858*, 2025.

Samuel Marks and Max Tegmark. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. *arXiv preprint arXiv:2310.06824*, 2023.

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

Kevin Meng, Arnab Sen Sharma, Alex Andonian, Yonatan Belinkov, and David Bau. Mass-editing memory in a transformer. In *The Eleventh International Conference on Learning Representations*, 2023.

OpenCompass. Aime 2025, February 2025. URL https://huggingface.co/datasets/opencompass/AIME2025.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. Gpqa: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*, 2024.

Qwen Team. Qwq-32b: Embracing the power of reinforcement learning, March 2025. URL https://qwenlm.github.io/blog/qwq-32b/.

Alexander Matt Turner, Lisa Thiergart, David Udell, Gavin Leech, Ulisse Mini, and Monte MacDiarmid Tong. Activation addition: Steering language models without optimization. *arXiv preprint arXiv:2308.10248*, 2024.

Chenlong Wang, Yuanning Feng, Dongping Chen, Zhaoyang Chu, Ranjay Krishna, and Tianyi Zhou. Wait, we don't need to "wait"! removing thinking tokens improves reasoning efficiency. *arXiv preprint arXiv:2506.08343*, 2025a.

Peiyi Wang, Lei Li, Zhihong Shao, Runxin Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. Math-shepherd: Verify and reinforce llms step-by-step without human annotations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9426–9439, 2024.

Yiming Wang, Pei Zhang, Baosong Yang, Derek Wong, and Rui Wang. Latent space chain-of-embedding enables output-free llm self-evaluation. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 70938–70970, 2025b. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/b0b1cfc8ede53f452cabf8b9cf4eef76-Paper-Conference.pdf.

Silei Xu, Wenhao Xie, Lingxiao Zhao, and Pengcheng He. Chain of draft: Thinking faster by writing less. *arXiv preprint arXiv:2502.18600*, 2025.

Chenxu Yang, Qingyi Si, Yongjie Duan, Zheliang Zhu, Chenyu Zhu, Qiaowei Li, Zheng Lin, Li Cao, and Weiping Wang. Dynamic early exit in reasoning models. *arXiv preprint arXiv:2504.15895*, 2025.

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

A NOTATION

Table 7 summarizes the symbols used throughout the paper, grouped by their role in the method: reasoning trace representation, stability index computation, editing direction extraction, adaptive control, and hyperparameters.

Table 7: Notation used throughout the paper.

Symbol	Definition	Description
Reasoning trace and model		
s	Reasoning step index	Steps segmented by `\n\n` delimiters
$h_s \in \mathbb{R}^D$	Layer-ℓ hidden state at first token of step s	Input to the probe and target of editing
D	Hidden dimension	Model-dependent
ℓ	Intermediate layer index	Selected via Ridge regression R^2
$f : \mathbb{R}^D \to [0, 1]$	Semantic confidence probe (MLP)	Maps hidden state to correctness probability
Stability index		
$\text{correct}(s) \in \{0, 1\}$	Step correctness label	Running answer matches gold via `math_equal`
θ	Stabilization threshold (default 0.5)	Fraction of subsequent correct steps required
s^*	Stability index (Eq. 1)	Earliest step where answer predominantly stabilizes
$\mathcal{S}_i^>$	$\{s : s > s_i^*\}$	Post-s^* steps for problem i
Editing direction		
δ_i	$h_{s_i^*} - \text{mean}_{s \in \mathcal{S}_i^>}(h_s)$	Per-problem difference vector (Eq. 3)
$v \in \mathbb{R}^D$	$\mathbb{E}_i[\delta_i] / \lVert\mathbb{E}_i[\delta_i]\rVert$	Global unit editing direction
$\tilde{h}_s$	$h_s + \alpha_s \cdot v$	Edited hidden state (Eq. 2)
Adaptive control		
c_s	$f(h_s)$	Probe confidence score at step s
$\bar{c}_s$	$\frac{1}{k}\sum_{j=s-k+1}^{s} c_j$	Windowed mean confidence
σ_s^2	$\frac{1}{k}\sum_{j=s-k+1}^{s}(c_j - \bar{c}_s)^2$	Local variance of probe confidence
α_s	$F_1(c_s) \cdot F_2(\sigma_s^2)$	Editing magnitude (Eq. 4)
$F_1(c_s)$	Eq. 5	Confidence-dependent component
$F_2(\sigma_s^2)$	Eq. 6	Variance-aware amplification
Hyperparameters		
$\beta = 3.0$	Maximum editing magnitude	Scales the low-confidence correction
$\lambda = 0.01$	High-confidence editing strength	Scales the reinforcement nudge
τ_l, τ_h	25th/75th percentiles of training-set c_s	Confidence thresholds for F_1
$\sigma_0 = 0.01$	Reference variance scale	Normalizes σ_s^2 in F_2
$\gamma = 0.005$	Variance activation threshold	Triggers amplification in F_2
$k = 5$	Sliding window size	Window for $\bar{c}_s$ and σ_s^2

B ALGORITHM

Algorithm 1 presents the full SHEAR pipeline. The offline stage (Lines 2–10) runs once per model to extract the editing direction and train the probe. The online stage (Lines 12–20) edits hidden states at each reasoning step during inference, with the editing magnitude determined adaptively by the probe and the control function.

C IMPLEMENTATION DETAILS

Offline pipeline (once per model). (1) Run 500 MATH training problems with greedy decoding, dumping layer-ℓ hidden states ($\sim$3h, 8$\times$NVIDIA H200 GPUs). Probe training, direction extraction, and s^* statistics use all 500 problems; full hidden states for a representative subset (up to 300 problems attempted, 266 retained after filtering for minimum step count) are cached (`records_*_300.npz`, $\sim$100–150 MB for 1.5B/7B) for visualization and figure generation; the 32B cache stores only scalar signals (probe scores, token confidence, labels) due to memory constraints. (2) Select the optimal editing layer via Ridge regression R^2 across candidate layers. (3) Extract the stability-anchored direction v (Eq. 3). (4) Train the semantic confidence probe ($<$1 min).

Algorithm 1 SHEAR: Semantic Hidden-layer Editing with Anchored Reasoning

Require: Model $\mathcal{M}$ with layer ℓ; training set $\mathcal{D}_{\text{train}}$ with gold answers; test problem x; hyperparameters $\beta, \lambda, k, \sigma_0, \gamma$
Ensure: Edited reasoning trace for x

// — Offline: Direction Extraction & Probe Training (once per model) —
for each problem $i \in \mathcal{D}_{\text{train}}$ **do**
  Generate reasoning trace; extract $h_s^{(i)} \in \mathbb{R}^D$ at layer ℓ for each step s
  $\text{correct}(s) \leftarrow \mathbf{1}[\texttt{math_equal}(\text{answer}(s),\ \text{gold}_i)]$
  Compute stability index s_i^* via Eq. 1
end for
$\delta_i \leftarrow h_{s_i^*} - \frac{1}{|\mathcal{S}_i^>|} \sum_{s \in \mathcal{S}_i^>} h_s; \quad v \leftarrow \mathbb{E}_i[\delta_i] / \|\mathbb{E}_i[\delta_i]\|$ ▷ Eq. 3
Train MLP probe $f : \mathbb{R}^D \to [0, 1]$ on $\{(h_s, \text{correct}(s))\}$ with BCE loss
$\tau_l \leftarrow P_{25}(\{f(h_s)\}); \quad \tau_h \leftarrow P_{75}(\{f(h_s)\})$ ▷ Confidence thresholds

// — Online: Inference-Time Editing —
for each reasoning step s during generation of x **do**
  $h_s \leftarrow$ layer-ℓ hidden state at first token of step s
  $c_s \leftarrow f(h_s)$ ▷ Semantic confidence
  $\bar{c}_s \leftarrow \frac{1}{k} \sum_{j=s-k+1}^{s} c_j; \quad \sigma_s^2 \leftarrow \frac{1}{k} \sum_{j=s-k+1}^{s} (c_j - \bar{c}_s)^2$
  $\alpha_s \leftarrow F_1(c_s) \cdot F_2(\sigma_s^2)$ ▷ Eqs. 5–6
  $\tilde{h}_s \leftarrow h_s + \alpha_s \cdot v$ ▷ Additive editing
  Continue generation conditioned on $\tilde{h}_s$
end for

Table 8: Probe quality metrics (%).

Metric	1.5B (L19)	7B (L22)	32B (L58)
Val Accuracy	95.4	93.4	92.7
Predicted Positive Rate	10.8	8.4	4.8
Actual Positive Rate	13.3	10.5	8.5

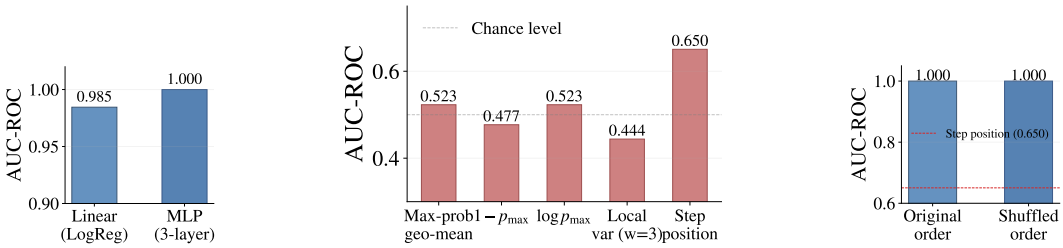
Probe architecture. MLP: Linear(D, 256) $\to$ ReLU $\to$ LayerNorm $\to$ Linear(256, 64) $\to$ ReLU $\to$ Linear(64, 1) $\to$ Sigmoid. Training: BCE loss, Adam optimizer, early stopping. Inference cost: one probe evaluation per reasoning step costs $2(256D + 16{,}448)$ FLOPs, i.e. 0.8, 1.9, and 2.7 MFLOPs at D=1536, 3584, and 5120 respectively. Generating a single token from the corresponding model costs $\approx 2N_{\text{params}}$ FLOPs, i.e. 3.0, 15, and 64 GFLOPs, so one probe evaluation stays below 3×10^{-4} of the cost of a *single* token, and a reasoning step spans many tokens. The edit $\tilde{h}_s = h_s + \alpha_s v$ adds a further $2D$ FLOPs. The arithmetic overhead of SHEAR is therefore negligible relative to decoding.

Control function parameters. τ_l, τ_h: 25th/75th percentiles of training-set probe confidence. $\beta = 3.0$: maximum editing magnitude. $\lambda = 0.01$: high-confidence editing strength, the parameter most sensitive to tuning. $\sigma_0 = 0.01$: reference variance scale. $\gamma = 0.005$: variance activation threshold. Sliding window size $k = 5$. These values are shared by all three model scales.

Probe calibration bias. Table 8 shows that the predicted positive rate is systematically lower than the actual positive rate at all three scales, indicating a conservative probe that slightly under-predicts correctness. Since τ_l and τ_h are defined as percentiles of the probe's own output distribution, the thresholds automatically absorb this bias, ensuring a consistent fraction of steps receives editing regardless of the probe's absolute calibration.

Table 9: Layer sensitivity on DEEPSEEK-R1-7B, AMC23 (Pass@1 %). **Bold**: best.

Layer	18	20	22	24
Accuracy	90.0	92.5	**95.0**	87.5



(a) Linear probe **(b)** Token variants **(c)** Shuffle control

Figure 7: Robustness checks on DEEPSEEK-R1-7B. (a) Linear separability test: AUC evaluated with probe-derived binary labels (c_s>0.5); GT-label AUC is reported in Table 1. (b) Token signal variants. (c) Shuffling step order does not affect probe AUC, validating that the probe captures per-step content rather than positional patterns.

D STEP CORRECTNESS DEFINITION

Step correctness is defined as follows: step s is labeled correct if and only if the running answer extracted from the model's output at step s matches the gold answer via symbolic equivalence (`math_equal`). This answer-based definition may occasionally mislabel a step whose reasoning is flawed but whose running answer happens to be correct, or vice versa. In practice, these edge cases are uncommon in chain-of-thought reasoning, and the definition aligns with the operational goal of SHEAR: preserving and accelerating *answer stabilization* indexed by s^*. The strong discriminative performance of the probe and the catastrophic collapse under reverse editing together confirm that the labeled signal captures genuine reasoning progress.

Boundary cases. Eq. 1 computes the fraction of subsequent steps that are correct: $|\{s' > s : \text{correct}(s') = 1\}|/|\{s' > s\}|$. When s is the final reasoning step, $|\{s' > s\}| = 0$ and this fraction is undefined. By convention, we treat the empty ratio as 0, so the last step never satisfies the condition $\geq \theta$ and therefore never qualifies as s^*. In practice, all valid s^* values in our dataset occur at least two steps before the end of the reasoning trace.

E ROBUSTNESS ANALYSIS

Layer sensitivity. Table 9 shows that performance is robust across adjacent layers on DEEPSEEK-R1-7B AMC23, with layers 20–24 all exceeding 87%.

Robustness checks. Figure 7 presents additional robustness checks including: linear probe performance, token signal variants (entropy, variance), and shuffle controls.

s^* threshold sensitivity. The 50% threshold in the s^* definition (Eq. 1) is a heuristic choice. Table 10 shows that the extracted direction is robust to threshold variation: for $\theta \in [0.4, 0.7]$, pairwise cosine similarity with $v_{\theta=0.5}$ exceeds 0.96.

F DIRECTION ANALYSIS

Direction comparison. Table 11 compares the stability-anchored direction against a confidence-mixed direction (averaging hidden states by confidence bins rather than stability index). On GSM8K, the stability-anchored direction improves by $+4.5$ percentage points.

Table 10: s^* threshold sensitivity on DEEPSEEK-R1-7B, restricted to the 345 problems with ≥ 3 reasoning steps and ≥ 1 correct step. "Valid Problems" denotes the fraction of this subset yielding a valid s^*. The editing direction is highly stable across thresholds. **Bold**: default setting.

Threshold θ	Valid Problems	$\overline{\cos(\delta_i, v_\theta)}$	$\cos(v_\theta, v_{0.5})$
0.3	79.5%	0.282	0.845
0.4	77.5%	0.343	0.982
0.5	**76.5%**	**0.399**	**1.000**
0.6	74.5%	0.469	0.994
0.7	65.0%	0.533	0.988

Table 11: Editing direction comparison on DEEPSEEK-R1-7B (Pass@1 %). **Bold**: best.

Direction	AMC23	GSM8K
Confidence-mixed	92.5	91.0
Stability-anchored	**95.0**	**95.5**

Random direction baseline. Table 12 shows that a random direction combined with the probe achieves only 87.5%, compared to 95.0% for the full system, a $+7.5$ percentage point gap that shows the direction carries meaningful semantic content.

Component replacement. Table 13 shows that replacing either the direction (stability→random) or the signal (probe→token) degrades performance by -7.5%, confirming both components are essential.

Direction consistency. Despite the high dimensionality, 87.6% of individual δ_i vectors point in the same hemisphere as v. At the default threshold θ=0.5, the mean cosine similarity reaches 0.399 (Table 10), well above the random baseline.

G STATISTICAL SIGNIFICANCE

The 5 non-significant cases involve small-sample benchmarks where N=30 limits statistical power, the near-ceiling 32B GSM8K setting at 97.2%, or 32B Olympiad where the effect size is only $+0.4\%$.

H SUPERVISION LEVEL COMPARISON

SHEAR uses GT step-level correctness labels from 500 problems, while all baselines are unsupervised or use only token-level signals. Two observations contextualize this difference: (i) 20 problems suffice for AUC = 0.933 (Appendix I), placing the supervision requirement on par with few-shot methods; (ii) the probe and direction transfer to GPQA-Diamond and LiveCodeBench with zero additional training (Table 4), ruling out overfitting to the MATH distribution.

I PROBE DATA-SIZE ABLATION

Table 16 and Figure 8a show that AUC exceeds 0.92 with as few as 20 training problems, far exceeding the best token-level probe at 0.716. The correctness signal in hidden states is near-linearly separable (full-dimensional linear probe: AUC = 0.945; MLP: AUC = 0.983), so the probe saturates rapidly. The high linear-probe AUC indicates a large-margin separation in representation space; under such conditions, standard sample complexity results for linear classifiers predict saturation with far fewer labeled examples than the feature dimension D.

J SCALING AND MULTI-LAYER ANALYSIS

The readability of reasoning correctness from hidden states is consistent from 1.5B to 32B (probe val accuracy: 95.4%, 93.4%, 92.7% respectively).

Table 12: Random direction baseline on DEEPSEEK-R1-7B, AMC23 (Pass@1 %). **Bold**: best.

Method	AMC23
Vanilla	85.0
Stability direction only	87.5
Random direction + probe	87.5
SHEAR	**95.0**

Table 13: Component replacement on DEEPSEEK-R1-7B, AMC23 (Pass@1 %). **Bold**: best.

Intervention	AMC23	Meaning
Direction: stability → random	87.5	Direction matters
Signal: probe → token	87.5	Signal quality matters
SHEAR	**95.0**	

Layer-wise probe AUC (logistic regression, DEEPSEEK-R1-7B) increases monotonically from layer 0 to layer 22 (Figure 8b), indicating that reasoning correctness information accumulates through the transformer's depth. Even the earliest layers outperform token max-probability (AUC = 0.563).

K CONTROL FUNCTION DESIGN RATIONALE

The control function (Eqs. 5–6) contains several design choices whose rationale we elaborate here.

Asymmetric editing strength. The ratio $\beta/\lambda = 300$ reflects a fundamental asymmetry in the editing task: correcting wrong reasoning requires a strong perturbation to shift the model's trajectory away from an incorrect path, while reinforcing correct reasoning needs only a gentle nudge to avoid disrupting an already successful computation. Empirically, high-confidence steps already produce correct answers and benefit little from intervention; aggressive editing at this regime risks destabilizing the model. Low-confidence steps, by contrast, are on an incorrect trajectory where the default generation would continue producing errors, so a stronger signal is needed to redirect computation.

Cross-scale parameter sharing. Three factors enable the same hyperparameters to work across 1.5B, 7B, and 32B scales. First, the editing direction v is a unit vector, so α_s controls the absolute displacement regardless of the hidden dimension D. Second, the confidence thresholds τ_l and τ_h are defined as the 25th and 75th percentiles of the training-set probe distribution, which automatically adapts to each model's confidence range. Third, the probe is retrained per model, so c_s values are calibrated to each scale before the control function is applied. The remaining parameters (β, σ_0, γ, k) control relative magnitudes and variance sensitivity, which empirically transfer well because the probe's output semantics are consistent across scales.

Variance term F_2. When the probe output oscillates between high and low values over consecutive steps, the model is uncertain about its reasoning direction. F_2 detects this oscillation: the threshold γ=0.005 corresponds to a standard deviation of approximately 0.07 in probe confidence, which is the typical fluctuation amplitude in the overlapping region between correct and incorrect step distributions. Amplifying the editing magnitude during such episodes helps the model escape the unstable state, while F_2=1 during stable periods ensures no unnecessary amplification.

L PROBE RELIABILITY UNDER INFERENCE-TIME EDITING

A natural concern is whether the probe remains calibrated when the model's representations have been modified by prior editing steps. We address this through three complementary arguments.

Propagation pathway. At step s, editing modifies h_s to $\tilde{h}_s = h_s + \alpha_s v$. The model then generates step $s{+}1$ conditioned on $\tilde{h}_s$. The probe at step $s{+}1$ is applied to h_{s+1}, which is the model's *own*

Table 14: SHEAR mean±std over 10 independent runs with T=0.7 sampling (Pass@1 %). Δ: improvement over the strongest baseline mean under the same setting. $^{**}\,p < 0.01$, $^{*}\,p < 0.05$ (Welch's t-test). 13 of 18 configurations reach statistical significance; 9 of those reach $p < 0.01$.

Benchmark	N	SHEAR	Δ	Sig.
		DEEPSEEK-R1-7B		
GSM8K	1319	95.3 ± 0.4	$+3.8$	**
MATH-500	500	94.0 ± 0.7	$+1.2$	**
Olympiad	675	59.5 ± 1.5	$+3.0$	**
AMC23	40	97.2 ± 3.0	$+4.0$	**
AIME24	30	59.3 ± 4.1	$+2.6$	
AIME25	30	41.3 ± 5.5	$+1.6$	
		DEEPSEEK-R1-1.5B		
GSM8K	1319	79.3 ± 0.7	$+1.0$	*
MATH-500	500	84.7 ± 0.7	$+1.5$	**
Olympiad	675	45.7 ± 1.6	$+2.3$	**
AMC23	40	76.0 ± 5.0	$+5.0$	*
AIME24	30	42.3 ± 6.5	$+9.3$	**
AIME25	30	35.7 ± 3.2	$+11.7$	**
		QWQ-32B		
MATH-500	500	95.9 ± 0.6	$+1.1$	**
AIME25	30	68.0 ± 5.9	$+6.3$	*
AMC23	40	95.5 ± 2.6	$+2.5$	*
AIME24	30	74.7 ± 6.1	$+3.4$	
GSM8K	1319	97.2 ± 0.3	$+0.2$	
Olympiad	675	68.8 ± 1.0	$+0.4$	

Table 15: Supervision levels of compared methods. "Supervision" denotes the type of labels required to construct each method.

Method	Supervision	Description
CoD, NoThinking	Unsupervised	Prompt engineering
NoWait	Unsupervised	Token suppression at decoding time
DEER	Unsupervised	Token-level early exit
Dynasor-CoT	Unsupervised	Internal-state probing + early termination
SEAL, Manifold Steering	Unsupervised	Activation steering (hidden-state space)
FlashThink	Answer-level	Early exit via trained verification model
TrimR	Unsupervised	Pretrained verifier for thought trimming
ReBalance	Token-level	Hidden-state steering + token-confidence control
SHEAR (ours)	**Step-level labels (500)**	Semantic confidence probe + stability-anchored direction

representation of the new step, not the edited $\tilde{h}_s$ directly. The probe thus evaluates a naturally generated representation rather than an artificially perturbed one.

Orthogonality guarantee. The near-zero cosine similarity between the probe gradient and the editing direction ($\cos(\nabla f, v) \approx -0.02$) ensures that editing does not shift representations along the probe's most sensitive axis. While this is a first-order argument, the relative perturbation magnitude ($\|\alpha_s v\|/\|h_s\| \approx 22.8\%$) is moderate, and higher-order effects decay rapidly for well-conditioned networks.

Empirical evidence. The case studies in Figure 1 show the probe responding appropriately throughout edited trajectories: it stays low during incorrect reasoning, rises sharply at the moment of answer stabilization, and remains high afterward. This behavior is consistent across both DEEPSEEK-R1-7B and QWQ-32B, confirming that editing does not disrupt probe calibration in practice.

M ARCHITECTURAL GENERALITY

We note three factors supporting the architectural generality of SHEAR's findings.

First, the two core assumptions, namely that hidden states encode reasoning progress and that additive editing can steer subsequent computation, are standard properties of transformer-based LLMs

Table 16: Probe AUC by number of training problems on DEEPSEEK-R1-7B.

Training Problems	Steps	Probe AUC
20	1,662	0.933
50	4,261	0.924
100	9,713	0.922
150	14,295	0.958



(a) Probe data efficiency

(b) Layer-wise probe AUC

Figure 8: Probe analysis on DEEPSEEK-R1-7B. (a) AUC by number of training problems: performance saturates rapidly, reaching 0.933 with only 20 problems. (b) Layer-wise AUC (logistic regression): reasoning correctness information increases monotonically through the transformer's depth, with all layers outperforming token max-probability (dashed line).

demonstrated across diverse architectures including GPT (Meng et al., 2022) and Llama (Turner et al., 2024; Li et al., 2023).

Second, the three evaluated models span substantially different training procedures: DEEPSEEK-R1-1.5B and DEEPSEEK-R1-7B are obtained via knowledge distillation from DeepSeek-R1 (Guo et al., 2025), while QWQ-32B is trained independently through reinforcement learning (Team, 2025). The consistency of SHEAR's improvements across these distinct training regimes provides evidence that the method captures properties of the transformer computation itself rather than artifacts of a specific training pipeline.

Third, the probe and editing direction trained on MATH transfer to GPQA-Diamond and LiveCodeBench with zero additional adaptation, as Table 4 shows. This cross-domain generalization further supports the generality of the underlying signal.

N DISCUSSION AND LIMITATIONS

Answering the title question. *Do LLMs already know the answer before they finish thinking?* Our evidence indicates yes. One might object that "a probe can decode it" does not entail "the model knows it." The reverse editing experiment in §5.5 addresses this directly: editing along $+v$ improves reasoning; flipping the sign collapses accuracy to near-zero. This extreme directional sensitivity demonstrates that reasoning correctness information is not a passive artifact in the hidden states but is functionally entangled with the model's generation dynamics, directly influencing downstream computation.

Why, then, do LRMs still overthink? We hypothesize a *utilization bottleneck*: the information is encoded and causally relevant, but standard autoregressive decoding lacks a mechanism to "stop

Table 17: Stability index distribution across model scales. "Anchor Pairs" counts the total number of $(h_{s_i^*}, h_s)$ pairs with $s > s_i^*$ used to compute the editing direction v via Eq. 3.

Model	Layer	Anchor Pairs	s^* Mean	s^* Median
DEEPSEEK-R1-1.5B	19	2,618	43.1	29.5
DEEPSEEK-R1-7B	22	2,635	42.9	27.0
QWQ-32B	58	1,865	61.7	37.5

when correct" or "redirect when stuck." Independent probing work reaches a compatible conclusion, reporting that reasoning models encode answer correctness yet do not exploit this internal knowledge during inference (Zhang et al., 2025). SHEAR bypasses this bottleneck through external reading (semantic confidence probe) and direct editing (stability-anchored editing direction).

Broader implications. The signal accessibility gap may extend beyond reasoning control to other tasks requiring content quality assessment, such as factuality checking or code correctness, though further investigation is needed.

Limitations. (i) Generalization to architectures beyond the evaluated models has not yet been tested, though the consistency across three models with distinct training regimes is encouraging. (ii) The mean relative perturbation $\|\alpha_s v\|/\|h_s\| \approx 22.8\%$ is non-negligible. Probe-editing orthogonality, with $\cos(\nabla f, v) \approx -0.02$, ensures the probe remains calibrated, and case studies confirm no degradation in reasoning coherence. (iii) Probe training uses 500 MATH problems with GT labels, though 20 suffice as shown in Appendix I, and the direction transfers across domains without retraining. (iv) Probe reliability under severe distribution shift remains uncharacterized; behavior on non-CoT or non-English tasks is an open question.