# SELECTTTRL: REPRESENTATION-GUIDED DATA SELECTION FOR EFFICIENT TEST-TIME REINFORCEMENT LEARNING

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Test-Time Reinforcement Learning (TTRL) enables large language models to adapt directly on unlabeled test problems by constructing pseudo-labels from multiple sampled responses. However, existing TTRL methods typically apply the same expensive sampling and optimization procedure to all test problems, despite substantial differences in their training utility. In this paper, we study whether such problems can be identified before costly test-time training, without access to ground-truth answers. We uncover a stable relationship between internal response representations and problem-level empirical accuracy. Specifically, we define COS as the cosine similarity between a model's response representation and a correctness-related direction constructed from an independent question-answering dataset. COS consistently correlates with problem solvability across different models, datasets, and experimental settings, and we further provide a theoretical analysis explaining this relationship. Building on this finding, we propose SelectTTRL, a representation-guided data selection method that ranks test problems using COS and applies TTRL only to a selected percentile interval. SelectTTRL leaves the standard TTRL pipeline unchanged for the selected problems, while avoiding unnecessary multi-response sampling and reinforcement learning updates on the remaining ones. Experiments across multiple reasoning benchmarks demonstrate that SelectTTRL substantially reduces test-time computation while maintaining or improving the performance of standard TTRL. Code is available at `https://anonymous.4open.science/r/SelectTTRL-7310/`.

## 1 INTRODUCTION

Large language models (LLMs) have demonstrated strong capabilities on complex reasoning tasks (Qwen et al., 2024; Yang et al., 2025), yet further improving their reasoning ability often relies on additional supervision signals (Shao et al., 2024; Guo et al., 2025). Test-Time Reinforcement Learning (TTRL) (Zuo et al., 2025) provides a new paradigm for adapting models directly on unlabeled test problems. Specifically, TTRL samples multiple responses for each test problem, derives pseudo-labels through majority voting, and constructs reward signals for reinforcement learning without access to the ground-truth answers. In this way, TTRL enables models to further improve their reasoning ability by exploiting the test data themselves (Hosseini et al., 2026; Yan et al., 2026).

However, TTRL typically incurs substantial test-time computational overhead (Liu et al., 2025). Specifically, to obtain reliable training signals, each problem requires multiple inference samples followed by reinforcement learning updates. When the test set is large, repeatedly sampling and training on every problem introduces substantial additional cost. More importantly, existing TTRL methods typically allocate similar computation across problems, overlooking differences in their value for test-time training. Not every problem is equally worth optimizing. For problems the model can already solve reliably, additional sampling and updates may provide only marginal gains. In contrast, for difficult problems, repeated sampling may produce highly consistent yet incorrect answers, causing erroneous pseudo-labels to be further reinforced during training (Yu et al., 2026; Lin et al., 2026). Therefore, uniformly applying the full TTRL procedure to all test problems may waste computation and allocate the limited test-time budget to problems with low training utility. This motivates a natural question:



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.

*Can we identify which test problems are worth training on before performing expensive multi-response sampling and reinforcement learning, without access to ground-truth answers?*

To answer this question, we investigate whether internal model representations can provide a prior signal of test-problem solvability (Zou et al., 2023; Azaria & Mitchell, 2023). Specifically, we construct a correctness-related direction from an independent question-answering dataset (Rimsky et al., 2024). For each test problem, we then compute the cosine similarity between this direction and the model's internal response representation. We denote this similarity as **COS**. We observe a stable relationship between COS and problem-level empirical accuracy. Problems with higher COS generally achieve higher overall accuracy, while those with lower COS are solved less reliably. This relationship holds consistently across different models, datasets, and experimental settings. We further provide a theoretical analysis to explain why this correlation emerges. Our analysis suggests that internal representations contain information related to the relative solvability of test problems. Therefore, COS provides a representation-level signal for estimating and ranking test problems without access to their ground-truth answers.

Building on this finding, we propose **SelectTTRL**, a representation-guided data selection method for efficient test-time reinforcement learning. Before expensive multi-response sampling, SelectTTRL ranks test problems according to COS. It then selects only problems within a prescribed percentile interval for subsequent training. For the selected problems, the standard TTRL pipeline remains unchanged, including multi-response sampling, pseudo-label construction, and policy optimization. In this way, SelectTTRL avoids performing costly sampling and reinforcement learning updates over the entire test set. Our main contributions are summarized as follows:

- In this paper, we uncover a stable relationship between internal response representations and problem-level empirical accuracy, and formalize it as **COS**, a representation-level signal for test-problem solvability. Besides, we provide a theoretical analysis explaining why COS reflects the relative solvability of test problems.
- Building on COS, we propose SelectTTRL. It ranks and selects test problems before the expensive multi-response sampling stage of standard TTRL, and performs subsequent reinforcement learning only on the selected subset. This reduces test-time training cost without modifying the underlying TTRL optimization pipeline.
- Experiments across multiple reasoning benchmarks show that SelectTTRL substantially reduces test-time computation while maintaining or even improving the performance of standard TTRL.

## 2 Related Work

**Test-time reinforcement learning.** Test-time reinforcement learning (TTRL) adapts language models to unlabeled test problems by constructing pseudo-supervision from their own sampled responses, typically through majority-voted answers (Zuo et al., 2025). Existing work mainly improves TTRL from two perspectives. The first focuses on improving the learning signal within each problem. Entropy-guided sampling and contextual guidance regulate exploration during rollout generation (Liu et al., 2025; Hosseini et al., 2026), while selective labeling and confidence- or consistency-based consensus aim to mitigate noisy pseudo-supervision (Yan et al., 2026; Wang et al., 2026b; Yu et al., 2026; Wang et al., 2026a). These methods improve how sampled responses are generated or converted into training signals, rather than selecting a fixed subset of test problems before adaptation. A second line of work adaptively controls whether or how a problem contributes to test-time learning. TTRL-Guard adjusts or suspends updates according to consensus dynamics (Lin et al., 2026), while DiSCTT routes problems to pseudo-label SFT or RL based on response agreement (Moradi & Mudur, 2026). However, such decisions are made only after multiple responses have already been sampled, meaning that much of the test-time computation has already been incurred. Moreover, response agreement does not necessarily indicate that a problem is informative for subsequent optimization. In contrast, we consider *pre-selection*: before multi-response sampling and policy updates, we rank the full set of test problems using representations obtained from an initial response, and apply standard TTRL only to the selected subset. This separates the question of *which problems to optimize on* from modifications to the TTRL learning rule itself.

**Data selection for LLM training.** Data selection reduces post-training cost by identifying compact subsets that preserve or improve learning effectiveness (Albalak et al., 2024). A broad family of

methods scores examples according to properties such as quality, difficulty, uncertainty, and diversity (Liu et al., 2024; Li et al., 2024; Zhao et al., 2025a). Other approaches estimate sample utility through gradients or influence functions. For example, LESS selects instruction-tuning examples according to their gradient influence on target tasks, but requires warm-up training and per-example gradient computation (Xia et al., 2024). Related methods similarly exploit gradient alignment or influence estimates to identify useful training subsets (Zhang et al., 2025; Li et al., 2026). More recent work explores selection signals that can be obtained without additional trial training. AIR estimates example utility through loss changes induced by retrieval-head masking (Liu et al., 2026), while SHIFT combines hidden-state dynamics from a single rollout with diversity-aware selection (Wu et al., 2026). These methods substantially reduce the overhead of estimating data utility. However, their downstream optimization still relies on externally available supervision, such as supplied reasoning traces or verifiable rewards. Our setting is different: in TTRL, neither selection nor subsequent learning has access to ground-truth correctness. We therefore study whether an internal representation signal can identify useful test problems *before* pseudo-label construction, under fully unlabeled test-time learning.

**Representation engineering.** Representation engineering studies how semantic or behavioral properties are encoded in model activations and how such representations can be read or manipulated (Zou et al., 2023). Prior work has shown that hidden states contain predictive signals associated with properties such as truthfulness and correctness (Azaria & Mitchell, 2023), while contrastive activation addition constructs semantic directions for controlling model behavior (Rimsky et al., 2024). Related activation-based methods intervene on internal representations to improve reasoning or elicit different reasoning behaviors (Højer et al., 2025; Tang et al., 2025; Li et al., 2025). Sparse-autoencoder-based approaches further identify and manipulate interpretable features associated with knowledge use, semantic attributes, and reasoning strategies (Zhao et al., 2025b; He et al., 2025; Fang et al., 2026). Our use of representations is complementary to these intervention-oriented approaches. Rather than modifying activations during generation, we use a separately calibrated correctness-related direction as a semantic reference for measuring the representation of an initial response. The resulting cosine-similarity signal, COS, is then used to rank unlabeled test problems before TTRL begins. This extends representation-based analysis from predicting or steering model behavior to allocating test-time optimization effort across problems.

## 3 Framework of SelectTTRL

Given an unlabeled target set $\mathcal{Q} = \{q_i\}_{i=1}^{M}$ and an initial policy $\pi_{\theta_0}$, our goal is to identify which target questions are most suitable for test-time reinforcement learning before expensive multi-rollout sampling and policy optimization are performed. The key challenge is that ground-truth correctness is unavailable on the target set. We therefore seek a lightweight internal signal that can be computed from an initial model response and used to estimate the relative training value of different questions.

SelectTTRL addresses this problem with a representation-based score, which we call **COS**. COS measures how strongly the representation of a target response aligns with a separately calibrated correctness-related direction. We first construct this direction from a small labeled calibration set and then keep it fixed throughout target-set selection. For each target question, we generate only an initial response, compute its COS score, and rank questions accordingly. We retain a prescribed percentile interval and apply standard TTRL reward construction and policy optimization only to the selected subset. Because selection is performed before TTRL rollout collection, excluded questions incur neither multi-response sampling nor policy updates.

### 3.1 Extracting a correctness-related direction

To obtain a semantic reference for correctness, we construct a correctness-related direction using the mean-difference procedure of CAA (Rimsky et al., 2024). Let $\mathcal{D}_{\mathrm{cal}} = (x_i, y_i^+, y_i^-)_{i=1}^{n}$ denote a separate labeled calibration set derived from TruthfulQA (Lin et al., 2022), where $y_i^+$ and $y_i^-$ are a correct and an incorrect answer to question $x_i$, respectively. For each calibration example, we place the two candidate answers in randomized option positions within the same prompt. Keeping the prompt and option order fixed, we separately teacher-force the answer letter corresponding to the correct and incorrect candidates. We then extract the layer-$\ell$ residual-stream activations at the answer-letter position, denoted by $\mathbf{h}_{\ell,i}^{+}$ and $\mathbf{h}_{\ell,i}^{-}$. The correctness-related direction is defined as the

normalized mean difference between these activations:

$$\mathbf{d}_\ell = \frac{1}{n}\sum_{i=1}^{n}\big(\mathbf{h}^+_{\ell,i} - \mathbf{h}^-_{\ell,i}\big), \qquad \mathbf{v}_\ell = \frac{\mathbf{d}_\ell}{|\mathbf{d}_\ell|_2}. \tag{1}$$

Intuitively, $\mathbf{v}_\ell$ captures a representation axis along which the model internally distinguishes more correct-like from less correct-like answers. Prior work has shown that language models encode confidence- and correctness-related information in their internal representations during generation (Kumaran et al., 2026). This motivates using $\mathbf{v}_\ell$ as a fixed reference direction for reading correctness-related evidence from responses to unseen target questions.

Importantly, the labeled calibration set is used only once to estimate $\mathbf{v}_\ell$. The direction is then frozen and no target labels are used during question selection or subsequent TTRL adaptation. In the next section, we use this direction to define COS for each target response and rank the target questions before rollout-intensive TTRL begins.

### 3.2 Question-level COS scoring

For a response $y$ generated by $\pi_{\theta_0}$ for question $q$, let $\mathbf{h}_\ell(q, y)$ denote the layer-$\ell$ residual-stream activation at the fixed answer position. Its COS score is

$$X(q, y) = \frac{\mathbf{h}_\ell(q, y)^\top \mathbf{v}_\ell}{\|\mathbf{h}_\ell(q, y)\|_2 \, \|\mathbf{v}_\ell\|_2}. \tag{2}$$

We obtain a question-level score either from a single greedy response $y^{\mathrm{g}}_q$ or by averaging COS over $K$ sampled responses under a fixed decoding protocol:

$$c_{\mathrm{g}}(q) = X(q, y^{\mathrm{g}}_q), \qquad c_K(q) = \frac{1}{K}\sum_{j=1}^{K} X(q, y_{qj}). \tag{3}$$

Both scoring variants are positively correlated with question-level accuracy in our experiments.

### 3.3 Selecting questions for TTRL

COS provides an ordering of target questions according to the model's internal correctness-related signal. However, our goal is not simply to select questions with the highest or lowest COS. Questions that the model already solves reliably may provide limited room for improvement, whereas questions at the opposite extreme are more likely to yield unreliable pseudo-labels during TTRL. We therefore focus the optimization budget on an intermediate region of the COS ranking, where sufficient learning opportunity and pseudo-label reliability are more likely to coexist.

Let $q_{(1)}, \ldots, q_{(M)}$ denote the target questions sorted in ascending order of $c(q)$, with ties broken by a fixed question identifier. For prescribed percentile endpoints $0 \le a < b \le 1$, we select

$$\mathcal{S}_{a,b} = \big\{q_{(i)} : \lfloor aM \rfloor < i \le \lfloor bM \rfloor\big\}, \qquad \rho = \frac{|\mathcal{S}_{a,b}|}{M}. \tag{4}$$

The percentile interval is fixed before target adaptation and is selected without access to target ground-truth answers. Only questions in $\mathcal{S}_{a,b}$ undergo subsequent TTRL rollout collection and policy updates, whereas all questions in $\mathcal{Q}$ remain in the evaluation set. Thus, excluded questions incur only the cost of the initial response used for COS scoring.

### 3.4 TTRL on the selected subset

After obtaining the selected subset $\mathcal{S}_{a,b}$, we apply the standard TTRL procedure ($Zuo et al.$, 2025) without modifying its reward construction or policy optimization. For each selected question $q$, the rollout policy $\pi_{\theta_{\mathrm{old}}}$ generates a vote pool $\{y_{q,j}\}_{j=1}^{N_{\mathrm{vote}}}$. Let $f(\cdot)$ denote the function that extracts and normalizes the final answer. The majority-voted answer is used as the pseudo-label, and each rollout receives a binary reward according to its agreement with this pseudo-label:

$$\hat{a}_q = \arg\max_a \sum_{j=1}^{N_{\mathrm{vote}}} \mathbf{1}\{f(y_{q,j}) = a\}, \qquad r_q(y) = \mathbf{1}\{f(y) = \hat{a}_q\}. \tag{5}$$

Following the standard vote-then-sample protocol, the pseudo-label is computed from the full vote pool, after which $G \leq N_{\text{vote}}$ responses are sampled for optimization. Their rewards are converted into group-relative advantages using GRPO (Shao et al., 2024), where $\hat{A}_{q,i} = \frac{r_{q,i} - \bar{r}_q}{s_q + \epsilon_{\text{num}}}$, $\bar{r}_q = \frac{1}{G}\sum_{i=1}^{G} r_{q,i}$, and $s_q$ is the reward standard deviation within the update group. Let $\ell_{\text{TTRL}}(q;\theta)$ denote the standard TTRL policy loss for question $q$. SelectTTRL optimizes this loss only over the selected subset:

$$\mathcal{L}_{\text{SelectTTRL}}(\theta) = \mathbb{E}_{q\sim\text{Unif}(\mathcal{S}_{a,b})}[\,\ell_{\text{TTRL}}(q;\theta)\,]. \tag{6}$$

Thus, SelectTTRL leaves the underlying TTRL learning rule unchanged and only determines which target questions receive rollout and update budgets.

# 4 Theoretical Analysis of COS

Our goal is to explain why COS can transfer across tasks. We assume that model representations share a correctness-related component across tasks, which COS reads through a contrastively extracted direction. This direction can capture multiple latent features, further analyzed in RQ3. The model, layer, and readout protocol are fixed. Let $U = \mathbf{h}_\ell/\|\mathbf{h}_\ell\|_2$, $v = \mathbf{v}_\ell$, and $X = v^\top U$.

## 4.1 Shared evidence geometry

**Assumption 4.1 (Spherical evidence model)** *For each target question $q$ in domain $D$, the normalized response representation follows $U \mid q \sim p_D(\cdot \mid t_q)$, where*

$$p_D(U \mid t) = b_D(U)\exp\{t\,u^\top U - A_D(tu)\}, \qquad \|u\|_2 = 1. \tag{7}$$

*Here $u$ is a shared evidence direction, $t_q$ is a question's evidence state, $b_D$ is a domain-specific background density on the unit sphere, and $A_D$ is the log-normalizer.*

Increasing $t$ favors representations with larger evidence projection $T = u^\top U$, while retaining the background geometry. The uniform-background case reduces to a von Mises–Fisher distribution; a general $b_D$ permits anisotropy. This model concerns evidence states rather than two classes with equal concentration. Let $m_D(t) = \mathbb{E}_t[U]$ and $\Sigma_D(t) = \text{Cov}_t(U)$. The exponential-family identities give (Wainwright & Jordan, 2008)

$$\frac{dm_D(t)}{dt} = \Sigma_D(t)u, \qquad \frac{d\,\mathbb{E}_t[X]}{dt} = v^\top\Sigma_D(t)u. \tag{8}$$

Thus, the response of COS depends on the extracted direction and the target representation geometry. The key transfer question is when source extraction yields $v^\top\Sigma_D(t)u > 0$ on the target domain.

## 4.2 Contrastive extraction and positive transfer

Eq. (1) maximizes the average signed projection of source contrasts. Under the controlled source construction in Appendix A.2, the population raw-activation contrast is:

$$d_S^\star = M_S u + e_S, \qquad M_S = \bar{R}_S \int_{t_-}^{t_+} \Sigma_S(r)\,dr \succeq 0, \tag{9}$$

where $t_+ > t_-$, $\bar{R}_S > 0$, and $e_S$ captures radial and unmatched source effects. Let $\hat{d}_S$ be the empirical contrast and $v = \hat{d}_S/\|\hat{d}_S\|_2$. We consider three hypotheses for how this direction reads shared evidence.

**$\mathsf{H}_1$: Uniform source gain.** Source accuracy $a_S$ changes only the overall contrast strength:

$$\hat{d}_S = \gamma(a_S)(d_0 + \xi_S), \qquad \gamma(a_S) > 0, \quad \xi_S \perp\!\!\!\perp a_S. \tag{10}$$

Here $d_0$ is fixed. Normalization removes $\gamma$, predicting source-accuracy-independent readout quality on fixed target representations. The overall positive trend in Table 4 argues against this prediction.

**$\mathsf{H}_2$: Single-factor linear response.** A single factor linearly mediates changes in both COS and correctness. Across feature interventions $j$, this predicts

$$\Delta P_j = b\,\Delta C_j, \qquad b > 0, \tag{11}$$

where $C$ and $P$ denote mean COS and accuracy. Table 3 does not follow this proportionality: ablations 31368 and 48965 produce the same accuracy decrease but markedly different COS changes.

$\mathsf{H}_3$**: Multicomponent positive compatibility.** Source reliability affects the relative strength of shared evidence and source-specific residuals. Multiple components may contribute to the readout; positive transfer requires their combined response to remain positive. We formalize this hypothesis for the actual extracted contrast.

**Condition 4.2 (Positive compatibility)** *Write $\hat{d}_S = \alpha_S u + r_S$, where $r_S \perp u$ and $\alpha_S > 0$. On an interval containing the question states, let $\beta_D(t) = u^\top \Sigma_D(t) u > 0$. For some $\delta \in (0, 1]$, assume*

$$r_S^\top \Sigma_D(t) u \geq -(1-\delta)\alpha_S \beta_D(t). \tag{12}$$

This condition allows rotated directions and nonzero residuals: their adverse response must not cancel the shared-evidence contribution. Tables 3 and 4 motivate this multicomponent account. Stronger evidence can improve directional stability by reducing relative extraction error (Appendix A.3).

### 4.3 From evidence to question accuracy

**Lemma 4.3 (Positive response margin)** *Under Assumption 4.1 and Condition 4.2,*

$$F_D'(t) = v^\top \Sigma_D(t) u \geq \frac{\delta \alpha_S \beta_D(t)}{\|\hat{d}_S\|_2} > 0, \qquad F_D(t) = \mathbb{E}_t[X]. \tag{13}$$

**Condition 4.4 (Reliability link)** *For $Y \in \{0, 1\}$ denoting objective correctness, assume $P(Y = 1 \mid T, Q = q) = g_D(T)$ for each question, with $g_D$ strictly increasing on the support of $T$.*

Define $G_D(t) = \mathbb{E}_t[g_D(T)]$. The response identity yields

$$G_D'(t) = \mathrm{Cov}_t(g_D(T), T) > 0. \tag{14}$$

Thus both expected COS and objective accuracy increase with evidence.

**Theorem 4.5 (Question-level correctness transfer)** *Under Assumption 4.1 and Conditions 4.2 and 4.4, let the question state $t_Q$ be non-degenerate. For $K$ responses sampled under the scoring protocol, and an additional response with correctness $Y'$ independent of them conditional on $Q$,*

$$\mathrm{Cov}(c_K(Q), Y') = \mathrm{Cov}_Q\big(F_D(t_Q), G_D(t_Q)\big) > 0, \quad I(c_K(Q); Y') > 0. \tag{15}$$

*The same positive covariance holds with $Y'$ replaced by $p_Q = P(Y' = 1 \mid Q)$.*

Proofs are in Appendix A.

*Remark.* Under these conditions, average COS predicts an independent answer's correctness without identifying every incorrect response. RQ2 complements this readout analysis by examining how steering direction and strength affect reasoning and answer completion. Different domains need not share a COS threshold. Averaging reduces sampling variation, while departures from the ideal model can produce local ranking reversals (Appendix A.6).

## 5 Experiments

We evaluate SelectTTRL and analyze the COS signal through four research questions: **RQ1:** How does SelectTTRL affect model performance and training efficiency in test-time reinforcement learning? **RQ2:** Can interventions along the extracted direction influence the correctness of model responses? **RQ3:** To what extent is COS explained by known response representations, and is it simply a combination of interpretable features? **RQ4:** What types of source datasets yield directions for which COS positively correlates with question-level accuracy?

### 5.1 Experimental Setup

**Benchmarks.** We evaluate SelectTTRL on six reasoning benchmarks: GPQA-Diamond (Rein et al., 2024), a challenging and high-quality subset of the Graduate-Level Google-Proof Question Answering benchmark; MMLU-Pro STEM, the STEM subset of MMLU-Pro (Wang et al., 2024); and four mathematical reasoning benchmarks, AIME 2024 (30 questions) (AI-MO, 2024a), AIME 2025 (Zhang & Math-AI Team, 2025), AMC (83 questions from 2022–2023) (AI-MO, 2024b), and MATH-500 (Hendrycks et al., 2021).

Table 1: Comparison of TTRL methods on Qwen3-8B across six reasoning benchmarks.

| | MATH-500 | | | GPQA-Diamond | | | MMLU-Pro STEM | | |
|---|---|---|---|---|---|---|---|---|---|
| Method | Greedy | Mean1 | Maj.@16 | Greedy | Mean1 | Maj.@16 | Greedy | Mean1 | Maj.@16 |
| Matched-TTRL | 86.20 | 87.18 | 92.40 | 51.01 | 51.33 | 57.07 | 75.30 | 76.61 | 79.52 |
| Random | 86.00 | 86.71 | 91.20 | 49.49 | 50.95 | 54.55 | 75.70 | 75.41 | 79.32 |
| A-PPL | 87.40 | 86.84 | 91.00 | 53.54 | 50.69 | 56.06 | 75.10 | 76.58 | 81.53 |
| SelectTTRL | 86.20 | 86.55 | 92.60 | 51.52 | 50.73 | 58.08 | 74.10 | 75.77 | 80.12 |
| | AIME 2024 | | | AIME 2025 | | | AMC | | |
| Method | Greedy | Mean1 | Maj.@16 | Greedy | Mean1 | Maj.@16 | Greedy | Mean1 | Maj.@16 |
| Matched-TTRL | 36.67 | 40.42 | 50.00 | 23.33 | 23.54 | 23.33 | 69.88 | 67.85 | 78.31 |
| Random | 33.33 | 36.04 | 46.67 | 16.67 | 17.71 | 16.67 | 65.06 | 63.86 | 73.49 |
| A-PPL | 33.33 | 37.29 | 46.67 | 16.67 | 17.50 | 16.67 | 60.24 | 65.74 | 73.49 |
| SelectTTRL | 43.33 | 38.75 | 50.00 | 23.33 | 19.79 | 20.00 | 66.27 | 66.04 | 74.70 |

Table 2: Comparison of TTRL methods on Qwen2.5-7B-Instruct across six reasoning benchmarks.

| | MATH-500 | | | GPQA-Diamond | | | MMLU-Pro STEM | | |
|---|---|---|---|---|---|---|---|---|---|
| Method | Greedy | Mean1 | Maj.@16 | Greedy | Mean1 | Maj.@16 | Greedy | Mean1 | Maj.@16 |
| Matched-TTRL | 78.80 | 78.34 | 82.00 | 35.35 | 35.04 | 35.86 | 64.66 | 65.31 | 69.28 |
| Random | 79.80 | 77.68 | 80.80 | 36.67 | 36.81 | 40.40 | 64.46 | 63.45 | 67.27 |
| A-PPL | 78.40 | 78.04 | 82.40 | 32.83 | 35.86 | 37.88 | 64.66 | 62.79 | 66.47 |
| SelectTTRL | 79.60 | 78.61 | 83.00 | 36.87 | 35.61 | 38.89 | 62.65 | 64.22 | 68.47 |
| | AIME 2024 | | | AIME 2025 | | | AMC | | |
| Method | Greedy | Mean1 | Maj.@16 | Greedy | Mean1 | Maj.@16 | Greedy | Mean1 | Maj.@16 |
| Matched-TTRL | 23.33 | 21.04 | 23.33 | 16.67 | 16.46 | 16.67 | 51.81 | 50.23 | 53.01 |
| Random | 16.67 | 15.21 | 20.00 | 6.67 | 8.96 | 13.33 | 45.78 | 46.08 | 51.81 |
| A-PPL | 6.67 | 12.92 | 20.00 | 6.67 | 9.17 | 16.67 | 42.17 | 44.95 | 54.22 |
| SelectTTRL | 16.67 | 18.33 | 23.33 | 13.33 | 12.08 | 13.33 | 48.19 | 48.27 | 53.01 |

**Baselines.** We use Qwen3-8B (Yang et al., 2025) and Qwen2.5-7B-Instruct (Qwen et al., 2024) as backbones and compare SelectTTRL with three baselines: (1) **Matched-TTRL** (Zuo et al., 2025), which applies TTRL to the full unlabeled test set; (2) **A-PPL**, which selects questions with the highest perplexity of model-generated answers; and (3) **Random**, which uniformly samples questions without replacement. A-PPL and Random retain 35% of the question pool, matching SelectTTRL. For all subset-selection methods, only the selection rule varies, while the TTRL procedure, training budget, and hyperparameters are kept fixed.

**Hyperparameter Settings.** All experiments use non-thinking mode. For SelectTTRL, we set $(a, b) = (0.25, 0.60)$, selecting the 25th–60th percentile interval of the ascending COS ranking. We train with AdamW and cosine learning-rate decay, with a peak learning rate of $5 \times 10^{-7}$. The maximum context and response lengths are 8,192 and 4,096 tokens, respectively. For each question, we sample 32 responses at temperature $0.6$ for pseudo-label voting and use 16 for policy optimization. Experiments are conducted on 4 NVIDIA A800 80GB GPUs.

**Evaluation Metrics.** We report greedy-decoding accuracy (Greedy), sampled pass@1 (Mean1), and majority-vote accuracy (Maj.@16). Among them, following DeepSeek-R1 (Guo et al., 2025), we evaluate sampled pass@1 using 16 responses per question at temperature $0.6$. Maj.@16 measures the accuracy of the most frequent answer among the same 16 responses.

### 5.2 Overall Performance (RQ1)

SelectTTRL performs TTRL on only 35% of the target questions using pretrained LLMs such as Qwen3-8B and Qwen2.5-7B-Instruct. Tables 1 and 2 compare it with full-data TTRL and alternative selection strategies. Overall, SelectTTRL retains most of the gains of full-data TTRL while using substantially fewer questions for adaptation. On MATH-500 and MMLU-Pro STEM, it performs close to Matched-TTRL despite training on only a subset of the target set. The benefit of selective training is more pronounced on the harder AIME benchmarks. With Qwen3-8B, SelectTTRL outperforms Random and A-PPL in Greedy accuracy on both AIME 2024 and AIME 2025 while remaining competitive with Matched-TTRL. On AIME 2024, it achieves 43.33% Greedy accuracy, compared with 36.67% for Matched-TTRL and 33.33% for both Random and A-PPL. These results suggest that TTRL performance depends not only on the amount of test-time training data, but also on which questions receive the optimization budget.

We attribute this behavior to the quality of COS-based ranking. Questions with intermediate COS values tend to provide a better trade-off between optimization headroom and pseudo-label reliability: very easy questions offer limited additional learning signal, whereas extremely difficult questions are more prone to unreliable consensus rewards. SelectTTRL therefore concentrates TTRL updates on questions that are more likely to provide useful training signals. The trend is strongest for Qwen3-8B, where COS exhibits a clearer association with question-level correctness. For Qwen2.5-7B-Instruct, the ranking signal is somewhat less discriminative, and SelectTTRL performs closer to full-data TTRL. Nevertheless, the method remains competitive across both backbone LLMs, suggesting that its effectiveness does not rely on a single model-specific ranking behavior.

Notably, all results use the correctness direction calibrated from TruthfulQA. Although alternative calibration sources can yield stronger COS–accuracy correlations (Section 5.5), the fixed TruthfulQA direction is sufficient to support effective question selection across models and benchmarks.

### 5.3 Activation Steering and Answer Correctness (RQ2)

To examine whether the direction used by COS is merely correlational or can actively influence model behavior, we perform CAA steering (Rimsky et al., 2024) with Qwen3-8B on MATH-500 and AIME24. We inject the raw CAA vector at layer 21 during generation with $\alpha \in \{-1, -0.5, +0.5, +1\}$, and recompute COS from the generated responses without intervention.

As shown in Figure 1, steering along the extracted direction produces systematic changes in both accuracy and reasoning behavior. On AIME24, moderate positive steering ($\alpha = +0.5$) improves accuracy from 30.00% to 33.33%, outperforming norm-matched random directions, whereas stronger intervention ($\alpha = +1$) degrades performance. Negative steering consistently lowers COS and tends to lengthen responses, while positive steering increases COS on AIME24 and generally shortens generation. On MATH-500, steering is less beneficial overall, with larger intervention magnitudes causing clearer accuracy degradation.

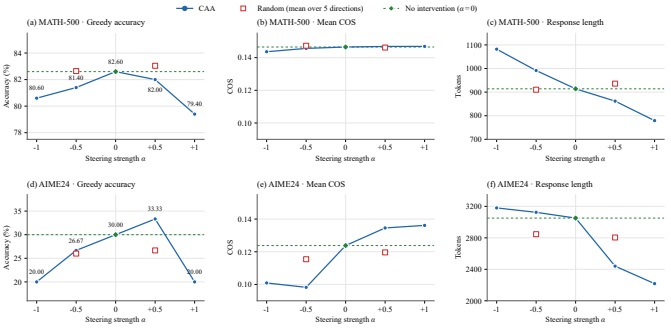


Figure 1: CAA and norm-matched random steering results on Qwen3-8B. Random results are means across five directions.

Response inspection shows that steering affects both calculations and answer completion. For example, on MATH-500, $\alpha = +0.5$ turns a repetitive, truncated response into a correct 578-token solution, whereas $\alpha = -0.5$ causes a previously correct solution to repeat an unfinished step until truncation. Steering can also alter intermediate calculations, such as selecting an incorrect triangle vertex. These cases suggest that the COS direction is behaviorally relevant to correctness, although its effect depends on intervention strength and task.

### 5.4 Interpreting COS through Sparse Features (RQ3)

To better understand what COS captures internally, we analyze its sparse-feature composition and test whether high-contribution features are behaviorally relevant to answer correctness. Following Marks et al. (2025), we decompose the residual-stream activation as:

$$\mathbf{h} = \underbrace{\sum_j z_j(\mathbf{h})\mathbf{d}_j}_{\text{SAE feature components}} + \underbrace{\boldsymbol{\epsilon}(\mathbf{h})}_{\text{reconstruction residual}} + \underbrace{\mathbf{b}_{\text{dec}}}_{\text{decoder bias}}, \tag{16}$$

where $z_j(\mathbf{h})$ and $\mathbf{d}_j$ denote the activation and decoder direction of feature $j$, and $\boldsymbol{\epsilon}(\mathbf{h}) = \mathbf{h} - \hat{\mathbf{h}}$ is the SAE reconstruction residual. We quantify the contribution of each component to COS. Among the 50 most activated SAE features, we select the five with the largest COS contributions. Their maximally activating tokens are associated with reasoning transitions, intermediate conclusions, and final-answer expressions (Appendix B). We then intervene on each feature using feature clamping (Templeton et al., 2024), setting it to zero at every generated-token position while preserving all other feature activations, decoder bias, and reconstruction residual.

Table 3: SAE feature ablation on AIME24. `Numbers` denote SAE feature IDs.

| Condition | Accuracy (%) | Mean COS | Δ Accuracy (pp) | Δ COS |
|---|---|---|---|---|
| Baseline | 30.00 | 0.1238 | — | — |
| Ablate `31368` | 20.00 | 0.1115 | −10.00 | −0.0123 |
| Ablate `52642` | 26.67 | 0.1239 | −3.33 | +0.0001 |
| Ablate `41030` | 23.33 | 0.1212 | −6.67 | −0.0026 |
| Ablate `3447` | 23.33 | 0.1194 | −6.67 | −0.0045 |
| Ablate `48965` | 20.00 | 0.1220 | −10.00 | −0.0018 |

Table 4: Comparison of direction-extraction datasets. Source Greedy is the model's accuracy on the source dataset. Pearson $r$ and Spearman $\rho$ between COS and question-level accuracy are averaged across MATH-500, AIME24, AIME25, and AMC.

| Direction source | Source Greedy (%) | Mean Pearson $r$ | Mean Spearman $\rho$ |
|---|---|---|---|
| PIQA | 87.65 | 0.6420 | 0.6273 |
| CommonsenseQA | 81.08 | 0.6340 | 0.6116 |
| MMLU-Pro | 72.69 | 0.6260 | 0.6071 |
| LogiQA 2.0 | 74.95 | 0.5942 | 0.5692 |
| GPQA | 44.95 | 0.5385 | 0.5126 |
| TruthfulQA | 71.97 | 0.5328 | 0.4519 |

Table 3 shows that ablating all five selected features reduces accuracy on AIME24, with four interventions also lowering mean COS. Feature `31368` has the strongest effect, reducing accuracy from 30.00% to 20.00% and mean COS from 0.1238 to 0.1115. The decomposition further shows that COS ranking is not explained by sparse features alone. Approximately 57% of the ranking contribution comes from the reconstruction residual, 48% from SAE features, and $-5\%$ from the decoder bias (Appendix B). Thus, COS combines interpretable reasoning- and answer-related features with substantial information not resolved by the current SAE.

### 5.5 Dependence on the Direction-Extraction Dataset (RQ4)

To evaluate COS sensitivity to the source dataset, we extract directions from PIQA (Bisk et al., 2020), CommonsenseQA (Talmor et al., 2019), MMLU-Pro, LogiQA 2.0 (Liu et al., 2023), GPQA, and TruthfulQA using the same CAA procedure, and measure Pearson and Spearman correlations between COS and question-level accuracy on MATH-500, AIME24, AIME25, and AMC. Table 4 shows positive correlations for all six source datasets. PIQA and CommonsenseQA, on which the model achieves 87.65% and 81.08% source accuracy, yield the strongest average Spearman correlations of 0.6273 and 0.6116, respectively. GPQA, with a lower source accuracy of 44.95%, yields a weaker average Spearman correlation of 0.5126.

We interpret this pattern as evidence that COS captures correctness-related internal representations. When the model reliably distinguishes correct from incorrect answers on the source task, the contrastive direction better aligns with objective correctness. On harder source tasks, teacher-forcing a correct answer does not guarantee that the model internally recognizes it as correct, weakening the extracted direction and its correlation with target-task accuracy.

These observations connect to the evidence-readout analysis in Section 4. When the model achieves higher source-task accuracy, the activation contrast between correct and incorrect answers is more likely to reflect differences in internal answer evidence. The resulting CAA direction can read this evidence in target tasks more effectively, improving COS ranking of question-level accuracy. This interpretation motivates choosing a source dataset the model solves with reasonably high accuracy.

## 6 Conclusion

In this paper, we investigated how to improve the efficiency of Test-Time Reinforcement Learning by selecting test problems before training-time rollout collection and reinforcement learning. We identified COS, a representation-level signal that exhibited a stable relationship with problem-level empirical accuracy, and provided a theoretical analysis of this relationship. Based on this finding, we proposed SelectTTRL, which selected a subset of test problems for subsequent test-time training according to COS. Experiments across multiple reasoning benchmarks showed that SelectTTRL substantially reduced test-time computation while maintaining or improving the performance of standard TTRL. Overall, our results demonstrated the potential of internal model representations for more selective and efficient test-time adaptation.

## AI USE STATEMENT

In this work, we used generative AI tools to assist with developing theoretical models, deriving and writing mathematical proofs, obtaining feedback on research methodology and experimental design, and implementing methods. AI agents also executed the experimental code under the authors' direction. Additionally, we used generative AI tools for literature search and manuscript polishing, including improving clarity, language, and organization.

The authors conducted the initial literature search, formulated the initial research idea, and developed the early experimental implementation to evaluate the idea. They subsequently refined the idea and extended the literature review. The authors designed all experiments and independently performed the statistical analyses and interpreted the experimental results. For manuscript preparation, the authors wrote most of the initial draft. AI tools provided a first round of language polishing, followed by detailed revisions by the authors and a second round of AI-assisted language polishing. We take responsibility for the final content of this work, including text, claims, proofs, code, and other artifacts produced with the aid of generative AI.

## LIMITATIONS

Although SelectTTRL improves the efficiency of TTRL while maintaining strong performance, two limitations remain. First, since COS has no universal threshold, we select questions based on a fixed percentile interval. While this strategy works well across the evaluated datasets, it is not adaptive and may be suboptimal when the difficulty distribution changes substantially. Second, although SelectTTRL is designed to avoid low-utility or potentially harmful updates, we have not yet fully quantified the negative effect of training on difficult questions or the extent to which our selection strategy mitigates such updates. Future work could therefore explore adaptive selection criteria and a more systematic analysis of harmful test-time updates.

## REFERENCES

AI-MO. AIMO validation AIME. Hugging Face dataset, 2024a. URL https://huggingface.co/datasets/AI-MO/aimo-validation-aime.

AI-MO. AIMO validation AMC. Hugging Face dataset, 2024b. URL https://huggingface.co/datasets/AI-MO/aimo-validation-amc.

Alon Albalak, Yanai Elazar, Sang Michael Xie, Shayne Longpre, Nathan Lambert, Xinyi Wang, Niklas Muennighoff, Bairu Hou, Liangming Pan, Haewon Jeong, Colin Raffel, Shiyu Chang, Tatsunori Hashimoto, and William Yang Wang. A survey on data selection for language models. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=XfHWcNTSHp.

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. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.findings-emnlp.68. URL https://aclanthology.org/2023.findings-emnlp.68/.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. PIQA: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 7432–7439, 2020. doi: 10.1609/aaai.v34i05.6239. URL https://ojs.aaai.org/index.php/AAAI/article/view/6239.

Yi Fang, Wenjie Wang, Mingfeng Xue, Boyi Deng, Fengli Xu, Dayiheng Liu, and Fuli Feng. Controllable LLM reasoning via sparse autoencoder-based steering. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 21290–21305, 2026. doi: 10.18653/v1/2026.acl-long.974. URL https://aclanthology.org/2026.acl-long.974/.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645(8081):633–638, 2025. doi: 10.1038/s41586-025-09422-z. URL https://www.nature.com/articles/s41586-025-09422-z.

Zirui He, Mingyu Jin, Bo Shen, Ali Payani, Yongfeng Zhang, and Mengnan Du. SAE-SSV: Supervised steering in sparse representation spaces for reliable control of language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 2207–2236, 2025. doi: 10.18653/v1/2025.emnlp-main.112. URL https://aclanthology.org/2025.emnlp-main.112/.

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. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, 2021. URL https://arxiv.org/abs/2103.03874.

Bertram Højer, Oliver Jarvis, and Stefan Heinrich. Improving reasoning performance in large language models via representation engineering. In *The Thirteenth International Conference on Learning Representations*, pp. 44746–44763, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/6e73c39cc428c7d264d9820319f31e79-Abstract-Conference.html.

Peyman Hosseini, Ondrej Bohdal, Taha Ceritli, Ignacio Castro, Matthew Purver, Mete Ozay, and Umberto Michieli. CG-TTRL: Context-guided test-time reinforcement learning for on-device large language models. *Transactions of the Association for Computational Linguistics*, 14: 1883–1898, 2026. doi: 10.1162/tacl.a.783. URL https://aclanthology.org/2026.tacl-1.85/.

Dharshan Kumaran, Arthur Conmy, Federico Barbero, Simon Osindero, Viorica Patraucean, and Petar Veličković. How do LLMs compute verbal confidence? In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://icml.cc/virtual/2026/poster/61163.

Ming Li, Yong Zhang, Zhitao Li, Jiuhai Chen, Lichang Chen, Ning Cheng, Jianzong Wang, Tianyi Zhou, and Jing Xiao. From quantity to quality: Boosting LLM performance with self-guided data selection for instruction tuning. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 7602–7635. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.naacl-long.421. URL https://aclanthology.org/2024.naacl-long.421/.

Shipeng Li, Zhiqin Yang, Shikun Li, Xiaobo Xia, Hengyu Liu, Xinghua Zhang, Gaode Chen, Dong Fang, Ying Tai, and Zhe Peng. LearnAlign: Data selection for LLM reinforcement learning with improved gradient alignment. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 40417–40434, 2026. doi: 10.18653/v1/2026.findings-acl.2009. URL https://aclanthology.org/2026.findings-acl.2009/.

Zihao Li, Xu Wang, Yuzhe Yang, Ziyu Yao, Haoyi Xiong, and Mengnan Du. Feature extraction and steering for enhanced chain-of-thought reasoning in language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 10893–10913, 2025. doi: 10.18653/v1/2025.emnlp-main.552. URL https://aclanthology.org/2025.emnlp-main.552/.

Hongxiang Lin, Zhirui Kuai, Erpeng Xue, and Lei Wang. Detecting and mitigating the correct-answer extinction window in test-time reinforcement learning with majority voting. *arXiv preprint arXiv:2605.19444*, 2026. URL https://arxiv.org/abs/2605.19444.

Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3214–3252, 2022. doi: 10.18653/v1/2022.acl-long.229. URL https://aclanthology.org/2022.acl-long.229/.

Hanmeng Liu, Jian Liu, Leyang Cui, Zhiyang Teng, Nan Duan, Ming Zhou, and Yue Zhang. LogiQA 2.0—an improved dataset for logical reasoning in natural language understanding. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:2947–2962, 2023. doi: 10.1109/TASLP.2023.3293046. URL https://doi.org/10.1109/TASLP.2023.3293046.

Jia Liu, ChangYi He, YingQiao Lin, MingMin Yang, FeiYang Shen, and ShaoGuo Liu. ETTRL: Balancing exploration and exploitation in LLM test-time reinforcement learning via entropy mechanism. *arXiv preprint arXiv:2508.11356*, 2025. URL https://arxiv.org/abs/2508.11356.

Jinrui Liu, Kai Hua, Xuanguang Pan, Ge Zhang, Yong Wang, Shuai Ma, and Chongyang Tao. AIR: Post-training data selection for reasoning via attention head influence. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://icml.cc/virtual/2026/poster/62115.

Wei Liu, Weihao Zeng, Keqing He, Yong Jiang, and Junxian He. What makes good data for alignment? a comprehensive study of automatic data selection in instruction tuning. In *The Twelfth International Conference on Learning Representations*, pp. 22353–22373, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/6091f2bb355e960600f62566ac0e2862-Abstract-Conference.html.

Samuel Marks, Can Rager, Eric J. Michaud, Yonatan Belinkov, David Bau, and Aaron Mueller. Sparse feature circuits: Discovering and editing interpretable causal graphs in language models. In *The Thirteenth International Conference on Learning Representations*, pp. 23888–23923, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/3ba4d47a83e498c2b1a0868cba20f6de-Paper-Conference.pdf.

Mohammad Mahdi Moradi and Sudhir Mudur. DiSCTT: Consensus-guided self-curriculum for efficient test-time adaptation in reasoning. *arXiv preprint arXiv:2603.05357*, 2026. URL https://arxiv.org/abs/2603.05357.

Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. doi: 10.48550/arXiv.2412.15115. URL https://arxiv.org/abs/2412.15115.

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. URL https://openreview.net/forum?id=Ti67584b98.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. Steering Llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15504–15522. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.828. URL https://aclanthology.org/2024.acl-long.828/.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024. doi: 10.48550/arXiv.2402.03300. URL https://arxiv.org/abs/2402.03300.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4149–4158, 2019. doi: 10.18653/v1/N19-1421. URL https://aclanthology.org/N19-1421/.

Xinyu Tang, Xiaolei Wang, Zhihao Lv, Yingqian Min, Wayne Xin Zhao, Binbin Hu, Ziqi Liu, and Zhiqiang Zhang. Unlocking general long chain-of-thought reasoning capabilities of large language models via representation engineering. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6832–6849, 2025. doi: 10.18653/v1/2025.acl-long.339. URL https://aclanthology.org/2025.acl-long.339/.

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, Hoagy Cunningham, Nicholas L. Turner, Callum McDougall, Monte MacDiarmid, Alex Tamkin, Esin Durmus, Tristan Hume, Francesco Mosconi, C. Daniel Freeman, Theodore R. Sumers, Edward Rees, Joshua Batson, Adam Jermyn, Shan Carter, Chris Olah, and Tom Henighan. Scaling monosemanticity: Extracting interpretable features from Claude 3 Sonnet. *Transformer Circuits Thread*, 2024. URL https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html.

Martin J. Wainwright and Michael I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1–2):1–305, 2008. doi: 10.1561/2200000001.

Ru Wang, Wei Huang, Qi Cao, Yusuke Iwasawa, Yutaka Matsuo, and Jiaxian Guo. Self-Harmony: Learning to harmonize self-supervision and self-play in test-time reinforcement learning. In *The Fourteenth International Conference on Learning Representations*, pp. 83112–83144, 2026a. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/86178c06fd59ebc58dc31f8edb3c54af-Abstract-Conference.html.

Weiqin Wang, Yile Wang, Kehao Chen, and Hui Huang. Beyond majority voting: Towards fine-grained and more reliable reward signal for test-time reinforcement learning. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 37251–37265. Association for Computational Linguistics, 2026b. doi: 10.18653/v1/2026.acl-long.1727. URL https://aclanthology.org/2026.acl-long.1727/.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. MMLU-Pro: A more robust and challenging multi-task language understanding benchmark. In *Advances in Neural Information Processing Systems*, volume 37, pp. 95266–95290, 2024. doi: 10.52202/079017-3018. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/ad236edc564f3e3156e1b2feafb99a24-Abstract.html.

Jianghao Wu, Jianfei Cai, Weiqiang Wang, Jin Ye, Daniel F. Schmidt, and Yasmeen George. Single-rollout hidden-state dynamics for training-free RLVR data selection. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://icml.cc/virtual/2026/poster/61458.

Mengzhou Xia, Sadhika Malladi, Suchin Gururangan, Sanjeev Arora, and Danqi Chen. LESS: Selecting influential data for targeted instruction tuning. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 54104–54132. PMLR, 2024. URL https://proceedings.mlr.press/v235/xia24c.html.

Dong Yan, Jian Liang, Yanbo Wang, Shuo Lu, Ran He, and Tieniu Tan. What if consensus lies? selective-complementary reinforcement learning at test time. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 28957–28970. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.1337. URL https://aclanthology.org/2026.acl-long.1337/.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. doi: 10.48550/arXiv.2505.09388. URL https://arxiv.org/abs/2505.09388.

Yongcan Yu, Lingxiao He, Jian Liang, Kuangpu Guo, Meng Wang, Qianlong Xie, Xingxing Wang, and Ran He. Understanding and mitigating spurious signal amplification in test-time reinforcement learning for math reasoning. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 37424–37436, 2026. doi: 10.18653/v1/2026.findings-acl.1865. URL https://aclanthology.org/2026.findings-acl.1865/.

Jipeng Zhang, Yaxuan Qin, Renjie Pi, Weizhong Zhang, Rui Pan, and Tong Zhang. TAGCOS: Task-agnostic gradient clustered coreset selection for instruction tuning data. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 4686–4701, 2025. doi: 10.18653/v1/2025.findings-naacl.264. URL https://aclanthology.org/2025.findings-naacl.264/.

Yifan Zhang and Math-AI Team. American invitational mathematics examination (AIME) 2025. Hugging Face dataset, 2025. URL https://huggingface.co/datasets/math-ai/aime25.

Yang Zhao, Kai Xiong, Xiao Ding, Li Du, Yangou Ouyang, Zhouhao Sun, Jiannan Guan, Wenbin Zhang, Bin Liu, Dong Hu, Bing Qin, and Ting Liu. UFO-RL: Uncertainty-focused optimization for efficient reinforcement learning data selection. In *Advances in Neural Information Processing Systems*, volume 38, pp. 577–596, 2025a. URL https://papers.neurips.cc/paper_files/paper/2025/hash/010a19805b0159022a8b29d735bb545f-Abstract-Conference.html.

Yu Zhao, Alessio Devoto, Giwon Hong, Xiaotang Du, Aryo Pradipta Gema, Hongru Wang, Xuanli He, Kam-Fai Wong, and Pasquale Minervini. Steering knowledge selection behaviours in LLMs via SAE-based representation engineering. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 5117–5136, 2025b. doi: 10.18653/v1/2025.naacl-long.264. URL https://aclanthology.org/2025.naacl-long.264/.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, J. Zico Kolter, and Dan Hendrycks. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*, 2023. URL https://arxiv.org/abs/2310.01405.

Yuxin Zuo, Kaiyan Zhang, Li Sheng, Shang Qu, Ganqu Cui, Xuekai Zhu, Haozhan Li, Yuchen Zhang, Xinwei Long, Ermo Hua, Biqing Qi, Youbang Sun, Zhiyuan Ma, Lifan Yuan, Ning Ding, and Bowen Zhou. TTRL: Test-time reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 38, pp. 131459–131483, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/be690ea16f005c174f6c4102a5970e67-Abstract-Conference.html.

# A PROOFS AND EXTENSIONS OF THE COS ANALYSIS

## A.1 SPHERICAL MODEL AND RESPONSE IDENTITY

Fix the model, layer, direction, prompt template, and reading protocol. Let $b_D$ be a probability density with respect to spherical surface measure $\omega$, with support independent of $t$. For finite natural parameters, define

$$A_D(\eta) = \log \int_{\mathbb{S}^{d-1}} b_D(z) \exp(\eta^\top z)\, d\omega(z). \tag{17}$$

Boundedness of $U$ permits differentiation under the integral. Hence

$$\nabla A_D(\eta) = \mathbb{E}_\eta[U], \qquad \nabla^2 A_D(\eta) = \mathrm{Cov}_\eta(U). \tag{18}$$

Applying the chain rule along $\eta = tu$ proves Eq. (8). More generally, for any bounded measurable $f$,

$$\frac{d}{dt}\mathbb{E}_t[f(U)] = \mathrm{Cov}_t(f(U), u^\top U). \tag{19}$$

These are standard exponential-family identities (Wainwright & Jordan, 2008).

The model also arises by minimizing $D_{\mathrm{KL}}(p\|b_D)$ subject to normalization and a specified expectation of $u^\top U$. A Lagrange multiplier for the latter constraint yields the exponential tilt in Eq. (7). With uniform $b_D$, the distribution is vMF with concentration $|t|$ and direction $\mathrm{sign}(t)u$ for $t \neq 0$. If $b_D(U) \propto \exp(a_D^\top U)$, its vMF natural parameter is $a_D + tu$, so both direction and concentration may vary.

## A.2 FROM RAW SOURCE CONTRASTS TO A SPHERICAL RESPONSE

For the empirical raw mean difference $\hat{d}_S \neq 0$, Cauchy–Schwarz gives

$$\frac{\hat{d}_S}{\|\hat{d}_S\|_2} = \arg\max_{\|w\|_2=1} \frac{1}{n}\sum_{i=1}^{n} w^\top (h_i^+ - h_i^-). \tag{20}$$

Thus the extracted direction maximizes the mean signed raw-activation contrast.

For its evidence interpretation, consider the source construction $h^\pm = c + R^\pm V^\pm + \zeta^\pm$, where $c$ is a paired common component, $R^\pm \geq 0$ have finite means, and $V^\pm$ are unit vectors following the source spherical family at $t_+ > t_-$. Here $V^\pm$ describe the varying source component. Set $r_\pm = \mathbb{E}[R^\pm]$, $m_\pm = \mathbb{E}[V^\pm]$, and $\bar{R}_S = (r_+ + r_-)/2 > 0$. Define the vector covariances $k_\pm = \mathbb{E}[(R^\pm - r_\pm)(V^\pm - m_\pm)]$. Then

$$\begin{aligned} d_S^\star &= \mathbb{E}[h^+ - h^-] = \bar{R}_S(m_+ - m_-) + e_S, \\ e_S &= (r_+ - \bar{R}_S)m_+ - (r_- - \bar{R}_S)m_- + k_+ - k_- + \mathbb{E}[\zeta^+ - \zeta^-]. \end{aligned} \tag{21}$$

Integrating Eq. (8) in the source domain proves Eq. (9). Each $\Sigma_S(r)$ is positive semidefinite. Thus $u^\top M_S u > 0$ when $\mathrm{Var}_r(u^\top V) > 0$ on a set of positive integration measure. If $e_S = 0$, this gives $u^\top d_S^\star > 0$. Axis symmetry of $b_S$ about $u$ further gives $d_S^\star = \gamma_S u$ with $\gamma_S > 0$.

Equal radial means, radial independence, and zero mean unmatched contrast give $e_S = 0$. More generally, the transfer analysis incorporates nonzero source bias and finite-sample error through the actual extracted direction.

## A.3 STRUCTURAL HYPOTHESES AND SOURCE RELIABILITY

**Uniform source gain.** Under $\mathsf{H}_1$, for $d_0 + \xi_S \neq 0$,

$$v_S = \frac{d_0 + \xi_S}{\|d_0 + \xi_S\|_2}. \tag{22}$$

Since $\xi_S$ is independent of source accuracy $a_S$, so is $v_S$. On a fixed target representation cache, any readout-quality statistic $\mathcal{R}_D(v_S)$ computed with fixed correctness labels is therefore independent of $a_S$, whenever defined. This includes Pearson and Spearman correlations. The overall trend in Table 4 argues against this independence prediction.

**Single-factor linear response.** One realization of $\mathsf{H}_2$ assumes a common intervention direction $w$, $\mathbb{E}_j[U] - \mathbb{E}_0[U] = \tau_j w$, and a shared affine reliability function $P_j = p_0 + \eta \mathbb{E}_j[T]$, with $\eta > 0$ and values in $[0, 1]$. Then

$$\Delta C_j = \tau_j v^\top w, \qquad \Delta P_j = \eta \tau_j u^\top w. \tag{23}$$

When $v^\top w$ and $u^\top w$ have the same nonzero sign, eliminating $\tau_j$ gives Eq. (11) with $b = \eta(u^\top w)/(v^\top w) > 0$. On the natural evidence path, the analogous closure $\Sigma_D(t)u = \lambda(t)w$, $\lambda(t) > 0$, gives a positive COS response whenever $v^\top w > 0$. Table 3 departs from the predicted proportionality: ablations 31368 and 48965 both reduce accuracy by 10 percentage points, whereas their mean COS reductions are 0.0123 and 0.0018.

**Positive compatibility and proof of Lemma 4.3.** The orthogonal decomposition of the realized contrast is exact: $\alpha_S = u^\top \hat{d}_S$ and $r_S = (I - uu^\top)\hat{d}_S$. Consequently,

$$v^\top \Sigma_D(t) u = \frac{\alpha_S \beta_D(t) + r_S^\top \Sigma_D(t) u}{\|\hat{d}_S\|_2} \geq \frac{\delta \alpha_S \beta_D(t)}{\|\hat{d}_S\|_2} > 0. \tag{24}$$

The inequality is precisely Condition 4.2. Together with Eq. (8), it proves the lemma. If $\Sigma_D(t) = \sum_k \lambda_k w_k w_k^\top$, the numerator before normalization is $\sum_k \lambda_k (\hat{d}_S^\top w_k)(u^\top w_k)$. Here $w_k$ are covariance eigenvectors, and compatibility constrains their aggregate contribution.

**A finite-sample sufficient condition.** Write the population contrast in Eq. (9) as $d_S^\star = s_S u + r_S^\star$, with $r_S^\star \perp u$ and $s_S > 0$. In terms of the source geometry, $s_S = u^\top M_S u + u^\top e_S$ and $r_S^\star = (I - uu^\top)(M_S u + e_S)$. Let $\hat{d}_S = d_S^\star + \epsilon_S$. For $n$ i.i.d. source contrasts with mean $d_S^\star$ and finite covariance $\Omega_S$,

$$\mathbb{E}[\epsilon_S] = 0, \qquad \mathbb{E}\|\epsilon_S\|_2^2 = \frac{\operatorname{tr} \Omega_S}{n}. \tag{25}$$

Define the target anisotropy ratio

$$\chi_* = \sup_t \frac{\|(I - uu^\top)\Sigma_D(t)u\|_2}{\beta_D(t)}. \tag{26}$$

For any $0 < \tau < s_S$, Markov's inequality gives $P(\|\epsilon_S\|_2 \leq \tau) \geq 1 - \operatorname{tr}(\Omega_S)/(n\tau^2)$. On this event, $\alpha_S \geq s_S - \tau$ and $\|r_S\|_2 \leq \|r_S^\star\|_2 + \tau$. Thus the sufficient condition

$$\chi_*\big(\|r_S^\star\|_2 + \tau\big) \leq (1 - \delta)(s_S - \tau) \tag{27}$$

implies positive compatibility with the same probability bound, since $r_S^\top \Sigma_D(t) u \geq -\|r_S\|_2 \chi_* \beta_D(t)$. Holding residual bias, sample size, and noise fixed, stronger evidence $s_S$ increases this margin. This provides a mechanism by which source reliability can improve transfer through smaller relative extraction error. Source accuracy provides an empirical proxy for source reliability; the margin also depends on residual bias and extraction noise.

### A.4 Evidence retained by a projection

**Lemma A.1 (Evidence retained by a projection)** *Under Assumption 4.1, $T = u^\top U$ is sufficient for the evidence-state parameter t. Let $J_U(t)$ and $J_X(t)$ denote the Fisher information about t in $U$ and $X$, respectively. Whenever the denominator is positive,*

$$\frac{J_X(t)}{J_U(t)} \geq \frac{\big(v^\top \Sigma_D(t) u\big)^2}{\big(v^\top \Sigma_D(t) v\big)\big(u^\top \Sigma_D(t) u\big)}. \tag{28}$$

**Proof.** The density factorizes through $T = u^\top U$, so $T$ is sufficient for $t$ within the model. The likelihood score for the full representation is

$$\partial_t \log p_D(U \mid t) = T - \mathbb{E}_t[T], \qquad J_U(t) = \operatorname{Var}_t(T). \tag{29}$$

For the induced distribution of $X = v^\top U$, differentiation of the marginal gives the score $\mathbb{E}_t[T - \mathbb{E}_t T \mid X]$. Therefore

$$J_X(t) = \operatorname{Var}_t\big(\mathbb{E}_t[T \mid X]\big). \tag{30}$$

Conditional expectation and Cauchy–Schwarz imply

$$\begin{aligned}\mathrm{Cov}_t(X,T)^2 &= \mathrm{Cov}_t\big(X,\mathbb{E}_t[T\mid X]\big)^2\\ &\le \mathrm{Var}_t(X)J_X(t).\end{aligned} \tag{31}$$

Substituting the projected covariances proves Eq. (28), quantifying information retained about the evidence-state parameter $t$. Condition 4.2 additionally establishes the positive response direction.

### A.5 Proof of Theorem 4.5

Under the reliability condition, $p_q = G_D(t_q)$, while $\mu_q = \mathbb{E}[X \mid q] = F_D(t_q)$. Lemma 4.3 gives $F_D'(t) > 0$. For independent $T, T'$ from the same state,

$$G_D'(t) = \mathrm{Cov}_t(g_D(T), T) = \tfrac{1}{2}\mathbb{E}_t[(g_D(T) - g_D(T'))(T - T')] > 0. \tag{32}$$

Strict positivity follows from strict increase of $g_D$ and nonzero variance of $T$. For independent question states $t_Q, t_{Q'}$,

$$\begin{aligned}C_0 &:= \mathrm{Cov}_Q(F_D(t_Q), G_D(t_Q))\\ &= \tfrac{1}{2}\mathbb{E}\big[(F_D(t_Q) - F_D(t_{Q'}))(G_D(t_Q) - G_D(t_{Q'}))\big] > 0.\end{aligned} \tag{33}$$

Each scoring response has conditional mean $\mu_q$, so $\mathbb{E}[c_K \mid Q] = \mu_Q$. Conditional independence of $Y'$ and the scoring responses then gives

$$\mathrm{Cov}(c_K, Y') = \mathrm{Cov}_Q(\mu_Q, p_Q) = C_0 = \mathrm{Cov}(c_K, p_Q). \tag{34}$$

Nonzero covariance implies dependence, hence $I(c_K; Y') > 0$. All variables in this covariance calculation are bounded.

### A.6 Angular tolerance and departures from the model

**Angular tolerance.** Write $v = \cos\theta\, u + \sin\theta\, w$ with $w \perp u$ and $\|w\|_2 = 1$. For $0 \le \theta < \pi/2$, define

$$\chi_D(t) = \frac{\|(I - uu^\top)\Sigma_D(t)u\|_2}{u^\top \Sigma_D(t) u}. \tag{35}$$

Projection and Cauchy–Schwarz yield

$$F_D'(t) \ge \big(\cos\theta - \chi_D(t)\sin\theta\big) u^\top \Sigma_D(t) u. \tag{36}$$

Thus $\cos\theta > \sup_t \chi_D(t)\sin\theta$ suffices for positive transfer, allowing angular mismatch under anisotropic target geometry.

**Question heterogeneity.** Let actual question means and success probabilities deviate from the ideal family as $\mu_q = F_D(t_q) + e_q$ and $p_q = G_D(t_q) + r_q$. With $\delta_x = \mathrm{sd}(e_Q)$, $\delta_p = \mathrm{sd}(r_Q)$, $\sigma_F = \mathrm{sd}(F_D(t_Q))$, and $\sigma_G = \mathrm{sd}(G_D(t_Q))$, expansion gives

$$\mathrm{Cov}_Q(\mu_Q, p_Q) \ge C_0 - \delta_x\sigma_G - \delta_p\sigma_F - \delta_x\delta_p. \tag{37}$$

Positive association survives whenever the right-hand side is positive, even if individual question pairs reverse order. This lower bound allows arbitrary signs of the heterogeneity cross terms.

**Averaging.** For conditionally independent, identically distributed scoring responses, write $C = \mathrm{Cov}_Q(\mu_Q, p_Q)$ and $\sigma_q^2 = \mathrm{Var}(X \mid q)$. Then

$$\mathrm{Corr}(c_K, p_Q) = \frac{C}{\sqrt{[\mathrm{Var}_Q(\mu_Q) + \mathbb{E}_Q\sigma_Q^2/K]\,\mathrm{Var}_Q(p_Q)}}. \tag{38}$$

For $C > 0$, averaging increases this population correlation by reducing sampling variance.

**Greedy scores.** For the greedy score, define $d_q = c_{\mathrm{g}}(q) - \mu_q$,

$$\mathrm{Cov}_Q(c_{\mathrm{g}}(Q), p_Q) = C + \mathrm{Cov}_Q(d_Q, p_Q). \tag{39}$$

Its association remains positive if $\mathrm{Cov}_Q(d_Q, p_Q) > -C$.

Table 5: Class-projection gaps and cross-half COS–accuracy correlations. Confidence intervals for $\widehat{\beta}$ use 1,000 question-level bootstrap resamples, retaining all responses of each sampled question.

| Dataset | $\widehat{\beta}$ | 95% CI | $\rho(c_A, \widehat{p}_B)$ | $\rho(c_B, \widehat{p}_A)$ |
|---|---|---|---|---|
| MATH-500 | 0.02271 | [0.01815, 0.02821] | 0.296 | 0.338 |
| GPQA-Diamond | 0.01602 | [0.01125, 0.02081] | 0.249 | 0.244 |
| MMLU-Pro STEM | 0.02131 | [0.01803, 0.02460] | 0.434 | 0.464 |
| AIME 2024 | 0.05077 | [0.03646, 0.06377] | 0.536 | 0.610 |

Table 6: Cross-half covariance decomposition. All values are multiplied by $10^3$; brackets give 95% bootstrap confidence intervals.

| Dataset | $\widehat{C}_{\text{total}}$ | $\widehat{C}_{\text{proj}}$ | $\widehat{C}_{\text{res}}$ |
|---|---|---|---|
| MATH-500 | 2.158<br>[1.535, 2.834] | 1.752<br>[1.276, 2.267] | 0.407<br>[0.136, 0.673] |
| GPQA-Diamond | 1.884<br>[0.993, 2.842] | 1.879<br>[1.302, 2.425] | 0.005<br>[−0.591, 0.627] |
| MMLU-Pro STEM | 2.909<br>[2.349, 3.491] | 2.471<br>[2.017, 2.927] | 0.438<br>[0.169, 0.708] |
| AIME 2024 | 7.533<br>[2.509, 9.894] | 6.738<br>[2.290, 8.874] | 0.796<br>[−1.364, 2.678] |

### A.7 Empirical support for question-level association

We examine the positive association in Theorem 4.5 and its persistence under question heterogeneity with Qwen3-8B on MATH-500 (500 questions), GPQA-Diamond (198), MMLU-Pro STEM (498), and AIME 2024 (30). We use the layer-21 TruthfulQA direction and 16 sampled responses per question. COS is read at the closing token of the final complete boxed answer, or the response endpoint when no complete box is available. Splitting responses by sample index into two groups of eight gives $(c_A, \widehat{p}_A)$ and $(c_B, \widehat{p}_B)$. Cross-half estimates use disjoint responses for COS and accuracy, matching the theorem's conditional-independence construction.

**Projection gap and cross-response association.** The class-projection gap $\beta = \mathbb{E}[X \mid Y = 1] - \mathbb{E}[X \mid Y = 0]$ is positive on all four datasets, with 95% confidence intervals above zero. Both cross-half Spearman correlations are also positive (Table 5).

**Association under question heterogeneity.** With $\alpha = \mathbb{E}[X \mid Y = 0]$, define $\varepsilon_q = \mu_q - \alpha - \beta p_q$. Then

$$C = \mathrm{Cov}_Q(\mu_Q, p_Q) = \beta \,\mathrm{Var}_Q(p_Q) + \mathrm{Cov}_Q(\varepsilon_Q, p_Q). \tag{40}$$

We estimate $\alpha, \beta$ from class-conditional projection means using five-fold cross-fitting over questions. Within each held-out fold,

$$\begin{aligned} \widehat{C}_{\text{total}} &= \tfrac{1}{2}\big[\widehat{\mathrm{Cov}}(c_A, \widehat{p}_B) + \widehat{\mathrm{Cov}}(c_B, \widehat{p}_A)\big], \\ \widehat{C}_{\text{proj}} &= \widehat{\beta}\,\widehat{\mathrm{Cov}}(\widehat{p}_A, \widehat{p}_B), \qquad \widehat{C}_{\text{res}} = \widehat{C}_{\text{total}} - \widehat{C}_{\text{proj}}. \end{aligned} \tag{41}$$

Sample covariances are averaged with fold-size weights. Confidence intervals use 1,000 within-fold question bootstrap resamples, refitting $\widehat{\alpha}, \widehat{\beta}$ each time.

The total covariance is positive on all four datasets, with confidence intervals above zero, and the projection term accounts for most of it (Table 6). The heterogeneity term is positive on MATH-500 and MMLU-Pro STEM; its intervals include zero on GPQA-Diamond and AIME 2024. Thus positive question-level association persists alongside heterogeneous response representations.

Table 7: Component contributions to the high–low COS difference on AIME24. Each group contains five questions. Values are rounded.

| Component | High COS | Low COS | Difference | Share (%) |
|---|---|---|---|---|
| SAE features | 0.07968 | 0.05503 | +0.02465 | 47.6 |
| Decoder bias | 0.05289 | 0.05538 | −0.00249 | −4.8 |
| Reconstruction residual | 0.04675 | 0.01714 | +0.02962 | 57.2 |
| Total COS | 0.17933 | 0.12755 | +0.05178 | 100.0 |

# B SAE Decomposition and Feature Interventions for COS

## B.1 Exact decomposition

Let $\mathbf{h}$ be the nonzero residual-stream representation used to compute COS, and let $z_j(\mathbf{h})$ and $\mathbf{d}_j$ denote the activation and decoder direction of SAE feature $j$. Following Marks et al. (2025), write

$$\mathbf{h} = \sum_j z_j(\mathbf{h})\mathbf{d}_j + \mathbf{b}_{\text{dec}} + \boldsymbol{\epsilon}(\mathbf{h}), \qquad \boldsymbol{\epsilon}(\mathbf{h}) = \mathbf{h} - \text{SAE}(\mathbf{h}). \tag{42}$$

All terms are expressed in the same activation coordinates. For the unit correctness-related direction $\mathbf{v}_\ell$, define

$$\begin{aligned} c_j(\mathbf{h}) &= \frac{z_j(\mathbf{h})\mathbf{d}_j^\top \mathbf{v}_\ell}{\|\mathbf{h}\|_2}, \\ c_{\text{bias}}(\mathbf{h}) = \frac{\mathbf{b}_{\text{dec}}^\top \mathbf{v}_\ell}{\|\mathbf{h}\|_2}, &\qquad c_{\text{res}}(\mathbf{h}) = \frac{\boldsymbol{\epsilon}(\mathbf{h})^\top \mathbf{v}_\ell}{\|\mathbf{h}\|_2}. \end{aligned} \tag{43}$$

Then $\text{COS}(\mathbf{h}) = \sum_j c_j(\mathbf{h}) + c_{\text{bias}}(\mathbf{h}) + c_{\text{res}}(\mathbf{h})$ exactly. The residual uses the full SAE reconstruction, and all contributions share the norm of the original representation as their denominator.

## B.2 Ranking contribution accounting

We quantify ranking contributions through the COS difference between high- and low-scoring questions on AIME24. After excluding truncated greedy responses, we rank questions by answer-position COS and select the five highest and five lowest as groups $H$ and $L$, respectively. For each question, let $c_{\text{SAE}} = \sum_j c_j$, with $c_{\text{bias}}$ and $c_{\text{res}}$ defined in Eq. (43). For each component $k \in \{\text{SAE}, \text{bias}, \text{res}\}$, define

$$\begin{aligned} \Delta_k &= \frac{1}{5}\sum_{q\in H} c_k(\mathbf{h}_q) - \frac{1}{5}\sum_{q\in L} c_k(\mathbf{h}_q), \\ w_k &= \frac{\Delta_k}{\Delta_{\text{COS}}}, \qquad \Delta_{\text{COS}} = \sum_k \Delta_k. \end{aligned} \tag{44}$$

Here $\mathbf{h}_q$ is the answer-position representation of the greedy response. The signed shares sum to one; a negative share indicates that the component offsets the high–low COS difference.

As shown in Table 7, SAE features account for 47.6% of the COS difference and the reconstruction residual for 57.2%, while the decoder bias offsets 4.8%. Thus, both sparse features and information retained in the reconstruction residual contribute to the observed COS ranking.

## B.3 Feature clamping

We select the five features with the largest absolute differences in mean COS contribution between the five highest- and five lowest-COS questions. Following Templeton et al. (2024), at each generated-token position $t$ we replace the activation of a selected feature $j$ by zero while retaining the other local SAE activations, decoder bias, and original reconstruction residual. The modified activation is

$$\mathbf{h}'_t = \sum_{k\neq j} z_k(\mathbf{h}_t)\mathbf{d}_k + \mathbf{b}_{\text{dec}} + \boldsymbol{\epsilon}(\mathbf{h}_t) = \mathbf{h}_t - z_j(\mathbf{h}_t)\mathbf{d}_j. \tag{45}$$

The model continues its forward pass and autoregressive generation using this modified activation.

Table 8: Examples of high-activation locations for COS-related SAE features in MATH-500 responses.

| Feature ID | Observed text or symbol | Location |
|---|---|---|
| 32531, 47984 | $$ following \boxed{6} or \boxed{3} | Formula-ending delimiter after the final answer |
| 3447 | So; This; exact expression is | Connective or result-statement expression |
| 41030 | 58500; 2000; 4 inches per second | Numerical results and units |
| 31368 | find the sum; how many; ordered quadruple | Expression of the quantity or object to be found |
| 605 | The digit 7 in \boxed{27} | Numerical token within the final answer |

Table 9: Training-efficiency reductions and time savings across six benchmarks.

| | Reduction (%) | | | Time saved | |
|---|---|---|---|---|---|
| Dataset | Generated tokens | Update response tokens | Training-step time | Wall-clock (min) | GPU-hours |
| MATH-500 | 17.07 | 17.14 | 5.43 | 14.23 | 0.949 |
| MMLU-Pro STEM | 19.56 | 18.93 | 6.68 | 18.02 | 1.202 |
| GPQA-Diamond | 2.88 | 3.24 | 0.66 | 1.43 | 0.095 |
| AIME 2024 | 10.74 | 10.73 | 3.73 | 12.96 | 0.864 |
| AIME 2025 | 26.12 | 25.93 | 10.86 | 36.66 | 2.444 |
| AMC | 27.83 | 27.91 | 8.91 | 35.47 | 2.365 |
| **Aggregate** | **18.38** | **18.33** | **6.48** | **118.77** | **7.918** |

### B.4 Feature activation examples

Given the small number of AIME24 questions, we use the larger MATH-500 set to illustrate token-wise SAE activations in original responses. Table 8 presents representative high-activation locations, including goal expressions, result statements, numerical outputs, and formula boundaries.

## C Training Efficiency

Table 9 reports the training-cost reductions of SelectTTRL relative to TTRL with Qwen3-8B across six benchmarks.