LET'S FAIL STEP BY STEP: BETTER CONTRASTIVE DISTILLATION WITH RELIABLE WORSE TEACHERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Contrastive Distillation, which trains a student policy using a reward defined by the log-probability gap between two teacher policies, has recently gained attention as an effective alternative to On-Policy Distillation. However, prior approaches typically contrast a task-specific, RL-trained teacher with a base policy, limiting scalability to larger teacher models and broader task domains. We instead propose to contrast a strong teacher with a negatively steered version of itself, using a natural-language description encouraging undesirable behaviors. This removes the need for task-specific RL training of the teacher while providing a simple and interpretable way to control contrastive rewards. On mathematical reasoning tasks, we show that steering the negative teacher toward incorrect or non-rigorous reasoning substantially improves the contrastive reward's effectiveness for student training, whereas positively steering the teacher using privileged information or behavioral descriptions yields only marginal gains. We explain this asymmetry through the Steering-Success Rate, defined as the probability that a completion from the steered policy is preferred over one from the base policy. We theoretically show that this quantity directly affects the converged student policy's win rate over its KL-regularized reference policy under reasonable conditions. Through a series of ablations, we demonstrate that the final performance of the student policy closely follows the Steering-Success Rate of its corresponding contrastive reward. In particular, negative steering consistently achieves higher success rates than positive steering, and improving the reliability of negative steering further enhances the resulting reward's training effectiveness. Together, our results establish natural-language steering as a simple and intuitive approach to contrastive distillation, while highlighting Steering-Success Rate as a key factor in designing effective contrastive rewards.

1 INTRODUCTION

Distilling the capabilities of large language models into smaller ones remains an important approach for deploying language models in resource-constrained settings (Hinton et al., 2015). On-Policy Distillation (OPD), which minimizes the KL-divergence between the student and teacher on student-generated trajectories, has become a popular choice due to its stable and strong empirical performance (DeepSeek-AI et al., 2025; Yang et al., 2025; Lu & Lab, 2025; Song & Zheng, 2026).

However, because OPD treats the teacher policy as the optimum of the training objective, it provides no explicit mechanism for filtering task-irrelevant discriminative features during distillation. When the two models differ substantially, such spurious signals can dominate training, so that the student learns the stylistic elements of the teacher's reasoning rather than the core capabilities that drive performance, such as exploration and self-revision (Zhao et al., 2026; Armandpour et al., 2026).

Contrastive Distillation aims to address this issue by optimizing a reward defined as the log-probability gap between two teachers, ordered such that the gap aligns with the task's reward structure (Feng et al., 2026; Heo et al., 2026). In this light, the spurious signals shared by both teachers cancel out in the contrast, yielding a reward that focuses on task-relevant features. However, prior approaches typically construct this contrast using a task-specific, RL-trained teacher against its base

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.

policy, which introduces several limitations. First, the resource intensive nature of RL training limits the range of teachers that can be used for contrastive distillation. Second, keeping two additional separate teacher weights during student training incurs substantial overhead (Yang et al., 2026). Finally, due to the black-box nature of RL training (Lambert et al., 2025; Jiang et al., 2025), the contrastive reward derived from the RL-trained policies can be difficult to interpret or adjust.

To address these issues of contrastive distillation, we propose CRISP (**C**ontrastive **R**ewards from **I**nstruction-**S**teered **P**olicies), which contrasts a strong teacher with a *negatively steered* version of itself (Section 2). We steer the teacher by adding a natural-language description that encourages *undesirable* behavior to its prompt (e.g., incorrect or non-rigorous reasoning). The student is then rewarded for completions that are more likely under the original teacher than under its negatively steered version. Because the two policies share the same weights and differ only in their prompts, CRISP requires no additional task-specific teacher training and keeps only a single set of teacher parameters in memory, reducing computational costs. The reward can also be easily adjusted by editing the steering instruction, providing an interpretable way to control the contrastive reward.

In Section 3, we empirically compare CRISP with recent methods in on-policy and contrastive distillation on mathematical reasoning tasks. We find that CRISP substantially outperforms these baselines in both peak student performance and training stability. To better understand this gap, we examine a common design choice in prior methods which typically construct rewards by contrasting a superior policy against a base policy. We compare contrastive rewards derived from positively and negatively steered policies and find that contrasting a positively steered teacher and its original policy yields only marginal gains, even with privileged external information or explicit descriptions of desirable behavior. Moreover, when increasing the size of steered teachers from 14B to 32B, the performance gains of CRISP scale reliably, whereas positive steering does not (Section 3.3).

To explain this asymmetry, we introduce the *Steering-Success Rate* (SSR), the probability that the steered policy generates completions in the intended direction relative to completions generated by the unsteered policy (Definition 4.2, Section 4). We theoretically show that, under comparable conditions, a higher SSR of the steered policy leads to a higher final win rate of the converged student over its KL-regularized reference policy (Section 4.1). Our ablations also confirm that student performance closely tracks the SSR of its contrastive reward (Section 4.2). Positive steering fails to reliably increase the teacher's probability of producing a correct answer, whereas negative steering reliably decreases it, and this gap in SSR matches the gap in downstream gains. In addition, improving the SSR of negative steering further increases the performance of the converged student policy (Section 4.2.2).

To summarize, our contributions are as follows:

- We propose CRISP, which derives rewards by contrasting a strong teacher with a negatively steered version of itself via behavioral descriptions, eliminating task-specific RL training of large teachers.
- On mathematical reasoning, CRISP substantially outperforms existing contrastive distillation methods, while positive steering counterparts yield only marginal gains.
- We show formally that the Steering-Success Rate (SSR) determines the converged student's win rate over its KL-regularized reference. Ablations confirm that distillation performance consistently follows the trend of SSR, which demonstrates the effectiveness of using SSR to help design better contrastive rewards.

2 Contrastive Distillation

2.1 Problem Setup

Let $x \in \mathcal{X}$ denote a prompt and $y \in \mathcal{Y}$ a completion generated auto-regressively; we omit the dependence on x when it is clear from context. We train a student policy π_θ, initialized from a reference policy π_{ref}, on token completions generated by π_θ, using the teacher's predicted log-probabilities as feedback. The quality or preference of y is defined by the ground-truth task reward $r^* : \mathcal{Y} \to \mathbb{R}$, such as answer correctness, which is not observed during training and used only at evaluation time. Our goal is to derive a contrastive reward from the teachers whose maximization improves the student with respect to r^*.

2.2 CONTRASTIVE DISTILLATION

Contrastive distillation rewards the student for completions that a target teacher π^* assigns higher probability to than a reference teacher π^*_{ref}, while keeping the student close to π_{ref}, commonly set to the student's initial policy (Yang et al., 2026; Rafailov et al., 2024):

$$\max_{\pi_\theta} \mathbb{E}_{y\sim\pi_\theta}\left[R(y)\right] - \beta \cdot \mathrm{KL}\left[\pi_\theta \parallel \pi_{\text{ref}}\right], \qquad R(y) = \log \frac{\pi^*(y)}{\pi^*_{\text{ref}}(y)}, \tag{1}$$

where $\beta > 0$ controls the strength of the KL regularization. Because both teachers are autoregressive, R decomposes into a sum of per-token log-ratios, providing token-level dense credit assignment along the student's trajectory. Thus, the student is trained to increase the probability of tokens where π^* assigns higher probability than π^*_{ref}, without diverging too far away from π_{ref}.

Existing methods can be viewed as specific instances of this KL-regularized RL formulation. OPD sets the reference teacher as the student's reference policy, $\pi^*_{\text{ref}} = \pi_{\text{ref}}$, and fixes $\beta = 1$, which reduces the training objective to minimizing $\mathrm{KL}\left[\pi_\theta \parallel \pi^*\right]$ (Lu & Lab, 2025). G-OPD (Yang et al., 2026) instead trains the student policy without restricting $\beta = 1$.[1] In either formulation, however, the contrastive reward reflects all differences between π^* and the student's initial policy π_{ref}, including those unrelated to the task reward r^*.

To address this issue, Direct-OPD (Feng et al., 2026) and OP^{2}D (Heo et al., 2026) instead contrasts an RL-trained teacher with its pre-RL base policy, with OP^{2}D additionally applying advantage restriction. The main intuition is that if the two teachers share the same level of task irrelevant features, then these features would cancel in the log-ratio reward. However, this comes at the additional cost of RL-training the teacher and incurs additional memory overhead from maintaining both teacher models during student training.

2.3 STEERED CONTRASTIVE DISTILLATION

We instead construct both teachers from a single policy π^*_o by steering it with *natural language*. A steering prefix C is a description of an intended behavior that is prepended only to the teacher's context. The student generates $y \sim \pi_\theta(\cdot \mid x)$ from the original prompt, and the teacher scores this same completion both with and without the prefix. We consider two kinds of prefixes: a negative prefix C_- encouraging undesirable behavior, such as producing plausible but logically invalid reasoning, while a positive prefix C_+ encourages desirable behavior or provides privileged external information, such as the reference solution for a math problem (Figure 1). Writing $\pi^*_- \coloneqq \pi^*_o(\cdot \mid C_-)$ and $\pi^*_+ \coloneqq \pi^*_o(\cdot \mid C_+)$, each prefix yields a contrastive reward:

$$R_+(y) = \log \frac{\pi^*_+(y)}{\pi^*_o(y)}, \qquad R_-(y) = \log \frac{\pi^*_o(y)}{\pi^*_-(y)}. \tag{2}$$

CRISP trains the student with R_-, while R_+ serves as a natural alternative which we compare against in Section 3. Unlike prior methods which require task-specific RL training of the teacher (Feng et al., 2026; Heo et al., 2026), CRISP requires no additional teacher training. Furthermore, only a single teacher needs to be kept in memory during student training, and the contrastive reward can be adjusted by simply editing the steering prefix.

Optimization. Following Feng et al. (2026), we train all methods with a Rao-Blackwellized per-token estimator computed over the student's top-k next-token distribution:

$$\nabla_\theta \mathcal{L} = \mathbb{E}_{y\sim\pi_\theta}\left[\sum_t \sum_{v\in V_t} \mathrm{sg}(\bar{p}_t(v)\cdot R(v))\nabla_\theta \log \pi_\theta(v \mid s_t) - \beta \cdot \mathrm{KL}\left[\pi_\theta(\cdot \mid s_t) \parallel \pi_{\text{ref}}(\cdot \mid s_t)\right]\right],$$

where V_t is the set of top-k most likely next tokens under π_θ given the partial completion $s_t = y_{<t}$, and $\bar{p}_t(v) = \frac{\pi_\theta(v|s_t)}{\sum_{u\in V_t}\pi_\theta(u|s_t)}$ is the student's next-token probability renormalized over V_t.

[1] The commonly used name Ex-OPD refers specifically to the setting $0 < \beta < 1$. Since we consider both $0 < \beta < 1$ and $\beta \geq 1$, we use the more general term G-OPD throughout this work.

Default Chat Template (No Steering)

```
Problem: {problem}
Please reason step by step, and put your final answer within \boxed{}.
```

Positive Steering Chat Template (Not-Negative)

```
Problem: {problem}
Provide a correct approach to the problem above using valid and
rigorous reasoning. Present the reasoning straightforwardly and
accurately. Reason step by step, and put your final answer within
\boxed{}.
```

Negative Steering Chat Template (Ours)

```
Problem: {problem}
Provide an incorrect approach to the problem above using invalid
or non-rigorous reasoning. Present the reasoning straightforwardly
without stating or implying that it is intentionally incorrect. Reason
(incorrectly) step by step, and put your final answer within \boxed{}.
```

Figure 1: Overview of different steering prompts used in our experiments. We provide the details of other steering methods in Appendix A.1.

3 Experiments

We evaluate CRISP on mathematical reasoning in two stages. We first compare it against existing on-policy and contrastive distillation methods (Section 3.2). We then isolate the role of steering direction by comparing negative steering with a range of positive steering prompts, both at equal teacher sizes and at the presence of a larger steered teacher (Section 3.3).

3.1 Setup

We use the Qwen3 model family (Yang et al., 2025), with Qwen3-14B as the teacher policy and Qwen3-4B and Qwen3-1.7B as student policies. All models are trained on the mathematics subset of OpenThoughts (Guha et al., 2026). We evaluate on five mathematical reasoning benchmarks: AMC23 (math-ai, 2023), AIME24 (Zhang & Math-AI, 2024), AIME25 (Zhang & Math-AI, 2025), and HMMT25-Feb&Nov (Dekoninck et al., 2026). For all tasks, we report the accuracy averaged over 12 independently sampled completions per problem (Accuracy@12). We provide the full details of training and evaluation in Appendix A.

Baselines. We summarize the teacher pair used by each method in the appendix (Table 2). OPD (Lu & Lab, 2025) and G-OPD (Yang et al., 2026) contrasts Qwen3-14B with the student's initial policy. Direct-OPD (Feng et al., 2026) and OP^{2}D (Heo et al., 2026) contrast Qwen3-14B with its base model, Qwen3-14B-Base. OPSD (Zhao et al., 2026) and RLCSD (Pan et al., 2026) form a contrastive reward using privileged information, where a teacher that is shown the reference answer or solution is contrasted with the same teacher without it. Thu, OPSD uses the initial student policy for the contrastive reward, whereas RLCSD uses the larger teacher model Qwen3-14B. CRISP contrasts Qwen3-14B with a steered version of itself, using the negative prompt in Figure 1.

3.2 Main Results

Figure 2 shows the training curves, and Table 3 reports the detailed benchmark results at the selected best checkpoint. CRISP approximately doubles the performance gains of the strongest baseline. With Qwen3-4B as the student, CRISP raises average accuracy by +17.3%p, about twice the gain of Direct-OPD (+8.4%p). With Qwen3-1.7B, it raises average accuracy by +12.3%p, about twice the gain of OP^{2}D (+6.1%p). **For both students, CRISP achieves the best score on all five benchmarks (Table 3) and exhibits faster and more stable convergence (Figure 2).**

3.3 Negative vs. Positive Steering

The results so far show that negative steering produces an effective reward while privileged information does not. This naturally motivates the question on whether the performance gap stems from the

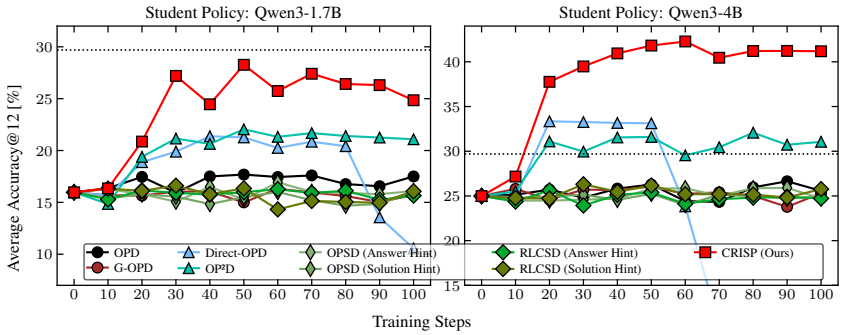

Figure 2: Average accuracy@12 across five mathematical reasoning tasks over the course of training. The dashed line represents the performance of Qwen3-14B. For both student policies, CRISP outperforms all other distillation methods and maintains its performance gains throughout training.

direction of steering or merely from the particular positive signals tested. To this end, we compare CRISP with four positive steering prompts that describe desirable behavior instead of utilizing external privileged information. Each prompt defines a reward R_+ (Eq. 2) that contrasts the positively steered teacher with the unsteered teacher:

1. **Not-Negative Steering** inverts our negative prompt, asking for a correct approach with valid and rigorous reasoning (Figure 1).
2. **Error-Revision** asks the teacher to check intermediate steps for errors and revise them when necessary (Figure 6).
3. **Exploration** asks the teacher to explore multiple strategies and to switch approaches when the current one becomes unproductive (Figure 6).
4. **Exploration & Error-Revision** combines both instructions (Figure 6).

The last three prompts target behaviors such as self-verification and exploration, which are widely regarded as central to strong reasoning (Lee et al., 2025). We provide the full training procedure in Appendix A.

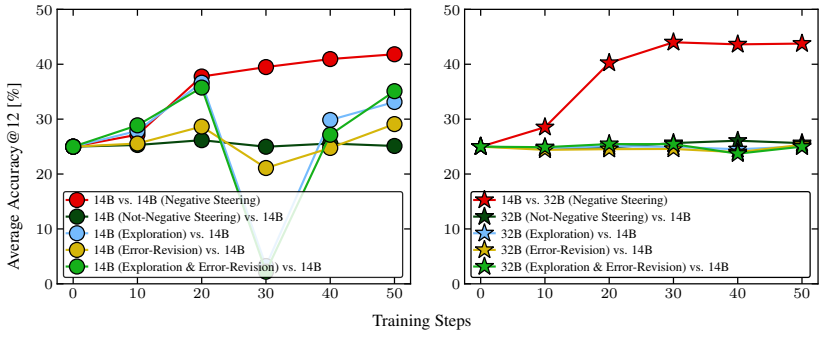

Figure 3: Negative vs. positive natural-language steering results for Qwen3-4B training, labeled by the numerator vs. the denominator of the contrastive reward. **Left:** both sides of the contrast use Qwen3-14B. Positive steering yields limited or unstable gains, while negative steering substantially improves the student. **Right:** the steered policy is scaled up to Qwen3-32B. Only negative steering benefits from the larger steered teacher, whereas, with positive steering, the student shows less improvements.

Figure 3 shows the student training curves under different steering strategies and teacher sizes. When both sides of the contrast use the same teacher Qwen3-14B (Figure 3, left), positive steering improves the student to some degree but consistently falls short of negative steering. Not-Negative Steering, despite differing from our prompt primarily in its semantic direction, yields little to no

performance gains, while behavior-specific prompts such as Error-Revision, Exploration, and their combination induce rewards that lead to unstable training.

We next replace the steered policy with Qwen3-32B while keeping the unsteered policy fixed at Qwen3-14B (Figure 3, right). Interestingly, increasing the steered teacher size has opposite effects depending on the steering direction. For positive steering, it results in an overall decrease in student training performance, whereas for negative steering, it further improves learning beyond the contrastive reward derived from equally sized Qwen3-14B teacher policies (Figure 3, left). **These results suggest that negative steering benefits from increased steering-teacher capacity, while positive steering provides weaker and less stable gains which fail to scale with teacher size.**

4 Analysis

In this section, we investigate what makes a "good" contrastive reward through the lens of the Steering-Success Rate (SSR). In Section 4.1, we formalize how the reliability of the steered policy affects the downstream training of the student policy, which explains why SSR matters for the student's final performance. In Section 4.2, we examine whether this explanation agrees with our empirical results. We find that negative steering consistently achieves a higher SSR than positive steering, which helps explain the performance gap observed in Section 3.3. We further verify that optimizing the negative steering method for higher SSR is key to improving the training performance of the induced contrastive reward.

4.1 Theoretical Analysis

We begin by formalizing why SSR is critical to the downstream performance of the trained student policy. The main intuition is that training involves a (regularized) improvement of the contrastive reward, so this reward must faithfully capture the structure of the task.

Consider the log-ratio form of the reward, $R(y) = \log \frac{\pi^*_+(y)}{\pi^*_o(y)}$ for positive steering, or $R(y) = \log \frac{\pi^*_o(y)}{\pi^*_-(y)}$ for negative steering. For the reward to be informative, the positively steered policy π^*_+ should assign higher probability than π^*_o to tokens that should be encouraged, and lower probability to tokens that should be discouraged. Conversely, the negatively steered policy π^*_- should *decrease* the probability of encouraged tokens and *increase* the probability of discouraged ones. Such reliability of these log-probability shifts directly corresponds to our intuition of reliable policy steering.

To quantify how reliably a steered policy moves in the intended direction, we first define the ground-truth win rate between two policies under a Bradley–Terry reward (Bradley & Terry, 1952).

Definition 4.1 (Win Rate; Zhang & Ranganath)**.** *Given a Bradley–Terry reward $r^* : \mathcal{Y} \to \mathbb{R}$, the **win rate** of a policy π_1 against a policy π_0 is defined as*

$$W(\pi_1 \succ \pi_0) \coloneqq \mathbb{E}_{y_1 \overset{\text{iid}}{\sim} \pi_1,\, y_0 \overset{\text{iid}}{\sim} \pi_0} \left[\mathbb{1}_{\{r^*(y_1) > r^*(y_0)\}} + \frac{1}{2} \cdot \mathbb{1}_{\{r^*(y_1) = r^*(y_0)\}} \right].$$

Using the win rate, we define the Steering-Success Rate, which measures how often steering improves (or, for negative steering, degrades) the base policy.

Definition 4.2 (Steering-Success Rate)**.** *Given a Bradley–Terry reward $r^* : \mathcal{Y} \to \mathbb{R}$ and a base policy π_o, let C_+ be a positive-steering prefix that induces the positively steered policy $\pi_+ \coloneqq \pi_o(\cdot \mid C_+)$. The **Positive Steering-Success Rate** of C_+ at π_o is the win rate of π_+ against π_o:*

$$S_{\pi_o}(C_+) \coloneqq W(\pi_+ \succ \pi_o).$$

*Similarly, let C_- be a negative-steering prefix that induces the negatively steered policy $\pi_- \coloneqq \pi_o(\cdot \mid C_-)$. Its **Negative Steering-Success Rate** is the win rate of π_o against π_-:*

$$S^-_{\pi_o}(C_-) \coloneqq W(\pi_o \succ \pi_-).$$

Next, we characterize how the contrastive reward determines the change in the converged student policy's log-probabilities.

Lemma 4.1. *Let π_θ^* be the optimal policy of the following KL-regularized contrastive distillation objective:*

$$\pi_\theta^* \coloneqq \arg\max_{\pi_\theta} \Big[\mathbb{E}_{y\sim\pi_\theta}\left[R(y)\right] - \beta \cdot \mathrm{KL}\left[\pi_\theta \parallel \pi_{\mathrm{ref}}\right]\Big],$$

where $R(y) = \log \frac{\pi_+^*(y)}{\pi_o^*(y)}$ is the contrastive reward and $\beta > 0$ is the KL-penalty weight. Then the student policy's pairwise log-probability change preserves the ordering induced by the contrastive reward; that is, for all $(y_0, y_1) \in \mathcal{Y} \times \mathcal{Y}$,

$$\log \frac{\pi_+^*(y_1)}{\pi_o^*(y_1)} > \log \frac{\pi_+^*(y_0)}{\pi_o^*(y_0)} \iff \log \frac{\pi_\theta^*(y_1)}{\pi_{\mathrm{ref}}(y_1)} > \log \frac{\pi_\theta^*(y_0)}{\pi_{\mathrm{ref}}(y_0)}.$$

Analogously, for the contrastive reward derived from a negatively steered policy, $R(y) = \log \frac{\pi_o^*(y)}{\pi_-^*(y)}$, the student preserves the ordering induced by this reward, which is the reverse of the ordering induced by the steered policy: for all $(y_0, y_1) \in \mathcal{Y} \times \mathcal{Y}$,

$$\log \frac{\pi_o^*(y_1)}{\pi_-^*(y_1)} > \log \frac{\pi_o^*(y_0)}{\pi_-^*(y_0)} \iff \log \frac{\pi_\theta^*(y_1)}{\pi_{\mathrm{ref}}(y_1)} > \log \frac{\pi_\theta^*(y_0)}{\pi_{\mathrm{ref}}(y_0)}.$$

The proof follows from applying the closed-form solution of the KL-regularized RL objective, $\pi_\theta^*(y) \propto \pi_{\mathrm{ref}}(y)\exp(R(y)/\beta)$ (Rafailov et al., 2024). Lemma 4.1 shows that the converged student shifts its log-probabilities relative to the reference policy in the direction determined by the contrastive reward. Consequently, whenever the steered policy ranks a pair of completions incorrectly (e.g., π_+^* favors a low-quality completion over a high-quality one, more strongly than π_o^*), the student also shifts its output distribution according to this incorrect order. Therefore, the reliability of steering, as measured by SSR, directly impacts the performance of the resulting student policy. The following proposition formalizes this observation.

Proposition 4.2 (Informal)**.** *Consider two positive-steering prefixes C_1, C_2 that induce the steered policies $\pi_{C_1}^* = \pi_o^*(\cdot \mid C_1)$ and $\pi_{C_2}^* = \pi_o^*(\cdot \mid C_2)$. Suppose the two steering methods have the same average magnitude of probability-ratio changes on correctly ranked pairs, and likewise on incorrectly ranked pairs. Let m_0 denote this common average on correctly ranked pairs.*

Each steered policy defines a contrastive reward $R_i(y) = \log \frac{\pi_{C_i}^*(y)}{\pi_o^*(y)}$ for $i \in \{1, 2\}$, which is used to train a student policy π_{θ,C_i}^* with KL penalty β_i toward a common reference policy π_{ref}. Suppose that β_1, β_2 are chosen so that the two KL-regularized student objectives share the same normalization factor:

$$\mathbb{E}_{y\sim\pi_{\mathrm{ref}}}\left[\left(\frac{\pi_{C_1}^*(y)}{\pi_o^*(y)}\right)^{1/\beta_1}\right] = \mathbb{E}_{y\sim\pi_{\mathrm{ref}}}\left[\left(\frac{\pi_{C_2}^*(y)}{\pi_o^*(y)}\right)^{1/\beta_2}\right].$$

Then, ignoring score ties, if the steering success of C_1 exceeds that of C_2 by a sufficient margin,

$$S_{\pi_o^*}(C_1) - S_{\pi_o^*}(C_2) > m_0 D_{\mathrm{TV}}(\pi_o^* \otimes \pi_o^*,\ \pi_{\mathrm{ref}} \otimes \pi_{\mathrm{ref}}),$$

the student trained with C_1 achieves a strictly higher downstream win rate against the common reference policy:

$$W(\pi_{\theta,C_1}^* \succ \pi_{\mathrm{ref}}) > W(\pi_{\theta,C_2}^* \succ \pi_{\mathrm{ref}}).$$

In other words, consider two steering prefixes that differ mainly in how often their induced probability changes point in the correct direction, rather than in the average magnitude of those changes. The prefix with the higher SSR then induces a reward which yields a student with a higher win rate, provided that its SSR advantage is large enough to overcome the distribution shift between π_o^* and π_{ref}. While Proposition 4.2 is stated for positive steering, the same result holds for negative steering by setting $R_i(y) = \log \frac{\pi_o^*(y)}{\pi_{C_i}^*(y)}$. We provide the formal statement and its proof in Appendix C.

4.2 Empirical Analysis

We now return to the experimental setup of Section 3 to verify whether SSR indeed explains the observed trends.

4.2.1 WHY DOES POSITIVE STEERING UNDERPERFORM NEGATIVE STEERING?

We sample 2,048 prompts from the training data, and independently sample $n = 4$ completions from each teacher policy (Qwen3-14B and Qwen3-32B) under four steering modes: No Steering, Not-Negative Steering, Exploration & Error-Revision, and Negative Steering. For each steering mode, we compare its average accuracy@4 against that of No Steering. For the positive steering modes (Not-Negative and Exploration & Error-Revision), we use the increase in average accuracy as a proxy for SSR; for Negative Steering, we use the decrease. We consider final-answer accuracy as a reasonable proxy to the quality of the reasoning trajectory, under the assumption that better reasoning correlates with higher accuracy (Armandpour et al., 2026; Lin et al., 2025)

	Steering Mode			
Steered Policy	No Steering	Not-Negative	Exploration & Error-Revision	Negative
Qwen3-14B	75.95	75.32 (-0.63%p)	75.67 (-0.28%p)	51.89 (**-24.06**%p)
Qwen3-32B	73.93	73.12 (-0.81%p)	73.99 (+0.06%p)	45.26 (**-28.67**%p)

Table 1: Average accuracy@4 [%] of each steering mode on 2,048 randomly sampled training prompts. Values in parentheses denote the change relative to No Steering. The positive steering modes (Not-Negative and Exploration & Error-Revision) fail to meaningfully increase accuracy, whereas negative steering substantially decreases it, producing a clear contrast.

As shown in Table 1, neither positive steering mode meaningfully improves accuracy. In contrast, negative steering reduces accuracy by 24.06%p and 28.67%p for Qwen3-14B and Qwen3-32B, respectively. **Negative steering thus achieves a substantially higher SSR than either positive steering method.**

We next show that this gap in SSR translates into the quality of the induced contrastive rewards. We use three models (Qwen3-0.6B, 1.7B, and 4B) to generate reasoning trajectories without steering on the same training subset ($N = 2{,}048$), and measure the average accuracy@4 of each model. We then compute the length-normalized contrastive reward of every trajectory under each steering method. Since a reliable reward should assign higher scores to trajectories from more accurate models, we examine whether the reward derived from the higher-SSR steering method ranks trajectories more consistently with the accuracy of the models that generated them.

Figure 4 shows the resulting reward distributions. The contrastive reward induced by negative steering orders the three models consistently with their accuracy, assigning the lowest rewards to trajectories from the least accurate model, Qwen3-0.6B. In contrast, the rewards induced by the other steering methods fail to recover this ordering. **Together with Table 1, this result suggests that negative steering, which achieves a higher SSR than positive steering, induces a contrastive reward that more faithfully captures trajectory quality and consequently yields a stronger student policy (Figure 3), consistent with Proposition 4.2.**

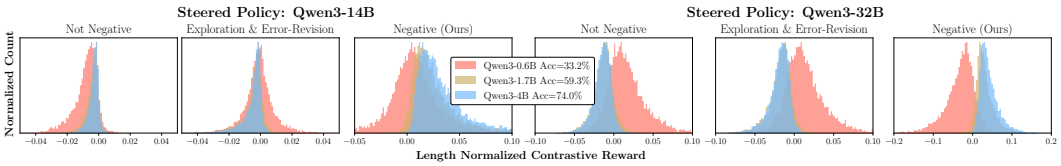

Figure 4: Distribution of length-normalized contrastive rewards on completions generated by Qwen3-0.6B, 1.7B, and 4B, whose average accuracy@4 increases in that order. Only the reward induced by negative steering assigns lower rewards to completions from the least accurate model (Qwen3-0.6B), while other rewards induced by positive steering methods fail to do so.

4.2.2 DOES SSR ALSO MATTER WITHIN NEGATIVE STEERING?

Having seen that SSR helps explain why positive steering underperforms negative steering in contrastive distillation, we next ask whether SSR can also bring improvements to negative steering.

We follow the experimental setup of Section 3.3 and train Qwen3-4B for 50 steps. We fix the numerator of the contrastive reward to the unsteered Qwen3-14B and vary the negatively steered

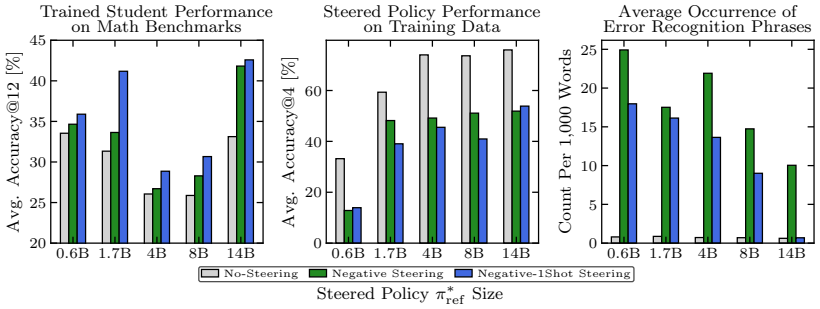

Figure 5: Ablation over the size of the negatively steered policy. **Left:** final student accuracy averaged over five math benchmarks. **Center:** training accuracy of the steered policy. **Right:** frequency of error-recognition phrases (Appendix A), which measures failures to follow the negative-steering instruction. Negative-1Shot Steering lowers both the steered policy's accuracy and its error-recognition rate, improving downstream student performance across all policy sizes.

policy from Qwen3-0.6B to Qwen3-14B. To isolate the effect of negative steering, we also train students with the same contrastive reward but without the negative-steering prompt. Finally, to examine how controlling SSR affects final training performance within negative steering, we include a single example of proper negative steering in the prompt (Negative-1Shot Steering; Figure 7).

Figure 5 presents the ablation results. Negative steering is most effective at both ends of the policy-size range (left plot). For the smallest policy, Qwen3-0.6B, however, the student trained without the negative-steering prompt also improves substantially. This can be explained based on the teacher policies' base accuracy (center): even without steering, Qwen3-0.6B has the lowest training accuracy, suggesting that small, under-performing models are already inclined toward negative behavior. Consistent with Figure 4, the policies whose training accuracy drops most under negative steering also yield the most consistent contrastive-reward rankings (Figure 8).

Beyond final-answer accuracy, we investigate another aspect in which a negatively steered policy can fail: *error recognition*. As discussed in Section 4.1, a negatively steered policy should ideally lower the probability of task-preferred tokens. Since recognizing errors during reasoning is a requirement for deep reasoning capabilities (Lee et al., 2025), the negative policy must be explicitly instructed *not* to recognize its errors. Under negative steering, we observe that smaller models fail to ignore their errors more often than larger models (Figure 5, right). This explains why the student trained with negatively steered Qwen3-14B outperforms the one trained with unsteered Qwen3-0.6B, despite the latter's lower average training accuracy.

Finally, adding a single example of proper negative steering to the prompt (Negative-1Shot; Figure 7) consistently lowers both the steered policy's training accuracy and its error-recognition rate (Figure 5, center and right), which in turn consistently improves student performance (left). Overall, these results show that SSR remains a useful criterion for designing better negative-steering prompts.

5 Conclusion

We introduced CRISP, a contrastive distillation method that derives dense rewards by contrasting a strong teacher with a negatively steered version of itself using natural-language instructions. CRISP requires neither task-specific RL training nor multiple teachers in memory, while providing interpretable control over the reward. On mathematical reasoning, it roughly doubles the gains of strong on-policy and contrastive distillation baselines for Qwen3-4B and Qwen3-1.7B students and continues to improve with teacher scale, unlike positive steering. We explain this asymmetry through the Steering-Success Rate (SSR), theoretically linking higher SSR to improved student win rate and empirically showing that negative steering achieves substantially higher SSR and better trajectory ranking accuracy. We also find that a single negative-steering demonstration helps improve SSR and downstream performance of student training. Overall, our results establish steering success as a practical and theoretically grounded principle for constructing effective contrastive rewards from a single strong teacher.

AI USE STATEMENT

In this work, we used generative AI tools for minor writing assistance, e.g., revising the text of the manuscript. We have not used generative AI tools for generating research ideas, designing experiments, or producing or analyzing experimental data. We have reviewed all AI-assisted work, where all AI-edited text was individually checked for factual accuracy against our experimental results. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work does not involve human subjects, personal data, or newly collected datasets. All experiments use publicly available open-weight models (Yang et al., 2025) and public mathematical reasoning datasets (Guha et al., 2026) and benchmarks (math-ai, 2023; Zhang & Math-AI, 2024; 2025; Dekoninck et al., 2026), used in accordance with their licenses.

REPRODUCIBILITY STATEMENT

We provide the specific details regarding training and evaluation in the Appendix A. We also provide the specific prompts used for each steering methods in Figures 1, 6, and 7. For our theoretical results, the formal statement of Proposition 4.2, its assumptions, and complete proofs are provided in Appendix C. We plan to release our source code for training and evaluation upon acceptance.

REFERENCES

Mohammadreza Armandpour, Fatih Ilhan, David Harrison, Ajay Jaiswal, Duc N. M. Hoang, Fartash Faghri, Yizhe Zhang, Minsik Cho, and Mehrdad Farajtabar. Unmasking On-Policy Distillation: Where It Helps, Where It Hurts, and Why, May 2026. URL http://arxiv.org/abs/2605.10889. arXiv:2605.10889 [cs.LG].

Ralph Allan Bradley and Milton E. Terry. Rank Analysis of Incomplete Block Designs: I. The Method of Paired Comparisons. *Biometrika*, 39(3/4):324–345, 1952. ISSN 0006-3444. doi: 10.2307/2334029. URL https://www.jstor.org/stable/2334029.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda

Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. *Nature*, 645(8081):633–638, September 2025. ISSN 0028-0836, 1476-4687. doi: 10.1038/s41586-025-09422-z. URL http://arxiv.org/abs/2501.12948. arXiv:2501.12948 [cs.CL].

Jasper Dekoninck, Nikola Jovanović, Tim Gehrunger, Kári Rögnvaldsson, Ivo Petrov, Chenhao Sun, and Martin Vechev. Beyond Benchmarks: MathArena as an Evaluation Platform for Mathematics with LLMs. 2026. URL https://arxiv.org/abs/2605.00674.

Shiyuan Feng, Huan-ang Gao, Haohan Chi, Hanlin Wu, Zhilong Zhang, Zheng Jiang, Bingxiang He, Wei-Ying Ma, Ya-Qin Zhang, and Hao Zhou. Weak-to-Strong Generalization via Direct On-Policy Distillation, July 2026. URL http://arxiv.org/abs/2607.05394. arXiv:2607.05394 [cs.LG].

Etash Guha, Ryan Marten, Sedrick Keh, Negin Raoof, Georgios Smyrnis, Hritik Bansal, Marianna Nezhurina, Jean Mercat, Trung Vu, Zayne Sprague, Ashima Suvarna, Benjamin Feuer, Leon Liangyu Chen, Zaid Khan, Eric Frankel, Sachin Grover, Caroline Choi, Niklas Muennighoff, Shiye Su, Wanjia Zhao, John Yang, Shreyas Pimpalgaonkar, Kartik sharma, Charlie Ji, Yichuan Deng, Sarah Pratt, Vivek Ramanujan, Jon Saad-Falcon, Stutee Acharya, Jeffrey Li, Achal Dave, Alon Albalak, Kushal Arora, Blake Wulfe, Chinmay Hegde, Greg Durrett, Sewoong Oh, Mohit Bansal, Saadia Gabriel, Aditya Grover, Kai-Wei Chang, Vaishaal Shankar, Aaron Gokaslan, Mike Merrill, Tatsunori Hashimoto, Yejin Choi, Jenia Jitsev, Reinhard Heckel, Maheswaran Sathiamoorthy, Alex Dimakis, and Ludwig Schmidt. OpenThoughts: Data Recipes for Reasoning Models. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 108059–108130, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/b010241b9f1cdfc7d4c392db899cef86-Paper-Conference.pdf.

Byeongho Heo, Jaehui Hwang, Sangdoo Yun, and Dongyoon Han. On-Policy Delta Distillation, July 2026. URL http://arxiv.org/abs/2607.15161. arXiv:2607.15161 [cs.LG].

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the Knowledge in a Neural Network, March 2015. URL http://arxiv.org/abs/1503.02531. arXiv:1503.02531 [stat.ML].

Junqi Jiang, Tom Bewley, Saumitra Mishra, Freddy Lecue, and Manuela Veloso. INTERPRETING LANGUAGE REWARD MODELS VIA CONTRASTIVE EXPLANATIONS. 2025.

Nathan Lambert, Valentina Pyatkin, Jacob Morrison, LJ Miranda, Bill Yuchen Lin, Khyathi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, Noah A. Smith, and Hannaneh Hajishirzi. RewardBench: Evaluating Reward Models for Language Modeling. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 1755–1797, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-195-7. doi: 10.18653/v1/2025.findings-naacl.96. URL https://aclanthology.org/2025.findings-naacl.96/.

Seongyun Lee, Seungone Kim, Minju Seo, Yongrae Jo, Dongyoung Go, Hyeonbin Hwang, Jinho Park, Xiang Yue, Sean Welleck, Graham Neubig, Moontae Lee, and Minjoon Seo. The CoT Encyclopedia: Analyzing, Predicting, and Controlling how a Reasoning Model will Think, May 2025. URL http://arxiv.org/abs/2505.10185. arXiv:2505.10185 [cs.CL].

Zicheng Lin, Tian Liang, Jiahao Xu, Qiuzhi Lin, Xing Wang, Ruilin Luo, Chufan Shi, Siheng Li, Yujiu Yang, and Zhaopeng Tu. Critical Tokens Matter: Token-Level Contrastive Estimation Enhances LLM's Reasoning Capability, January 2025. URL http://arxiv.org/abs/2411.19943. arXiv:2411.19943 [cs.CL].

Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization, January 2019. URL http://arxiv.org/abs/1711.05101. arXiv:1711.05101 [cs.LG].

Kevin Lu and Thinking Machines Lab. On-Policy Distillation. *Thinking Machines Lab: Connectionism*, 2025. doi: 10.64434/tml.20251026.

math-ai. AMC23, 2023. URL https://huggingface.co/datasets/math-ai/amc23.

Leyi Pan, Shuchang Tao, Yunpeng Zhai, Lingzhe Zhang, Zhaoyang Liu, Bolin Ding, Aiwei Liu, and Lijie Wen. RLCSD: Reinforcement Learning with Contrastive On-Policy Self-Distillation, June 2026. URL http://arxiv.org/abs/2606.11709. arXiv:2606.11709 [cs.LG].

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct Preference Optimization: Your Language Model is Secretly a Reward Model, July 2024. URL http://arxiv.org/abs/2305.18290. arXiv:2305.18290 [cs.LG].

Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. ZeRO: Memory Optimizations Toward Training Trillion Parameter Models, May 2020. URL http://arxiv.org/abs/1910.02054. arXiv:1910.02054 [cs.LG].

Mingyang Song and Mao Zheng. A Survey of On-Policy Distillation for Large Language Models, June 2026. URL http://arxiv.org/abs/2604.00626. arXiv:2604.00626 [cs.LG].

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 Technical Report, May 2025. URL http://arxiv.org/abs/2505.09388. arXiv:2505.09388 [cs.CL].

Wenkai Yang, Weijie Liu, Ruobing Xie, Kai Yang, Saiyong Yang, and Yankai Lin. Learning beyond Teacher: Generalized On-Policy Distillation with Reward Extrapolation, February 2026. URL http://arxiv.org/abs/2602.12125. arXiv:2602.12125 [cs.LG].

Fangxu Yu, Weijia Xu, Michael Xu, Tianyi Zhou, and Zinan Lin. Weak-to-Strong On-Policy Distillation, August 2026. URL http://arxiv.org/abs/2607.26246. arXiv:2607.26246 [cs.LG].

Lily H Zhang and Rajesh Ranganath. Preference learning made easy:Everything should be understood through win rate.

Yifan Zhang and Team Math-AI. American Invitational Mathematics Examination (AIME) 2024, 2024.

Yifan Zhang and Team Math-AI. American Invitational Mathematics Examination (AIME) 2025, 2025.

Siyan Zhao, Zhihui Xie, Mengchen Liu, Jing Huang, Guan Pang, Feiyu Chen, and Aditya Grover. Self-Distilled Reasoner: On-Policy Self-Distillation for Large Language Models, January 2026. URL https://arxiv.org/abs/2601.18734v3.

A Training and Evaluation Details

We train all models using a global batch size of 64 and using DeepSpeed-Zero 3 (Rajbhandari et al., 2020) with full fine-tuning. We utilize the AdamW optimizer (Loshchilov & Hutter, 2019) with weight-decay disabled and use a constant learning rate of 2e-6 with a linear warmup for the first 10 steps. During training, we set the maximum prompt length to 2,048 tokens and limit the completion length to 2,048 tokens. For all methods except OP^{2}D, we sweep the KL coefficient $\beta \in \{0.05, 0.1, 1, 5, 10\}$ and select the coefficient with the highest AIME24 Accuracy@12 value after 50 training steps. We then continue training the selected configuration for another 50 steps, evaluating every 10 steps, and report the checkpoint with the highest average accuracy across the five benchmarks. For the experiments in Section 3.3 through Section 4, we limit training to a total of 50 steps.

For both training and evaluation, we disable thinking mode and sample tokens with a temperature 1.0. During evaluation, we sample up to 16,384 tokens for all mathematical benchmarks, and utilize the no steering chat template in Figure 1 to format requests.

We utilize the following words as error-recognition phrases in Section 4.2.2:

- incorrect
- wrong
- error
- mistake
- fail
- invalid
- false

Table 2: Overview of different teacher pairs used by the compared methods for the Qwen3-4B student in Section 3.2. For the Qwen3-1.7B student experiment, Qwen3-4B is replaced by Qwen3-1.7B (Section 3.1). Each method trains the student using the contrastive reward $R(y) = \log \frac{\pi^*(y)}{\pi^*_{\mathrm{ref}}(y)}$ in equation 1, with KL regularization toward π_{ref}. OP^{2}D does not use KL regularization and instead uses an advantage-scaling term as its main hyper-parameter.

Method	π^*	π^*_{ref}	π_{ref}
OPD (Lu & Lab, 2025)	Qwen3-14B	Qwen3-4B	Qwen3-4B
G-OPD (Yang et al., 2026)	Qwen3-14B	Qwen3-4B	Qwen3-4B
Direct-OPD (Feng et al., 2026)	Qwen3-14B	Qwen3-14B-Base	Qwen3-4B
OP^{2}D (Heo et al., 2026)	Qwen3-14B	Qwen3-14B-Base	N/A
OPSD (Zhao et al., 2026)	Qwen3-4B (+Answer / Solution)	Qwen3-4B	Qwen3-4B
RLCSD (Pan et al., 2026)	Qwen3-14B (+Answer / Solution)	Qwen3-14B	Qwen3-4B
CRISP (*Ours*)	Qwen3-14B	Qwen3-14B (+Negative Steering)	Qwen3-4B

A.1 Steering Prompts

Positive Steering Chat Template (Error-Revision)

```
Problem: {problem}
Solve the problem above. Check intermediate steps for errors and
revise them when necessary. Please reason step by step, and put
your final answer within \boxed{}.
```

Positive Steering Chat Template (Exploration)

```
Problem: {problem}
Solve the problem above. Explore multiple plausible strategies,
and if an approach becomes unproductive or leads to dead ends,
reconsider your assumptions and try a different method. Please
reason step by step, and put your final answer within \boxed{}.
```

Positive Steering Chat Template (Exploration & Error-Revision)

```
Problem: {problem}
Solve the problem above. Explore multiple plausible strategies,
and if an approach becomes unproductive or leads to dead ends,
reconsider your assumptions and try a different method. Check
intermediate steps for errors and revise them when necessary.
Please reason step by step, and put your final answer within
\boxed{}.
```

Privileged Information Chat Template (Solution; Zhao et al. (2026))

```
Problem: {problem}
Here is a reference solution to this problem:
=== Reference Solution Begin ===
{solution}
=== Reference Solution End ===

After reading the reference solution above, make sure you
truly understand the reasoning behind each step - do not copy
or paraphrase it. Now, using your own words and independent
reasoning, derive the same final answer to the problem above.
Think step by step, explore different approaches, and don't be
afraid to backtrack or reconsider if something doesn't work out.
Put your final answer within \boxed{}.
```

Positive Steering Chat Template (Answer; Yu et al. (2026))

```
Problem: {problem}
Answer: {answer}
Please reason step by step, and put your final answer within
\boxed{}.
```

Figure 6: Overview of different steering prompts used in our experiments.

Negative-1 Shot Steering Chat Template

```
Provide an incorrect approach to a math problem using
invalid or non-rigorous reasoning.  Present the reasoning
straightforwardly without stating or implying that it is
intentionally incorrect.
Use the following example as a reference:
[Example Problem]
The value of \(\frac{2}{5} + \frac{1}{3}\) is:
(A) \(\frac{3}{8}\)
(B) \(\frac{2}{15}\)
(C) \(\frac{11}{15}\)
(D) \(\frac{13}{15}\)
(E) \(\frac{3}{15}\)

[Example Response]
To find the value of \(\frac{2}{5} + \frac{1}{3}\), we can add
the numerators and the denominators separately.
First, add the numerators:
\[ 2 + 1 = 3 \]
Next, add the denominators:
\[ 5 + 3 = 8 \]
Combining these gives the fraction:
\[ \frac{3}{8} \]
This corresponds to option (A).
\boxed{A}

Now, provide a similarly incorrect approach to the problem
below.  Reason (incorrectly) step by step, and put your final
answer within \boxed{}.
Problem:  {problem}
```

Figure 7: Prompt used for Negative-1Shot Steering.

B Detailed Experimental Results

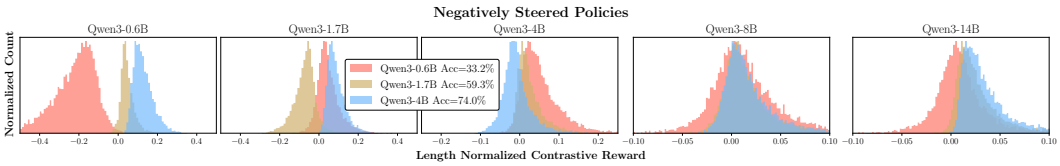

Figure 8: Distribution of length-normalized contrastive rewards on completions generated by Qwen3-0.6B, 1.7B, and 4B. All rewards are induced by the same negative-steering prompt, applied to steered policies of different sizes ranging from 0.6B to 32B. The steered policies that yield the best training performance under negative steering (Qwen3-0.6B and Qwen3-14B) also produce the most consistent rewards, assigning the lowest scores to completions from Qwen3-0.6B and higher scores to those from Qwen3-1.7B and Qwen3-4B.

Table 3: Detailed benchmark results for the main experiments in Section 3.2. We highlight the best performance in **bold** for each benchmarks.

Method	AMC23	AIME24	AIME25	HMMT25-Feb	HMMT25-Nov	Avg.
Qwen3-14B (Teacher)	72.92	27.78	22.50	11.94	13.33	29.69
Qwen3-4B (Student)	65.21	24.72	18.06	11.38	5.56	24.99
OPD	69.58	22.22	18.89	12.22	10.28	26.64
G-OPD	65.00	25.00	20.28	11.94	8.06	26.06
Direct-OPD	75.62	31.94	24.72	16.11	18.33	33.35
OP²D	77.29	29.44	26.11	13.89	13.61	32.07
OPSD (Answer Hint)	65.62	22.22	21.67	12.50	7.50	25.90
OPSD (Solution Hint)	66.46	22.78	17.22	10.83	8.89	25.24
RLCSD (Answer Hint)	66.46	22.22	18.06	10.56	10.83	25.62
RLCSD (Solution Hint)	67.08	24.72	20.83	11.39	7.50	26.31
CRISP (*Ours*)	**85.83**	**40.56**	**33.61**	**23.06**	**28.33**	**42.28**
Qwen3-1.7B (Student)	47.92	11.67	11.39	4.44	4.44	15.97
OPD	48.96	15.56	11.94	7.50	4.44	17.68
G-OPD	47.92	11.94	10.83	5.56	5.28	16.31
Direct-OPD	56.04	22.22	14.44	6.11	8.06	21.38
OP²D	56.04	21.67	16.11	8.61	7.78	22.04
OPSD (Answer Hint)	49.38	13.61	11.39	5.00	5.28	16.93
OPSD (Solution Hint)	44.79	15.28	9.72	4.72	5.56	16.01
RLCSD (Answer Hint)	47.71	14.44	10.28	3.89	5.00	16.26
RLCSD (Solution Hint)	49.17	13.89	10.00	5.00	5.00	16.61
CRISP (*Ours*)	**67.71**	**29.17**	**20.56**	**12.78**	**11.11**	**28.26**

C PROOFS

Proposition C.1 (Formal). *Consider two steering prefixes C_1, C_2 which induce*

$$\pi^*_{C_i} = \pi^*_o(\cdot \mid C_i), \qquad i \in \{1, 2\},$$

and define

$$q_i(y) \coloneqq \frac{\pi^*_{C_i}(y)}{\pi^*_o(y)}.$$

Let π^*_{θ,C_i} be the optimal student trained with the contrastive reward

$$R_i(y) = \log q_i(y)$$

and KL penalty $\beta_i > 0$.

Assume score ties do not occur and that the relevant conditional events below have nonzero probability. Suppose the two steering modes have the same mean probability change on correctly ranked positive pairs,

$$\begin{aligned} m_0 &\coloneqq \mathbb{E}_{y_1,y_0 \overset{\text{iid}}{\sim} \pi^*_o} \left[q_1(y_1) - q_1(y_0) \mid r^*(y_1) > r^*(y_0),\ q_1(y_1) > q_1(y_0)\right] \\ &= \mathbb{E}_{y_1,y_0 \overset{\text{iid}}{\sim} \pi^*_o} \left[q_2(y_1) - q_2(y_0) \mid r^*(y_1) > r^*(y_0),\ q_2(y_1) > q_2(y_0)\right], \end{aligned} \tag{3}$$

and the same mean probability change on incorrectly ranked positive pairs,

$$\begin{aligned} m_0^- &\coloneqq \mathbb{E}_{y_1,y_0 \overset{\text{iid}}{\sim} \pi^*_o} \left[q_1(y_1) - q_1(y_0) \mid r^*(y_1) < r^*(y_0),\ q_1(y_1) > q_1(y_0)\right] \\ &= \mathbb{E}_{y_1,y_0 \overset{\text{iid}}{\sim} \pi^*_o} \left[q_2(y_1) - q_2(y_0) \mid r^*(y_1) < r^*(y_0),\ q_2(y_1) > q_2(y_0)\right] \le m_0. \end{aligned} \tag{4}$$

For

$$u_i(y) \coloneqq q_i(y)^{1/\beta_i}, \qquad Z_i \coloneqq \mathbb{E}_{y \sim \pi_{\text{ref}}} \left[u_i(y)\right],$$

assume further that $Z_1 = Z_2$ and that the corresponding downstream probability changes are matched:

$$\begin{aligned} m_\beta^+ &\coloneqq \mathbb{E}_{y_1,y_0 \overset{\text{iid}}{\sim} \pi_{\text{ref}}} \left[u_1(y_1) - u_1(y_0) \mid r^*(y_1) > r^*(y_0),\ u_1(y_1) > u_1(y_0)\right] \\ &= \mathbb{E}_{y_1,y_0 \overset{\text{iid}}{\sim} \pi_{\text{ref}}} \left[u_2(y_1) - u_2(y_0) \mid r^*(y_1) > r^*(y_0),\ u_2(y_1) > u_2(y_0)\right], \end{aligned} \tag{5}$$

and

$$\begin{aligned} m_\beta^- &\coloneqq \mathbb{E}_{y_1,y_0 \overset{\text{iid}}{\sim} \pi_{\text{ref}}} \left[u_1(y_1) - u_1(y_0) \mid r^*(y_1) < r^*(y_0),\ u_1(y_1) > u_1(y_0)\right] \\ &= \mathbb{E}_{y_1,y_0 \overset{\text{iid}}{\sim} \pi_{\text{ref}}} \left[u_2(y_1) - u_2(y_0) \mid r^*(y_1) < r^*(y_0),\ u_2(y_1) > u_2(y_0)\right]. \end{aligned} \tag{6}$$

Then

$$\begin{aligned} S_{\pi^*_o}(C_1) - S_{\pi^*_o}(C_2) > m_0 D_{\text{TV}} \left(\pi^*_o \otimes \pi^*_o,\ \pi_{\text{ref}} \otimes \pi_{\text{ref}}\right) \\ \implies \quad W(\pi^*_{\theta,C_1} \succ \pi_{\text{ref}}) > W(\pi^*_{\theta,C_2} \succ \pi_{\text{ref}}). \end{aligned} \tag{7}$$

Proof. For $i \in \{1, 2\}$, let

$$S_i \coloneqq S_{\pi^*_o}(C_i).$$

We first express S_i in terms of the hard ranking induced by q_i.

Define

$$G_{0,i} \coloneqq \mathbb{E}_{y_1,y_0 \overset{\text{iid}}{\sim} \pi^*_o} \left(\mathbb{1}_{\{r^*(y_1) > r^*(y_0),\ q_i(y_1) > q_i(y_0)\}}\right), \tag{8}$$

$$B_{0,i} \coloneqq \mathbb{E}_{y_1,y_0 \overset{\text{iid}}{\sim} \pi^*_o} \left(\mathbb{1}_{\{r^*(y_1) < r^*(y_0),\ q_i(y_1) > q_i(y_0)\}}\right), \tag{9}$$

and their signed difference

$$A_{0,i} \coloneqq G_{0,i} - B_{0,i}.$$

Using

$$\pi^*_{C_i}(y) = \pi^*_o(y) q_i(y),$$

the definition of the win rate (equation 4.1)gives

$$2S_i - 1 = \sum_{y_1, y_0 \in \mathcal{Y}} \pi^*_{C_i}(y_1)\pi^*_o(y_0)\text{sign}(r^*(y_1) - r^*(y_0)) \tag{10}$$

$$= \sum_{y_1, y_0 \in \mathcal{Y}} \pi^*_o(y_1)\pi^*_o(y_0)q_i(y_1)\text{sign}(r^*(y_1) - r^*(y_0))\,. \tag{11}$$

Swapping y_1 and y_0 in equation 11 gives

$$2S_i - 1 = -\sum_{y_1, y_0 \in \mathcal{Y}} \pi^*_o(y_1)\pi^*_o(y_0)q_i(y_0)\text{sign}(r^*(y_1) - r^*(y_0))\,.$$

Averaging the two expressions,

$$2S_i - 1 = \frac{1}{2}\sum_{y_1, y_0 \in \mathcal{Y}} \pi^*_o(y_1)\pi^*_o(y_0)\left[q_i(y_1) - q_i(y_0)\right] \cdot \text{sign}(r^*(y_1) - r^*(y_0))\,. \tag{12}$$

Now, the contribution from pairs satisfying $q_i(y_1) < q_i(y_0)$ is identical to the contribution from their swapped pairs satisfying $q_i(y_1) > q_i(y_0)$. Since pairs with $r^*(y_1) = r^*(y_0)$ contribute zero to equation 12, we have

$$2S_i - 1 = G_{0,i}m_0 - B_{0,i}m_0^-\,. \tag{13}$$

Let

$$\tau_0 \coloneqq \mathbb{E}_{y_1, y_0 \overset{\text{iid}}{\sim} \pi^*_o}\left(\mathbb{1}_{\{r^*(y_1) = r^*(y_0)\}}\right).$$

Because score ties do not occur and the pair distribution is symmetric, exactly one orientation of each non-tied score pair satisfies $q_i(y_1) > q_i(y_0)$. Moreover, among reward-tied pairs exactly one of the two orientations also has positive score difference. Hence

$$G_{0,i} + B_{0,i} = \frac{1 - \tau_0}{2},$$

which is independent of i.

Define

$$\bar{m}_0 \coloneqq \frac{m_0 + m_0^-}{2}.$$

Since

$$G_{0,i} = \frac{1}{2}\left(\frac{1 - \tau_0}{2} + A_{0,i}\right), \qquad B_{0,i} = \frac{1}{2}\left(\frac{1 - \tau_0}{2} - A_{0,i}\right),$$

Equation equation 13 becomes

$$2S_i - 1 = \bar{m}_0 A_{0,i} + \frac{m_0 - m_0^-}{2}\frac{1 - \tau_0}{2}. \tag{14}$$

The second term is common to both steering modes. Therefore,

$$2(S_1 - S_2) = \bar{m}_0(A_{0,1} - A_{0,2}). \tag{15}$$

Since $m_0^- \leq m_0$,

$$\bar{m}_0 = \frac{m_0 + m_0^-}{2} \leq m_0. \tag{15a}$$

We next transfer this ranking advantage from $\pi^*_o \otimes \pi^*_o$ to $\pi_{\text{ref}} \otimes \pi_{\text{ref}}$.

For each mode, define the bounded pairwise ranking score

$$a_i(y_1, y_0) \coloneqq \mathbb{1}_{\{(r^*(y_1) - r^*(y_0))(q_i(y_1) - q_i(y_0)) > 0\}} + \frac{1}{2}\mathbb{1}_{\{r^*(y_1) = r^*(y_0)\}}.$$

Because score ties do not occur,

$$0 \leq a_i(y_1, y_0) \leq 1.$$

By symmetry under swapping pairs,

$$A_{0,i} = \mathbb{E}_{y_1, y_0 \overset{\text{iid}}{\sim} \pi_o^*} [a_i(y_1, y_0)] - \frac{1}{2}. \tag{16}$$

Analogously, define

$$A_{\text{ref},i} \coloneqq \mathbb{E}_{y_1, y_0 \overset{\text{iid}}{\sim} \pi_{\text{ref}}} [a_i(y_1, y_0)] - \frac{1}{2}.$$

Since the same function $a_i : \mathcal{Y} \times \mathcal{Y} \to [0, 1]$ is evaluated under the two pair distributions, the variational characterization of total variation gives

$$|A_{\text{ref},i} - A_{0,i}| \leq D_{\text{TV}} \left(\pi_o^* \otimes \pi_o^*,\, \pi_{\text{ref}} \otimes \pi_{\text{ref}}\right). \tag{17}$$

Let

$$t \coloneqq D_{\text{TV}} \left(\pi_o^* \otimes \pi_o^*,\, \pi_{\text{ref}} \otimes \pi_{\text{ref}}\right).$$

Then

$$A_{\text{ref},1} - A_{\text{ref},2} \geq (A_{0,1} - t) - (A_{0,2} + t) \tag{18}$$

$$= A_{0,1} - A_{0,2} - 2t. \tag{19}$$

Now suppose

$$S_1 - S_2 > m_0 t.$$

Using equation 15 and $\bar{m}_0 \leq m_0$,

$$A_{0,1} - A_{0,2} = \frac{2(S_1 - S_2)}{\bar{m}_0} \tag{20}$$

$$> \frac{2m_0 t}{\bar{m}_0} \tag{21}$$

$$\geq 2t. \tag{22}$$

Due to equation 19, we have

$$A_{\text{ref},1} > A_{\text{ref},2}. \tag{23}$$

It remains to show that this ordering is preserved by the converged student policies. From Lemma 4.1,

$$\pi^*_{\theta, C_i}(y) = \frac{\pi_{\text{ref}}(y) q_i(y)^{1/\beta_i}}{Z_i} = \frac{\pi_{\text{ref}}(y) u_i(y)}{Z_i}.$$

Since $\beta_i > 0$,

$$u_i(y_1) > u_i(y_0) \iff q_i(y_1) > q_i(y_0),$$

so the student and the corresponding contrastive reward induce the same ordering.

Write

$$Z_1 = Z_2 =: Z.$$

Under $y_1, y_0 \overset{\text{iid}}{\sim} \pi_{\text{ref}}$, define

$$G_{\text{ref},i} \coloneqq \mathbb{E}_{y_1, y_0 \overset{\text{iid}}{\sim} \pi_{\text{ref}}} \left(\mathbb{1}_{\{r^*(y_1) > r^*(y_0),\, u_i(y_1) > u_i(y_0)\}}\right), \tag{24}$$

$$B_{\text{ref},i} \coloneqq \mathbb{E}_{y_1, y_0 \overset{\text{iid}}{\sim} \pi_{\text{ref}}} \left(\mathbb{1}_{\{r^*(y_1) < r^*(y_0),\, u_i(y_1) > u_i(y_0)\}}\right). \tag{25}$$

By construction,

$$A_{\text{ref},i} = G_{\text{ref},i} - B_{\text{ref},i}.$$

The same symmetrization used in equation 12 now gives

$$2W(\pi^*_{\theta, C_i} \succ \pi_{\text{ref}}) - 1 = \frac{1}{Z} \left[G_{\text{ref},i} m_\beta^+ - B_{\text{ref},i} m_\beta^-\right]. \tag{26}$$

Let

$$\bar{m}_\beta \coloneqq \frac{m_\beta^+ + m_\beta^-}{2} > 0.$$

If

$$\tau_{\text{ref}} \coloneqq \mathbb{E}_{y_1, y_0 \overset{\text{iid}}{\sim} \pi_{\text{ref}}} \left(\mathbb{1}_{\{r^*(y_1) = r^*(y_0)\}} \right),$$

then, we have

$$G_{\text{ref},i} + B_{\text{ref},i} = \frac{1 - \tau_{\text{ref}}}{2},$$

for both $i \in \{1, 2\}$. Hence equation 26 can be written as

$$2W(\pi^*_{\theta,C_i} \succ \pi_{\text{ref}}) - 1 = \frac{1}{Z} \left[\bar{m}_\beta A_{\text{ref},i} + \frac{m^+_\beta - m^-_\beta}{2} \frac{1 - \tau_{\text{ref}}}{2} \right]. \tag{27}$$

Subtracting the expressions for $i = 1$ and $i = 2$, we have:

$$2\Big[W(\pi^*_{\theta,C_1} \succ \pi_{\text{ref}}) - W(\pi^*_{\theta,C_2} \succ \pi_{\text{ref}}) \Big] = \frac{\bar{m}_\beta}{Z} \left(A_{\text{ref},1} - A_{\text{ref},2} \right). \tag{28}$$

Since $\bar{m}_\beta > 0$, $Z > 0$, and $A_{\text{ref},1} > A_{\text{ref},2}$ by equation 23,

$$W(\pi^*_{\theta,C_1} \succ \pi_{\text{ref}}) > W(\pi^*_{\theta,C_2} \succ \pi_{\text{ref}}).$$

Therefore,

$$S_{\pi^*_o}(C_1) - S_{\pi^*_o}(C_2) > m_0 D_{\text{TV}} \left(\pi^*_o \otimes \pi^*_o,\, \pi_{\text{ref}} \otimes \pi_{\text{ref}} \right)$$

is sufficient for

$$W(\pi^*_{\theta,C_1} \succ \pi_{\text{ref}}) > W(\pi^*_{\theta,C_2} \succ \pi_{\text{ref}}).$$

□