CoTD-PO: Chain-of-Thought Distillation with Preference Optimization

Lujie Niu[1], **Haochen Sun**[1], **Fangkun Zhao**[1], **Sheng Chen**[1], **Zimeng Bai**[1], **Jiawei Zhang**[1], **Caixia Yuan**[1*], **Xiaojie Wang**[1]

[1] Beijing University of Posts and Telecommunications
{lujien, haochen_sun, yuancx, xjwang}@bupt.edu.cn

Abstract

Chain-of-Thought (CoT) distillation has emerged as a promising paradigm to enhance the reasoning ability of small language models by imitating the reasoning and outputs of larger teacher models. However, existing approaches suffer from a critical limitation: a distribution mismatch between teacher-generated training trajectories and the student model's own generative distribution. This mismatch leads to exposure bias during inference and often induces mode collapse or mode averaging, thereby degrading the student model's generative diversity and robustness. To address these issues, we propose CoTD-PO (Chain-of-Thought Distillation with Preference Optimization) [1], a reinforcement learning framework that shifts the training paradigm from passive imitation to active trajectory exploration. Instead of forcing the student to imitate exact teacher traces, our method enables the student to sample its own answer paths. To support training with non-open-source teacher models, we approximate the teacher's output distribution through preference-based scoring. Furthermore, we adopt an offline iterative training procedure that enables stable and efficient optimization. Experiments on diverse open-ended generation tasks demonstrate that CoTD-PO significantly outperforms standard CoT distillation baselines, achieving higher output quality while mitigating mode collapse and preserving semantic diversity.

1 Introduction

Recently, the capabilities of large language models (LLMs) have been shown to scale not only with an increase in model size but also through techniques such as chain-of-thought (CoT) reasoning (Wei et al., 2022b) and model distillation (Tunstall et al., 2023). CoT reasoning has gained significant attention for its ability to enhance the reasoning ability of LLMs by encouraging them to explicitly articulate intermediate reasoning steps before arriving at a final answer (Kojima et al., 2022; Wei et al., 2022a). However, empirical evidence demonstrates a strong scale-dependence in CoT effectiveness, with significant performance gains confined to models exceeding 100 billion parameters. Smaller models often fail to perform well with CoT, as their limited parameter space hinders their ability to generate intermediate reasoning steps effectively.

This limitation has motivated research into chain-of-thought distillation (CoTD), which seeks to transfer multi-step reasoning processes and outputs from large teacher models to compact student architectures (Wang et al., 2023; Li et al., 2023). Unlike traditional knowledge distillation (KD) that primarily aligns final answers (Hinton et al., 2015), CoTD requires the student to imitate both the intermediate rationales and final answers of the teacher. This strict trajectory-level imitation introduces :

(1) Exposure biass (Arora et al., 2022; Ranzato et al., 2016): During training, CoTD relies on teacher forcing, where gold-standard intermediate steps are provided at each stage. However, during inference, the model must autoregressively generate reasoning steps based on its own previous outputs. This train-test discrepancy leads to exposure bias, where early errors propagate through sequential dependencies, resulting in compounded mistakes in downstream reasoning.

(2) Distribution mismatch: As large teacher models possess stronger reasoning abilities, their CoT traces are often more diverse and coherent. However, student models have limited capacity to capture such rich patterns, which creates a severe mismatch because the teacher's sampled answers often receive near-zero probability under the student model. Minimizing forward or reverse KL

* Corresponding author.

[1] https://github.com/little-mushroom0/COTD_PO

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 19975–19986
November 4-9, 2025

divergence further amplifies this gap. Forward KL forces the student to spread probability mass over regions unsupported by the teacher, while reverse KL encourages over-concentration on a few modes, which leads to mode averaging or mode collapse. Thereby undermining generalization from the perspective of distributional coverage and output diversity (Wang et al., a; Gu et al.).

While recent advances have sought to avoid exposure bias and distribution mismatch by focusing on structured tasks (e.g., closed-form arithmetic problems with deterministic solutions and multiple-choice QA (Feng et al., 2024; Lee et al., 2024)), where the output space usually consists of a finite number of classes. However, these approaches often struggle in open-ended generation settings (Fu et al., 2023a), which are characterized by high output diversity and semantic ambiguity. Meanwhile, they typically assume access to full teacher distributions—an impractical requirement for black-box models, struggling to generalize to open-ended settings with high diversity and ambiguity (Feng et al., 2024; Lee et al., 2024).

To address these challenges, we introduce Chain-of-Thought Distillation with Preference Optimization (CoTD-PO). Our method shifts the paradigm from strict teacher-forcing imitation to student-led exploration. Specifically, we leverage reinforcement learning to empower the student model to discover its own answer trajectories, using the teacher's rationales as high-level guidance rather than a fixed path to be mimicked. This self-exploration directly mitigates exposure bias by aligning the training process with the autoregressive nature of inference. Furthermore, to enable distillation from black-box teachers, CoTD-PO incorporates a preference alignment strategy that replaces explicit probability matching with reward-based relative quality scoring. This allows effective supervision without requiring access to the full teacher distribution. We evaluate our approach on multiple instruction-following benchmarks, and experimental results demonstrate that CoTD-PO not only achieves superior performance but also effectively mitigates the aforementioned issues of exposure bias and distribution mismatch.

To summarize, our contributions are as follows:

1. We introduce CoTD-PO, a novel framework that transitions from teacher-forcing imitation to student-led generation guided by teacher rationales, reducing exposure bias and distribution mismatch and enhancing continual learning.

2. We introduce a preference alignment strategy that replaces explicit probability matching with reward-based relative quality scoring, enabling effective distillation from black-box teacher models without access to their full output distributions.

3. We extend CoTD-PO to instruction-following tasks by leveraging teacher rationales as implicit reasoning guidance, achieving consistent performance improvements across multiple benchmarks.

2 Methodology

2.1 Preliminaries

Building upon established chain-of-thought (CoT) distillation frameworks (Ho et al., 2023; Fu et al., 2023a), the student model is trained to mimic the reasoning capability of the teacher through probability matching. The vanilla CoT distillation framework consists of two interdependent stages: **(1) Teacher Model CoT Generation Step:** For each question-answer pair $\{x, y\} \in \mathcal{D}$, where y is the correct answer, the teacher model M_t is prompted to generate a rationale CoT. **(2) Student Imitation Learning:** Then, the student M_s is trained to replicate both the teacher's rationale CoT and answer tokens y, given the question x as input through probability matching formalized as Kullback-Leibler divergence minimization. The standard objective decomposes into reasoning and answer alignment components:

$$\begin{aligned}&\mathrm{KL}(p_t(y, CoT \mid x) \parallel p_s(y, CoT \mid x)) \\ &= \mathbf{E}_{CoT \sim p_t} \underbrace{\mathrm{KL}(p_t(CoT \mid x) \parallel p_s(CoT \mid x))}_{\text{Reasoning Alignment}} \\ &+ \mathbf{E}_{y \sim p_t} \underbrace{\mathrm{KL}(p_t(y \mid CoT, x) \parallel p_s(y \mid CoT, x))}_{\text{Answer Alignment}},\end{aligned} \tag{1}$$

leading to the standard negative log-likelihood loss:

$$\begin{aligned}\mathcal{L}_{\text{CoTD}} &= \lambda_{\text{ans}} \cdot \mathbf{E}\left[-\log M_s(y \mid q, CoT)\right] \\ &+ \lambda_{\text{cot}} \cdot \mathbf{E}\left[\sum_{t=1}^{T} -\log M_s(CoT_t \mid q, CoT_{<t})\right].\end{aligned} \tag{2}$$

where $\lambda_{\text{ans}}, \lambda_{\text{cot}} \in [0, 1]$ control task focus. This joint training enforces strict trajectory alignment between teacher and student.

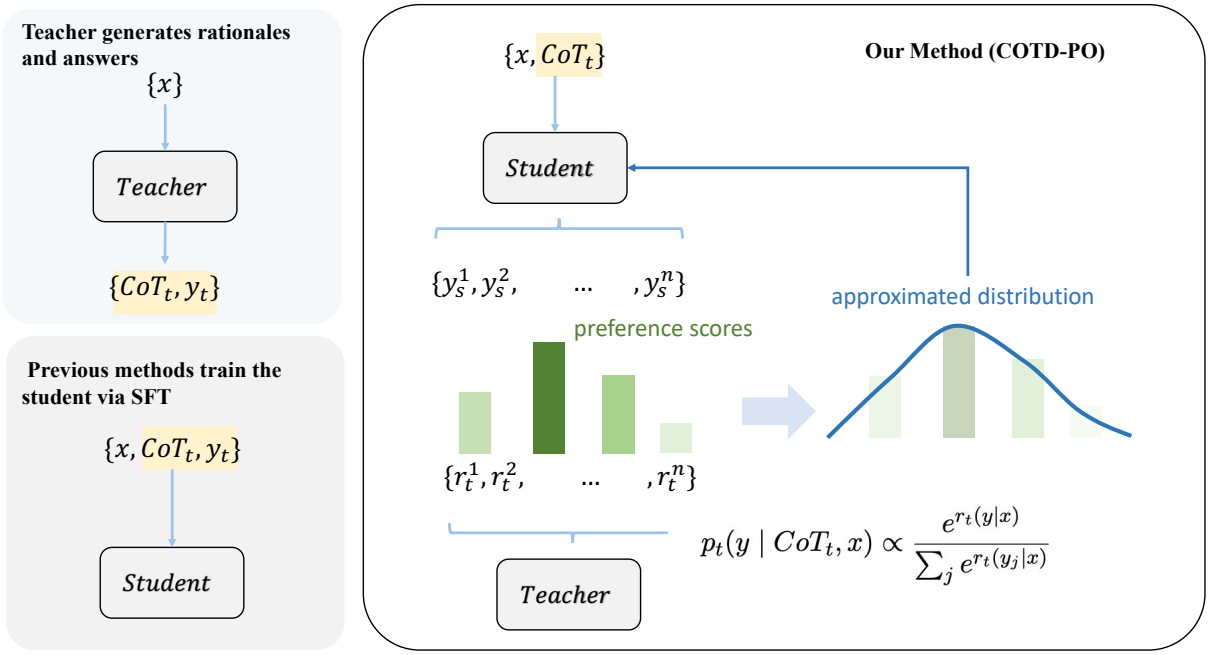


Figure 1: The COTD-PO training procedure. Unlike standard SFT (left) which performs direct imitation of a single teacher answer, our framework (right) operates on a distribution of student-generated answers. The student first samples multiple candidates y_s. A teacher model then provides preference scores r_t for these candidates, which are converted into a target distribution. This distribution acts as a soft label to align the student's policy with the teacher's preferences.

2.2 Chain-of-Thought Distillation with Preference Optimization

Prior approaches to Chain-of-Thought Distillation (CoTD) often face a critical trade-off where gains in specialized reasoning come at the cost of degraded generalizability (Fu et al., 2023a). This issue stems from a significant distributional mismatch between the teacher's training data and the student model's own distribution, leading to problems like mode collapse and mode averaging. Consequently, the student model's ability for continual learning is often impaired. Furthermore, the challenge is exacerbated when employing black-box teacher models. The inability to access the internal states and probability distributions of these models significantly hinders effective probabilistic alignment between the teacher and the student.

To overcome these challenges, we propose **CoTD-PO**, a reinforcement learning framework that redefines CoTD through three key innovations: (1) Active Trajectory Exploration: Shift from passive cloning to self-exploration of answers guided by teacher reasoning; (2) Black-Box Preference Alignment: Bridge teacher-student probability via reward-based preference scoring, bypassing probability access; (3) Entropy-Regularized Optimization: Balance answer quality and diversity through principled exploration.

Figure 1 illustrates the key differences between standard CoT distillation and our proposed CoTD-PO framework.

Answer Alignment in Student Generation Space Conventional CoTD performs answer alignment within the teacher's output space. This induces mode collapse or mode averaging when the teacher's and student's distributions diverge. To circumvent this issue, we propose reformulating the alignment objective to operate within the student's own generation space. Specifically, we encourage the student model to explore its own answer trajectories, conditioned on the teacher's high-quality chain-of-thought rationales. We formally redefine the Answer Alignment objective from Equation (1) as follows:

$$\mathbf{E}_{y\sim p_s,\, CoT\sim p_t} \\ \mathrm{KL}\left(p_t(y \mid CoT, x) \,\|\, p_s(y \mid CoT, x)\right), \tag{3}$$

Where x denotes the instruction, CoT denotes the rationale from the teacher, and y denotes the student-generated answer. This formulation allows the student to continuously learn the reasoning component from the teacher, while simultaneously exploring how to interpret and utilize that reasoning

for its own generation. In doing so, the student mitigates distribution mismatch during training and better generalizes to auto-regressive inference scenarios.

Black-Box Teacher Approximation via Preference Scores In many real-world scenarios, the teacher model functions as a black box, making it infeasible to directly access the conditional probability $p_t(y \mid x)$ for arbitrary student-generated outputs $y \sim p_s$. To overcome this limitation, we propose approximating the teacher's distribution using preference scores. These scores can be sourced either from a dedicated reward model, $r_t(y \mid x)$, trained to emulate the teacher's judgments, or directly from the teacher if it can provide scalar quality ratings. This scalar scoring mechanism provides a bridge between discrete judgments and probabilistic alignment. Specifically, we replace the intractable probability distributions with softmax-normalized teacher-assigned preference scores:

$$p_t(y \mid CoT_t, x) \propto \frac{e^{r_t(y|x)}}{\sum_j e^{r_t(y_j|x)}}, \tag{4}$$

where y_j are candidate answers. This formulation provides a monotonic mapping from scores to probabilities. For example, if the teacher model deems an answer y_1 to be of higher quality than another answer y_2, this is reflected in their assigned scores $r_t(y_1 \mid x) > r_t(y_2 \mid x)$.The softmax function ensures that this preference translates directly into the probability space, such that the approximated probability for the superior answer is strictly greater $p_s(y_1 \mid CoT_t, x) > (p_s(y_2 \mid CoT_t, x)$. Similarly, the student's distribution $p_s(y \mid x)$ is parameterized by its own preference scores $r_s(y \mid x)$. Substituting Equation (4) into the answer alignment term in the KL divergence into Equation (3) as:

$$\begin{aligned}&\mathrm{KL}\left(p_t(y \mid CoT_t, x) \,\|\, p_s(y \mid CoT_t, x)\right) \\ &\propto -\sum_i \frac{e^{r_t(y_i|x)}}{\sum_j e^{r_t(y_j|x)}} \log\left(\frac{e^{r_s(y_i|x)}}{\sum_j e^{r_s(y_j|x)}}\right),\end{aligned} \tag{5}$$

The detailed derivation is provided in Appendix A. where $r_t(y_i \mid x)$ and $r_s(y_i \mid x)$ represent the teacher's and student's preference scores for candidate output y_i, respectively. Minimizing this divergence encourages the student to match the teacher's implicit preferences over multiple candidates, without requiring explicit teacher distributions.

Diversity-Preserving Optimization To prevent the student from overfitting to high-reward but low-diversity outputs, we augment the training objective with entropy regularization:

$$\max_{p_s}; \mathbf{E}_{y\sim p_s}\left[r_s(y \mid x)\right] + \lambda H_s(y \mid CoT_t, x), \tag{6}$$

$H_s(y \mid CoT_t, x) = -\mathbf{E}_{y\sim p_s}\left[\log p_s(y \mid CoT_t, x)\right]$ is the entropy of the student's distribution and λ controls the strength of entropy regularization. The optimal solution to this objective satisfies:

$$r_s(y \mid x) \propto \lambda \log p_s(y \mid CoT_t, x) + \lambda \log Z, \tag{7}$$

where $Z = \sum_y e^{\lambda r_s(y)}$ is the partition function. This establishes a direct link between the log-probabilities of the student policy and its internal preference scores, enabling tractable optimization. The full derivation of this optimal policy is provided in Appendix B.

Unified Training Objective Combining the answer alignment term of KL divergence formulation in (5) with preference scores in (7), we derive the final Answer Preference Alignment Loss:

$$\begin{aligned}&\mathcal{L}_{\text{apa}} = \\ &-\sum_i \frac{e^{r_t(y_i|x)}}{\sum_j e^{r_t(y_j|x)}} \log\left(\frac{e^{\lambda \log p_s(y_i|CoT_t,x)}}{\sum_j e^{\lambda \log p_s(y_j|CoT_t,x)}}\right).\end{aligned} \tag{8}$$

The core training objective explicitly aligns the student's answer distribution with the teacher's preferences, encouraging the student to respect the relative rankings imposed by the teacher's reward model while preserving exploration flexibility through entropy-regularized sampling. To jointly optimize both rationale generation and answer synthesis, we introduce a unified training objective. We adopt offloading techniques and iteratively refine the model over multiple training rounds to reduce memory consumption during large-batch optimization. As detailed in Algorithm 1, we implement an iterative offline procedure:*(1) Student Answer Sampling:* For each instruction x and corresponding teacher rationale CoT_t, the student samples K candidate answers $\{y_j\}_{j=1}^K$ from $p_s(y \mid CoT_t, x)$; *(2) Teacher Preference Scoring:* A frozen teacher-aligned reward model $r_t(y \mid x)$ scores each answer; *(3) Distribution Alignment Update:* To jointly optimize rationale generation and answer synthesis, we introduce a composite training objective:

Algorithm 1 Chain-of-Thought Distillation with Preference Optimization (CoTD-PO)

Input: Instruction $x \in \mathcal{D}$, teacher model M_t, reward model r_t, student model p_s, number of iterations N, number of candidates K

for $i = 1$ **to** N **do**

(1) Teacher Rationale Generation:
Generate rationale $CoT_t \sim M_t(CoT \mid x)$

(2) Student Answer Sampling:
Sample K candidate answers $\{y_j\}_{j=1}^{K} \sim p_s(y \mid CoT_t, x)$

(3) Teacher Scoring:
Compute preference scores $r_t(y_j \mid x)$ for each $j = 1, \ldots, K$

(4) Student Update:
Compute total loss $\mathcal{L}_{\text{total}}$ using Eq. (9) Back-propagate gradient $\nabla_\theta \mathcal{L}_{\text{total}}$ and update p_s

$$\mathcal{L}_{\text{total}} = \underbrace{\mathbf{E}\left[-\sum_{i=1}^{I} \log p_s(CoT_t^i | CoT_t^{<i}, x)\right]}_{\text{Rationale NLL Loss}} \underbrace{-\alpha \sum_i \frac{e^{r_t(y_i|x)}}{\sum_j e^{r_t(y_j|x)}} \log\left(\frac{e^{\lambda \log p_s(y_i|CoT_t,x)}}{\sum_j e^{\lambda \log p_s(y_j|CoT_t,x)}}\right)}_{\mathcal{L}_{\text{apa}}} . \tag{9}$$

where the first term enforces faithful rationale generation via teacher-forced NLL, and the second aligns the answer distribution with teacher preferences. The balancing coefficient α controls the trade-off between reasoning fidelity and answer quality.

3 Experiments

3.1 Experimental Setup

Model In this study, we use GPT-4o (OpenAI, 2024) as the teacher model to generate rationales, which are then employed to guide the self-exploration process of the student models. The student models are Llama-3.1-8B-Instruct (Dubey et al., 2024) and Ministral-8B-Instruct-2410[2].To obtain the teacher's preference scores, we fine-tune the QRM-LLaMA3.1-8B reward model (Dorka, 2024) on teacher-labeled preference data, resulting in a dedicated reward model aligned with the teacher's implicit preferences. Detailed training settings and reward alignment procedures are provided in Appendix C.

Datasets Unlike prior CoT distillation works that primarily focus on classification tasks, we conduct our experiments on instruction-following tasks, which are more representative of real-world LLM usage scenarios. Specifically, we use training sets of UltraFeedback (Cui et al., 2023) as our dataset, which is a large-scale, fine-grained, diverse preference dataset collected about 64k prompts from diverse resources. During training, we randomly sampled 8,000 Ultra-Feedback prompts in each iteration.

Training Details All experiments were conducted on two NVIDIA A800 GPUs. For each training iteration, we generated K = 8 responses per prompt using temperature in [0.6, 1.2] and top-p sampling with p = 0.95, which ensured sufficient diversity for the student model to learn from multiple candidate outputs. To ensure the stability and convergence of our experimental results, we performed more than five iterations for each setting. We trained the models using the AdamW optimizer with a learning rate of 5e-5. The batch size was set to 2 per GPU, and gradient accumulation was applied over 32 steps to simulate a larger effective batch size. We used a linear learning rate scheduler with warm-up over the first 10% of training steps.

Evaluation To rigorously assess instruction-following capabilities, we evaluate models on three benchmarks: AlpacaEval 2.0 (Dubois et al., 2024) measures real-world task fidelity through 805 diverse user queries spanning multi-turn dialogues and creative tasks, employing GPT-4-turbo for automated win rate calculation; MT-Bench (Zheng et al., 2023) provides fine-grained analysis across seven categories (Writing, Roleplay, Reasoning, etc.) using 160 multi-turn prompts with GPT-4-based rubric scoring (0-10 scale) to quantify specialized competencies; Arena-Hard-Auto (Li et al., 2024), a benchmark consisting 500 challenging prompts curated by BenchBuilder. All benchmarks employ auto-evaluation using their respective default judge models.

Beyond task performance, we systematically evaluate output diversity through lexical (Distinct-N scores, N=1, 2 (Li et al., 2016)) and semantic

[2] https://huggingface.co/mistralai/Ministral-8B-Instruct-2410

Method	AlpacaEval 2 (LC)	Arena-Hard	MT-Bench
Baselines			
Specialized KD (Fu et al., 2023a)	18.59	26.3	7.03
MCCKD (Chen et al., 2023)	10.11	9.4	4.00
CasCoD (Dai et al., 2024)	20.53	21.1	7.71
Zephyr (Tunstall et al., 2023)	24.25	27.0	7.89
Initial Models			
Llama-3.1-8B-Instruct	20.85	21.3	6.94
Mistral-8B-Instruct-2410	32.34	23.4	7.25
Our Method (CoTD-PO)			
CoTD-PO + Llama-3.1-8B-Instruct	**53.80**	49.5	**8.54**
CoTD-PO + Mistral-8B-Instruct-2410	53.18	**52.4**	8.31
Black-Box Models			
GPT-4o (2024-05-13) (teacher-model)	**57.5**	**79.2**	-
GPT-4o-mini	50.7	74.9	-
GPT-4-0613	30.2	37.9	**9.18**

Table 1: Comparison of different methods on AlpacaEval 2 (LC), Arena-Hard, and MT-Bench. CoTD-PO demonstrates strong performance across all metrics.

(NLI diversity (Stasaski and Hearst, 2022)). For all diversity metrics, higher values indicate greater diversity. Using AlpacaEval 2.0 prompts, we generate 16 responses per prompt and assess diversity from dual perspectives: **(1) Intra-Prompt Diversity:** Measures variability across multiple responses to the same prompt. **(2) Inter-Prompt Diversity:** Evaluates output uniqueness across different prompts to detect global pattern overfitting. This dual-axis analysis ensures comprehensive characterization of both local creativity and global generalization.

Baselines As baselines, we compare our method with three representative CoT distillation approaches: Specialized KD (Fu et al., 2023a), MCCKD (Chen et al., 2023), and CasCoD (Dai et al., 2024). Although these methods were originally designed for classification tasks, we adapt them to generative instruction-following tasks using the same student models and dataset as in our setup to ensure a fair comparison. In addition, we include Zephyr (Tunstall et al., 2023), distilling via direct preference optimization (dDPO), as a preference-based distillation baseline.

3.2 Main Results

Generation Performance Our experimental results demonstrate the consistent effectiveness of our proposed method across various student models and evaluation benchmarks, as summarized in Table 1. Our approach achieves a peak win rate of 53.80% on AlpacaEval 2.0 LC, surpassing GPT-4o-mini by a margin of +3.1%. In Arena-Hard, our method not only significantly outperforms all distillation-based approaches but also exceeds the performance of GPT-4 (0613). While GPT-4 (0613) still leads on MT-Bench with a score of 9.18, our method narrows the gap considerably, achieving 8.54.

In our baseline methods, the training data for the answer component is sourced exclusively from the teacher model's output space. Specifically, Specialized KD trains the student via direct supervised fine-tuning (SFT) on the teacher's outputs, while Zephyr first uses a reward model to select preferred teacher-generated responses and then applies DPO. In contrast, our method shifts the training paradigm to the student model's own output space. This approach effectively mitigates the distribution mismatch between the training data and the student model, thereby addressing the issue of catastrophic forgetting.

In contrast to previous approaches that may improve performance on specific tasks at the cost of general capabilities (e.g., (Fu et al., 2023a)), our method achieves a better balance. As shown in the fine-grained breakdown in Figure 2 (a), CoTD-PO improves reasoning skills without degrading

performance on general tasks.

Diversity Analysis Figures 2 (b) and 2 (c) illustrate the model performance across various diversity metrics. We observe that compared to the original LLaMA-3.1-8B-Instruct model, our method (CoTD-PO) significantly improves lexical diversity (as indicated by Distinct-1/2) within the same prompt, while showing a slight reduction in semantic diversity (NLI Diversity). However, in terms of inter-prompt diversity, our method consistently improves across all metrics. This suggests that CoTD-PO not only enhances output diversity but also encourages greater consistency in generation.

We find that CasCoD boosts all diversity metrics significantly. Combined with its relatively weaker performance on instruction-following tasks, this implies a mode-averaging effect that the student distribution becomes overly smooth and fails to adequately capture the teacher's intent. This commonly occurs when aligning distributions with large support mismatches using forward KL divergence, resulting in underfitting.

Overall, our method addresses the issue faced by prior CoT distillation methods: the distribution mismatch between teacher outputs and the student's generative distribution. By leveraging both preference modeling and entropy regularization, CoTD-PO achieves superior performance on open-ended instruction-following tasks.

3.3 Ablation Study

We conduct rigorous ablation studies to validate hypotheses: (1) Aligning student outputs in their own generation space is superior to teacher-space alignment; (2) Teacher CoT rationales provide essential inductive biases beyond standard fine-tuning; (3) Our new preference alignment loss, by combining KL divergence and entropy regularization, not only ensures robust alignment with preference signals but also preserves output diversity.

Student-Space vs. Teacher-Space Answer Alignment One of the core contributions of CoTD-PO is aligning the student's answer distribution in its own generation space rather than in the teacher's space (Eq. 3). We ablate this component by switching back to teacher-space KL divergence, where answers are sampled from the teacher model. As shown in Table 2, this setup performs poorly on generation tasks. This verifies that forcing students to mimic teacher distributions induces distribution drift when their generation spaces diverge.

Teacher CoT Guidance Analysis When ablating teacher rationales requiring the student to autonomously generate both reasoning paths and final answers, with feedback only on outputs. As shown in Table 2, removing explicit teacher CoT guidance leads to significantly lower performance ceilings during optimization. This highlights the effectiveness of leveraging teacher rationales to guide the student's own reasoning process, demonstrating the feasibility and necessity of CoT based supervision.

KL-Based Preference Loss vs. Pairwise Loss Our main loss function is derived by combining KL divergence with softmax-normalized teacher preferences, incorporating entropy regularization to preserve diversity. To isolate its effectiveness, we compare it to the widely-used pairwise SimPO loss, which only optimizes over best/worst pairs. Our loss outperforms the pairwise alternative on semantic diversity (see figure 2(b,c)) while maintaining strong task performance Table 3, demonstrating that it better alleviates mode collapse and promotes diverse generation.

3.4 Gradient Analysis

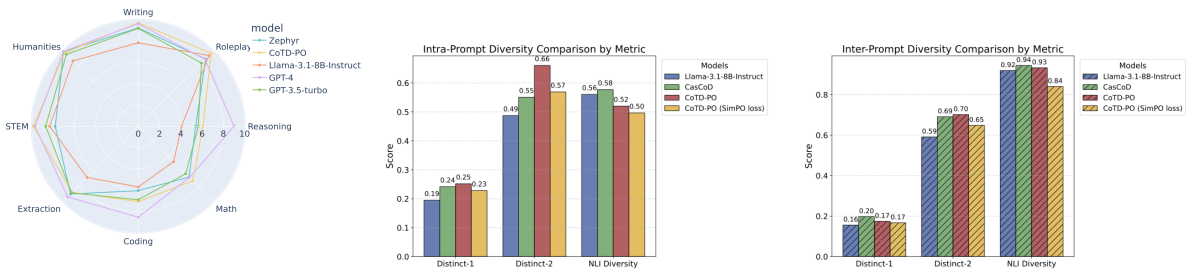
To deconstruct the learning dynamics of our alignment objective, we analyze the gradient of the answer-focused loss (Eq. (8)):

$$\nabla_\theta \mathcal{L}_{apa} = \lambda \Big(\underbrace{\mathbf{E}[\sum_i \nabla_\theta \log p_s(y_i)]}_{\text{Student Policy Calibration}} - \underbrace{\mathbf{E}[\nabla_\theta \tilde{p}_t \log p_s(y)]}_{\text{Teacher Signal}} \Big) \tag{10}$$

Component Analysis

- **Teacher Signal Term**: Maximizes likelihood of teacher-preferred outputs through $\tilde{p}_t$-weighted gradients.
- **Student Policy Calibration Term**: Subtracts expected gradient under student's sharpened distribution $p_s(y)$, acting as a self-normalizing control variate.

This gradient structure enables conservative policy updates that student distribution evolves smoothly toward teacher preferences while preserving diversity through entropy regularization.



(a) fine-grained capability analysis (b) Intra-prompt diversity metrics (c) Cross-prompt diversity profiles

Figure 2: Comprehensive evaluation of CoTD-PO: (a) fine-grained capability analysis, (b) Intra-prompt diversity metrics, (c) Cross-prompt diversity profiles.

Benchmark	Student-Space (Ours)	Teacher-Space	No Teacher CoT
AlpacaEval2(LC)	**53.80**	34.30	46.67
Arena-Hard	**49.5**	35.6	42.5
MT-Bench	**8.54**	7.64	8.28

Table 2: Ablation study comparing (1) output space alignment strategies, (2) with/without teacher CoT guidance.

Pairwise Preference Loss

While our primary objective in Equation (8) provides theoretically optimal distribution alignment, its computational complexity grows quadratically with candidate set size. We therefore analyze a lightweight alternative that distills only extremal preferences:

$$\mathcal{L}_{\text{pairwise}} = -\mathbf{E}_{(y^c,y^r)} \left[\log \sigma\left(\frac{\beta}{|y^c|}\log \pi_s(y^c) - \frac{\beta}{|y^r|}\log \pi_s(y^r) - \gamma\right)\right], \tag{11}$$

where $\pi_s(y)$ is the student's generation probability. This matches SimPO's objective (Meng et al., 2024) but reveals critical differences in gradient behavior:

$$\nabla_\theta \mathcal{L}_{\text{pairwise}} \propto -\sigma(-\Delta)\left(\frac{\nabla \log \pi_s(y^c)}{|y^c|} - \frac{\nabla \log \pi_s(y^r)}{|y^r|}\right) \tag{12}$$

where $\Delta = \frac{\beta \log \pi_s(y^c)}{|y^c|} - \frac{\beta \log \pi_s(y^r)}{|y^r|} - \gamma$. Compared to our full-distribution gradient (Eq. 10), the pairwise loss suffers from high-variance updates and often encourages convergence to narrow response patterns (i.e., mode collapse). In practice, we observe both the log-likelihoods of chosen and rejected responses tend to decrease during pairwise training. This mirrors reported in prior work. We addition NLL loss on the chosen samples, which stabilizes training and improves performance.

As shown in Table 3, pairwise optimization yields comparable results on generation performance. However, it significantly reduces semantic diversity, shown in Figure 2(b,c). This supports that our method effectively alleviates mode collapse and promotes output diversity.

Benchmark	CoTD-PO	Pairwise
AlpacaEval2(LC)	**53.80**	52.04
NLI Diversity	**0.520**	0.496

Table 3: Full vs. pairwise preference optimization. Despite comparable task performance, CoTD-PO better preserves semantic diversity (NLI).

4 Related Work

4.1 Chain-of-Thought Reasoning

The chain-of-thought (CoT) paradigm, first formalized by (Wei et al., 2022b), enables language models to decompose complex reasoning tasks into intermediate rationales before generating final answers. Subsequent work has explored diverse extensions of this paradigm, including prompting strategies for eliciting reasoning steps (Kojima et al., 2022), self-consistency mechanisms that aggregate multiple reasoning paths (Wang et al., b), and iterative self-improvement frameworks like STaR (Zelikman et al., 2022). The STaR method exemplifies a closed-loop refinement process: it initially generates rationales for batches of questions

using few-shot exemplars, re-generates erroneous rationales conditioned on known correct answers, and fine-tunes the model exclusively on verified reasoning traces.

However, empirical analyses reveal critical limitations in its applicability. While most studies (Zhou et al., 2023; Fu et al., 2023b; Hosseini et al., 2024), (Fu et al., 2023b), (Hosseini et al., 2024) and (Trung et al., 2024) focus on explicit reasoning domains (e.g., arithmetic and logic puzzles), recent evidence challenges its broader utility. A systematic meta-analysis by (Sprague et al., 2024) found that CoT substantially benefits only tasks requiring mathematical, logical, or algorithmic problem-solving. TPO (Wu et al., 2024) further confirms that simply prompting the model to articulate its thought process actually hurts performance on general instruction-following tasks. Yet, when integrated into training objectives, CoT principles not only enhance reasoning and problem-solving performance but also improve capabilities in non-traditional reasoning domains.

4.2 Knowledge Distillation with Rationales

Traditional knowledge distillation (Hinton et al., 2015) primarily focuses on transferring output-level knowledge from teacher models to student models. More recent work has extended this to rationale-aware distillation, aiming to transfer not only the final prediction but also the reasoning process behind it. Prior research (Wei et al., 2022b) has demonstrated that Chain-of-Thought (CoT) reasoning requires large models for optimal performance. Studies by (Ho et al., 2023), (Magister et al., 2023), and (Li et al., 2023) have shown that smaller models may not inherently generate CoT reasoning chains, but they can be trained to do so through the use of augmented training sets.

However, large language models (LLMs) are prone to producing hallucinations. These inconsistent rationales can be inherited by the student models. and the student model may treat rationale generation and answer prediction as two independent tasks, which can hinder performance. To address these issues, (Wang et al., 2023) proposed a self-consistent CoT method, and (Chen et al., 2024) employed Mutual Information to better align the reasoning process between teacher and student models. Previous work mainly focused on determined tasks (e.g., label classification tasks or tasks with definite answers, such as mathematical problems), where the output resides in a discrete space, while the generated rationales lie in a continuous space. This discrepancy can lead to inconsistency.

5 Conclusion

We present CoTD-PO, a novel RL framework that transforms chain-of-thought distillation from passive imitation to active exploration. By aligning student-teacher preferences instead of probabilities and enforcing entropy-regularized policy updates, our method effectively mitigates distribution mismatch while preserving generation diversity. Experiments across instruction-following benchmarks demonstrate state-of-the-art performance without mode collapse degradation.

Limitations

Although the method proposed in this study has produced promising outcomes, several limitations persist. Notably, the teacher models utilized are general-purpose language models, such as GPT-4, which are not specifically tailored for reasoning tasks. This dependency imposes a theoretical upper limit on performance, as general models may lack the logical rigor and complexity inherent in the reasoning steps generated by specialized chain-of-thought (CoT) optimizers like DeepSeek-R1 when producing CoT trajectories. Furthermore, it remains to be verified whether our distillation method is effective in distilling models akin to DeepSeek-R1.

Current implementation lacks formal verification of reasoning chain validity. This allows logical missteps (e.g., arithmetic errors in derivations) to persist undetected during distillation.

The proposed method, CoTD-PO, and its associated implementation are intended solely for research purposes.

Acknowledgments

We thank the anonymous reviewers for their insightful comments. The research is supported by the Natural Science Foundation of Beijing, China (Grant No. L247010).

References

Kushal Arora, Layla El Asri, Hareesh Bahuleyan, and Jackie Cheung. 2022. Why exposure bias matters: An imitation learning perspective of error accumulation in language generation. In *Findings of the Association for Computational Linguistics: ACL*

2022, pages 700–710, Dublin, Ireland. Association for Computational Linguistics.

Hongzhan Chen, Siyue Wu, Xiaojun Quan, Rui Wang, Ming Yan, and Ji Zhang. 2023. Mcc-kd: Multi-cot consistent knowledge distillation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6805–6820.

Xin Chen, Hanxian Huang, Yanjun Gao, Yi Wang, Jishen Zhao, and Ke Ding. 2024. Learning to maximize mutual information for chain-of-thought distillation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6857–6868, Bangkok, Thailand. Association for Computational Linguistics.

Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Wei Zhu, Yuan Ni, Guotong Xie, Zhiyuan Liu, and Maosong Sun. 2023. Ultrafeedback: Boosting language models with high-quality feedback.

Chengwei Dai, Kun Li, Wei Zhou, and Songlin Hu. 2024. Improve student's reasoning generalizability through cascading decomposed cots distillation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15623–15643.

Nicolai Dorka. 2024. Quantile regression for distributional reward models in rlhf. *Preprint*, arXiv:2409.10164.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B Hashimoto. 2024. Length-controlled alpacaeval: A simple way to debias automatic evaluators. *arXiv preprint arXiv:2404.04475*.

Kaituo Feng, Changsheng Li, Xiaolu Zhang, Jun Zhou, Ye Yuan, and Guoren Wang. 2024. Keypoint-based progressive chain-of-thought distillation for llms. In *Proceedings of the 41st International Conference on Machine Learning*, pages 13241–13255.

Yao Fu, Hao Peng, Litu Ou, Ashish Sabharwal, and Tushar Khot. 2023a. Specializing smaller language models towards multi-step reasoning. In *International Conference on Machine Learning*, pages 10421–10430. PMLR.

Yao Fu, Hao Peng, Ashish Sabharwal, Peter Clark, and Tushar Khot. 2023b. Complexity-based prompting for multi-step reasoning. In *The Eleventh International Conference on Learning Representations*.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. Minillm: Knowledge distillation of large language models. In *The Twelfth International Conference on Learning Representations*.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *Preprint*, arXiv:1503.02531.

Namgyu Ho, Laura Schmid, and Se-Young Yun. 2023. Large language models are reasoning teachers. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14852–14882, Toronto, Canada. Association for Computational Linguistics.

Arian Hosseini, Xingdi Yuan, Nikolay Malkin, Aaron Courville, Alessandro Sordoni, and Rishabh Agarwal. 2024. V-STar: Training verifiers for self-taught reasoners. In *First Conference on Language Modeling*.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Hojae Lee, Junho Kim, and SangKeun Lee. 2024. Mentor-kd: Making small language models better multi-step reasoners. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17643–17658.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and William B Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119.

Liunian Harold Li, Jack Hessel, Youngjae Yu, Xiang Ren, Kai-Wei Chang, and Yejin Choi. 2023. Symbolic chain-of-thought distillation: Small models can also "think" step-by-step. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2665–2679, Toronto, Canada. Association for Computational Linguistics.

Tianle Li, Wei-Lin Chiang, Evan Frick, Lisa Dunlap, Tianhao Wu, Banghua Zhu, Joseph E Gonzalez, and Ion Stoica. 2024. From crowdsourced data to high-quality benchmarks: Arena-hard and benchbuilder pipeline. *arXiv preprint arXiv:2406.11939*.

Chris Yuhao Liu, Liang Zeng, Jiacai Liu, Rui Yan, Jujie He, Chaojie Wang, Shuicheng Yan, Yang Liu, and Yahui Zhou. 2024. Skywork-reward: Bag of tricks for reward modeling in llms. *Preprint*, arXiv:2410.18451.

Lucie Charlotte Magister, Jonathan Mallinson, Jakub Adamek, Eric Malmi, and Aliaksei Severyn. 2023. Teaching small language models to reason. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1773–1781, Toronto, Canada. Association for Computational Linguistics.

Yu Meng, Mengzhou Xia, and Danqi Chen. 2024. SimPO: Simple preference optimization with a reference-free reward. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

OpenAI. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. 2016. Sequence level training with recurrent neural networks. *Preprint*, arXiv:1511.06732.

Zayne Sprague, Fangcong Yin, Juan Diego Rodriguez, Dongwei Jiang, Manya Wadhwa, Prasann Singhal, Xinyu Zhao, Xi Ye, Kyle Mahowald, and Greg Durrett. 2024. To cot or not to cot? chain-of-thought helps mainly on math and symbolic reasoning. *Preprint*, arXiv:2409.12183.

Katherine Stasaski and Marti A Hearst. 2022. Semantic diversity in dialogue with natural language inference. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 85–98.

Luong Trung, Xinbo Zhang, Zhanming Jie, Peng Sun, Xiaoran Jin, and Hang Li. 2024. ReFT: Reasoning with reinforced fine-tuning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7601–7614, Bangkok, Thailand. Association for Computational Linguistics.

Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, et al. 2023. Zephyr: Direct distillation of lm alignment. *arXiv preprint arXiv:2310.16944*.

Chaoqi Wang, Yibo Jiang, Chenghao Yang, Han Liu, and Yuxin Chen. a. Beyond reverse kl: Generalizing direct preference optimization with diverse divergence constraints. In *NeurIPS 2023 Workshop on Instruction Tuning and Instruction Following*.

Peifeng Wang, Zhengyang Wang, Zheng Li, Yifan Gao, Bing Yin, and Xiang Ren. 2023. SCOTT: Self-consistent chain-of-thought distillation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5546–5558, Toronto, Canada. Association for Computational Linguistics.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. b. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022a. Emergent abilities of large language models. *Preprint*, arXiv:2206.07682.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022b. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Tianhao Wu, Janice Lan, Weizhe Yuan, Jiantao Jiao, Jason Weston, and Sainbayar Sukhbaatar. 2024. Thinking llms: General instruction following with thought generation. *Preprint*, arXiv:2410.10630.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah Goodman. 2022. Star: Bootstrapping reasoning with reasoning. *Advances in Neural Information Processing Systems*, 35:15476–15488.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc V Le, and Ed H. Chi. 2023. Least-to-most prompting enables complex reasoning in large language models. In *The Eleventh International Conference on Learning Representations*.

A Derivation of KL Divergence Simplification

We provide the detailed derivation for Equation (5), starting from the KL alignment objective in Equation (3):

$$\mathop{\mathbf{E}}_{\substack{y\sim p_s \\ CoT\sim p_t}} \mathrm{KL}\left(p_t(y \mid CoT, x) \,\|\, p_s(y \mid CoT, x)\right)$$

In our framework, both the teacher and student answer distributions are modeled as softmax over their preference scores:

$$p_t(y_i \mid x) = \frac{e^{r_t(y_i|x)}}{\sum_j e^{r_t(y_j|x)}},$$

$$p_s(y_i \mid x) = \frac{e^{r_s(y_i|x)}}{\sum_j e^{r_s(y_j|x)}}.$$

the KL divergence between them is:

$$\mathrm{KL}(p_t \| p_s) = \sum_i p_t(y_i \mid x) \log \frac{p_t(y_i \mid x)}{p_s(y_i \mid x)}.$$

Expanding the logarithm, we have:

$$\begin{aligned}&\mathrm{KL}(p_t \| p_s) = \\ &\underbrace{\sum_i p_t(y_i \mid x) \log p_t(y_i \mid x)}_{-\mathcal{H}(p_t)} - \\ &\underbrace{\sum_i p_t(y_i \mid x) \log p_s(y_i \mid x)}_{\text{Cross-Entropy } \mathcal{H}(p_t, p_s)},\end{aligned}$$

where $\mathcal{H}(p_t)$ is the entropy of the teacher distribution, and $\mathcal{H}(p_t, p_s)$ is the cross-entropy.

During optimization, the teacher distribution p_t is fixed (i.e., parameters of r_t are frozen). Thus, $\mathcal{H}(p_t)$ becomes a constant term independent of the student's parameters θ. Consequently, minimizing $\mathrm{KL}(p_t \| p_s)$ is equivalent to minimizing the cross-entropy term:

$$\mathrm{KL}(p_t \| p_s) \propto -\sum_i p_t(y_i \mid x) \log p_s(y_i \mid x).$$

Substituting p_t and p_s with their softmax parameterizations:

$$\begin{aligned}&\mathrm{KL}(p_t \| p_s) \propto \\ &-\sum_i \frac{e^{r_t(y_i|x)}}{\sum_j e^{r_t(y_j|x)}} \log \left(\frac{e^{r_s(y_i|x)}}{\sum_j e^{r_s(y_j|x)}} \right)\end{aligned}$$

which matches Equation (5) in the main text.

B Derivation of Entropy-Regularized Optimal Policy

Given the entropy-regularized objective:

$$\max_{p_s} \mathbf{E}_{y \sim p_s} [r_s(y \mid x)] + \lambda H_s(y \mid CoT_t, x),$$

where $H_s(y \mid CoT_t, x) = -\mathbf{E}_{y \sim p_s} \log p_s(y \mid CoT_t, x)$, we reformulate it as a constrained optimization problem:

$$\begin{aligned}&\max_{p_s} \mathbf{E}_{y \sim p_s} [r_s(y \mid x) - \lambda \log p_s(y \mid CoT_t, x)] \\ &\text{s.t.} \quad \sum_y p_s(y \mid CoT_t, x) = 1.\end{aligned}$$

Introducing a Lagrangian multiplier η for the constraint:

$$\begin{aligned}\mathcal{L} = \mathbf{E}_{y \sim p_s} [r_s(y \mid x) - \lambda \log p_s(y \mid CoT_t, x)] \\ + \eta \left(1 - \sum_y p_s(y \mid CoT_t, x) \right).\end{aligned}$$

Taking the derivative with respect to $p_s(y \mid CoT_t, x)$ and setting it to zero:

$$\begin{aligned}&\frac{\partial \mathcal{L}}{\partial p_s(y \mid CoT_t, x)} = \\ &r_s(y \mid x) - \lambda \left(\log p_s(y \mid CoT_t, x) + 1 \right) - \eta = 0.\end{aligned}$$

Rearranging terms:

$$\log p_s(y \mid CoT_t, x) = \frac{r_s(y \mid x) - \eta - \lambda}{\lambda}.$$

Exponentiating both sides:

$$p_s(y \mid CoT_t, x) = \frac{e^{r_s(y|x)/\lambda}}{e^{(\eta+\lambda)/\lambda}} = \frac{e^{r_s(y|x)/\lambda}}{Z},$$

where $Z = \sum_y e^{r_s(y|x)/\lambda}$ is the partition function. Substituting back into the definition of r_s:

$$r_s(y \mid x) = \lambda \log p_s(y \mid CoT_t, x) + \lambda \log Z,$$

which matches Equation (7) in the main text.

C Reward Model Training Details

Base Model. We adopt **QRM-LLaMA3.1-8B** (Dorka, 2024) as our base reward model, which is a value head model initialized from the instruction-tuned LLaMA3.1-8B. The model is designed to score generated answers with scalar rewards, serving as the supervision signal for student model optimization.

Teacher-Aligned Preference Labels. To ensure that the reward model reflects the teacher model's preferences rather than general crowd sourced annotations, we relabel an existing preference dataset Skywork-Reward-Preference-80K-v0.2 (Liu et al., 2024), using the teacher model M_t as an automatic preference judge. For each prompt, we present M_t with two candidate completions and ask it to select the preferred one. This yields teacher-aligned preference pairs of the form (x, y^w, y^l).

Training Procedure. We fine-tune the reward model using a standard pairwise ranking loss:

$$\mathcal{L}_{\text{RM}} = -\log \sigma \left(r(y^w \mid x) - r(y^l \mid x) \right),$$

where $r(\cdot)$ is the scalar output of the reward model.