TeacherGRPO: Closing the Capacity Gap in Reasoning Distillation via Teacher Alignment

Anonymous ACL submission

Abstract

Reasoning distillation from powerful teacher models to smaller students faces the Gap Curse: as teachers grow more sophisticated, their complex distributions increasingly diverge from what students can approximate, causing performance degradation. Existing mitigation strategies either filter out challenging examples through data selection or introduce weaker intermediate assistant models, inherently compromising supervision coverage or quality. We propose Teacher Alignment, which directly adapts the teacher toward the student's distribution without discarding data or degrading reasoning quality. However, naive alignment through standard knowledge distillation triggers catastrophic collapse of the teacher's reasoning capabilities. To address this, we reformulate teacher alignment as reinforcement learning and introduce TeacherGRPO, built on Group Relative Policy Optimization with two key innovations: (i) Curriculum Selective Alignment applies dual token- and distribution-level curricula to focus rewards on high-signal reasoning gaps while filtering noise from trivial tokens and uncertain tail distributions, and (ii) Importance-Adaptive Length Regularization selectively penalizes verbose redundancy while preserving pedagogically critical reasoning steps. The aligned teacher then distills knowledge to students via standard pipelines. Extensive experiments show TeacherGRPO significantly outperforms baselines across diverse reasoning benchmarks and distillation methods. Our code is available at `https://anonymous.4open.science/r/TeacherGRPO-290F/README.md`.

1 Introduction

Large Language Models (LLMs) have achieved remarkable reasoning milestones (Xu et al., 2025), yet their computational cost remains a significant barrier for resource-constrained deployment (Bai et al., 2024). While reasoning distillation offers a promising path to transfer these capabilities to smaller models (Kang et al., 2023; Fu et al., 2023), it faces a counterintuitive hurdle known as the "Gap Curse" (Zhang et al., 2025; Zhu et al., 2022; Jiang et al., 2025b; Zhan et al.). As teacher models become increasingly sophisticated, their complex reasoning patterns diverge distributionally from what capacity-limited students can approximate, which leads to degraded student performance. Current approaches to mitigating this gap fall into two primary paradigms. The first employs *Data-Centric Selection* strategies, utilizing difficulty-aware metrics to filter samples with excessive teacher-student divergence (Liu and Zhang, 2025; Huang et al., 2025; Li et al., 2025). However, these methods often discard valuable challenging examples that could benefit learning. The second paradigm focuses on *Intermediate Supervision*, introducing intermediate teacher assistant models to bridge the complexity gap between teacher and student (Zhou and Ai, 2024; Lv et al., 2024; Ding et al., 2025). However, these assistants typically exhibit inferior reasoning

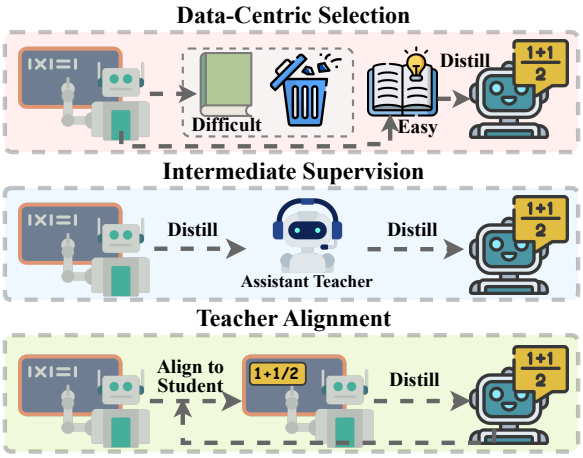

Figure 1: **Approaches to address the Gap Curse.** Data-Centric Selection discards potentially valuable examples while Intermediate Supervision dilutes supervision quality. Teacher Alignment adapts the teacher to student capacity, preserving both data and teaching quality.

capabilities compared to the original teacher, potentially propagating errors or suboptimal reasoning patterns to the student (Budiman et al., 2025).

As both paradigms inherently compromise supervision quality, recent knowledge distillation work suggests a third path (Qian et al., 2025): adapting the teacher toward the student's distribution via standard KD (Hinton et al., 2015), which we term *Teacher Alignment*. By shifting the teacher's distribution closer to the student without discarding high-quality data or replacing the original teacher with a weaker surrogate, this paradigm produces more accessible supervision signals while preserving the supervision quality at the source. However, naively applying Teacher Alignment to reasoning distillation introduces a critical challenge: reasoning ability is fragile and uniquely sensitive to distribution shift, making the teacher prone to collapse under aggressive fine-tuning (Li et al., 2024) – paradoxically *worsening* the supervision quality it was designed to protect. Our preliminary experiments (Section 2) confirm this fatal flaw: as the teacher is aligned toward the student's distribution, its reasoning capabilities rapidly deteriorate, rendering it an even less effective knowledge source.

Inspired by recent findings demonstrating that reinforcement learning (RL) can mitigate model collapse in LLMs (Kumar et al., 2024; Korbak et al., 2022), we reformulate Teacher Alignment as an RL problem. Unlike direct distillation, which forces rigid distribution mimicry, our approach treats the student's capacity as a reward signal. This allows the teacher to actively explore and discover reasoning paths that are digestible for the student, while maintaining its own logical correctness (Yue et al., 2025). However, directly applying RL to Teacher Alignment introduces two specific challenges:

(1) Diluted Alignment Signals. Standard KD rewards obscure effective supervision. (i) *Token-level*: The reward objective indiscriminately incentivizes the matching of trivial tokens (e.g. "The answer is") that are irrelevant to reasoning capability (Ye et al., 2025; Gao et al., 2024), preventing the teacher from understanding the genuine reasoning gaps where the student struggles. (ii) *Distribution-level*: The reward is contaminated by the long tail regions where both models are uncertain (Zhai et al., 2023). Considering rewards from these irrelevant low-probability tokens injects noise rather than meaningful knowledge gap.

(2) Length Exploitation. Capacity-constrained students are prone to error propagation in long chains (Ji et al., 2025), necessitating concise reasoning (Jiang et al., 2025a). However, naive alignment creates perverse incentives for verbosity: teachers can trivially minimize divergence by generating repetitive, low-information chains that mirror the student's uncertainty (Wang et al., 2025b). This effectively hacks the reward without improving pedagogical quality.

To address these challenges, we introduce TeacherGRPO, a framework based on Group Relative Policy Optimization (Shao et al., 2024) with two core innovations: First, to eliminate reward dilution, we propose **Curriculum Selective Alignment (CSA)**. Rather than optimizing against all tokens and full vocabulary, we employ a dual curriculum strategy: (i) *token-level curriculum* initially targeting the top-k tokens with highest teacher-student gaps instead of trivial ones, and (ii) *distribution-level curriculum* beginning with the top-K probability-divergence entries that capture core distributional differences, then gradually expanding coverage. This hierarchical approach filters out statistical noise, focusing reward solely on high-signal reasoning gaps. Second, to prevent length exploitation, we design **Importance-Adaptive Length Regularization (IALR)** to penalize over verbose reasoning paths. Unlike uniform penalties that indiscriminately penalize all steps that risks truncating critical reasoning steps (Jin et al., 2024), IALR scales penalty by each step's estimated importance. This selectively suppresses redundant reasoning steps while preserving pedagogically essential reasoning, enabling more effective knowledge transfer.

After Teacher Alignment with TeacherGRPO, we employ the optimized teacher to distill knowledge into student models using standard distillation techniques. Extensive experiments demonstrate that TeacherGRPO provides substantially superior supervision signals compared to both vanilla and KD-aligned teachers, yielding significant improvements in student performance across diverse reasoning benchmarks.

2 Preliminary Study

To motivate our approach, we investigate the feasibility and limitations of *Teacher Alignment*. Specifically, we aim to answer two questions: (1) whether aligning the teacher to the student's distribution improves downstream distillation, and (2) whether naive alignment strategies are sufficient.

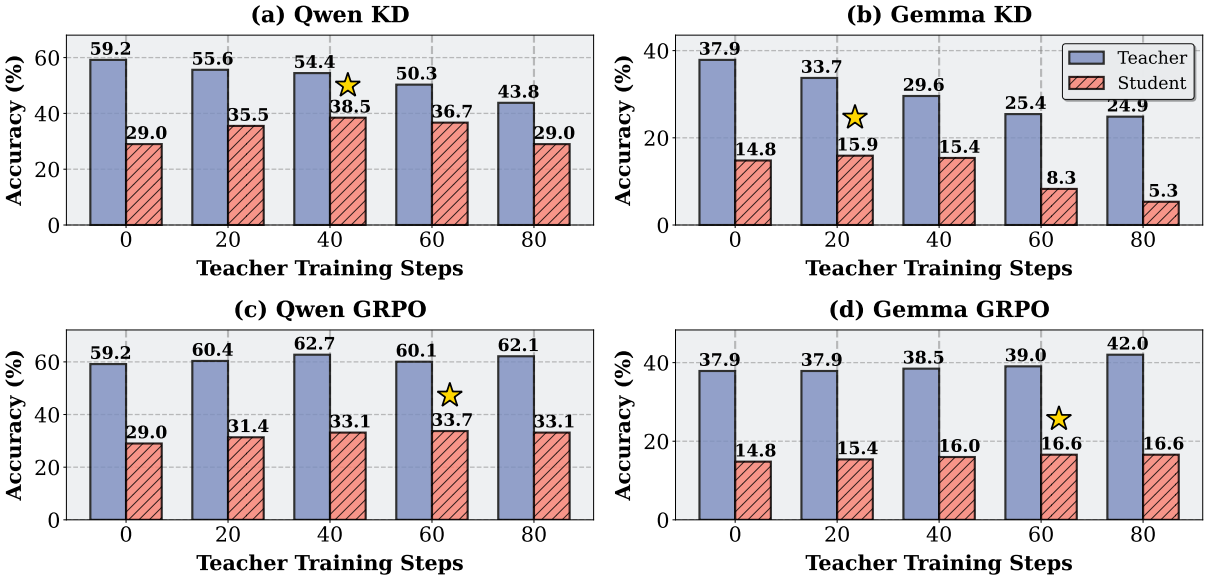

Figure 2: **Teacher and student performance during alignment.** Stars indicate peak student performance. KD-based alignment initially improves distillation but causes catastrophic teacher collapse. GRPO maintains teacher capability but achieves limited alignment effectiveness, motivating TeacherGRPO.

Experimental Setup. We compare two alignment methods on teacher models (`Qwen2.5-3B-Instruct` (Yang et al., 2025) and `Gemma-3-1b-it` (Team et al., 2025)) to align them with their student counterparts (`Qwen2.5-0.5B-Instruct` and `Gemma-3-270m-it`): (1) **KD**: minimizes KL divergence via logit-based knowledge distillation, and (2) **GRPO**: optimizes teacher distribution using GRPO with reverse KL divergence as rewards. Both methods train on Date Understanding dataset (Srivastava et al., 2023) for 80 steps (checkpoints at steps 0, 20, 40, 60, 80) with learning rate 5e-6 and 2e-5. We evaluate: (1) **Teacher Capability**—teacher validation accuracy, and (2) **Distillation Efficacy**—accuracy of students distilled from each teacher checkpoint.

Observation 1: Teacher Alignment Improves Distillation. Figure 2 shows that both KD and GRPO reduce teacher-student divergence and improve distillation. Specifically, Qwen student accuracy increases from 30.2% (baseline) to 38.5% (KD, step 40) and 33.1% (GRPO, step 80), confirming that teachers aligned to students' distributional characteristics provide better supervision.

Observation 2: KD Causes Catastrophic Capability Collapse. Beyond the optimal checkpoint (marked with stars), KD-aligned teachers suffer severe performance degradation. Specifically, Qwen teacher accuracy drops from 59.2% to 43.8% by step 80, while student's plummets from peak 38.5% to 29.0%. By forcing exact distribution mimicry, KD causes the teacher to overfit to the student's errors and inferior reasoning patterns, corrupting it as a knowledge source and degrading late-stage distillation performance.

Observation 3: GRPO Maintains Capability but Limits Alignment Effectiveness. GRPO-aligned teachers maintain stable reasoning capabilities throughout training. Qwen teacher accuracy remains at 58.5% at step 80, and Gemma maintains 36.2%. By treating alignment as reward optimization rather than distribution mimicry, GRPO explores reasoning paths accessible to students without sacrificing logical correctness. However, GRPO-distilled students achieve lower peak performance than KD. This limitation motivates TeacherGRPO to enhance GRPO's alignment effectiveness while preserving capability stability.

Conclusion. This study reveals a critical dilemma: we must bridge the distributional gap while preserving teacher reasoning capabilities. KD achieves strong alignment but causes catastrophic capability degradation. GRPO maintains capability stability but achieves insufficient alignment effectiveness. This motivates TeacherGRPO to enhance GRPO's alignment while preserving its capability stability.

3 Methodology

We propose **TeacherGRPO**, a reinforcement learning framework that aligns teacher reasoning ca-

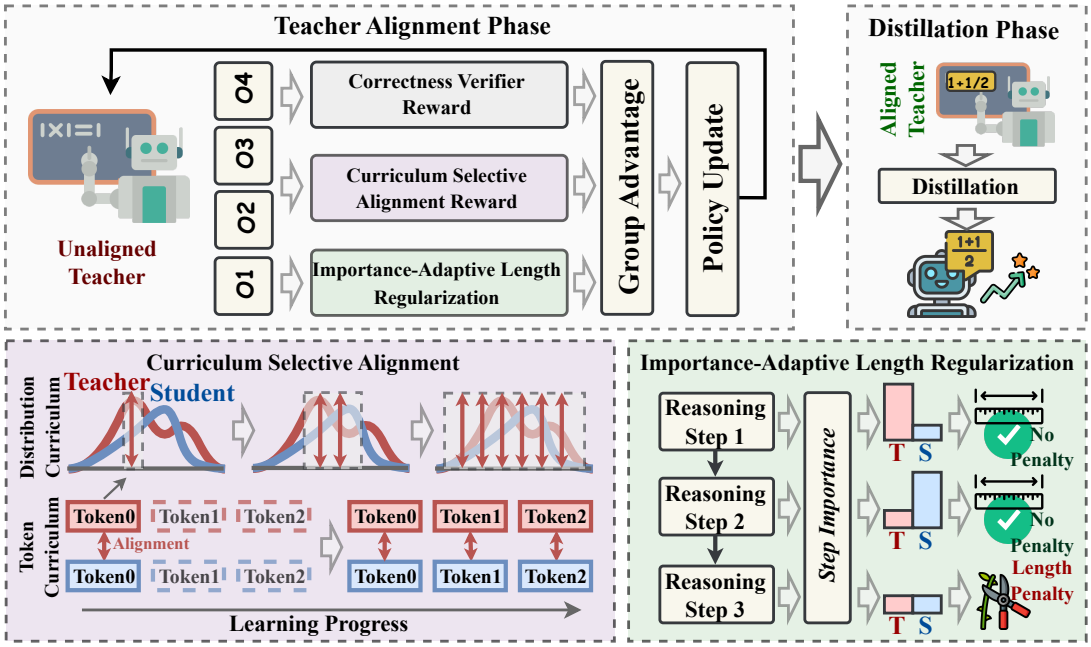

Figure 3: **TeacherGRPO framework overview.** The Alignment phase optimizes the teacher via GRPO with two novel reward CSA and IALR, plus a correctness verification. The aligned teacher then supervises student distillation.

pabilities with student capacity while preserving the teacher's superior reasoning abilities. Our approach addresses two fundamental challenges in GRPO-based Teacher Alignment: (1) diluted alignment signals from token-level and distribution-level noise, and (2) length exploitation that produces verbose, pedagogically ineffective outputs. We tackle these through two novel reward mechanisms: *Curriculum Selective Alignment (CSA)* and *Importance-Adaptive Length Regularization (IALR)*.

3.1 Problem Formulation

Given a teacher model with parameters θ generating outputs according to policy π_θ, and a student model with fixed policy π_S, our objective is to adapt π_θ to align with the reasoning ability of π_S within GRPO. Formally, we seek to optimize:

$$\theta^* = \arg\max_\theta \mathbb{E}_{x\sim\mathcal{D},\, y\sim\pi_\theta(\cdot|x)} \left[R_{align}(y;\pi_\theta,\pi_S)\right] - \lambda D_{KL}(\pi_\theta \,\|\, \pi_{ref})$$

where $\mathcal{D}$ denotes the dataset and D_{KL} denotes the KL penalty against reference policy π_{ref} that prevents capability collapse. The core innovation of TeacherGRPO lies in the design of R_{align}, which we decompose into CSA and IALR components.

3.2 Curriculum Selective Alignment (CSA)

Standard logit-based reward often suffer from signal dilution due to two sources of noise: *token-level noise*, arising from trivial tokens that obscure genuine reasoning gap, and *distribution-level noise*, stemming from non-informative probability mass in the vocabulary tail. To address this, we introduce Curriculum Selective Alignment, a dual-curriculum strategy that progressively filters these noise to focus on genuine reasoning discrepancies.

3.2.1 Token-Level Curriculum

Building on prior research indicating that distributional divergence often signal underlying reasoning disparities (Wang et al., 2025a), we focus the alignment on critical tokens with high KL divergence. For each token position t, we compute $D_t = D_{KL}(\pi_\theta(\cdot|y_{<t})\|\pi_S(\cdot|y_{<t}))$ where $y_{<t}$ denotes the teacher-generated history. To prioritize learning on these pivots, we generate a binary selection mask M_t, defined as $M_t = 1$ if the rank of D_t falls within the top k percentile of the sequence, and 0 otherwise. However, a static k is suboptimal, as it risks overlooking subtle reasoning gaps that become relevant as alignment improves. Therefore, we implement a dynamic curriculum where the selection threshold $k(\tau)$ expands linearly over training steps τ:

$$k(\tau) = k_{\min} + (k_{\max} - k_{\min}) \cdot \frac{\tau}{T} \qquad (1)$$

where T represents the total training steps. This progression ensures the teacher resolves fundamen-

tal reasoning disconnects (high D_t) before gradually fine-tuning on more subtle reasoning logic.

3.2.2 Distribution-Level Curriculum

Beyond token selection, the alignment signal is susceptible to noise from the vocabulary tail, where both models are uncertain and assign small probability. To mitigate high-variance gradients from these regions, we constrain the divergence calculation to $\mathcal{V}_p(y_t) = \{w \in \mathcal{V} \mid \text{rank}(|\pi_\theta(w|y_{<t}) - \pi_S(w|y_{<t})|) \leq p\}$, the subset of the top-$p$ entries with highest probability gap. Combined with token-level mask M_t, the final truncated alignment reward at position t is defined as:

$$R_{\text{vocab}}(y_t; p) = -\sum_{w \in \mathcal{V}_p(y_t)} \pi_\theta(w|y_{<t}) \log\left(\frac{\pi_\theta(w|y_{<t})}{\pi_S(w|y_{<t})}\right) * M \tag{2}$$

Analogous to the token-level strategy, we implement a progressive curriculum for the vocabulary coverage $p(\tau)$. This threshold linearly interpolates from a focused core to the full vocabulary $|\mathcal{V}|$ over the training course τ. This schedule ensures the teacher initially aligns with the student's dominant probability modes before adapting to the subtle uncertainties in the distributional tail.

3.3 Importance-Adaptive Length Regularization

Teacher models often exhibit a tendency toward verbosity. A common countermeasure is a uniform length penalty in the reward function, but this can inadvertently penalize essential reasoning steps necessary for effective student comprehension. To address this limitation, we propose **Importance-Adaptive Length Regularization (IALR)**, which selectively penalizes low-information verbosity while preserving critical reasoning pivots.

3.3.1 Step Importance Quantification

To identify and preserve critical steps, we firstly decompose the reasoning chain y into logical steps $S = \{s_1, s_2, \ldots, s_N\}$ based on structural delimiters (e.g., newlines). In reasoning tasks, step importance is commonly quantified through predictive uncertainty (Zeng et al., 2025): high-uncertainty steps represent critical decision points requiring careful reasoning, while low-uncertainty steps indicate routine continuations. In addition, for teacher alignment, a step's pedagogical value depends on both teacher and student: a step is essential if it represents (i) a critical reasoning pivot for the teacher, preserving logical coherence, or (ii) a challenging transition where the student must follow complex reasoning. Therefore, we define the *joint importance score* for step s_i as

$$\mathcal{I}_i = \sigma\left(\frac{\mathcal{H}\left(\pi_\theta(\cdot|y_{<t_i})\right) + \mathcal{H}\left(\pi_S(\cdot|y_{<t_i})\right)}{\eta}\right) \tag{3}$$

where $\mathcal{H}(\pi) = -\sum_w \pi(w) \log \pi(w)$ denotes entropy, t_i denotes the token index marking the start of step s_i, σ is the sigmoid function, and τ controls sensitivity. High joint entropy indicates pivotal reasoning points, whereas low entropy suggests routine elaboration.

3.3.2 Importance-Weighted Length Penalty

We then apply a length penalty that is inversely proportional to semantic importance, producing the IALR reward:

$$R_{\text{IALR}}(y) = -\sum_{i=1}^{N} \beta\, |s_i| \left(1 - \mathcal{I}_i\right)$$

where $|s_i|$ denotes the token count of step s_i and β controls the penalty's strength. This formulation penalizes long, low-entropy steps more heavily while imposing minimal penalties on concise or high-entropy reasoning, thereby yielding concise yet complete reasoning chains optimized for student learning.

3.4 TeacherGRPO Optimization

We integrate CSA and IALR into GRPO for stable and effective alignment. For each query x, we sample G outputs $\{y_1, \ldots, y_G\}$ from policy $\pi_{\theta_{old}}$. The total reward combines:

$$R_{total}(y_j) = \alpha_1 R_{CSA}(y_j) + \alpha_2 R_{SALR}(y_j) + \alpha_3 R_{ver}(y_j)$$

where $R_{ver}(y_j) \in \{0, 1\}$ is a correctness verifier, and $\{\alpha_1, \alpha_2, \alpha_3\}$ balance alignment, compression, and correctness. We compute group-normalized advantages and optimize:

$$\begin{aligned}\mathcal{L}(\theta) = -\frac{1}{G}\sum_{j=1}^{G}\Bigg[\frac{1}{|y_j|}\sum_{t=1}^{|y_j|}\min\Big(r_t(\theta)A_j, \text{clip}(r_t(\theta),\\ 1-\epsilon_{clip}, 1+\epsilon_{clip})A_j\Big)\Bigg] + \gamma D_{KL}(\pi_\theta||\pi_{ref})\end{aligned} \tag{4}$$

where $r_t(\theta) = \frac{\pi_\theta(y_{j,t}|y_{j,<t})}{\pi_{\theta_{old}}(y_{j,t}|y_{j,<t})}$, $A_j = \frac{R_{total}(y_j) - \mu_G}{\sigma_G + \epsilon}$ is the advantage normalized by the mean μ_G and standard deviation σ_G of R_{total} with a small $\epsilon > 0$ for numerical stability, and π_{ref} is the original teacher. The clip prevents destructive updates, while the KL preserves teacher capabilities.

Table 1: Performance comparison across four reasoning benchmarks with two model families: Qwen and Gemma. Best average values are in bold.

Method	Qwen2.5 (3B → 0.5B)					Gemma3 (1B → 270M)				
	Date	**SQA**	**ARC**	**CQA**	**Avg**	**Date**	**SQA**	**ARC**	**CQA**	**Avg**
Teacher										
Base	58.6	59.8	81.7	71.5	67.9	37.9	51.1	47.9	47.3	46.1
+ *TeacherKD*	53.9	55.5	79.2	61.2	62.5	31.5	47.2	46.3	44.3	43.8
+ ***TeacherGRPO***	59.2	64.6	81.2	72.7	**69.4**	40.2	51.5	48.3	47.1	**46.8**
Student										
Base	20.1	35.8	37.4	39.1	33.1	13.6	28.8	12.5	13.9	17.2
KD	29.0	39.7	40.7	47.3	39.2	14.8	35.4	24.7	21.5	24.1
+ *TeacherKD*	38.5	47.6	45.4	26.8	39.6	15.4	38.4	24.1	21.0	24.7
+ ***TeacherGRPO***	40.2	44.5	45.2	51.9	**45.5**	17.8	42.8	25.7	22.1	**27.1**
SeqKD	34.6	38.4	36.4	40.3	37.4	15.4	32.8	23.1	20.5	23.0
+ *TeacherKD*	38.5	43.3	34.1	41.4	39.3	14.8	36.2	23.4	20.7	23.8
+ ***TeacherGRPO***	38.8	42.7	41.3	41.7	**41.1**	17.8	36.2	25.4	23.1	**25.6**
Adaptive Off-Policy	20.7	38.0	41.9	46.4	36.8	16.6	40.2	22.3	16.8	24.0
+ *TeacherKD*	26.0	39.7	38.1	28.3	33.0	8.9	47.2	19.9	19.7	23.9
+ ***TeacherGRPO***	30.2	42.4	44.6	48.3	**41.4**	20.7	46.3	23.9	18.8	**27.4**

Once the teacher is aligned using this procedure, we distill its knowledge into the student using standard distillation techniques. The resulting teacher provides concise, pedagogically valuable supervision, leading to improved student performance.

4 Experiments

We conduct experiments to answer five research questions: **RQ1:** How does TeacherGRPO perform across different distillation methods? **RQ2:** What is the contribution of each component in TeacherGRPO? **RQ3:** How does TeacherGRPO compare to alternative paradigms for addressing the Gap Curse? **RQ4:** How does CSA improve alignment effectiveness? **RQ5:** How does IALR balance alignment and reasoning quality?

4.1 Experimental Setup

Unless otherwise stated, we evaluate TeacherGRPO across four reasoning benchmarks: Date Understanding (Date) (Srivastava et al., 2023), StrategyQA (SQA) (Geva et al., 2021), ARC-Challenge (ARC) (Clark et al., 2018), and CommonsenseQA (CQA) (Talmor et al., 2019). We experiment with two model families: Qwen (teacher: Qwen2.5-3B-Instruct, student: Qwen2.5-0.5B-Instruct) and Gemma (teacher: Gemma-3-1b-it, student: Gemma-3-270m-it). We benchmark our approach against baselines without Teacher Alignment and with KD-based alignment, alongside several standard distillation methods: Knowledge Distillation (KD) (Hinton et al., 2015), Sequence-Level KD (SeqKD) (Kim and Rush, 2016), and Adaptive Off-Policy Distillation (Ko et al., 2024). Comprehensive implementation details are provided in Appendix C.

4.2 Experiment Results and Analyses

Main Results. To address **RQ1**, we compare TeacherGRPO against vanilla distillation and KD-based Teacher Alignment (TeacherKD) across four reasoning benchmarks and four distillation methods using Qwen and Gemma model families. We have three key observations from Table 1:

(1) **TeacherGRPO consistently demonstrates robust performance gains.** For the Qwen family, TeacherGRPO achieves average improvements of 7.4 points over vanilla distillation methods, while the improvements average 3.6 points for the Gemma family. These consistent gains demonstrate TeacherGRPO's effectiveness at aligning teacher reasoning to student capacity, enabling smaller models to achieve substantially improved reasoning performance.

(2) **TeacherKD suffers from unstable performance.** While TeacherKD occasionally achieves strong results, where on Gemma with KD where it reaches 47.6 on StrategyQA (even 3.1 points above TeacherGRPO), it catastrophically fails in other settings. For instance, on Qwen with KD distilla-

Table 2: Ablation results on Date and ARC with Qwen. Best values are in bold.

Method	Date	ARC	Avg
KD			
TeacherGRPO	40.2	45.2	**42.7**
w/o CSA	34.9	43.6	39.3
w/o IALR	36.4	42.8	39.6
SeqKD			
TeacherGRPO	38.8	41.3	**40.1**
w/o CSA	36.7	37.8	37.3
w/o IALR	35.9	34.8	35.4
Adaptive Off-Policy			
TeacherGRPO	30.2	44.6	37.4
w/o CSA	28.7	44.1	36.4
w/o IALR	29.6	46.1	**37.9**

Table 3: Performance comparison of different paradigms on Date and ARC with Qwen.

Method	Date	ARC	Avg
KD			
+ *Data-Centric Selection*	32.5	43.9	38.2
+ *Intermediate Supervision*	34.9	41.2	38.1
+ ***TeacherGRPO***	40.2	45.2	**42.7**
SeqKD			
+ *Data-Centric Selection*	36.1	40.6	38.4
+ *Intermediate Supervision*	36.7	41.9	39.3
+ ***TeacherGRPO***	38.8	41.3	**40.1**
Adaptive Off-Policy			
+ *Data-Centric Selection*	26.6	42.2	34.4
+ *Intermediate Supervision*	27.2	41.8	34.5
+ ***TeacherGRPO***	30.2	44.6	**37.4**

tion, TeacherKD drops 20.5 point from the vanilla baseline of 47.3. This instability stems from the collapse phenomenon identified in Section 2.

(3) **TeacherGRPO preserves teacher reasoning capabilities while TeacherKD degrades them.** TeacherKD-aligned teachers consistently underperform the base teacher: Qwen accuracy drops 5.4 points, and Gemma drops 2.3 points. In contrast, TeacherGRPO-aligned teachers maintain or even improve performance. This preservation ensures the aligned teacher remains a high-quality knowledge source throughout distillation.

Ablation Study. To address **RQ2**, we systematically ablate CSA and IALR components and evaluate their individual contributions on Date Understanding and ARC-Challenge using the Qwen model family. We observe in Table 2 that both components are essential for TeacherGRPO's performance, where ablating CSA decreases performance by 2.4 on average, and removing IALR leads to drops of 2.4 on average. Notably, the impact of each component varies by distillation method: CSA proves more critical for logit-based methods (KD) where distribution-level alignment is paramount, while IALR shows greater importance for SeqKD where sequence length affects training data quality.

Paradigm Comparison. To address **RQ3**, we compare TeacherGRPO against two alternative paradigms for addressing the Gap Curse on Date Understanding and ARC-Challenge using the Qwen family. For Data-Centric Selection, we filter out 10% the most challenging examples (Liu and Zhang, 2025). For Intermediate Supervision, we introduce an assistant teacher (Qwen2.5-1.5B) (Zhou and Ai, 2024). From the results in Table 3, we observe that TeacherGRPO consistently outperforms both alternatives, with average improvements of 4.3 points over Data-Centric Selection and 3.4 points over Intermediate Supervision. These results demonstrate that well-designed Teacher Alignment without discarding valuable training examples or relying on weaker intermediate models, provides a more effective solution to the Gap Curse.

CSA Analysis. To address **RQ4**, we compare TeacherGRPO's curriculum-based alignment against static strategies that use only the top 30% or bottom 30% of tokens and distribution entries ranked by teacher-student divergence. Table 4 presents results on Date Understanding with Qwen using KD distillation. The curriculum approach substantially outperforms both static strategies, demonstrating its effectiveness at progressively reducing noise while preserving information. Notably, the bottom 30% performs dramatically worse, confirming that low-divergence regions hardly contain beneficial alignment signals. To illustrate CSA's mechanism, Figure 4 visualizes token-level KL divergence for a sample reasoning path. We observe that high-divergence tokens consistently correspond to reasoning-critical elements. For example, "current" marks a pivotal intermediate state, while connectives like "but" and "and" signal logical transitions between reasoning steps.

IALR Analysis. To address **RQ5**, we compare TeacherGRPO's importance-adaptive length regularization against two baselines: uniform length penalty and no length penalty. We track both student performance and mean generation length during Teacher Alignment on Date Understanding with Qwen using KD distillation. Figure 5(a)

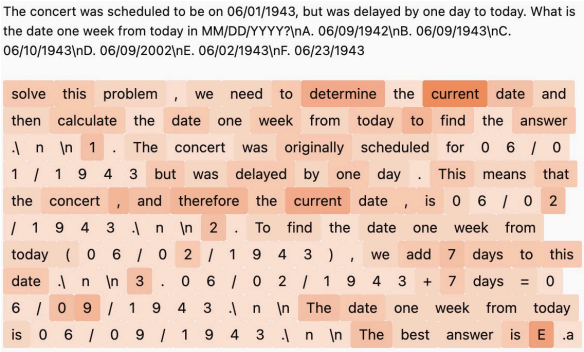

Figure 4: **KL divergence visualization for a sample path.** Darker tokens correspond to high-divergence.

Table 4: Student performance using top 30%, bottom 30%, or curriculum-based token and distribution selection on Date Understanding with Qwen.

Method	Token	Distribution
Curriculum	40.2	40.2
Top 30%	36.1	35.5
Bottom 30%	30.2	27.2

shows that TeacherGRPO outperforms both uniform penalty and no penalty, demonstrating IALR's effectiveness at preventing length exploitation without eliminating essential reasoning steps. The alignment dynamics in Figure 5(b) provide further validation, where IALR achieves gradual, controlled reduction, while no penalty increases verbosity and uniform penalty causes abrupt compression.

5 Related Work

Gap Curse in Knowledge Distillation. Knowledge distillation (Hinton et al., 2015) transfers capabilities from large language models to smaller ones by minimizing KL divergence between output distributions (Xu et al., 2024). Recent advances include white-box methods such as On-Policy Distillation (Agarwal et al., 2024) and Adaptive Off-Policy Distillation (Ko et al., 2024), alongside black-box methods like SeqKD (Kim and Rush, 2016). Despite these innovations, distillation methods fail to escape the Gap Curse: increasingly powerful teachers degrade rather than improve student performance (Zhang et al., 2025; Zhu et al., 2022; Jiang et al., 2025b), stemming from capacity mismatch (Ji et al., 2025). Current solutions employ either *Data-Centric Selection*, utilizing easier examples (Liu and Zhang, 2025; Huang et al., 2025; Li et al., 2025; Just et al., 2025), or *Intermediate Supervision*, introducing assistant models (Zhou and Ai, 2024; Lv et al., 2024; Ding et al., 2025). Both paradigms sacrifice supervision breadth or quality (Budiman et al., 2025).

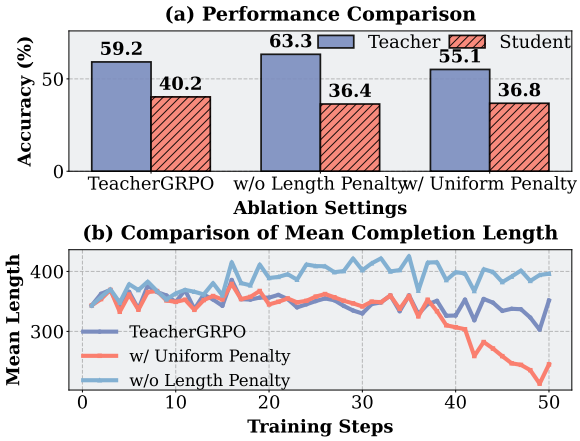

Figure 5: **Impact of length penalty.** (a) Student performance and (b) generation length during Alignment.

Teacher-Side Solution for Gap Curse. In vision tasks, an alternative perspective treats the gap curse as a teacher-side problem requiring adaptation. Methods such as reverse knowledge distillation (Nasser et al., 2024), pruning (Park and No, 2022), and student-friendly distillation (Yuan et al., 2024) create "easier" teachers in vision tasks. However, these methods are not suitable for LLM reasoning distillation as they risk catastrophic collapse (Li et al., 2024), degrading logical correctness as teachers overfit to student errors. In this work, TeacherGRPO reframes Teacher Alignment as an RL problem with specialized rewards that preserve pedagogical value, reviving alignment-based methods for reasoning distillation.

6 Conclusion

We introduced Teacher Alignment, a paradigm that adapts teachers to match student capacity rather than discarding data or relying on weaker intermediaries. To address the catastrophic collapse of naive alignment methods, we presented TeacherGRPO, a reinforcement learning framework built on GRPO with two key innovations: CSA filters token- and distribution-level noise to focus on high-signal reasoning gaps, while IALR penalizes verbosity without truncating essential reasoning. Extensive experiments across diverse benchmarks and distillation methods demonstrate the effectiveness of TeacherGRPO. Our work establishes Teacher Alignment as an effective solution to the Gap Curse, enabling more effective knowledge transfer from powerful reasoning models.

Limitations

Model Scale. Our work focuses on the small-model reasoning distillation regime, where the Gap Curse is most pronounced and efficient deployment matters most in practice. Within this scope, we validate TeacherGRPO on two distinct model families (Qwen and Gemma) with different architectures and compression ratios (6× and 3.7×), providing evidence of cross-architecture generalizability. That said, extending experiments to larger teacher-student pairs (e.g., 7B→1.5B or beyond) would further characterize how TeacherGRPO scales with model capacity, which we leave to future work given current computational constraints.

Task Coverage. Following standard evaluation protocols in recent reasoning distillation work (Liu and Zhang, 2025; Zhou and Ai, 2024; Ko et al., 2024), we evaluate on four widely-adopted benchmarks spanning temporal, multi-hop, scientific, and commonsense reasoning. The Gap Curse phenomenon we address is most clearly characterized in tasks with verifiable answers, making our chosen benchmarks an appropriate testbed for the core contribution. However, broader task coverage remains a natural direction for future work.

Computational Overhead. TeacherGRPO introduces an additional alignment phase before standard distillation. However, this cost is amortized across all downstream students distilled from the same aligned teacher, while our LoRA-based implementation keeps the overhead modest. Reducing the overhead through more efficient alignment objectives is an interesting direction for further optimization and can make it more practical.

References

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. 2024. On-policy distillation of language models: Learning from self-generated mistakes. In *The twelfth international conference on learning representations*.

Guangji Bai, Zheng Chai, Chen Ling, Shiyu Wang, Jiaying Lu, Nan Zhang, Tingwei Shi, Ziyang Yu, Mengdan Zhu, Yifei Zhang, and 1 others. 2024. Beyond efficiency: A systematic survey of resource-efficient large language models. *arXiv preprint arXiv:2401.00625*.

Setiawan Budiman, Siti Rochimah, and Daniel O Siahaan. 2025. The survey of model scale impact on knowledge distillation efficiency in llm. In *2025 International Electronics Symposium (IES)*, pages 771–776. IEEE.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

Dongyi Ding, Tiannan Wang, Chenghao Zhu, Meiling Tao, Yuchen Eleanor Jiang, and Wangchunshu Zhou. 2025. Micota: Bridging the learnability gap with intermediate cot and teacher assistants. *arXiv preprint arXiv:2507.01887*.

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

Jiaxuan Gao, Shusheng Xu, Wenjie Ye, Weilin Liu, Chuyi He, Wei Fu, Zhiyu Mei, Guangju Wang, and Yi Wu. 2024. On designing effective rl reward at training time for llm reasoning. *arXiv preprint arXiv:2410.15115*.

Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. 2021. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. *Transactions of the Association for Computational Linguistics*, 9:346–361.

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

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Haiduo Huang, Jiangcheng Song, Yadong Zhang, and Pengju Ren. 2025. Selectkd: Selective token-weighted knowledge distillation for llms. *arXiv preprint arXiv:2510.24021*.

Mingxin Ji, Kehui Song, Rize Jin, Xuguang Qiu, Hongbo Zhao, and Weiwei Yang. 2025. From chain to loop: Improving reasoning capability in small language models via loop-of-thought. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 294–306. Springer.

Wangyi Jiang, Yaojie Lu, Hongyu Lin, Xianpei Han, and Le Sun. 2025a. Teach small models to reason by curriculum distillation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 7423–7433.

Yuxuan Jiang, Dawei Li, and Frank Ferraro. 2025b. Drp: Distilled reasoning pruning with skill-aware step decomposition for efficient large reasoning models. *arXiv preprint arXiv:2505.13975*.

Mingyu Jin, Qinkai Yu, Dong Shu, Haiyan Zhao, Wenyue Hua, Yanda Meng, Yongfeng Zhang, and Mengnan Du. 2024. The impact of reasoning step length on large language models. *arXiv preprint arXiv:2401.04925*.

Hoang Anh Just, Myeongseob Ko, and Ruoxi Jia. 2025. Distilling reasoning into student llms: Local naturalness for selecting teacher data. *arXiv preprint arXiv:2510.03988*.

Minki Kang, Seanie Lee, Jinheon Baek, Kenji Kawaguchi, and Sung Ju Hwang. 2023. Knowledge-augmented reasoning distillation for small language models in knowledge-intensive tasks. *Advances in Neural Information Processing Systems*, 36:48573–48602.

Yoon Kim and Alexander M Rush. 2016. Sequence-level knowledge distillation. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pages 1317–1327.

Jongwoo Ko, Sungnyun Kim, Tianyi Chen, and Se-Young Yun. 2024. Distillm: Towards streamlined distillation for large language models. *arXiv preprint arXiv:2402.03898*.

Tomasz Korbak, Hady Elsahar, Germán Kruszewski, and Marc Dymetman. 2022. On reinforcement learning and distribution matching for fine-tuning language models with no catastrophic forgetting. *Advances in Neural Information Processing Systems*, 35:16203–16220.

Aviral Kumar, Vincent Zhuang, Rishabh Agarwal, Yi Su, John D Co-Reyes, Avi Singh, Kate Baumli, Shariq Iqbal, Colton Bishop, Rebecca Roelofs, and 1 others. 2024. Training language models to self-correct via reinforcement learning. *arXiv preprint arXiv:2409.12917*.

Ping Li, Chenhao Ping, Wenxiao Wang, and Mingli Song. 2025. Sample-level adaptive knowledge distillation for action recognition. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 544–552.

Songze Li, Tonghua Su, Xu-Yao Zhang, and Zhongjie Wang. 2024. Continual learning with knowledge distillation: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 36(6):9798–9818.

Lingyuan Liu and Mengxiang Zhang. 2025. Less is more: Selective reflection for compatible and efficient knowledge distillation in large language models. *arXiv preprint arXiv:2508.06135*.

Bo Lv, Xin Liu, Kaiwen Wei, Ping Luo, and Yue Yu. 2024. Taekd: Teacher assistant enhanced knowledge distillation for closed-source multilingual neural machine translation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15530–15541.

Sahar Almahfouz Nasser, Nihar Gupte, and Amit Sethi. 2024. Reverse knowledge distillation: Training a large model using a small one for retinal image matching on limited data. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 7778–7787.

Jinhyuk Park and Albert No. 2022. Prune your model before distill it. In *European Conference on Computer Vision*, pages 120–136. Springer.

Chengyao Qian, Trung Le, and Mehrtash Harandi. 2025. A good teacher adapts their knowledge for distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1239–1248.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, and 1 others. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on machine learning research*.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. Commonsenseqa: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, and 1 others. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Jingyuan Wang, Yankai Chen, Zhonghang Li, and Chao Huang. 2025a. Lightreasoner: Can small language models teach large language models reasoning? *arXiv preprint arXiv:2510.07962*.

Xinpeng Wang, Nitish Joshi, Barbara Plank, Rico Angell, and He He. 2025b. Is it thinking or cheating? detecting implicit reward hacking by measuring reasoning effort. *arXiv preprint arXiv:2510.01367*.

Fengli Xu, Qianyue Hao, Zefang Zong, Jingwei Wang, Yunke Zhang, Jingyi Wang, Xiaochong Lan, Jiahui Gong, Tianjian Ouyang, Fanjin Meng, and 1 others. 2025. Towards large reasoning models: A survey of reinforced reasoning with large language models. *arXiv preprint arXiv:2501.09686*.

Xiaohan Xu, Ming Li, Chongyang Tao, Tao Shen, Reynold Cheng, Jinyang Li, Can Xu, Dacheng Tao,

and Tianyi Zhou. 2024. A survey on knowledge distillation of large language models. *arXiv preprint arXiv:2402.13116*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Ziang Ye, Zhenru Zhang, Yang Zhang, Jianxin Ma, Junyang Lin, and Fuli Feng. 2025. Disentangling reasoning tokens and boilerplate tokens for language model fine-tuning. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 20939–20957.

Mengyang Yuan, Bo Lang, and Fengnan Quan. 2024. Student-friendly knowledge distillation. *Knowledge-Based Systems*, 296:111915.

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

Wenhao Zeng, Yaoning Wang, Chao Hu, Yuling Shi, Chengcheng Wan, Hongyu Zhang, and Xiaodong Gu. 2025. Pruning the unsurprising: Efficient code reasoning via first-token surprisal. *arXiv preprint arXiv:2508.05988*.

Yuanzhao Zhai, Han Zhang, Yu Lei, Yue Yu, Kele Xu, Dawei Feng, Bo Ding, and Huaimin Wang. 2023. Uncertainty-penalized reinforcement learning from human feedback with diverse reward lora ensembles. *arXiv preprint arXiv:2401.00243*.

Weixiao Zhan, Yongcheng Jing, Leszek Rutkowski, and Dacheng Tao. What makes large language models undistillable?

Chen Zhang, Qiuchi Li, Dawei Song, Zheyu Ye, Yan Gao, and Yao Hu. 2025. Towards the law of capacity gap in distilling language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 22504–22528.

Yuhang Zhou and Wei Ai. 2024. Teaching-assistant-in-the-loop: Improving knowledge distillation from imperfect teacher models in low-budget scenarios. *arXiv preprint arXiv:2406.05322*.

Yichen Zhu, Ning Liu, Zhiyuan Xu, Xin Liu, Weibin Meng, Louis Wang, Zhicai Ou, and Jian Tang. 2022. Teach less, learn more: On the undistillable classes in knowledge distillation. *Advances in Neural Information Processing Systems*, 35:32011–32024.

A Case Study for IALR

To illustrate the impact of IALR, we present a qualitative comparison of teacher generations across four configurations: (1) **Vanilla** (no teacher alignment), (2) **TeacherGRPO**, (3) **Uniform Penalty**, and (4) **w/o IALR**.

As shown in Figure 6, **TeacherGRPO**'s generation demonstrates optimal balance between conciseness and reasoning completeness. Compared to the vanilla teacher, TeacherGRPO produces responses that are: (1) more structured, explicitly formatting the answer for clarity (e.g. "convert this date into the MM/DD/YYYY format"), and (2) more direct in their reasoning chain (e.g., "subtract one year" versus "go back another year"). These characteristics make the reasoning path more accessible for student learning while preserving logical correctness.

In contrast, the **Uniform Penalty** approach produces overly compressed outputs that sacrifice essential reasoning steps. This aggressive truncation eliminates critical intermediate logic, ultimately leading to an incorrect answer.

The **w/o IALR** variant exhibits the opposite problem: excessive verbosity without pedagogical value. While it systematically examines each answer choice with detailed explanations, the resulting response is unnecessarily long and difficult for capacity-constrained students to process effectively. This demonstrates that length regularization without importance weighting fails to distinguish between essential reasoning steps and redundant elaboration.

These examples validate IALR's core mechanism: by adaptively penalizing based on step importance, TeacherGRPO produces reasoning that are concise yet complete, optimizing for student digestibility without compromising logical integrity.

B Training Stability Analysis

To evaluate the optimization stability of TeacherGRPO, we analyze the training dynamics during the Teacher Alignment phase on the Date Understanding dataset. Figure 7 presents the training reward (left) and its standard deviation (right), both smoothed using exponential moving average with weight 0.9.

We observe two key indicators of stable convergence: (1) **Steady reward improvement**: The mean reward increases consistently from approximately -2.6 to -1.5 over 50 training steps, demon-

This is the last day of 1899. What is the date one year ago from today in MM/DD/YYYY?\nA. 12/31/1898\nB. 02/11/1899\nC. 12/25/1898\nD. 12/03/1898\nE. 01/11/1899

Vanilla

To solve this problem, we need to determine the date that is one year before the last day of 1899, which is January 1st, 1900.\n\n1. The last day of 1899 is December 31, 1899.\n2. One year before this date will be December 31, 1898.\n3. To find the date one year before this date, we need to go back another year, making it December 31, 1897.\n4. However, the question asks for the date one year before the last day of 1899, which is December 31, 1898.\n\nGiven the options:\nA. 12/31/1898\nB. 02/11/1899\nC. 12/25/1898\nD. 12/03/1898\nE. 01/11/1899\n\nThe correct answer should be December 31, 1898, which corresponds to option A.\n\nThe best answer is A

TeacherGRPO

To solve this problem, we need to determine the date that is one year ago from the last day of 1899.\n\n1. The question states that today is the last day of 1899, which is December 31, 1899.\n2. To find the date one year ago, we need to subtract one year from December 31, 1899.\n3. One year ago from December 31, 1899, would be December 31, 1898.\n\nNow, let's convert this date into the MM/DD/YYYY format:\n\n- Month: December (12)\n- Day: 31\n- Year: 1898\n\nTherefore, December 31, 1898 in MM/DD/YYYY format is 12/31/1898.\n\nThe best answer is A.

Uniform Penalty

This is the last day of 1899. The date one year ago from today in MM/DD/YYYY [12/31/1898] The best answer is D

w/o IALR

To solve this problem, let's break it down step by step:\n\n1. The current date is given as the last day of 1899, which is December 31, 1899.\n2. We need to find the date one year ago from today, which means we need to go back to December 31, 1898.\n\nNow let's calculate the date one year ago:\n- From December 31, 1899 to December 31, 1898, we count back one full year.\n\nTherefore, one year ago from December 31, 1899, is December 31, 1898.\n\nNow, let's match this calculated date with the given options:\n- Option A: 12/31/1898 (This matches our calculated date)\n- Option B: 02/11/1899 (This is the date in February 11, 1899, which is not one year ago)\n- Option C: 12/25/1898 (This is December 25, 1898, which is not December 31)\n- Option D: 12/03/1898 (This is December 3, 1898, which is not December 31)\n- Option E: 01/11/1899 (This is January 11, 1899, which is not December 31)\n\nAfter carefully examining each option, the correct date one year ago from December 31, 1899, is December 31, 1898.\n\nTherefore, the best answer is:\nThe best answer is A.

Figure 6: **Qualitative comparison of teacher generations with different length penalties.** We compare reasoning outputs from four teacher configurations on a Date Understanding example: Vanilla (no alignment), TeacherGRPO, Uniform Penalty, and w/o IALR.

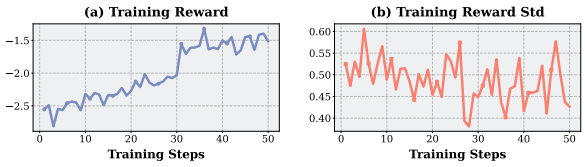

Figure 7: **Training dynamics during Teacher Alignment on the Date Understanding dataset.** Both curves are smoothed using exponential moving average (EMA weight = 0.9).

strating effective optimization toward the alignment objective. (2) **Decreasing reward variance**: The reward standard deviation exhibits a gradual declining trend from around 0.60 to 0.45, indicating that the policy becomes increasingly confident and consistent in its generations.

These training dynamics confirm that TeacherGRPO achieves stable optimization and validates the effectiveness of CSA in maintaining stable learning signals throughout the alignment process.

C Experiment Setup Details

C.1 Datasets

We evaluate our approach across four diverse reasoning benchmarks that require different types of reasoning capabilities:

- **Date Understanding (Date)** (Srivastava et al., 2023): Temporal reasoning tasks requiring models to understand and manipulate date-related information.
- **StrategyQA (SQA)** (Geva et al., 2021): Multi-hop reasoning questions requiring implicit strategy decomposition and commonsense knowledge integration.
- **ARC-Challenge (ARC)** (Clark et al., 2018): Science questions from standardized tests requiring commonsense and scientific reasoning.
- **CommonsenseQA (CQA)** (Talmor et al., 2019): Multiple-choice questions testing commonsense knowledge and reasoning capabilities across diverse everyday scenarios.

C.2 Model Selection

We conduct experiments on two distinct model families to demonstrate the generalizability of our approach across different architectures and compression ratios:

- **Qwen Family** (Yang et al., 2025): Teacher model is `Qwen2.5-3B-Instruct` (3B parameters), and student model is `Qwen2.5-0.5B-Instruct` (0.5B parameters), yielding a 6× compression ratio.
- **Gemma Family** (Team et al., 2025): Teacher model is `Gemma-3-1b-it` (1B parameters), and student model is `Gemma-3-270m-it` (270M parameters), achieving approximately 3.7× compression.

Both model families represent state-of-the-art open-source instruction-tuned models with strong baseline reasoning capabilities.

C.3 Distillation Methods

We apply TeacherGRPO to several state-of-the-art distillation approaches to demonstrate its compatibility and effectiveness across different distillation paradigms:

- **Knowledge Distillation (KD)** (Hinton et al., 2015): Logit-based knowledge distillation using responses from Gemini 2.5 Pro as ground truth labels. The student minimizes KL divergence between its output distribution and the teacher's softened probability distribution.

- **Sequence-Level KD (SeqKD)** (Kim and Rush, 2016): Token-level distillation where the teacher generates training sequences via sampling, and the student learns to reproduce these sequences through standard cross-entropy training.

- **Adaptive Off-Policy Distillation** (Ko et al., 2024): An efficient variant of iterative on-policy distillation (Agarwal et al., 2024) where the student first generates responses, which are then stored and replayed during training. This approach maintains effectiveness while significantly reducing computational overhead compared to standard on-policy methods.

C.4 Baseline Methods

We compare TeacherGRPO against the following baselines:

- **Vanilla Distillation**: Distillation without any teacher alignment, serving as the primary baseline.

- **TeacherKD**: Our implementation of knowledge distillation for teacher alignment, where the teacher is adapted toward the student's distribution via standard KL divergence minimization.

- **Data-Centric Selection** (Liu and Zhang, 2025): Filters out the 10% most challenging examples based on teacher-student divergence metrics, representing the data filtering paradigm.

- **Intermediate Supervision** (Zhou and Ai, 2024): Introduces an intermediate assistant teacher (`Qwen2.5-1.5B-Instruct` for the Qwen family) positioned between the teacher and student to bridge the capability gap.

C.5 Implementation Details

C.5.1 Teacher Alignment Phase

TeacherKD. We train the teacher to minimize KL divergence with the student using a learning rate of 5×10^{-6}, batch size 4, and AdamW optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$. We train for one complete epoch on Date Understanding and limit training to 100 steps maximum on other datasets to prevent catastrophic collapse.

TeacherGRPO. We employ GRPO with a learning rate of 2×10^{-5}, batch size 4, gradient accumulation over 32 steps, and sample 4 generations per query for group-based optimization. We use AdamW optimizer with the same β values. We train for five epochs on Date Understanding and limit training to 100 steps maximum on other datasets. The GRPO-specific hyperparameters are set as follows: clip ratio $\epsilon_{clip} = 0.2$, KL penalty coefficient $\gamma = 0.01$, and group size $G = 4$.

All teacher alignment is performed using LoRA (Hu et al., 2022) with rank $r = 32$, $\alpha = 64$, and dropout $p = 0.1$ for memory efficiency, adapting only the attention projection matrices (query, key, value, and output).

C.5.2 Student Distillation Phase

Following teacher alignment, we distill knowledge into student models using a learning rate of 2×10^{-5}, batch size 4, and AdamW optimizer. We train for two epochs on Date Understanding and ARC-Challenge, and one epoch on StrategyQA and CommonsenseQA. For KD-based distillation, we use a temperature of $T = 2.0$ and balance coefficient $\alpha = 0.5$ between the distillation loss and task loss. All student distillation is performed with full fine-tuning.

C.5.3 TeacherGRPO Hyperparameters

For Curriculum Selective Alignment (CSA), we set token-level curriculum bounds to $k_{\text{min}} = 0.3$ and $k_{\text{max}} = 1.0$, and distribution-level curriculum bounds to $p_{\text{min}} = 0.3$ and $p_{\text{max}} = 1.0$. For Importance-Adaptive Length Regularization (IALR), we use temperature $\eta = 4$ for importance score computation and penalty strength $\beta = 0.01$. The reward combination weights are set to $\alpha_1 = 0.6$ for CSA reward, $\alpha_2 = 0.6$ for IALR reward, and $\alpha_3 = 0.4$ for verification reward.

C.5.4 Evaluation Protocol

We evaluate all models using greedy decoding with maximum generation length of 1024 tokens. For multiple-choice questions (ARC-Challenge, CommonsenseQA, Date Understanding), we extract the model's final answer by pattern matching on the response format specified in the prompt. For StrategyQA, we extract binary True/False predictions. We report accuracy as the primary metric across all datasets.

C.5.5 Infrastructure

Our implementation is based on PyTorch 2.1.0 with Python 3.11.7. We use the Transformers library (version 4.36.0) for model loading and the TRL library (version 0.7.4) for GRPO implementation. All experiments are conducted on NVIDIA A100 GPUs with 80GB memory.

C.6 Prompts

We provide the exact prompts used for generation across all datasets below.

Prompt for ARC-Challenge, Date, and CommonsenseQA

Given the following question and {num_choices} candidate answers ({choice_str}), choose the best answer.
Question: {question}
{formatted_choices}
Please reason step by step, and conclude with your choice. Your response should end with "The best answer is []" where the [] is one of {choice_str}.

Prompt for StrategyQA

Question: {question}
Please reason step by step, and conclude with either "True" or "False".