

APEX: Adaptively Balancing Off-Policy Guidance and On-Policy Exploration for LLM Reasoning

Yuchen Wang^{1,2,3}, Zikang Zhang², Duo Xu¹, Kunpeng Xu¹
Qihui Zhu¹, Wei Huang², Siying Wu^{3,*}, Caihong Cheng²
Xin He², Yang Yang², Guanjun Jiang², Xiaoyan Sun^{1,3,*}

¹University of Science and Technology of China

²Qwen Business Unit of Alibaba

³Institute of Artificial Intelligence, Hefei Comprehensive National Science Center

*Corresponding authors

Abstract

In post-training large language models (LLMs), supervised fine-tuning (SFT) provides expert guidance but can overfit, while reinforcement learning (RL) enables exploration but often suffers from cold-start and sparse-reward issues. Although combining SFT with RL is intuitive, naively mixing off-policy expert data into on-policy RL can induce distribution shift and destabilize optimization. We propose **APEX**, a framework that harmonizes off-policy guidance and on-policy exploration via a Cognitive State Awareness mechanism. Specifically, APEX leverages a perplexity-based Student’s t -test to compare the sequence-level perplexity distributions of on-policy generations against candidate offline data. This comparison diagnoses whether failures reflect sampling stochasticity or missing knowledge. Guided by this diagnosis, APEX adaptively gates GRPO optimization and modulates SFT supervision. The framework preferentially selects offline samples that are statistically consistent with the current on-policy distribution to guide RL updates, while filtering out high-shift data and verifiably incorrect samples that could destabilize training. APEX also uses a dynamically weighted SFT loss when needed. Extensive experiments show that APEX improves over SFT-then-RL by +7.10 points in average pass@1 on six math benchmarks and by +15.50 points on three general reasoning benchmarks.

1 Introduction

Recently, the rapid evolution of Large Language Models (LLMs) (Achiam et al., 2023; Brown et al., 2020) and Multimodal Large Language Models (MLLMs) (Bai et al., 2025; Zhu et al., 2025; Liu et al., 2024, 2023) has significantly advanced artificial intelligence. In post-training, large-scale Reinforcement Learning with Verifiable Rewards

(RLVR) (Guo et al., 2025)—exemplified by Proximal Policy Optimization (PPO) (Schulman et al., 2017) and Group Relative Policy Optimization (GRPO) (Shao et al., 2024)—has substantially improved the reasoning capabilities of these models. RL enables LLMs to explore reasoning trajectories and optimize behavior through environmental feedback, but applying RL directly to base models is often constrained by limited initial capabilities, sparse rewards, and the vast search space of complex tasks. In contrast, SFT (Ouyang et al., 2022) allows models to absorb high-quality demonstrations and rapidly approach the target distribution, yet it can restrict exploration and overfit to demonstration data, resulting in degraded generalization in out-of-distribution scenarios. This tension makes it crucial to balance expert guidance from SFT with on-policy exploration from RL.

Recent efforts have explored combining SFT and RL through interleaved training, unified objectives, or off-policy guidance (Ma et al., 2025; Zhang et al., 2025c; Lv et al., 2025; Yan et al., 2025; Fu et al., 2025). However, these methods typically decide when and how to use expert data through fixed schedules, heuristic weights, or prefix-based constraints, without explicitly measuring whether the expert trajectory is compatible with the current policy. This omission is critical: incorporating high-shift off-policy data into on-policy RL can destabilize optimization, while overly strong supervision can suppress exploration. These limitations motivate a training strategy that adapts to the model’s current cognitive state.

To address the aforementioned issues, we propose **APEX** (Adaptively Balancing Off-Policy Guidance and On-Policy EXploration for LLM Reasoning), a novel framework that dynamically synergizes SFT and RL based on the model’s cognitive state. We posit that training strategies should

adaptively align with the model’s current capability boundaries. To this end, we categorize reasoning failures into two distinct scenarios:

- **Internal Knowledge Availability:** The model’s latent space inherently possesses the capability to solve the problem, but output failures occur due to sampling stochasticity.
- **External Knowledge Absence:** The model fundamentally lacks the underlying reasoning patterns or domain knowledge required to solve the problem.

When facing reasoning failures, APEX employs RL to elicit and refine internal knowledge, while utilizing SFT to inject external knowledge.

APEX introduces a statistical filtering mechanism based on **Perplexity (PPL)** (Jelinek et al., 1977) and the **Student’s t -distribution** (Student, 1908) to quantify the model’s familiarity with Expert Demonstrations. Specifically, when all trajectories generated by the policy model fail validation (i.e., all rollouts are incorrect), APEX leverages the statistical characteristics of these erroneous samples to construct a one-sided low-PPL compatibility gate. This gate determines whether the expert trajectory is sufficiently familiar to the policy model. As shown in Figure 1a, there is a distributional difference in the PPL evaluated by the on-policy model between off-policy (expert data) and on-policy data.

When the expert data passes the one-sided low-PPL gate, it indicates that the problem belongs to the Internal Knowledge Availability state. In this case, APEX treats the expert data as an in-distribution perfect sample and integrates it into the **GRPO** sampling group to rectify gradient estimation and guide model exploration. Conversely, when expert data is identified as a statistical outlier, APEX applies gating truncation to the RL loss to prevent high-variance updates caused by excessive distributional discrepancy.

Furthermore, APEX introduces a **dynamically weighted SFT loss mechanism**. This mechanism imposes stronger Teacher Forcing on unfamiliar concepts while progressively attenuating supervision signals for known concepts. This approach ensures effective knowledge injection while mitigating model degeneration caused by overfitting. As shown in Figure 1b, using SFT on the policy model to fit off-policy data can reduce the distributional shift between the two datasets.

We conducted extensive experiments to validate the efficacy of APEX across complex mathematical and logical reasoning tasks (Hugging Face H4, 2024; Hendrycks et al., 2021). Our evaluation encompasses models of varying parameter scales (e.g., Qwen2.5-Math-1.5B, Qwen2.5-Math-7B) (Yang et al., 2024b) and distinct architectures (e.g., Qwen (Yang et al., 2024a), Llama (Dubey et al., 2024)). Empirical results demonstrate that APEX consistently achieves state-of-the-art (SOTA) performance across diverse experimental settings.

Our Contributions. (1) *New Insight.* We dissect the distributional mismatch caused by indiscriminate data augmentation and reveal the heterogeneity of all-wrong RL rollouts: some failures are sampling errors around policy-compatible expert trajectories, while others reflect policy-distant knowledge gaps. This motivates compatibility-aware use of off-policy demonstrations instead of direct mixing with on-policy RL. (2) *Novel Approach.* We propose the **APEX** framework, which employs a PPL-based Student’s t -test statistic to dynamically categorize the model’s mastery of specific samples. By synergizing SFT and RL, APEX ensures the effective acquisition of external knowledge while maintaining the stability of on-policy optimization. (3) *Strong Performance.* We demonstrate through extensive evaluation that APEX achieves state-of-the-art performance across diverse experimental configurations and model architectures.

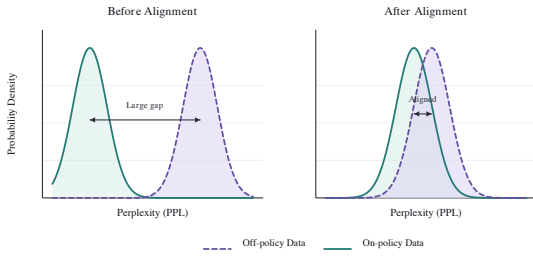
2 Related Work

2.1 LLM Post-training

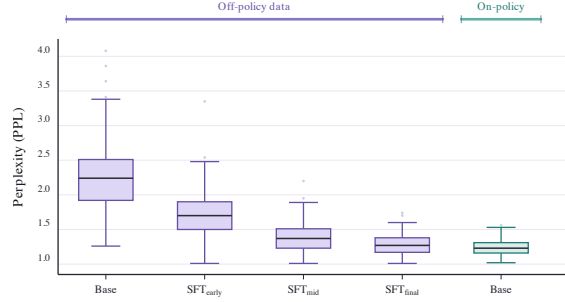
LLM post-training commonly combines supervised fine-tuning (SFT) and reinforcement learning (RL). SFT injects expert demonstrations efficiently but can inherit the limits of the demonstration distribution, while RL methods such as PPO and GRPO optimize policies with reward feedback and can improve reasoning through exploration. However, RL training is often sensitive to exploration quality, sparse rewards, and the initial capability of the base model. These complementary strengths and weaknesses motivate recent efforts to coordinate supervised and reinforcement objectives during post-training.

2.2 Unified Perspectives on SFT and RL

Beyond existing post-training paradigms, substantial research (Yan et al., 2025; Lv et al., 2025; Ma



(a) Illustration of distributional shift. The PPL discrepancy between external off-policy data and the base model’s on-policy rollouts reflects the initial mismatch.



(b) Validation on Qwen2.5-7B-Math fine-tuned with OpenR1-Math-46k-8192. As SFT progresses, the PPL of off-policy data decreases and converges toward that of on-policy responses.

Figure 1: Analysis of distributional alignment. The left panel illustrates the distributional gap between off-policy data and on-policy rollouts, while the right panel empirically shows that SFT reduces this gap by internalizing external knowledge.

et al., 2025) has explored the synergistic optimization of SFT and RL, aiming to combine their respective strengths: SFT introduces expert knowledge to break through the model’s performance ceiling, while RL facilitates the convergence of generation probabilities toward correct solutions via online rollouts. The core challenge is how to effectively integrate supervisory signals from expert knowledge into the reinforcement learning framework, which can be primarily categorized into three types.

The first category adopts staged or interleaved optimization strategies. For instance, (Ma et al., 2025) identifies hard samples that RL struggles to solve and performs targeted remediation via SFT. This alternating strategy iteratively optimizes the model by executing different post-training algorithms at different stages. However, it faces challenges in policy scheduling during practical application, specifically the difficulty in determining the optimal training ratios and the precise timing for switching between SFT and RL.

The second category uses offline data as hints or prefixes to constrain and guide the RL exploration space. Prior methods anneal hint lengths (Huang et al., 2025; Liu et al., 2025), adapt hint ratios by difficulty (Zhang et al., 2025a), mix rollouts with varying hint levels, naive rollouts, and ground-truth data (Zhang et al., 2025b), or introduce offline data when RL rollouts are entirely incorrect to ensure the validity of the advantage function (Yan et al., 2025). Although these methods effectively reduce exploration difficulty via prefix construction, they still face two primary risks: first, they fail to fully account for distribution shift between expert data

and the current policy; second, for highly challenging problems, simple prefix continuation is often insufficient to guide the model toward the correct reasoning path.

The third category coordinates on-policy RL and off-policy expert optimization through weighted loss functions. (Lv et al., 2025) theoretically unifies the perspectives of SFT and RL, providing a foundation for joint training. Building on this, (Fu et al., 2025) and (Zhang et al., 2025c) achieve weighted collaboration between the two objectives with fixed or dynamic weights. These approaches enable joint training and a smoother transition from imitation learning to autonomous exploration, but their weighting schedules remain heuristic and may not reflect the current policy’s familiarity with specific expert samples.

3 Method

3.1 Preliminaries

In LLM post-training, SFT fits the policy π_θ to static expert demonstrations. Its key advantage is knowledge transfer: high-quality demonstrations can introduce reasoning patterns that are absent or underrepresented in the base model, thereby expanding the model’s potential solution space. However, because SFT learns from a fixed off-policy distribution, excessive imitation can over-regularize the policy, constrain exploration, and reduce robustness under distribution shift.

In contrast, on-policy RL optimizes the policy using rewards from model-generated trajectories. For tasks with verifiable outcomes, RL can refine the probability mass toward successful reasoning

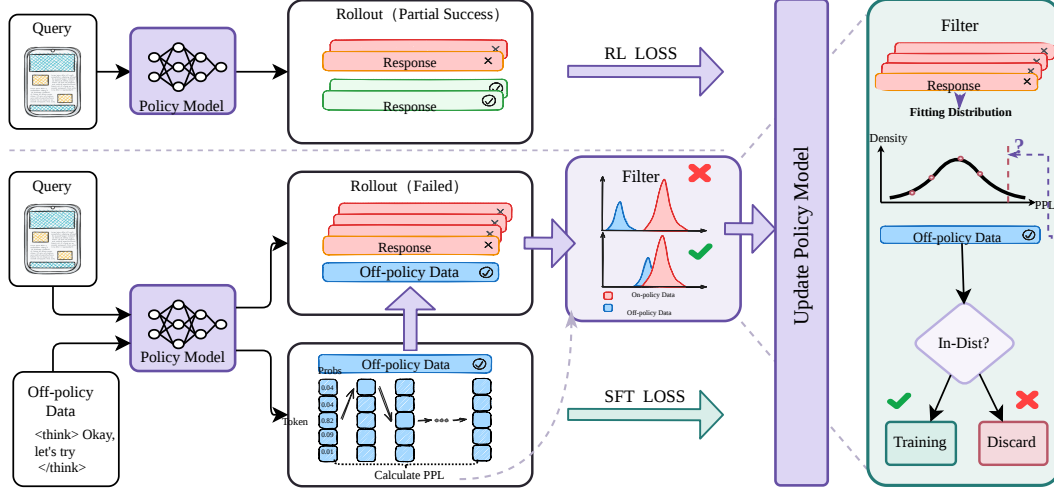


Figure 2: **Overview of APEX.** APEX performs inference rollouts for each query. For scenarios demonstrating partial success, the RL loss is applied directly. Conversely, in instances of complete failure, off-policy data is introduced and processed via two parallel pathways: it is unconditionally utilized to compute the SFT loss, while simultaneously undergoing PPL-based filtering. Specifically, off-policy data is integrated into the rollout set for RL optimization only if it falls within the current policy’s distribution ; otherwise, it is discarded from the RL pathway.

paths and improve single-pass accuracy. Yet this process depends on the model’s ability to sample at least partially correct trajectories; when all rollouts fail, the reward signal becomes sparse and the policy update may provide little useful guidance.

APEX builds on GRPO, which estimates relative advantages within a group of sampled responses and avoids training an additional value model. For a question q , GRPO samples responses $\{o_1, o_2, \dots, o_G\}$ and normalizes their rewards to obtain the advantage:

$$A_i = \frac{r_i - \text{mean}(r_1, r_2, \dots, r_G)}{\text{std}(r_1, r_2, \dots, r_G)} \quad (1)$$

where r_i is the reward of response o_i .

The study (Yue et al., 2025) indicates that under existing RLVR paradigms, the GRPO algorithm does not induce novel reasoning patterns that transcend the priors established by SFT. Its essence functions more as a sampling optimizer it drives the convergence of the model’s probability distribution toward correct paths that already exist within the prior space. While this mechanism significantly enhances single-pass accuracy ($\text{pass}@1$), it often comes at the expense of exploration breadth in reasoning paths, exhibiting a distinct phenomenon of distribution contraction.

Conversely, (Lv et al., 2025) shows that SFT and GRPO can be interpreted under a unified policy-gradient view rather than as opposing paradigms.

In this view, SFT resembles an RL update with consistently positive advantages over a fixed expert distribution, whereas on-policy GRPO estimates gradients from trajectories sampled by the current policy. This unified perspective motivates jointly optimizing supervised and reinforcement objectives:

$$\mathcal{L} = \alpha \cdot \mathcal{L}_{RL} + \beta \cdot \mathcal{L}_{SFT},$$

Here, α and β are weighting coefficients used to balance the RL loss ($\mathcal{L}_{RL}$) and SFT loss ($\mathcal{L}_{SFT}$). In existing studies, these hyperparameters are typically set as fixed constants or dynamically adjusted using a decay strategy. Strategies such as SFT followed by RL, RL followed by SFT, or synchronous parallel optimization can all be viewed as particular solutions of this general formula. Based on this unified perspective, we propose an adaptive SFT-RL collaborative algorithm designed to dynamically balance the weights of both components, thereby achieving superior post-training performance.

3.2 Overview of APEX

We propose **APEX**, which aims to dynamically bridge SFT and on-policy RL according to the model’s cognitive state. We categorize the model’s mastery of the ground truth into two distinct states:

- **Internal Knowledge Availability:** The model’s latent space contains the knowledge

required to solve the problem, but it fails to generate the correct output due to sampling stochasticity.

- **External Knowledge Absence:** The model lacks the underlying reasoning patterns necessary to solve the problem. In this state, the optimization should shift focus toward SFT to inject new knowledge via *Teacher Forcing*.

APEX quantifies this familiarity by introducing a *PPL*-based test statistic and leverages off-policy to guide the training process. The overall framework is depicted in Figure 2.

3.3 Cognitive State-Aware Demonstration Selection

GRPO relies on Intra-group Relative Advantage to estimate gradients. For out-of-capability problems, the entire sample set $\{o_1, \dots, o_G\}$ may be incorrect. In such cases, the intra-group advantage is derived solely from erroneous trajectories, becoming ineffective noise that fails to guide the model toward the correct solution.

To address the issue of gradient failure in GRPO caused by All-Wrong sampling when handling high-difficulty problems, we propose dynamically introducing offline expert demonstrations into the current rollout group. By constructing a Heterogeneous Training Batch, we combine the on-policy data generated by the current policy with selected off-policy expert data for training.

Specifically, we define the augmented sampling group G_{new} as the union of the current policy sampling group G_{roll} and the selected expert demonstration group $G_{\text{off-data}}$:

$$G_{\text{new}} = \{(x, y_i)\}_{i=1}^{|G_{\text{roll}}|} \cup \{(x, y_j^*)\}_{j=1}^{|G_{\text{off-data}}|} \quad (2)$$

where x denotes the input prompt, y_i represents the response generated via online rollout, and y_j^* is the offline data sample.

Based on this mixed group, we re-estimate the intra-group advantage. The advantage $\hat{A}_k$ of the k -th sample is determined by its relative standardized reward within the augmented group:

$$\hat{A}_k = \frac{r(x, y_k) - \text{mean}(\{r(x, y) \mid (x, y) \in G_{\text{new}}\})}{\text{std}(\{r(x, y) \mid (x, y) \in G_{\text{new}}\})} \quad (3)$$

However, directly introducing off-policy data presents challenges: the expert trajectory y^* is not

generated by the current policy π_θ , and its distribution may exhibit a significant distribution shift from the current model. Blindly introducing data with excessive deviation would violate the on-policy assumption of on-policy RL. To quantify this distribution discrepancy and filter out off-policy data that aligns with the model’s current cognitive state, we designed a statistical filtering mechanism based on PPL.

Perplexity quantifies the uncertainty of the policy model in predicting a sample sequence conditioned on the prompt. For a sequence y of length T paired with prompt x , its PPL is defined as:

$$\text{PPL}_\theta(y \mid x) = \exp \left(-\frac{1}{T} \sum_{t=1}^T \log \pi_\theta(y_t \mid x, y_{<t}) \right), \quad (4)$$

where $\pi_\theta(y_t \mid x, y_{<t})$ denotes the conditional probability of the t -th token y_t given the prompt x and its preceding context $y_{<t}$. PPL captures the model’s prediction uncertainty, with lower values indicating higher confidence.

By feeding off-policy data into the current policy model, we compute *PPL*, which quantifies the model’s difficulty in reconstructing these external samples. A low *PPL* suggests that the off-policy data resides within the high-probability region of the current policy’s distribution. In such cases, if the corresponding online rollouts within the same group fail to produce correct answers, the failure is attributed to sampling stochasticity rather than a deficient generative capability. Consequently, incorporating these in-distribution off-policy samples into the GRPO training group can effectively mitigate distributional shift. Conversely, when all on-policy responses fail, off-policy samples with high *PPL* are discarded to prevent gradient instability caused by out-of-distribution noise.

Since the group size G in GRPO is typically small and the population variance is unknown, we employ a **Student’s t -test** to assess the distributional consistency of the off-policy sample y^* relative to the current policy. We first compute the mean μ_{roll} and standard deviation σ_{roll} of the PPL

for the rollout set $\{o_j\}_{j=1}^G$:

$$\begin{aligned}\mu_{roll} &= \frac{1}{G} \sum_{j=1}^G \text{PPL}_{\theta}(o_j | x), \\ \sigma_{roll} &= \sqrt{\frac{1}{G-1} \sum_{j=1}^G (\text{PPL}_{\theta}(o_j | x) - \mu_{roll})^2}.\end{aligned}\quad (5)$$

To evaluate the off-policy sample y^* , we construct a t -statistic that accounts for the uncertainty in small-sample estimation:

$$t^* = \frac{\text{PPL}_{\theta}(y^* | x) - \mu_{roll}}{\sigma_{roll} \sqrt{1 + 1/G}}. \quad (6)$$

The alignment is quantified by the right-tail probability

$$u = 1 - \text{CDF}_t(t^*, G-1) \quad (7)$$

Since u is larger when the expert demonstration has lower PPL than the rollout distribution, a larger value indicates that the expert sample is more familiar to the current policy. We use this t -statistic as a practical compatibility score rather than a strict hypothesis test, since sequence-level PPL distributions may deviate from normality. To ensure high confidence in sample quality, we set a strict one-sided low-PPL threshold $\delta \approx 0.84$, corresponding approximately to a boundary at $t^* \approx -1$ (one standard deviation below the mean). The RL compatibility gate is formally defined as:

$$g_{\text{RL}} = \begin{cases} 1, & \text{if } u > \delta, \\ 0, & \text{otherwise.} \end{cases} \quad (8)$$

By setting $g_{\text{RL}} = 0$ when $u \leq \delta$, we prevent the policy model from undergoing detrimental RL updates driven by samples that are either out-of-distribution or lack significant relative advantage.

Upon constructing the augmented group G_{new} , we apply the advantage-weighted objective to update the policy:

$$\mathcal{L}_{\text{RL}} = \frac{1}{|G_{\text{new}}|} \sum_{(x,y) \in G_{\text{new}}} \left[\frac{\pi_{\theta}(y | x)}{\pi_{\text{old}}(y | x)} \hat{A}_k - \beta_{\text{KL}} \mathbb{D}_{\text{KL}}(\pi_{\theta} \parallel \pi_{\text{ref}}) \right]. \quad (9)$$

3.4 Adaptive Distribution Alignment via SFT

Concurrently, an auxiliary SFT objective is introduced alongside RL updates to bridge the distributional gap and inject external knowledge. Formally,

for a single offline expert pair (x, y^*) , the standard SFT loss is defined as the negative log-likelihood of the target sequence:

$$\mathcal{L}_{\text{SFT}}(x, y^*) = - \sum_{t=1}^{|y^*|} \log \pi_{\theta}(y_t^* | x, y_{<t}^*). \quad (10)$$

However, to prevent **Over-SFT**—which could lead to model collapse or hinder exploration—we employ an adaptive loss coefficient based on the probability u derived in the previous section. Specifically, we weight the sample-level loss by $(1 - u)$:

$$\mathcal{L}_{\text{ADA-SFT}}(x, y^*) = (1 - u) \cdot \mathcal{L}_{\text{SFT}}(x, y^*). \quad (11)$$

This mechanism dynamically modulates supervision intensity via the alignment probability u . For samples with high u , suggesting strong distributional consistency, the gradient contribution is attenuated to prevent overfitting. Conversely, samples with lower u reside at the distributional boundary and receive a stronger supervision signal.

3.5 Total Training Objective

In scenarios where the model encounters high-difficulty problems resulting in an **All-Wrong** sampling outcome, we propose a decoupled composite objective. Crucially, the PPL-based filtering gate is applied exclusively to the reinforcement learning component to ensure stability, while the auxiliary supervision remains active to bridge the distribution gap.

Formally, the total training objective is defined as:

$$\mathcal{L}_{\text{total}} = \underbrace{g_{\text{RL}} \cdot \mathcal{L}_{\text{RL}}}_{\text{gated RL}} + \underbrace{\mathcal{L}_{\text{ADA-SFT}}(x, y^*)}_{\text{adaptive SFT}}. \quad (12)$$

Here, g_{RL} acts as a hard gate for the RL update:

- **When $u \leq \delta$ (External Knowledge Absence):** The gate closes ($g_{\text{RL}} = 0$), effectively disabling the RL loss to prevent optimization on high-variance noise. Meanwhile, the SFT term dominates (since $1 - u$ is large), forcing the model to learn the correct reasoning patterns via strong supervision.
- **When $u > \delta$ (Internal Knowledge Availability):** The gate opens ($g_{\text{RL}} = 1$), allowing the filtered expert demonstrations to rectify the policy gradient. Concurrently, the SFT weight

Table 1: Performance comparison of different methods based on Qwen2.5-Math-7B across In-Distribution (Math) and Out-Of-Distribution benchmarks. The best results are highlighted in **bold**.

Model	In-Distribution (Math)							Out-Of-Distribution (General)			
	AIME 24	AIME 25	AMC	MATH	Minerva	Olym	Avg	ARC-c	GPQA	MMLU-Pro	Avg
Base Models											
Qwen2.5-Math-7B	12.3	4.7	33.0	43.6	8.8	13.6	19.3	30.9	28.3	16.9	25.4
SFT	25.1	22.8	56.1	84.2	33.8	44.7	44.5	67.4	25.3	42.7	45.1
GRPO	19.4	13.8	59.1	81.8	38.2	46.2	43.1	81.2	36.4	49.3	55.6
SFT → GRPO	25.7	21.6	62.2	84.6	38.2	46.8	46.5	67.7	30.8	37.7	45.4
RL Baselines											
SimpleRL-Zero	27.0	6.8	54.9	76.0	25.0	34.7	37.4	30.2	23.2	34.5	29.3
Oat-Zero	33.4	11.9	61.2	78.0	34.6	43.4	43.8	70.1	23.7	41.7	45.2
SOTA Methods											
LUFFY	26.1	21.8	66.2	88.4	41.9	54.1	49.8	80.8	39.4	53.0	57.7
SRFT	18.4	15.5	55.9	83.8	42.6	48.9	44.2	80.5	36.8	55.9	57.7
HPT	33.0	21.9	69.4	89.2	46.0	56.9	52.7	81.6	42.9	52.1	58.9
Prefix-RFT	31.8	26.4	68.2	88.4	37.5	57.2	51.6	84.0	39.1	52.1	58.4
Ours											
APEX	30.4	24.0	72.9	88.7	47.8	57.8	53.6	84.0	43.9	54.9	60.9

$(1 - u)$ naturally decays, preventing excessive regularization and allowing the RL objective to fine-tune the policy.

This decoupled design ensures a seamless transition from supervised initialization to reinforcement learning based on the model’s dynamic cognitive state.

4 Experiments

4.1 Experimental Setup

We use OpenR1-Math-46k-8192 (Yan et al., 2025) as the training corpus. It contains approximately 45k off-policy samples derived from NuminaMath 1.5 (LI et al., 2024) problems and DeepSeek-R1 (Guo et al., 2025) reasoning traces, and is used for rollout, reward calculation, and SFT. We evaluate Pass@1 on in-domain mathematical reasoning benchmarks and out-of-domain general reasoning benchmarks, comparing APEX against SFT-only, GRPO-only, SFT → GRPO, Zero-RL, and recent hybrid training methods including LUFFY, SRFT, HPT, and Prefix-RFT. Further details on evaluation metrics, baselines, and implementation settings are provided in Appendix.

4.2 Main Results

4.2.1 Performance on Reasoning Benchmark

As summarized in Table 1, our method achieves the best average performance across the six mathematical reasoning datasets, with a leading average score of **53.6**. This performance represents a substantial improvement over standard alignment pipelines,

surpassing the SFT → GRPO baseline by a margin of +7.1 points and the direct SFT baseline by +9.1 points. Notably, our approach outperforms strong concurrent methods such as HPT and Prefix-RFT, by +0.9 and +2.0 points respectively. The advantage is particularly evident in high-difficulty tasks; for instance, on the AMC and Olympiad benchmarks, our model achieves scores of 72.9 and 57.8, demonstrating exceptional robustness in solving complex competition-level problems where standard RL baselines (e.g., SimpleRL-Zero) often struggle.

Table 2 demonstrates the robust generalization of APEX across LLaMA and Qwen architectures. On LLaMA-3.1-8B-Instruct, APEX achieves a leading average of **18.8**, outperforming all baselines. For Qwen2.5-Math-1.5B, APEX achieves the best average score of **43.4**, notably surpassing GRPO by +9.4 points and HPT by +1.5 points.

4.2.2 Out-of-Distribution Generalization

Beyond mathematical domains, we evaluate the transfer capabilities of our model on general reasoning benchmarks (ARC-c, GPQA, MMLU-Pro). The results in the rightmost columns of Table 1 reveal that our method effectively mitigates the risk of overfitting to math-specific patterns. We achieve the highest average OOD score of **60.9**, outperforming the closest competitor, HPT, by +2.0 points, Prefix-RFT by +2.5 points, and the SRFT baseline by +3.2 points. Specifically, on the expert-level GPQA dataset, our method secures a decisive lead of +4.5 points over LUFFY. These results underscore that our training strategy not only enhances

Table 2: Performance comparison of different methods based on LLaMA-3.1-8B-Instruct and Qwen2.5-Math-1.5B. The best results are highlighted in **bold**.

Model	AIME 24	AIME 25	AMC	MATH-500	Minerva	Olym	Avg
LLaMA-3.1-8B-Instruct							
LLaMA-3.1-8B-Instruct	0.4	0.1	4.7	13.8	4.8	3.9	4.6
SFT	0.5	0.1	5.4	20.2	4.0	5.3	5.9
GRPO	0.3	0.5	9.4	23.4	17.6	6.1	9.6
LUFFY	1.9	0.1	13.5	39.0	15.1	9.6	13.2
HPT	2.1	1.2	18.6	47.8	18.8	20.4	18.2
APEX (Ours)	2.3	1.0	21.7	48.2	18.0	21.6	18.8
Qwen2.5-Math-1.5B							
Qwen2.5-Math-1.5B	2.8	6.1	24.5	32.8	11.0	16.4	15.6
SFT	14.7	17.6	45.4	78.4	29.4	35.7	36.9
GRPO	12.2	8.5	43.8	71.0	33.1	35.3	34.0
LUFFY	16.0	13.1	47.1	80.2	30.5	41.0	38.0
HPT	16.6	17.8	51.0	81.0	37.5	47.3	41.9
APEX (Ours)	15.9	15.0	60.9	81.8	38.6	48.0	43.4

Table 3: Ablation study of different components in APEX. The best results are highlighted in **bold**.

Model	AIME 24	AIME 25	AMC	MATH-500	Olym	Minerva	Avg
Baselines							
Qwen2.5-Math-1.5B	2.8	6.1	24.5	32.8	16.4	11.0	15.6
GRPO	12.2	8.5	43.8	71.0	35.3	33.1	34.0
Cognitive State-Aware Demonstration Selection							
w/o Off-policy RL	15.3	13.2	57.3	80.6	45.3	37.5	41.5
μ	20.1	16.1	59.4	81.4	46.8	35.3	43.2
$\mu + \sigma$	16.4	15.5	56.1	80.2	40.6	36.0	40.8
Adaptive Distribution Alignment via SFT							
w/o $(1 - u)$ coeff	19.2	17.2	58.4	80.2	45.2	37.5	43.0
w/o SFT Loss	12.6	10.6	52.7	76.4	39.9	32.7	37.5
Ours							
APEX ($\mu - \sigma$)	15.9	15.0	60.9	81.8	48.0	38.6	43.4

specialized problem-solving skills but also fosters broader cognitive abilities, enabling superior generalization to diverse reasoning tasks.

4.3 Ablation Study

Table 3 validates the two key components of APEX on Qwen2.5-Math-1.5B.

Effectiveness of Cognitive State-Aware Demonstration Selection Removing off-policy RL lowers the average score from 43.4 to 41.5, confirming the benefit of introducing expert demonstrations when the policy enters an all-wrong sampling state. Relaxing the filtering threshold from $\mu - \sigma$ to μ or $\mu + \sigma$ further degrades performance, indicating that loosely admitting high-shift off-policy samples can destabilize learning.

Effectiveness of Adaptive SFT Loss Removing the auxiliary SFT term $\mathcal{L}_{\text{ADA-SFT}} = (1 - u)\mathcal{L}_{\text{SFT}}$ causes the largest drop among the objective vari-

ants, showing that expert supervision remains important for stabilizing RL and bridging the distribution gap. Replacing the adaptive coefficient $(1 - u)$ with a fixed weight also slightly hurts performance, suggesting that supervision should decay as the expert sample becomes more compatible with the current policy.

5 Conclusion

In this work, we propose **APEX**, a framework that synergizes off-policy expert guidance and on-policy exploration in LLM post-training. By employing a perplexity-based Student’s t -test to distinguish Internal Knowledge Availability from External Knowledge Absence, APEX adaptively modulates the intensity of **GRPO** and **SFT**. Experiments on both in-domain and out-of-domain math benchmarks show that APEX consistently outperforms SFT-then-RL baselines.

Limitations

APEX is mainly evaluated on mathematical and general reasoning benchmarks, and its effectiveness on broader instruction-following or domain-specific tasks remains to be further explored. In addition, the method introduces extra computation for rollout generation and PPL estimation. Future work may extend APEX to more diverse tasks and improve its training efficiency.

Ethical Considerations

This work uses publicly available models, datasets, benchmarks, and software tools for research purposes. We do not collect personal data or conduct human-subject experiments. Models trained with APEX may inherit biases, factual errors, or unsafe behaviors from the base models and training data, and improved reasoning capabilities may be misused. We encourage responsible use and appropriate safety evaluation before real-world deployment.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant 62472399 and by the Open Fund of APKL of BIIP, IAI, Hefei Comprehensive National Science Center under Grant 24YGXT003.

References

- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.
- Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2025. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.
- 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:1803.05457v1*.
- Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407.
- Yuqian Fu, Tinghong Chen, Jiajun Chai, Xihuai Wang, Songjun Tu, Guojun Yin, Wei Lin, Qichao Zhang, Yuanheng Zhu, and Dongbin Zhao. 2025. *Srft: A single-stage method with supervised and reinforcement fine-tuning for reasoning*. *Preprint*, arXiv:2506.19767.
- Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, and 1 others. 2025. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638.
- Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Leng Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, Jie Liu, Lei Qi, Zhiyuan Liu, and Maosong Sun. 2024. *Olympiadbench: A challenging benchmark for promoting agi with olympiad-level bilingual multimodal scientific problems*. *Preprint*, arXiv:2402.14008.
- Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. *Measuring mathematical problem solving with the math dataset*. *Preprint*, arXiv:2103.03874.
- Zeyu Huang, Tianhao Cheng, Zihan Qiu, Zili Wang, Yinghui Xu, Edoardo M. Ponti, and Ivan Titov. 2025. *Blending supervised and reinforcement fine-tuning with prefix sampling*. *Preprint*, arXiv:2507.01679.
- Hugging Face H4. 2024. Aime 2024. https://huggingface.co/datasets/HuggingFaceH4/aime_2024.
- Fred Jelinek, Robert L Mercer, Lalit R Bahl, and James K Baker. 1977. Perplexity—a measure of the difficulty of speech recognition tasks. *The Journal of the Acoustical Society of America*, 62(S1):S63–S63.
- Hynek Kydlíček. 2025. *Math-Verify: Math Verification Library*. Version 0.6.1.
- Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, Yuhuai Wu, Behnam Neyshabur, Guy Gur-Ari, and Vedant Misra. 2022. *Solving quantitative reasoning problems with language models*. *Preprint*, arXiv:2206.14858.
- Jia LI, Edward Beeching, Lewis Tunstall, Ben Lipkin, Roman Soletskyi, Shengyi Costa Huang, Kashif Rasul, Longhui Yu, Albert Jiang, Ziju Shen, Zihan Qin, Bin Dong, Li Zhou, Yann Fleureau, Guillaume Lample, and Stanislas Polu. 2024. *Numinamath*.
- Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2023. Let’s verify step by step. *arXiv preprint arXiv:2305.20050*.

- Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 26296–26306.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. [Visual instruction tuning](#). In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.
- Mingyang Liu, Gabriele Farina, and Asuman Ozdaglar. 2025. [Uft: Unifying supervised and reinforcement fine-tuning](#). *Preprint*, arXiv:2505.16984.
- Xingtai Lv, Yuxin Zuo, Youbang Sun, Hongyi Liu, Yuntian Wei, Zhekai Chen, Xuekai Zhu, Kaiyan Zhang, Bingning Wang, Ning Ding, and Bowen Zhou. 2025. [Towards a unified view of large language model post-training](#). *Preprint*, arXiv:2509.04419.
- Lu Ma, Hao Liang, Meiyi Qiang, Lexiang Tang, Xiaochen Ma, Zhen Hao Wong, Junbo Niu, Chengyu Shen, Runming He, Yanhao Li, Bin Cui, and Wentao Zhang. 2025. [Learning what reinforcement learning can't: Interleaved online fine-tuning for hardest questions](#). *Preprint*, arXiv:2506.07527.
- Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. [Training language models to follow instructions with human feedback](#). *Preprint*, arXiv:2203.02155.
- David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dierani, Julian Michael, and Samuel R. Bowman. 2023. [Gpqa: A graduate-level google-proof q&a benchmark](#). *Preprint*, arXiv:2311.12022.
- John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. [Proximal policy optimization algorithms](#). *Preprint*, arXiv:1707.06347.
- Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. [Deepseekmath: Pushing the limits of mathematical reasoning in open language models](#). *Preprint*, arXiv:2402.03300.
- Student. 1908. The probable error of a mean. *Biometrika*, pages 1–25.
- Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. 2023. [Roformer: Enhanced transformer with rotary position embedding](#). *Preprint*, arXiv:2104.09864.
- Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyang Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhui Chen. 2024. [Mmlu-pro: A more robust and challenging multi-task language understanding benchmark](#). *Preprint*, arXiv:2406.01574.
- Jianhao Yan, Yafu Li, Zican Hu, Zhi Wang, Ganqu Cui, Xiaoye Qu, Yu Cheng, and Yue Zhang. 2025. [Learning to reason under off-policy guidance](#). *Preprint*, arXiv:2504.14945.
- An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, and 1 others. 2024a. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.
- An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu, Jingren Zhou, Junyang Lin, Keming Lu, Mingfeng Xue, Runji Lin, Tianyu Liu, Xingzhang Ren, and Zhenru Zhang. 2024b. Qwen2.5-math technical report: Toward mathematical expert model via self-improvement. *arXiv preprint arXiv:2409.12122*.
- Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Yang Yue, Shiji Song, and Gao Huang. 2025. [Does reinforcement learning really incentivize reasoning capacity in llms beyond the base model?](#) *Preprint*, arXiv:2504.13837.
- Feng Zhang, Zezhong Tan, Xinhong Ma, Ziqiang Dong, Xi Leng, Jianfei Zhao, Xin Sun, Yang Yang, and Guanjun Jiang. 2025a. [Adhint: Adaptive hints with difficulty priors for reinforcement learning](#). *Preprint*, arXiv:2512.13095.
- Kaiyi Zhang, Ang Lv, Jinpeng Li, Yongbo Wang, Feng Wang, Haoyuan Hu, and Rui Yan. 2025b. [Stephint: Multi-level stepwise hints enhance reinforcement learning to reason](#). *Preprint*, arXiv:2507.02841.
- Wenhao Zhang, Yuexiang Xie, Yuchang Sun, Yanxi Chen, Guoyin Wang, Yaliang Li, Bolin Ding, and Jingren Zhou. 2025c. [On-policy rl meets off-policy experts: Harmonizing supervised fine-tuning and reinforcement learning via dynamic weighting](#). *Preprint*, arXiv:2508.11408.
- Yifan Zhang and Team Math-AI. 2025. American invitational mathematics examination (aime) 2025. <https://huggingface.co/datasets/math-ai/aime25>.
- Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, and 1 others. 2025. InternV3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*.

A Appendix

A.1 Experimental Setup Details

A.1.1 Training Datasets

The training corpus is OpenR1-Math-46k-8192 (Yan et al., 2025). Its original problems are derived from NuminaMath 1.5 (LI et al., 2024), and its high-quality reasoning responses are generated by DeepSeek-R1 (Guo et al., 2025). The dataset is cleaned for mathematical logic: examples that cannot be verified by Math-Verify (Kydliček, 2025) or exceed 8,192 tokens are removed, yielding approximately 45k prompt-response pairs. In APEX, this corpus is used for policy rollout, ground-truth reward calculation, and SFT.

A.1.2 Evaluation

We employ Pass@k (He et al., 2024) as our evaluation metric and report Pass@1 results using non-zero temperature sampling with temperature 0.6 and top-p 0.95. Our evaluation covers two dimensions: in-domain mathematical reasoning and out-of-domain knowledge generalization. The in-domain evaluation includes AIME 24 (Hugging Face H4, 2024), AIME 25 (Zhang and Math-AI, 2025), AMC (Hendrycks et al., 2021), MATH-500 (Lightman et al., 2023), Minerva (Lewkowycz et al., 2022), and OlympiadBench (He et al., 2024); the out-of-domain evaluation uses ARC-c (Clark et al., 2018), GPQA (Rein et al., 2023), and MMLU-Pro (Wang et al., 2024). To mitigate variance on smaller datasets, we estimate Pass@1 for AIME 24, AIME 25, and AMC as the average accuracy over 32 independently sampled completions per question. For MATH-500, Minerva, OlympiadBench, and all out-of-domain datasets, we use a single-sample evaluation protocol.

A.1.3 Baseline Methods

Our main experiments are conducted on the Qwen2.5-Math backbone (Yang et al., 2024b). Since APEX integrates SFT signals into a GRPO-style on-policy RL framework, we first establish fundamental post-training baselines that isolate the contribution of each component: (1) SFT-only, which trains only on expert demonstrations; (2) GRPO-only, which relies solely on on-policy RL; and (3) the standard sequential SFT → GRPO pipeline, which represents the commonly used two-stage alignment recipe. We also include a Zero-RL model trained on the same backbone to compare against RL training without supervised initializa-

tion. To benchmark against recent hybrid SFT-RL strategies, we further compare with LUFFY (Yan et al., 2025), which introduces off-policy guidance when rollouts fail; SRFT (Fu et al., 2025), which combines supervised and reinforcement fine-tuning in a single stage; HPT (Lv et al., 2025), which studies unified post-training from a policy-gradient perspective; and Prefix-RFT (Huang et al., 2025), which blends SFT and RL through prefix sampling.

A.1.4 Implementation Settings

To verify the generalization capability of APEX, we conduct experiments across Qwen and LLaMA model families, including Qwen2.5-Math-1.5B, Qwen2.5-Math-7B (Yang et al., 2024b), and LLaMA-3.1-8B-Instruct (Dubey et al., 2024). For Qwen2.5-Math models, whose original context length is limited to 4096, we follow prior work and adjust the RoPE (Su et al., 2023) θ from 10000 to 40000, extending the context window to 16384. During rollout, we use a sampling temperature of 1.0, generate 8 samples per instance, and set the maximum response length to 8192 tokens. Training runs for 500 steps. Additional training hyperparameters are listed in Table 4.

A.2 Additional Implementation Details

A.2.1 System Prompt

We provide the exact system instruction used in our experiments in Figure 3. This prompt is designed to align the model with a robust Chain-of-Thought (CoT) reasoning behavior. By explicitly defining the `<think>` and `</think>` delimiters, we guide the model to traverse a thorough reasoning path—analyzing the problem, brainstorming approaches, and verifying steps—prior to generating the final response.

A.2.2 Reward Function Design

The reward function employed in this study consists of three components: a correctness score, a format reward, and a repetition penalty. First, the correctness score is calculated based on the consistency between the generated response and the ground truth. The format verification mechanism is activated solely when the correctness score is positive. Specifically, if the generated text contains a chain-of-thought segment enclosed within `<think>` tags that exceeds 64 characters in length and precedes the final `\boxed{}` answer, a format reward of 1.0 is added to the correctness score. Conversely, if the correctness score is zero, the base reward is

Your task is to follow a systematic, thorough reasoning process before providing the final solution. This involves analyzing, summarizing, exploring, reassessing, and refining your thought process through multiple iterations. Structure your response into two sections: Thought and Solution. In the Thought section, present your reasoning using the format: “<think>\n thoughts </think>\n”. Each thought should include detailed analysis, brainstorming, verification, and refinement of ideas. After “</think>\n” in the Solution section, provide the final, logical, and accurate answer, clearly derived from the exploration in the Thought section. If applicable, include the answer in \boxed{} for closed-form results like multiple choices or mathematical solutions.

User: This is the problem: {QUESTION}

Assistant: <think>

Figure 3: **Overview of the System Instruction.** The model is prompted to generate a systematic chain of thought within specific delimiters before producing the final output.

Table 4: Hyperparameters for Qwen2.5-Math-7B Training using APEX Algorithm

Hyperparameter	Value
Base Model	Qwen2.5-Math-7B
Algorithm	APEX
Learning Rate	5×10^{-6}
Global Batch Size	128
Max Response Length	8192
Rollout Samples (N)	8
Sampling Temperature	1.0
SFT Loss Coefficient	1.0
KL Coefficient	0.0
Max Gradient Norm	80.0
Optimizer	AdamW

reset to 0.0 regardless of format adherence. Finally, an N -gram-based repetition penalty ($N = 3$, with a maximum penalty of -0.5) is applied to the base score. The resulting scalar reward is assigned to the position of the last valid token in the response sequence.

A.2.3 Hyperparameters

We present the detailed hyperparameters for fine-tuning Qwen2.5-Math-7B via the APEX algorithm in Table 4. The model is trained with a learning rate of 5×10^{-6} and a global batch size of 128. To encourage extensive chain-of-thought reasoning, we set the maximum response length to 8192 tokens. Additionally, an auxiliary SFT loss (coefficient 1.0) is employed to maintain training stability, while the KL divergence penalty is disabled.

A.2.4 Artifact Licenses and Intended Use

All external artifacts used in this work, including pretrained models, training datasets, evaluation benchmarks, and software tools, are publicly available research artifacts. We use them only for research purposes and follow their respective licenses and terms of use. Specifically, we use OpenR1-Math-46k-8192 as the training corpus, Qwen2.5-Math and LLaMA-series models as base models, and standard mathematical and general reasoning benchmarks for evaluation. Our released code and data are intended to support reproducibility of the proposed method.