Toward Plasticity-Preserving KL Regularization for Capability Retention in LLM Reinforcement Learning

Li Wang[1,*] **Xiaodong Lu**[1,*] **Xiaohan Wang**[1,†] **Jiajun Chai**[1]
Wei Lin[1] **Tianhao Peng**[2,†] **Guojun Yin**[1,†]

[1]Meituan
[2]Nanyang Technological University

Abstract

Reinforcement learning (RL) has become a central paradigm for large language model (LLM) post-training, but optimization toward new objectives can degrade capabilities already present in the base model. KL regularization is widely used to mitigate such forgetting by constraining policy drift toward a reference model. However, standard full-policy KL regularization constrains the entire response distribution and may unnecessarily restrict exploration and target-task learning. This raises a natural question: can a more precise constraint preserve existing capabilities while minimizing interference with learning new tasks? To this end, we propose Correctness-Conditioned KL Regularization (CoKL), a conditional regularization framework that narrows the preservation constraint from the full output distribution to correctness-conditioned response distributions. We instantiate CoKL with forward KL divergence and derive a practical finite-group training objective for RL-based LLM post-training. At the population level, CoKL decouples the total probability assigned to correct responses from their correctness-conditioned distribution, thereby regularizing the relative probability allocation among reference-supported correct responses without directly anchoring incorrect outputs or total correctness mass. We further show that full-policy forward and reverse KL regularization induce a strict optimal correctness gap when the reference policy is imperfect, whereas CoKL avoids this limitation. Experiments in controlled multi-solution environments and continual post-training settings across multiple model scales demonstrate that CoKL achieves a more favorable balance between target-task improvement and prior-capability retention than existing regularization methods. Our code is available at https://github.com/Lumina04/CoKL.

Introduction

Reinforcement learning (RL) has become a central paradigm for large language model (LLM) post-training. Reinforcement Learning from Human Feedback (RLHF) (Bai et al. 2022; Ouyang et al. 2022) is widely used to align model behavior with human preferences, while Reinforcement Learning with Verifiable Rewards (RLVR) (Guo et al. 2025) enables scalable optimization of reasoning performance through automatically verifiable feedback (Lin et al. 2026b; Lu et al. 2026). Despite these advances, RL-based post-training can degrade capabilities already encoded in the base model. By repeatedly amplifying reward-favored outputs, RL may concentrate probability mass on a narrow set of high-reward behaviors while suppressing alternative response patterns and reasoning strategies that remain valid. This form of capability forgetting can manifest as degraded performance on previously learned or out-of-distribution tasks (Lin et al. 2024; Kotha, Springer, and Raghunathan 2024; Cai et al. 2026; Li et al. 2025a), indicating that adaptation to a new objective may interfere with capabilities acquired before post-training. This issue is particularly consequential for general-purpose LLMs, for which post-training should improve adaptation to new objectives without compromising previously acquired capabilities. Achieving effective adaptation while preserving existing capabilities therefore remains a fundamental challenge in RL-based LLM post-training.

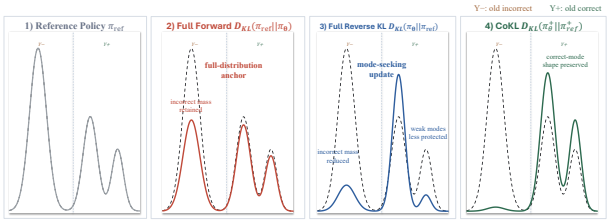

Figure 1: Schematic comparison of KL constraint scopes. Full-policy KL constrains correctness mass and the conditional distributions of both correct and incorrect responses, whereas CoKL constrains only the correctness-conditioned distribution without anchoring incorrect responses or total correctness mass.

KL regularization is widely used to preserve pretrained capabilities during LLM post-training by constraining policy drift from the base model. As shown in Figure 1, reverse KL tends to favor high-probability regions of the reference policy, whereas forward KL more strongly preserves its support. However, existing KL regularizers are typically applied to the full response distribution(Shao et al. 2024; Liu et al. 2025). Such global constraints may preserve useful behaviors, but can also retain ineffective or erroneous modes, thereby restricting exploration and adaptation to new tasks. To avoid

*Equal contribution.
†Corresponding authors: tianhao.peng@ntu.edu.sg; {wangxiaohan17,yinguojun02}@meituan.com.

this limitation, some approaches remove the KL penalty entirely to encourage exploration (Yu et al. 2026; Xue et al. 2025). Yet unconstrained policy updates can cause substantial drift from the base model and increase the risk of capability forgetting. This raises a natural question: can we impose constraints only where preservation is necessary, thereby retaining existing capabilities while minimizing interference with adaptation during post-training?

To preserve existing capabilities without unnecessarily restricting learning during post-training, we propose Correctness-Conditioned KL Regularization (CoKL), a more precise conditional regularization framework that constrains the response distribution according to correctness feedback rather than the full output distribution. In this work, we instantiate CoKL with forward KL divergence and apply it to the conditional distribution over verified-correct responses. CoKL preserves the relative probability allocation among correct response modes inherited from the reference policy while allowing the total probability assigned to correct responses to increase. Incorrect response modes are therefore not anchored to the reference policy and can be freely reshaped by the post-training objective. By narrowing the scope of regularization rather than uniformly reducing its strength, CoKL retains previously accessible correct behaviors while minimizing interference with the acquisition of new capabilities. We derive a practical finite-group objective for RL-based LLM post-training and provide a population-level analysis showing that CoKL decouples the total correctness probability from the conditional distribution over correct responses. In contrast, full-policy forward and reverse KL regularization couple these quantities and induce a strict optimal correctness gap when the reference policy is imperfect. Controlled simulations and continual post-training experiments across multiple model scales further demonstrate that CoKL achieves a favorable balance between capability retention and learning during post-training. Our contributions are summarized as follows:

- We introduce a more precise preservation principle for LLM post-training. Rather than constraining the full response distribution, it restricts regularization to the conditional distribution over verified-correct responses, avoiding direct constraints on incorrect outputs and total correctness mass.
- We instantiate this principle as CoKL, a practical conditional regularization method for RL-based LLM post-training, and derive its finite-group training objective. Our population-level analysis shows that CoKL decouples total correctness probability from the correctness-conditioned response distribution, whereas full-policy KL induces a strict optimal correctness gap under an imperfect reference policy.
- We evaluate the distributional behavior of CoKL in a controlled multi-solution environment and assess its practical retention–adaptation trade-off in continual LLM post-training across multiple model scales.

Related work

LLM RL and Capability Preservation

Reinforcement learning has become a central component of LLM post-training, enabling models to optimize task-specific objectives and improve complex reasoning capabilities. Building on policy optimization methods such as GRPO (Shao et al. 2024), recent algorithms including DAPO (Yu et al. 2026) and GSPO (Zheng et al. 2025) improve training stability and optimization efficiency, yielding substantial gains in mathematical reasoning (Lin et al. 2026b; Lu et al. 2026) and tool use (Lin et al. 2025; Wang et al. 2026b). However, optimization toward new objectives can shift the model away from its original policy distribution, degrading performance on previously learned tasks (Li et al. 2025c) or suppressing effective reasoning patterns already encoded in the base model (Chen et al. 2026). Prior work has mitigated such interference through regularization, parameter merging (Wang et al. 2026a), and multi-domain training (Cai et al. 2026). In contrast, we revisit KL regularization in RL-based post-training and introduce a simple yet more precise constraint that preserves existing capabilities while reducing interference with post-training learning.

KL Regularization in LLM RL

KL regularization is a standard component in RLHF and RLVR for stabilizing policy optimization and preventing excessive deviation from a reference policy (Ziegler et al. 2019; Stiennon et al. 2020; Ouyang et al. 2022). Existing methods typically apply KL regularization to the full policy distribution while adjusting its direction, strength, or token-level application to balance training stability and exploration (Yu et al. 2026; Xue et al. 2025; Deng et al. 2025; Li et al. 2025b; Lin et al. 2026a; Wang et al. 2025; Cui et al. 2025; Vassoyan, Beau, and Plaud 2025; Lee, Kang, and Hwang 2026; GX-Chen et al. 2025). However, full-policy constraints do not distinguish the total probability assigned to correct responses from the relative probability allocation within the correct-response set. Recent verifier-aware methods instead operate on correct responses to improve distributional coverage or diversity. Kruszewski et al. (Kruszewski et al. 2025) construct an explicit target distribution by filtering incorrect responses while preserving the relative reference probabilities among correct responses, UCPO (Lochab, Li, and Zhang 2026) encourages a uniform allocation over correct solutions, and DSDR (Wan et al. 2026) introduces trajectory- and token-level regularization to promote diversity among correct reasoning paths. In contrast, CoKL conditions both the reference and current policies on correctness and regularizes $D_{\mathrm{KL}}(\pi_{\mathrm{ref}}^{+} \| \pi_{\theta}^{+})$. Unlike correct-only target matching, this preserves the reference-induced allocation within the correct-response set without directly optimizing total correctness mass or constraining incorrect responses.

Preliminaries

Group Relative Policy Optimization

Group Relative Policy Optimization (GRPO) (Shao et al. 2024) eliminates the need for a learned value function by

estimating advantages from multiple responses generated for the same prompt. Given a prompt x_i, the behavior policy $\pi_{\theta_{\text{old}}}$ samples a group $\mathcal{O}_i = \{o_{i,g}\}_{g=1}^{G}$, where each response receives a verifier reward $r_{i,g}$. Its group-normalized advantage is

$$A_{i,g} = \frac{r_{i,g} - \mu_i}{\sigma_i + \epsilon}, \qquad \mu_i = \frac{1}{G}\sum_{g=1}^{G} r_{i,g}, \tag{1}$$

where σ_i is the standard deviation of rewards within $\mathcal{O}_i$, and ϵ is a small numerical constant. Since the reward is assigned at the response level, $A_{i,g}$ is shared by all tokens in $o_{i,g}$. Let

$$\rho_{i,g,t}(\theta) = \frac{\pi_\theta(o_{i,g,t} \mid x_i, o_{i,g,<t})}{\pi_{\theta_{\text{old}}}(o_{i,g,t} \mid x_i, o_{i,g,<t})}, \tag{2}$$

and define the clipped surrogate

$$\ell_{i,g,t}(\theta) = \min\left\{\rho_{i,g,t}A_{i,g}, \text{clip}\big(\rho_{i,g,t}, 1-\delta, 1+\delta\big)A_{i,g}\right\}. \tag{3}$$

The GRPO objective is then written as

$$\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}\left[\frac{1}{G}\sum_{g=1}^{G}\frac{1}{|o_{i,g}|}\sum_{t=1}^{|o_{i,g}|}\ell_{i,g,t}(\theta)\right], \tag{4}$$

where δ controls the clipping range.

Method

To preserve prior capabilities while minimizing interference with new-task learning, we propose Correctness-Conditioned KL Regularization (CoKL), which constrains only the distribution over verified-correct responses. We derive its forward-KL objective and finite-group surrogate and integrate it into GRPO.

Correctness Decomposition of Full-Policy KL

For clarity, we formulate CoKL using binary verifier feedback. For each prompt x, let $r(x, y) \in \{0, 1\}$ indicate whether response y is accepted as correct by the verifier, thereby inducing a binary partition of the response space. When the verifier provides a continuous score $s(x, y)$, an analogous partition can be obtained using a task-dependent threshold τ, with $r_\tau(x, y) = \mathbb{I}[s(x, y) \geq \tau]$. For any policy π, we define its total correctness probability as

$$Z_\pi(x) = \sum_y r(x, y)\pi(y \mid x), \tag{5}$$

and define the corresponding conditional distributions as

$$\begin{aligned}\pi^+(y \mid x) &= \frac{r(x, y)\pi(y \mid x)}{Z_\pi(x)},\\ \pi^-(y \mid x) &= \frac{\big(1 - r(x, y)\big)\pi(y \mid x)}{1 - Z_\pi(x)}.\end{aligned} \tag{6}$$

Here, π^+ and π^- denote the policy conditioned on producing a correct or incorrect response, respectively. Let π_{ref} denote the frozen reference policy and π_θ the current policy. For readability, we omit the prompt x below and write $Z_{\text{ref}} = Z_{\pi_{\text{ref}}}(x)$ and $Z_\theta = Z_{\pi_\theta}(x)$. The full-policy forward KL admits the following decomposition:

$$\begin{aligned}D_{\text{KL}}\left(\pi_{\text{ref}} \| \pi_\theta\right) = {} & D_{\text{KL}}\left(\text{Bern}(Z_{\text{ref}}) \| \text{Bern}(Z_\theta)\right)\\ & + Z_{\text{ref}} D_{\text{KL}}\left(\pi^+_{\text{ref}} \| \pi^+_\theta\right)\\ & + (1 - Z_{\text{ref}}) D_{\text{KL}}\left(\pi^-_{\text{ref}} \| \pi^-_\theta\right).\end{aligned} \tag{7}$$

The derivation is in Appendix A. Eq. 7 shows that full-policy KL simultaneously constrains the total probability of correctness, the relative distribution among correct responses, and the distribution among incorrect responses. The first term pulls the current correctness probability Z_θ toward that of π_{ref}, which may conflict with the RL-based post-training objective when the reference policy initially assigns limited probability mass to correct responses. The final term further anchors the policy to incorrect modes. These constraints are not necessary for preserving effective solutions and may restrict the transfer of probability mass from incorrect to correct responses. We therefore retain only the correctness-conditioned component as the preservation target.

CoKL Objective and Gradient Interpretation

Motivated by the decomposition above, we define CoKL as the forward KL divergence between the correctness-conditioned reference policy and the correctness-conditioned current policy:

$$\mathcal{L}_{\text{CoKL}}(x) = D_{\text{KL}}\left(\pi^+_{\text{ref}}(\cdot \mid x) \| \pi^+_\theta(\cdot \mid x)\right). \tag{8}$$

Unlike a full-policy KL regularizer, CoKL compares the two policies only after conditioning on correctness. It therefore penalizes changes in the relative probability allocation among correct responses, without explicitly anchoring incorrect responses under the reference policy. For a correct response y, we have $\log \pi^+_\theta(y \mid x) = \log \pi_\theta(y \mid x) - \log Z_\theta(x)$. Therefore, up to terms independent of θ,

$$\mathcal{L}_{\text{CoKL}}(x) \doteq -\mathbb{E}_{y \sim \pi^+_{\text{ref}}(\cdot|x)}\left[\log \pi_\theta(y \mid x)\right] + \log Z_\theta(x). \tag{9}$$

The $\log Z_\theta(x)$ term removes the implicit correctness-mass optimization in correct-only forward-KL replay:

$$D_{\text{KL}}\big(\pi^+_{\text{ref}} \| \pi_\theta\big) = D_{\text{KL}}\big(\pi^+_{\text{ref}} \| \pi^+_\theta\big) - \log Z_\theta.$$

Thus, CoKL preserves the relative distribution among correct responses while leaving the total correctness probability to be optimized by the RL-based post-training objective. Its gradient is

$$\begin{aligned}\nabla_\theta \mathcal{L}_{\text{CoKL}}(x) = {} & -\mathbb{E}_{y \sim \pi^+_{\text{ref}}(\cdot|x)}\left[\nabla_\theta \log \pi_\theta(y \mid x)\right]\\ & + \mathbb{E}_{y \sim \pi^+_\theta(\cdot|x)}\left[\nabla_\theta \log \pi_\theta(y \mid x)\right].\end{aligned} \tag{10}$$

The first term anchors the update toward reference-supported correct responses, while the second provides the normalization correction required by conditioning. The full derivation are provided in Appendix B.

Sample-based CoKL Surrogate

The distribution-level objective in Eq. 8 is approximated during RL-based LLM post-training using sampled responses

and verifier outcomes. For each prompt group, we construct self-normalized empirical weights over the verified-correct responses. We first present the on-policy finite-group surrogate, where both the reference-side and current-policy-side terms are estimated from fresh samples. Let $\{y_{ij}\}_{j=1}^{G} \sim \pi_{\mathrm{ref}}(\cdot \mid x_i)$ be reference samples with rewards $r_{ij} = r(x_i, y_{ij})$, and let $\{y'_{ij}\}_{j=1}^{G} \sim \pi_\theta(\cdot \mid x_i)$ be current-policy samples with rewards $r'_{ij} = r(x_i, y'_{ij})$. We define the empirical conditional weights as

$$\alpha_{ij} = \frac{r_{ij}}{\sum_{k=1}^{G} r_{ik}}, \qquad \delta_{ij} = \frac{r'_{ij}}{\sum_{k=1}^{G} r'_{ik}}. \tag{11}$$

with the convention that $\alpha_{ij} = 0$ if $\sum_{k=1}^{G} r_{ik} = 0$, and $\delta_{ij} = 0$ if $\sum_{k=1}^{G} r'_{ik} = 0$. The on-policy CoKL surrogate is

$$\begin{aligned}\widehat{\mathcal{L}}_{\mathrm{CoKL}} = &- \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{G} \mathrm{sg}(\alpha_{ij}) \log \pi_\theta(y_{ij} \mid x_i) \\ &+ \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{G} \mathrm{sg}(\delta_{ij}) \log \pi_\theta(y'_{ij} \mid x_i).\end{aligned} \tag{12}$$

where $\mathrm{sg}(\cdot)$ denotes stop-gradient and N is the number of prompt groups in the minibatch. When a current-policy group contains verified-correct responses, the two terms approximate the reference- and current-policy expectations in Eq. 10, respectively. If no verified-correct response is observed, the current-policy correction is unavailable and the surrogate reduces to the reference-correct term. Since such zero-correct groups occur more frequently when the current correctness probability is low, this fallback acts as an implicit adaptive gate that strengthens the recovery pressure toward correct responses, as formalized in Appendix B.

In the off-policy setting, the current-policy conditional term is estimated from responses generated by a behavior policy π_{old}. Let $\{\hat{y}_{ij}\}_{j=1}^{G} \sim \pi_{\mathrm{old}}(\cdot \mid x_i)$ denote the rollout group, with verifier rewards $\hat{r}_{ij} = r(x_i, \hat{y}_{ij})$. Since CoKL is defined over complete responses, we apply importance correction at the sequence level. For a response $\hat{y}_{ij}$ of length T_{ij}, we write its sequence log-likelihood as

$$\ell_\theta(\hat{y}_{ij} \mid x_i) = \sum_{t=1}^{T_{ij}} \log \pi_\theta(\hat{y}_{ij,t} \mid x_i, \hat{y}_{ij,<t}). \tag{13}$$

The sequence-level log-ratio is then

$$\Delta_{ij}(\theta) = \ell_\theta(\hat{y}_{ij} \mid x_i) - \ell_{\mathrm{old}}(\hat{y}_{ij} \mid x_i). \tag{14}$$

To control the variance of off-policy correction, we use the clipped sequence-level importance weight

$$\tilde{\rho}_{ij} = \exp\left[\mathrm{clip}\left(\Delta_{ij}(\theta), \log(1-\epsilon_{\mathrm{IS}}), \log(1+\epsilon_{\mathrm{IS}})\right)\right]. \tag{15}$$

The empirical current conditional distribution is approximated by the self-normalized weights

$$\gamma_{ij} = \frac{\tilde{\rho}_{ij}\hat{r}_{ij}}{\sum_{k=1}^{G} \tilde{\rho}_{ik}\hat{r}_{ik}}. \tag{16}$$

If the denominator is zero, we set $\gamma_{ij} = 0$ for all responses in that group. The clipped off-policy CoKL surrogate is then

$$\begin{aligned}\widehat{\mathcal{L}}_{\mathrm{CoKL}} = &- \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{G} \mathrm{sg}(\alpha_{ij}) \log \pi_\theta(y_{ij} \mid x_i) \\ &+ \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{G} \mathrm{sg}(\gamma_{ij}) \log \pi_\theta(\hat{y}_{ij} \mid x_i).\end{aligned} \tag{17}$$

Integration with RL-Based LLM Post-Training

We integrate CoKL into RL-based LLM post-training through a reference-correct buffer and a joint RL and regularization objective. Before training, π_{ref} is used to generate G_{ref} responses for each prompt in the regularization prompt set. The verifier is then applied to these responses, and prompts with no verified correct reference response are filtered out. The remaining prompt groups form a reference-correct buffer $\mathcal{D}^+$, which provides the empirical support for the reference-side term in Eq. 17. During training, we sample a current-task minibatch $\mathcal{B}_{\mathrm{RL}}$ from the current-task data and independently sample a regularization minibatch $\mathcal{B}_{\mathrm{KL}}$ from $\mathcal{D}^+$. Current-task rollouts are used to compute $\mathcal{L}_{\mathrm{GRPO}}$, whereas rollouts generated for $\mathcal{B}_{\mathrm{KL}}$ instantiate the current-policy conditional term in Eq. 17. The policy is optimized by minimizing

$$\begin{aligned}\mathcal{L}_{\mathrm{total}}(\theta) &= \mathcal{L}_{\mathrm{GRPO}}(\theta) + \beta\widehat{\mathcal{L}}_{\mathrm{CoKL}}(\theta), \\ \mathcal{L}_{\mathrm{GRPO}}(\theta) &= -J_{\mathrm{GRPO}}(\theta).\end{aligned} \tag{18}$$

where J_{GRPO} is defined in Eq. 4, $\widehat{\mathcal{L}}_{\mathrm{CoKL}}$ is the clipped off-policy CoKL surrogate, and β controls the strength of correctness-conditioned preservation. Minimizing $\mathcal{L}_{\mathrm{GRPO}}$ increases the probability of verified-correct responses, while $\widehat{\mathcal{L}}_{\mathrm{CoKL}}$ preserves the relative probability allocation among reference-supported correct modes. The combined objective improves the current task while preserving the relative allocation among reference-supported correct modes on the regularization prompts, without explicitly constraining their total correctness mass or incorrect responses. The full algorithm is summarized in Appendix I.

Theoretical Analysis

We next compare the population-level optima induced by CoKL and full-policy KL regularization. To isolate the objective-level bias introduced by different regularizers, we first consider the setting where correctness optimization and regularization are defined on the same prompt distribution.

Theorem 1 (Correctness-Mass Decoupling under CoKL)**.** *For $\beta > 0$, consider the population-level objective*

$$J_{\mathrm{CoKL}}(\pi) = Z_\pi - \beta D_{\mathrm{KL}}\left(\pi^+_{\mathrm{ref}} \| \pi^+\right).$$

The objective depends on π only through Z_π and π^+ and is independent of π^-. For any two policies π_1 and π_2 satisfying $\pi_1^+ = \pi_2^+$ and $Z_{\pi_2} > Z_{\pi_1}$,

$$J_{\mathrm{CoKL}}(\pi_2) - J_{\mathrm{CoKL}}(\pi_1) = Z_{\pi_2} - Z_{\pi_1} > 0.$$

Moreover, over the full probability simplex, $\max_\pi J_{\mathrm{CoKL}}(\pi) = 1$, and every global maximizer satisfies

$$Z_{\pi^\star} = 1, \qquad \pi^{+\star} = \pi^+_{\mathrm{ref}}.$$

The full proof is provided in Appendix C. Theorem 3 shows that, under the population-level CoKL objective, increasing the total probability assigned to correct responses does not increase the regularization penalty when their relative allocation is unchanged. The objective therefore permits probability mass to move from incorrect to correct responses while preserving the correct-response structure inherited from the reference policy. Full-policy KL behaves differently because its Bernoulli component directly constrains the probability allocation between the correct and incorrect response sets. The following theorem quantifies the resulting correctness gap at the population-level optimum.

Theorem 2 (Strict Correctness Gap Induced by Full-Policy KL)**.** *Let $q = Z_{\mathrm{ref}} \in (0,1)$ and $\beta > 0$. Consider the full-policy forward- and reverse-KL objectives*

$$J_{\mathrm{FKL}}(\pi) = Z_\pi - \beta D_{\mathrm{KL}}\left(\pi_{\mathrm{ref}} \| \pi\right),$$
$$J_{\mathrm{RKL}}(\pi) = Z_\pi - \beta D_{\mathrm{KL}}\left(\pi \| \pi_{\mathrm{ref}}\right).$$

Under the required absolute-continuity conditions, their global optimizers satisfy $\pi^{+\star} = \pi^{+}_{\mathrm{ref}}$ and $\pi^{-\star} = \pi^{-}_{\mathrm{ref}}$. The corresponding optimal correctness probabilities are

$$Z^\star_{\mathrm{FKL}} = \frac{1 - \beta + \sqrt{(\beta-1)^2 + 4\beta q}}{2},$$
$$Z^\star_{\mathrm{RKL}} = \sigma\left(\mathrm{logit}(q) + \frac{1}{\beta}\right),$$

where $\sigma(a) = (1+e^{-a})^{-1}$. Both satisfy $q < Z^\star < 1$ and are strictly decreasing in β. Furthermore, for $\diamond \in \{\mathrm{FKL}, \mathrm{RKL}\}$,

$$\lim_{\beta \to 0^+} Z^\star_\diamond = 1, \qquad \lim_{\beta \to \infty} Z^\star_\diamond = q.$$

The proof is provided in Appendix D. Together, Theorems 1 and 2 establish a fundamental distinction between the population-level objectives. The distribution-level CoKL objective preserves the conditional structure of verified-correct responses without penalizing their total probability, whereas full-policy KL shifts the global optimum away from the pure-reward solution $Z_\pi = 1$ whenever π_{ref} is imperfect. These results characterize the intrinsic objective-level bias of the regularizers, while their practical effects under shared neural parameterization are evaluated in the continual post-training experiments.

Experiments

We evaluate CoKL at two complementary levels. The controlled environment provides direct access to correctness-conditioned distributions and correct-mode coverage, while the continual LLM experiments assess the downstream trade-off between prior-task retention and new-task adaptation.

Controlled Multi-Solution RL Environment

Environment. We consider a controlled contextual bandit with multiple valid solutions. Each input is generated from a latent problem cluster z, for which a subset $\mathcal{C}_z$ of the finite action space receives binary reward. To model multiple reasoning modes, we construct a long-tailed oracle distribution over the correct actions, pretrain a shared-parameter MLP to approximate this distribution, and use the resulting frozen model as the common initialization for all methods. We compare unregularized RL with entropy, forward KL, reverse KL, JS, and CoKL regularization. All methods are trained using the exact expected-reward objective and exact regularization over the finite action space. Implementation details are provided in the Appendix E.1.

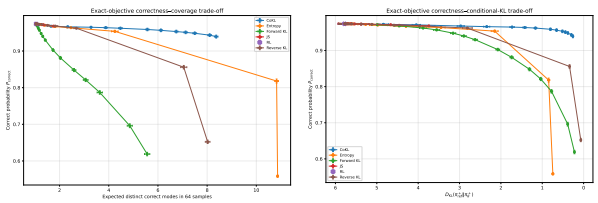

Figure 2: Trade-offs between correctness and coverage@64 (Left), and between correctness and conditional KL (Right).

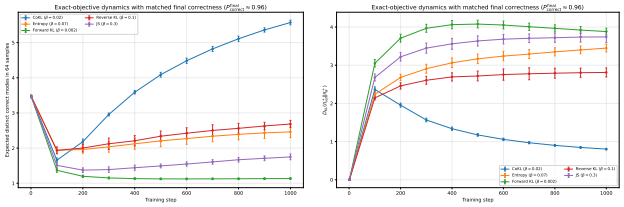

Figure 3: (Left) Dynamics of correct-mode coverage@64 under the same setting. (Right) Dynamics of $D_{\mathrm{KL}}(\pi^+_{\mathrm{ref}} \| \pi^+_\theta)$.

Results. For each regularized method, we sweep the regularization coefficient and evaluate the final policy using the total correct probability P_{corr}, correct-mode coverage@64, and the correctness-conditioned KL $D_{\mathrm{KL}}(\pi^+_{\mathrm{ref}} \| \pi^+_\theta)$, with all results averaged over five random seeds. Figure 2 shows the resulting trade-offs. CoKL achieves higher correct-mode coverage and a smaller correctness-conditioned KL while maintaining high correctness, whereas the other regularizers either sacrifice correctness or exhibit stronger deviation from the reference correct-mode distribution. These results indicate that CoKL better mitigates correct-mode collapse and provides a more favorable trade-off between correctness improvement and correct-mode preservation. To further examine the optimization dynamics, we select, for each method, the coefficient whose final correctness is closest to $P_{\mathrm{corr}} = 0.96$. Figure 3 shows the evolution of coverage@64 and the correctness-conditioned KL during training. Under comparable final correctness, CoKL consistently achieves higher correct-mode coverage and a substantially lower conditional KL, demonstrating a more favorable balance between correctness improvement and correct-mode preservation.

Continual Learning under Task Shift

Tasks and Datasets. We evaluate all methods under a two-stage sequential training protocol consisting of mathematical reasoning followed by general chat, assessing both the retention of previously acquired reasoning capabilities and continued adaptation to the subsequent task. For the math

Method	Normalized Mathematical Reasoning				Normalized Chat	Aggregate
	MATH-Val	MATH500	Olympiad	Math Avg.	Chat-Val	Overall
			Qwen3-0.6B			
Base (after Math training)	100.00	100.00	100.00	100.00	0.00	50.00
GRPO w/o KL	0.00	0.00	0.00	0.00	<u>100.00</u>	50.00
Full Reverse-KL	**66.79**	83.04	55.40	68.41	85.22	76.82
Full Forward-KL	59.66	88.95	64.14	70.92	92.60	81.76
Correct-only Reverse-KL	46.07	52.59	45.93	48.20	87.30	67.75
Correct-only Forward-KL	<u>64.72</u>	**90.86**	**65.08**	**73.55**	94.65	<u>84.10</u>
CoKL (ours)	61.96	<u>90.35</u>	<u>64.91</u>	<u>72.41</u>	**100.53**	**86.47**
			Qwen3-1.7B			
Base (after Math training)	100.00	100.00	100.00	100.00	0.00	50.00
GRPO w/o KL	0.00	0.00	0.00	0.00	100.00	50.00
Full Reverse-KL	33.17	27.55	18.37	26.36	<u>101.81</u>	64.09
Full Forward-KL	**88.19**	80.77	49.74	72.90	98.14	85.52
Correct-only Reverse-KL	5.48	3.42	6.90	5.27	**102.29**	53.78
Correct-only Forward-KL	86.68	<u>81.73</u>	**51.41**	<u>73.28</u>	98.44	<u>85.86</u>
CoKL (ours)	<u>88.11</u>	**86.63**	<u>50.58</u>	**75.11**	101.45	**88.28**
			Qwen3-4B			
Base (after Math training)	100.00	100.00	100.00	100.00	0.00	50.00
GRPO w/o KL	0.00	0.00	0.00	0.00	100.00	50.00
Full Reverse-KL	-186.44	-684.64	-149.25	-340.11	78.25	-130.93
Full Forward-KL	**84.11**	<u>39.31</u>	<u>12.37</u>	<u>45.26</u>	101.73	<u>73.50</u>
Correct-only Reverse-KL	14.32	-3.66	1.18	3.95	88.97	46.46
Correct-only Forward-KL	<u>83.88</u>	32.91	**12.47**	43.09	<u>103.65</u>	73.37
CoKL (ours)	83.06	**43.51**	10.75	**45.77**	**104.02**	**74.90**

Table 1: Dual-anchor normalized results. Math Avg. averages the three normalized math scores, while Overall equally averages Math Avg. and normalized Chat-Val. Best and second-best post-training results are **bolded** and <u>underlined</u>.

stage, we randomly sample 50K training examples from Nemotron-Math-v2 (Du et al. 2025) and use an additional 1K examples as the evaluation set. For the chat stage, we use WildChat (Bhaskar, Ye, and Chen 2025) and Skywork-Reward-Llama-3.1-8B-v0.2 (Liu et al. 2024) to score model responses. Dataset statistics are in Appendix E.2.

Baselines. We compare CoKL with the following GRPO-based baselines: (1) GRPO w/o KL removes the KL regularizer; (2) Full Reverse-KL applies reverse-KL regularization to the full response distribution; (3) Full Forward-KL applies forward-KL regularization to the full response distribution using all sampled responses; (4) Correct-only Reverse-KL applies reverse-KL regularization only to verified-correct responses; and (5) Correct-only Forward-KL uses only verified-correct responses for forward-KL-style supervised replay. All baselines use the same GRPO training setup and differ only in the auxiliary regularization objective.

Implementation. We use the Qwen3 family (Yang et al. 2025) as the base models, including Qwen3-0.6B, Qwen3-1.7B and Qwen3-4B. For all models, we enable thinking mode and set the maximum generation length to $8{,}192$ tokens. Further implementation details are provided in Appendix E.3, and the prompt templates in Appendix H.

Evaluation. We evaluate all methods on two task categories. Mathematical reasoning is evaluated on MATH-500 (Hendrycks et al. 2021), OlympiadBench (He et al. 2024), and the Math-Val validation set, while chat performance is evaluated on WildChat-Val (Bhaskar, Ye, and Chen 2025). We sample with temperature 0.7, top-$p = 1.0$, and top-$k = -1$, and report average@8 throughout. Mathematical reasoning is evaluated by exact match and chat performance by reward-model scores. We apply dual-anchor normalization before aggregation to align their scales; details and unnormalized results are provided in Appendix E.5.

Main Results. As shown in Table 1, GRPO w/o KL and the Reverse-KL variants often suffer from substantial forgetting of previously acquired mathematical capabilities. The Forward-KL variants generally provide stronger capability retention, but may constrain exploration and adaptation on the new task. CoKL achieves the most favorable balance between prior-capability retention and new-task learning, yielding the highest aggregate performance across all model scales. Compared with the strongest baseline, CoKL improves Overall by 2.37, 2.42, and 1.40 points on Qwen3-0.6B, Qwen3-1.7B, and Qwen3-4B, respectively, demonstrating the effectiveness of the proposed method.

Method	Math Avg.	Chat-Val	Overall
Models based on Qwen3-1.7B			
CoKL	**75.11**	**101.45**	**88.28**
w/o term1	3.85	100.83	52.34
w/o term2	73.28	98.44	85.86

Table 2: Ablation results under dual-anchor normalization.

Ablation Studies

Ablation Study of the Loss Structure. To examine the necessity of the two loss terms in CoKL, we conduct an ablation study on Qwen3-1.7B by removing each term in Eq. 17, denoted as *w/o term1* and *w/o term2*, respectively. Removing the second term reduces the objective to the Correct-only Forward-KL formulation. As shown in Table 2, w/o term1 leads to a substantial performance degradation, primarily due to pronounced forgetting on the mathematical reasoning tasks. This result confirms that the first term is essential for preserving the model's existing capabilities. In contrast, *w/o term2* underperforms CoKL on the chat task, demonstrating the importance of the normalization correction. This term removes the implicit correctness-mass component of correct-only reference replay, thereby isolating within-correct distribution preservation and reducing interference with subsequent-task adaptation.

Batch-Level Correctness Floor. Since CoKL does not directly constrain total correctness mass, we examine an explicit batch-level correctness floor. For $p \in \{\mathrm{ref}, \theta\}$, define

$$\widehat{Z}_p = \frac{1}{NG}\sum_{i=1}^{N}\sum_{j=1}^{G} r_{i,j}^{p}, \qquad g_{\mathcal{B}} = [\widehat{Z}_{\mathrm{ref}} - \widehat{Z}_{\theta}]_{+}. \tag{19}$$

The auxiliary loss and combined objective are

$$\begin{aligned} \mathcal{L}_{\mathrm{floor}} &= -\frac{\mathrm{sg}(g_{\mathcal{B}})}{NG}\sum_{i=1}^{N}\sum_{j=1}^{G} r_{i,j}\ell_{\theta}(y_{i,j} \mid x_i), \\ \mathcal{L}_{\mathrm{total}} &= \mathcal{L}_{\mathrm{GRPO}} + \beta\mathcal{L}_{\mathrm{CoKL}} + \lambda_{\mathrm{floor}}\mathcal{L}_{\mathrm{floor}}. \end{aligned} \tag{20}$$

We denote this variant as *CoKL* $+\mathcal{L}_{\mathrm{floor}}$ with $\lambda_{\mathrm{floor}} = 1.0$; since the loss is gated by $\mathrm{sg}(g_{\mathcal{B}}) \in [0, 1]$, its effective strength adapts to the correctness deficit rather than acting as a hard constraint. As shown in Table 3, this mild, one-sided mass constraint yields no consistent retention gains while restricting adaptation and reducing the overall score. This echoes our analysis: an explicit correctness-mass term reintroduces the coupling between retention and adaptation that CoKL avoids. Together with the w/o term2 ablation, it further suggests that, in our setting, forgetting acts more through drift within the correctness-conditioned distribution than through proportional loss of correctness mass.

KL Batch Analysis. We examine KL batch sizes of 8, 16, and 32, as shown in Figure 4 (Left). Increasing the batch size from 8 to 16 substantially improves mathematical retention, while the additional gain from 16 to 32 is modest. Chat performance varies only slightly across the three settings. These results suggest that larger KL batches provide a more reliable estimate of the regularization objective and can strengthen capability retention without substantially affecting new-task adaptation. To limit the additional rollout cost, we use a KL batch size of 8 in the main experiments.

Method	Math Avg.	Chat-Val	Overall
Models based on Qwen3-0.6B			
CoKL	**72.41**	**100.53**	**86.47**
CoKL + $\mathcal{L}_{\mathrm{floor}}$	71.65	92.49	82.07
Models based on Qwen3-1.7B			
CoKL	75.11	**101.45**	**88.28**
CoKL + $\mathcal{L}_{\mathrm{floor}}$	**75.79**	99.63	87.71
Models based on Qwen3-4B			
CoKL	**45.77**	**104.02**	**74.90**
CoKL + $\mathcal{L}_{\mathrm{floor}}$	43.02	100.26	71.64

Table 3: Ablation of the batch-level correctness floor under dual-anchor normalization. The best result is shown in **bold**.

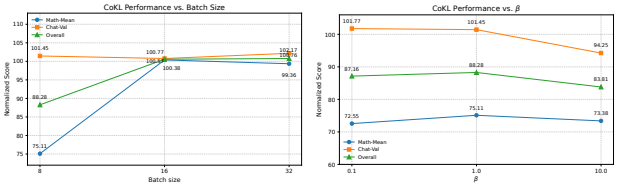

Figure 4: Effects of KL batch size (Left) and loss coefficient β (Right) on CoKL performance using Qwen3-1.7B.

Sensitivity Analysis of the Loss Coefficient β. We analyze $\beta \in \{0.1, 1, 10\}$. As shown in Figure 4 (Right), a large β restricts chat-task adaptation, whereas a small β weakens mathematical capability retention. We therefore set $\beta = 1.0$ to balance retention and adaptation.

Conclusion

In this paper, we proposed CoKL, a correctness-conditioned KL regularizer that narrows the preservation constraint from the full response distribution to the conditional distribution over verified-correct responses. At the population level under a shared prompt distribution, CoKL decouples total correctness probability from the relative probability allocation among correct responses and, unlike full-policy forward and reverse KL regularization, avoids a strict optimal correctness gap under an imperfect reference policy. Controlled multi-solution experiments directly validate its behavior in preserving the correctness-conditioned distribution under improved correctness, while continual LLM post-training experiments demonstrate a favorable practical trade-off between prior-capability retention and new-task adaptation.

References

Bai, Y.; Jones, A.; Ndousse, K.; Askell, A.; Chen, A.; DasSarma, N.; Drain, D.; Fort, S.; Ganguli, D.; Henighan, T.; et al. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Bhaskar, A.; Ye, X.; and Chen, D. 2025. Language models that think, chat better. *arXiv preprint arXiv:2509.20357*.

Cai, M.; Liang, Y.; Wang, L.; Wang, Y.; Zhang, Y.; Xia, L.; Sun, Z.; Ye, X.; and Shi, D. 2026. Advancing General-Purpose Reasoning Models with Modular Gradient Surgery. *arXiv preprint arXiv:2602.02301*.

Chen, Z.; Lu, R.; Zhao, A.; Wang, Z.; Yue, Y.; Song, S.; and Huang, G. 2026. Does reinforcement learning really incentivize reasoning capacity in llms beyond the base model? *Advances in Neural Information Processing Systems*, 38: 57654–57689.

Cui, G.; Zhang, Y.; Chen, J.; Yuan, L.; Wang, Z.; Zuo, Y.; Li, H.; Fan, Y.; Chen, H.; Chen, W.; et al. 2025. The entropy mechanism of reinforcement learning for reasoning language models. *arXiv preprint arXiv:2505.22617*.

Deng, W.; Wei, L.; Yu, C.; and Wu, T. 2025. Unlocking reasoning capabilities in llms via reinforcement learning exploration. *arXiv preprint arXiv:2510.03865*.

Du, W.; Toshniwal, S.; Kisacanin, B.; Mahdavi, S.; Moshkov, I.; Armstrong, G.; Ge, S.; Minasyan, E.; Chen, F.; and Gitman, I. 2025. Nemotron-Math: Efficient Long-Context Distillation of Mathematical Reasoning from Multi-Mode Supervision. *arXiv preprint arXiv:2512.15489*.

Guo, D.; Yang, D.; Zhang, H.; Song, J.; Wang, P.; Zhu, Q.; Xu, R.; Zhang, R.; Ma, S.; Bi, X.; et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

GX-Chen, A.; Prakash, J.; Guo, J.; Fergus, R.; and Ranganath, R. 2025. KL-Regularized Reinforcement Learning is Designed to Mode Collapse. *arXiv preprint arXiv:2510.20817*.

He, C.; Luo, R.; Bai, Y.; Hu, S.; Thai, Z.; Shen, J.; Hu, J.; Han, X.; Huang, Y.; Zhang, Y.; et al. 2024. Olympiadbench: A challenging benchmark for promoting agi with olympiad-level bilingual multimodal scientific problems. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 3828–3850.

Hendrycks, D.; Burns, C.; Kadavath, S.; Arora, A.; Basart, S.; Tang, E.; Song, D.; and Steinhardt, J. 2021. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*.

Kotha, S.; Springer, J.; and Raghunathan, A. 2024. Understanding catastrophic forgetting in language models via implicit inference. In *International Conference on Learning Representations*, volume 2024, 24110–24139.

Kruszewski, G.; Erbacher, P.; Rozen, J.; and Dymetman, M. 2025. Whatever Remains Must Be True: Filtering Drives Reasoning in LLMs, Shaping Diversity. *arXiv preprint arXiv:2512.05962*.

Kwon, W.; Li, Z.; Zhuang, S.; Sheng, Y.; Zheng, L.; Yu, C. H.; Gonzalez, J.; Zhang, H.; and Stoica, I. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th symposium on operating systems principles*, 611–626.

Lee, C.; Kang, M.; and Hwang, S. J. 2026. SAGE: Shaping Anchors for Guided Exploration in RLVR of LLMs. *arXiv preprint arXiv:2605.18864*.

Li, D.; Zhou, J.; Brunswic, L. M.; Ghaddar, A.; Sun, Q.; Ma, L.; Luo, Y.; Li, D.; Coates, M.; Hao, J.; et al. 2025a. Omni-Thinker: Scaling Multi-Task RL in LLMs with Hybrid Reward and Task Scheduling. *arXiv preprint arXiv:2507.14783*.

Li, L.; Zhou, Z.; Hao, J.; Liu, J. K.; Miao, Y.; Pang, W.; Tan, X.; Chu, W.; Wang, Z.; Pan, S.; et al. 2025b. The choice of divergence: A neglected key to mitigating diversity collapse in reinforcement learning with verifiable reward. *arXiv preprint arXiv:2509.07430*.

Li, Y.; Pan, Z.; Lin, H.; Sun, M.; He, C.; and Wu, L. 2025c. Can one domain help others? a data-centric study on multi-domain reasoning via reinforcement learning. *arXiv preprint arXiv:2507.17512*.

Lin, M.; Gong, Z.; Tang, M.; Li, Q.; Wang, C.; Ma, J.; Huang, S.; Tang, K.; and Lu, H. 2026a. expo: Exploration-prioritized policy optimization via adaptive kl regulation and gaussian curriculum sampling. *arXiv preprint arXiv:2605.09923*.

Lin, Y.; Lin, H.; Xiong, W.; Diao, S.; Liu, J.; Zhang, J.; Pan, R.; Wang, H.; Hu, W.; Zhang, H.; et al. 2024. Mitigating the alignment tax of rlhf. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 580–606.

Lin, Z.; Wang, X.; Cao, J.; Chai, J.; Wang, L.; Lu, X.; Lin, W.; He, R.; and Yin, G. 2026b. Resrl: Boosting llm reasoning via negative sample projection residual reinforcement learning. *arXiv preprint arXiv:2605.00380*.

Lin, Z.; Wang, X.; Cao, J.; Chai, J.; Yin, G.; Lin, W.; and He, R. 2025. ResT: Reshaping Token-Level Policy Gradients for Tool-Use Large Language Models. *arXiv preprint arXiv:2509.21826*.

Liu, C. Y.; Zeng, L.; Liu, J.; Yan, R.; He, J.; Wang, C.; Yan, S.; Liu, Y.; and Zhou, Y. 2024. Skywork-Reward: Bag of Tricks for Reward Modeling in LLMs. *arXiv preprint arXiv:2410.18451*.

Liu, Z.; Chen, C.; Li, W.; Qi, P.; Pang, T.; Du, C.; Lee, W. S.; and Lin, M. 2025. Understanding r1-zero-like training: A critical perspective. *arXiv preprint arXiv:2503.20783*.

Lochab, A.; Li, B.; and Zhang, R. 2026. Uniform-Correct Policy Optimization: Breaking RLVR's Indifference to Diversity. *arXiv preprint arXiv:2605.00365*.

Lu, X.; Wang, X.; Chai, J.; Yin, G.; Lin, W.; Chen, Z.; Luo, Y.; Zhuang, F.; Ban, Y.; and Wang, D. 2026. Contextual Rollout Bandits for Reinforcement Learning with Verifiable Rewards. *arXiv preprint arXiv:2602.08499*.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744.

Shao, Z.; Wang, P.; Zhu, Q.; Xu, R.; Song, J.; Bi, X.; Zhang, H.; Zhang, M.; Li, Y.; Wu, Y.; et al. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Sheng, G.; Zhang, C.; Ye, Z.; Wu, X.; Zhang, W.; Zhang, R.; Peng, Y.; Lin, H.; and Wu, C. 2024. HybridFlow: A Flexible and Efficient RLHF Framework. *arXiv preprint arXiv: 2409.19256*.

Stiennon, N.; Ouyang, L.; Wu, J.; Ziegler, D.; Lowe, R.; Voss, C.; Radford, A.; Amodei, D.; and Christiano, P. F. 2020. Learning to summarize with human feedback. *Advances in neural information processing systems*, 33: 3008–3021.

Vassoyan, J.; Beau, N.; and Plaud, R. 2025. Ignore the kl penalty! boosting exploration on critical tokens to enhance rl fine-tuning. In *Findings of the Association for Computational Linguistics: NAACL 2025*, 6108–6118.

Wan, Z.; Shen, Y.; Dou, Z.; Zhou, D.; Zhang, Y.; Wang, X.; Shen, H.; Xiong, J.; Tao, C.; Zhong, Z.; et al. 2026. Dsdr: Dual-scale diversity regularization for exploration in llm reasoning. *arXiv preprint arXiv:2602.19895*.

Wang, H.; Long, X.; Li, Z.; Xu, Y.; Li, T.; and Tang, Y. 2026a. To Mix or To Merge: Toward Multi-Domain Reinforcement Learning for Large Language Models. *arXiv preprint arXiv:2602.12566*.

Wang, J.; Liu, R.; Zhang, F.; Li, X.; Zhou, G.; and Pan, L. 2025. Stabilizing knowledge, promoting reasoning: Dual-token constraints for rlvr. *arXiv preprint arXiv:2507.15778*.

Wang, L.; Wang, X.; Lu, X.; Zhang, Z.; Wu, J.; Chai, J.; Lin, W.; and Yin, G. 2026b. Implicit Hierarchical GRPO: Decoupling Tool Invocation from Execution for Tool-Integrated Mathematical Reasoning. *arXiv preprint arXiv:2605.18500*.

Xue, Z.; Zheng, L.; Liu, Q.; Li, Y.; Zheng, X.; Ma, Z.; and An, B. 2025. Simpletir: End-to-end reinforcement learning for multi-turn tool-integrated reasoning. *arXiv preprint arXiv:2509.02479*.

Yang, A.; Li, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Gao, C.; Huang, C.; Lv, C.; et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Yu, Q.; Zhang, Z.; Zhu, R.; Yuan, Y.; Zuo, X.; Yue, Y.; Dai, W.; Fan, T.; Liu, G.; Liu, L.; et al. 2026. Dapo: An open-source llm reinforcement learning system at scale. *Advances in Neural Information Processing Systems*, 38: 113222–113244.

Zheng, C.; Liu, S.; Li, M.; Chen, X.-H.; Yu, B.; Gao, C.; Dang, K.; Liu, Y.; Men, R.; Yang, A.; et al. 2025. Group sequence policy optimization. *arXiv preprint arXiv:2507.18071*.

Ziegler, D. M.; Stiennon, N.; Wu, J.; Brown, T. B.; Radford, A.; Amodei, D.; Christiano, P.; and Irving, G. 2019. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*.

Appendix A: Derivation of the Correctness Decomposition

For a fixed prompt x, let $r(x, y) \in \{0, 1\}$ indicate whether response y is verified as correct. For continuous rewards, samples can be partitioned into positive and negative groups using a threshold, allowing the same analysis to apply. Define the correct and incorrect response sets as

$$\mathcal{Y}^+(x) = \{y : r(x, y) = 1\}, \qquad \mathcal{Y}^-(x) = \{y : r(x, y) = 0\}.$$

For any policy π, its probability of producing a correct response is

$$Z_\pi(x) = \sum_y r(x, y)\pi(y \mid x). \tag{21}$$

The corresponding correctness-conditioned and incorrectness-conditioned policies are

$$\begin{aligned} \pi^+(y \mid x) &= \frac{r(x, y)\pi(y \mid x)}{Z_\pi(x)}, \\ \pi^-(y \mid x) &= \frac{[1 - r(x, y)]\pi(y \mid x)}{1 - Z_\pi(x)}. \end{aligned} \tag{22}$$

Let π_{ref} denote the reference policy and π_θ the current policy. For readability, we omit the prompt x below and write

$$Z_{\text{ref}} = Z_{\pi_{\text{ref}}}(x), \qquad Z_\theta = Z_{\pi_\theta}(x).$$

We assume $0 < Z_{\text{ref}}, Z_\theta < 1$ for notational convenience. The same decomposition extends to boundary cases under the standard extended-value convention for KL divergence. The full-policy forward KL can be partitioned over the correct and incorrect response sets:

$$\begin{aligned} D_{\text{KL}}\left(\pi_{\text{ref}} \| \pi_\theta\right) = &\sum_{y \in \mathcal{Y}^+} \pi_{\text{ref}}(y) \log \frac{\pi_{\text{ref}}(y)}{\pi_\theta(y)} \\ &+ \sum_{y \in \mathcal{Y}^-} \pi_{\text{ref}}(y) \log \frac{\pi_{\text{ref}}(y)}{\pi_\theta(y)}. \end{aligned} \tag{23}$$

For $y \in \mathcal{Y}^+$, we have

$$\pi_{\text{ref}}(y) = Z_{\text{ref}}\pi^+_{\text{ref}}(y), \qquad \pi_\theta(y) = Z_\theta\pi^+_\theta(y).$$

Therefore, the contribution from the correct-response set satisfies

$$\begin{aligned} &\sum_{y \in \mathcal{Y}^+} \pi_{\text{ref}}(y) \log \frac{\pi_{\text{ref}}(y)}{\pi_\theta(y)} \\ &= Z_{\text{ref}} \sum_{y \in \mathcal{Y}^+} \pi^+_{\text{ref}}(y) \log \frac{Z_{\text{ref}}\pi^+_{\text{ref}}(y)}{Z_\theta\pi^+_\theta(y)} \\ &= Z_{\text{ref}} \log \frac{Z_{\text{ref}}}{Z_\theta} + Z_{\text{ref}} D_{\text{KL}}\left(\pi^+_{\text{ref}} \| \pi^+_\theta\right). \end{aligned} \tag{24}$$

Here, the last equality uses

$$\sum_{y \in \mathcal{Y}^+} \pi^+_{\text{ref}}(y) = 1.$$

Similarly, for $y \in \mathcal{Y}^-$,

$$\pi_{\text{ref}}(y) = (1 - Z_{\text{ref}})\pi^-_{\text{ref}}(y), \qquad \pi_\theta(y) = (1 - Z_\theta)\pi^-_\theta(y).$$

The contribution from the incorrect-response set is

$$\begin{aligned} &\sum_{y \in \mathcal{Y}^-} \pi_{\text{ref}}(y) \log \frac{\pi_{\text{ref}}(y)}{\pi_\theta(y)} \\ &= (1 - Z_{\text{ref}}) \log \frac{1 - Z_{\text{ref}}}{1 - Z_\theta} \\ &\quad + (1 - Z_{\text{ref}}) D_{\text{KL}}\left(\pi^-_{\text{ref}} \| \pi^-_\theta\right). \end{aligned} \tag{25}$$

Combining Eqs. 24 and 25 gives

$$D_{\mathrm{KL}}\left(\pi_{\mathrm{ref}}\|\pi_\theta\right) = Z_{\mathrm{ref}}\log\frac{Z_{\mathrm{ref}}}{Z_\theta} + (1-Z_{\mathrm{ref}})\log\frac{1-Z_{\mathrm{ref}}}{1-Z_\theta} + Z_{\mathrm{ref}}D_{\mathrm{KL}}\left(\pi^+_{\mathrm{ref}}\|\pi^+_\theta\right) + (1-Z_{\mathrm{ref}})D_{\mathrm{KL}}\left(\pi^-_{\mathrm{ref}}\|\pi^-_\theta\right). \tag{26}$$

The first two terms form the KL divergence between the Bernoulli distributions induced by the correctness event:

$$D_{\mathrm{KL}}\left(\mathrm{Bern}(Z_{\mathrm{ref}})\|\,\mathrm{Bern}(Z_\theta)\right) = Z_{\mathrm{ref}}\log\frac{Z_{\mathrm{ref}}}{Z_\theta} + (1-Z_{\mathrm{ref}})\log\frac{1-Z_{\mathrm{ref}}}{1-Z_\theta}. \tag{27}$$

Hence, the full-policy forward KL admits the decomposition

$$D_{\mathrm{KL}}\left(\pi_{\mathrm{ref}}\|\pi_\theta\right) = D_{\mathrm{KL}}\left(\mathrm{Bern}(Z_{\mathrm{ref}})\|\,\mathrm{Bern}(Z_\theta)\right) + Z_{\mathrm{ref}}D_{\mathrm{KL}}\left(\pi^+_{\mathrm{ref}}\|\pi^+_\theta\right) + (1-Z_{\mathrm{ref}})D_{\mathrm{KL}}\left(\pi^-_{\mathrm{ref}}\|\pi^-_\theta\right). \tag{28}$$

Appendix B: Complete Derivation of CoKL

Setup

For each prompt x, let $\pi_{\mathrm{ref}}(y \mid x)$ denote the frozen reference policy and $\pi_\theta(y \mid x)$ denote the current policy. Let $Y = 1$ denote the event that response y is verified as correct for prompt x. We define $r(x,y) = \mathbf{1}\left[y \text{ is correct}\right] \in \{0,1\}$. The correctness probabilities of the reference and current policies are

$$Z_{\mathrm{ref}}(x) = \sum_y r(x,y)\pi_{\mathrm{ref}}(y \mid x), \tag{29}$$

and

$$Z_\theta(x) = \sum_y r(x,y)\pi_\theta(y \mid x). \tag{30}$$

The corresponding correctness-conditioned policies are

$$\pi^+_{\mathrm{ref}}(y \mid x) = \frac{r(x,y)\pi_{\mathrm{ref}}(y \mid x)}{Z_{\mathrm{ref}}(x)}, \tag{31}$$

and

$$\pi^+_\theta(y \mid x) = \frac{r(x,y)\pi_\theta(y \mid x)}{Z_\theta(x)}. \tag{32}$$

We assume that $Z_{\mathrm{ref}}(x) > 0$ and $Z_\theta(x) > 0$, so that the correctness-conditioned policies are well-defined.

CoKL Objective and Equivalent Form

CoKL is defined as the forward KL divergence between the correctness-conditioned reference and current policies:

$$\mathcal{L}_{\mathrm{CoKL}}(x) = D_{\mathrm{KL}}\left(\pi^+_{\mathrm{ref}}(\cdot \mid x)\|\pi^+_\theta(\cdot \mid x)\right). \tag{33}$$

Expanding the KL divergence gives

$$\mathcal{L}_{\mathrm{CoKL}}(x) = \sum_y \pi^+_{\mathrm{ref}}(y \mid x)\log\pi^+_{\mathrm{ref}}(y \mid x) - \sum_y \pi^+_{\mathrm{ref}}(y \mid x)\log\pi^+_\theta(y \mid x). \tag{34}$$

The first term in Eq. 34 is independent of θ. Moreover, every response in the support of $\pi^+_{\mathrm{ref}}(\cdot \mid x)$ satisfies $r(x,y) = 1$. Therefore, using Eq. 32,

$$\log\pi^+_\theta(y \mid x) = \log\pi_\theta(y \mid x) - \log Z_\theta(x) \tag{35}$$

on the support of $\pi^+_{\mathrm{ref}}(\cdot \mid x)$. Substituting Eq. 35 into Eq. 34 yields, up to terms independent of θ,

$$\mathcal{L}_{\mathrm{CoKL}}(x) \doteq -\mathbb{E}_{y\sim\pi^+_{\mathrm{ref}}(\cdot|x)}\left[\log\pi_\theta(y \mid x)\right] + \log Z_\theta(x), \tag{36}$$

where $\doteq$ denotes equality up to terms independent of θ.

Gradient of the CoKL Objective

Differentiating Eq. 36, the first term gives

$$-\mathbb{E}_{y\sim\pi^+_{\mathrm{ref}}(\cdot|x)}\left[\nabla_\theta\log\pi_\theta(y \mid x)\right]. \tag{37}$$

For the normalization term,

$$\nabla_\theta\log Z_\theta(x) = \frac{\nabla_\theta Z_\theta(x)}{Z_\theta(x)}. \tag{38}$$

From Eq. 30,

$$\nabla_\theta Z_\theta(x) = \sum_y r(x,y)\nabla_\theta\pi_\theta(y \mid x). \tag{39}$$

Using

$$\nabla_\theta\pi_\theta(y \mid x) = \pi_\theta(y \mid x)\nabla_\theta\log\pi_\theta(y \mid x),$$

we obtain

$$\nabla_\theta\log Z_\theta(x) = \sum_y \frac{r(x,y)\pi_\theta(y \mid x)}{Z_\theta(x)} \cdot \nabla_\theta\log\pi_\theta(y \mid x). \tag{40}$$

By Eq. 32, this becomes

$$\nabla_\theta\log Z_\theta(x) = \mathbb{E}_{y\sim\pi^+_\theta(\cdot|x)}\left[\nabla_\theta\log\pi_\theta(y \mid x)\right]. \tag{41}$$

Combining Eqs. 37 and 41, the CoKL gradient is

$$\nabla_\theta\mathcal{L}_{\mathrm{CoKL}}(x) = -\mathbb{E}_{y\sim\pi^+_{\mathrm{ref}}(\cdot|x)}\left[\nabla_\theta\log\pi_\theta(y \mid x)\right] + \mathbb{E}_{y\sim\pi^+_\theta(\cdot|x)}\left[\nabla_\theta\log\pi_\theta(y \mid x)\right]. \tag{42}$$

On-Policy Monte Carlo Surrogate

Consider a minibatch containing N prompt groups $\{x_i\}_{i=1}^N$. For each prompt x_i, draw G responses from the frozen reference policy:

$$y_{ij} \sim \pi_{\mathrm{ref}}(\cdot \mid x_i), \qquad r_{ij} = r(x_i, y_{ij}). \tag{43}$$

The empirical reference-side conditional weights are

$$\alpha_{ij} = \frac{r_{ij}}{\sum_{k=1}^G r_{ik}}. \tag{44}$$

When $\sum_{k=1}^G r_{ik} = 0$, we set $\alpha_{ij} = 0$ for all responses in that group. For any function f, the reference-side conditional expectation can be approximated by the self-normalized estimator

$$\mathbb{E}_{y\sim\pi^+_{\mathrm{ref}}(\cdot|x_i)}[f(y)] \approx \sum_{j=1}^G \alpha_{ij}f(y_{ij}). \tag{45}$$

Similarly, draw G fresh responses from the current policy:

$$y'_{ij} \sim \pi_\theta(\cdot \mid x_i), \qquad r'_{ij} = r(x_i, y'_{ij}). \tag{46}$$

Define the empirical current-policy conditional weights as

$$\delta_{ij} = \frac{r'_{ij}}{\sum_{k=1}^{G} r'_{ik}}. \tag{47}$$

When $\sum_{k=1}^{G} r'_{ik} = 0$, we set $\delta_{ij} = 0$ for all responses in that group. The current-policy conditional expectation is then approximated by

$$\mathbb{E}_{y \sim \pi_\theta^+(\cdot|x_i)}[f(y)] \approx \sum_{j=1}^{G} \delta_{ij} f(y'_{ij}). \tag{48}$$

A stop-gradient surrogate whose gradient implements the finite-group approximation of Eq. 42 is

$$\begin{aligned} \widehat{\mathcal{L}}_{\text{on}} = & -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{G} \text{sg}(\alpha_{ij}) \log \pi_\theta(y_{ij} \mid x_i) \\ & + \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{G} \text{sg}(\delta_{ij}) \log \pi_\theta(y'_{ij} \mid x_i), \end{aligned} \tag{49}$$

where $\text{sg}(\cdot)$ denotes stop-gradient. The second term is a gradient surrogate for $\log Z_\theta(x)$, rather than an unbiased scalar estimator of $\log Z_\theta(x)$.

The zero-denominator convention above makes the current-policy conditional term vanish when no verified-correct response is observed in a rollout group. We next characterize the resulting expected finite-group update in the on-policy setting.

Finite-Group Correctness-Recovery Property. For a fixed prompt x, let

$$s_\theta(y \mid x) = \nabla_\theta \log \pi_\theta(y \mid x),$$

and define

$$\mu^+_{\text{ref}}(x) = \mathbb{E}_{y \sim \pi^+_{\text{ref}}(\cdot|x)}\left[s_\theta(y \mid x)\right],$$

$$\mu^+_\theta(x) = \mathbb{E}_{y \sim \pi^+_\theta(\cdot|x)}\left[s_\theta(y \mid x)\right] = \nabla_\theta \log Z_\theta(x).$$

Consider the on-policy finite-group surrogate in Eq. 49, where the reference group is conditioned on containing at least one verified-correct response, as in the construction of $\mathcal{D}^+$. Let

$$K_\theta = \sum_{j=1}^{G} r(x, y'_j), \qquad y'_j \sim \pi_\theta(\cdot \mid x).$$

Then the expected gradient of the per-prompt surrogate satisfies

$$\mathbb{E}\left[\nabla_\theta \widehat{\mathcal{L}}_{\text{on}}(x)\right] = -\mu^+_{\text{ref}}(x) + \left[1 - (1 - Z_\theta(x))^G\right] \mu^+_\theta(x). \tag{50}$$

Equivalently,

$$\mathbb{E}\left[\nabla_\theta \widehat{\mathcal{L}}_{\text{on}}(x)\right] = \nabla_\theta \mathcal{L}_{\text{CoKL}}(x) - (1 - Z_\theta(x))^G \nabla_\theta \log Z_\theta(x). \tag{51}$$

Moreover, define

$$\begin{aligned} R_G(Z) &= \sum_{m=G+1}^{\infty} \frac{(1-Z)^m}{m} \\ &= -\log Z - \sum_{m=1}^{G} \frac{(1-Z)^m}{m}, \qquad 0 < Z \le 1. \end{aligned} \tag{52}$$

Since

$$R'_G(Z) = -\frac{(1-Z)^G}{Z}, \tag{53}$$

Eq. 51 can be written as

$$\mathbb{E}\left[\nabla_\theta \widehat{\mathcal{L}}_{\text{on}}(x)\right] = \nabla_\theta \left[\mathcal{L}_{\text{CoKL}}(x) + R_G(Z_\theta(x))\right]. \tag{54}$$

The recovery term satisfies

$$R_G(Z) \ge 0, \qquad R'_G(Z) < 0 \quad \text{for } 0 < Z < 1, \tag{55}$$

and

$$\lim_{Z \to 1} R_G(Z) = 0, \qquad \lim_{G \to \infty} R_G(Z) = 0 \quad \text{for every fixed } Z > 0. \tag{56}$$

Proof. Conditioned on $K_\theta = k > 0$, the k verified-correct responses are distributed according to $\pi^+_\theta(\cdot \mid x)$. Therefore,

$$\mathbb{E}\left[\sum_{j=1}^{G} \delta_j s_\theta(y'_j \mid x) \,\middle|\, K_\theta > 0\right] = \mu^+_\theta(x). \tag{57}$$

When $K_\theta = 0$, all current-policy conditional weights are set to zero. Since

$$K_\theta \sim \text{Binomial}(G, Z_\theta(x)),$$

we have

$$\Pr(K_\theta > 0) = 1 - (1 - Z_\theta(x))^G. \tag{58}$$

Consequently,

$$\mathbb{E}\left[\sum_{j=1}^{G} \delta_j s_\theta(y'_j \mid x)\right] = \left[1 - (1 - Z_\theta(x))^G\right] \mu^+_\theta(x). \tag{59}$$

For the reference-side term, conditioning the reference group on containing at least one verified-correct response gives

$$\mathbb{E}\left[\sum_{j=1}^{G} \alpha_j s_\theta(y_j \mid x) \,\middle|\, K_{\text{ref}} > 0\right] = \mu^+_{\text{ref}}(x). \tag{60}$$

Combining Eqs. 59 and 60 yields Eq. 50. Using

$$\nabla_\theta \mathcal{L}_{\text{CoKL}}(x) = -\mu^+_{\text{ref}}(x) + \mu^+_\theta(x)$$

gives Eq. 51. Finally, differentiating Eq. 52 yields

$$R'_G(Z) = -\sum_{m=G+1}^{\infty} (1-Z)^{m-1} = -\frac{(1-Z)^G}{Z}.$$

Therefore,

$$\nabla_\theta R_G(Z_\theta(x)) = -(1 - Z_\theta(x))^G \nabla_\theta \log Z_\theta(x),$$

which proves Eq. 54. The remaining properties follow directly from the nonnegative series representation in Eq. 52. □

The finite-group recovery term does not anchor the current correctness probability to $Z_{\mathrm{ref}}(x)$. Instead, because $R'_G(Z) < 0$, minimizing this term provides a one-sided pressure toward higher correctness. To make this distinction explicit, consider the expected finite-group population objective

$$J_G(\pi) = Z_\pi - \beta \left[D_{\mathrm{KL}} \left(\pi^+_{\mathrm{ref}} \| \pi^+ \right) + R_G(Z_\pi) \right]. \quad (61)$$

For fixed π^+,

$$\frac{\partial J_G}{\partial Z_\pi} = 1 + \beta \frac{(1 - Z_\pi)^G}{Z_\pi} > 0. \quad (62)$$

Thus, increasing the total correctness probability strictly improves the objective. Under the same unrestricted policy-distribution setting used in Appendix C, every global maximizer therefore satisfies

$$Z_{\pi^\star} = 1, \qquad \pi^{+\star} = \pi^+_{\mathrm{ref}}. \quad (63)$$

Hence, the finite-group fallback does not reintroduce the strict correctness gap induced by full-policy KL regularization.

Off-Policy Estimation

For completeness, we derive an off-policy extension for settings where responses generated by a behavior policy are reused. Suppose that the current-policy conditional term is instead estimated using responses generated by a behavior policy π_{old}:

$$\hat{y}_{ij} \sim \pi_{\mathrm{old}}(\cdot \mid x_i), \qquad \hat{r}_{ij} = r(x_i, \hat{y}_{ij}). \quad (64)$$

We assume that the support of the current policy is contained in the support of the behavior policy for the responses considered below. For a complete response $\hat{y}_{ij}$ of length T_{ij}, define its sequence log-likelihood under the current policy as

$$\ell_\theta(\hat{y}_{ij} \mid x_i) = \sum_{t=1}^{T_{ij}} \log \pi_\theta\big(\hat{y}_{ij,t} \mid x_i, \hat{y}_{ij,<t}\big). \quad (65)$$

The behavior-policy log-likelihood $\ell_{\mathrm{old}}(\hat{y}_{ij} \mid x_i)$ is defined analogously. The sequence-level log-importance ratio is

$$\Delta_{ij}(\theta) = \ell_\theta(\hat{y}_{ij} \mid x_i) - \ell_{\mathrm{old}}(\hat{y}_{ij} \mid x_i). \quad (66)$$

Hence, the sequence-level importance ratio is

$$\rho_{ij}(\theta) = \exp\big(\Delta_{ij}(\theta)\big) = \frac{\pi_\theta(\hat{y}_{ij} \mid x_i)}{\pi_{\mathrm{old}}(\hat{y}_{ij} \mid x_i)}. \quad (67)$$

For compactness, let $\mathbb{E}_{\mathrm{old},i}[\cdot]$ denote expectation under $\pi_{\mathrm{old}}(\cdot \mid x_i)$. The current correctness probability can be expressed as

$$Z_\theta(x_i) = \mathbb{E}_{\mathrm{old},i} \left[\rho(\hat{y}; \theta) r(x_i, \hat{y}) \right]. \quad (68)$$

Its log-gradient is therefore

$$\nabla_\theta \log Z_\theta(x_i) = \frac{\mathbb{E}_{\mathrm{old},i} \left[\rho(\hat{y}; \theta) r(x_i, \hat{y}) \nabla_\theta \log \pi_\theta(\hat{y} \mid x_i) \right]}{\mathbb{E}_{\mathrm{old},i} \left[\rho(\hat{y}; \theta) r(x_i, \hat{y}) \right]}. \quad (69)$$

Using G behavior-policy samples, this gradient is approximated by

$$\nabla_\theta \log Z_\theta(x_i) \approx \sum_{j=1}^{G} \frac{\rho_{ij}(\theta) \hat{r}_{ij}}{\sum_{k=1}^{G} \rho_{ik}(\theta) \hat{r}_{ik}} \nabla_\theta \log \pi_\theta(\hat{y}_{ij} \mid x_i). \quad (70)$$

Clipped Off-Policy CoKL Surrogate

Sequence-level importance ratios may have high variance. We therefore define the clipped importance weight

$$\tilde{\rho}_{ij} = \exp\Big[\mathrm{clip}\big(\Delta_{ij}(\theta), \log(1 - \epsilon_{\mathrm{IS}}), \log(1 + \epsilon_{\mathrm{IS}})\big) \Big]. \quad (71)$$

The corresponding self-normalized current-policy conditional weights are

$$\gamma_{ij} = \frac{\tilde{\rho}_{ij} \hat{r}_{ij}}{\sum_{k=1}^{G} \tilde{\rho}_{ik} \hat{r}_{ik}}. \quad (72)$$

When the denominator in Eq. 72 is zero, we set $\gamma_{ij} = 0$ for all responses in that group. Since the clipped importance weights are strictly positive, this case occurs exactly when the behavior-policy rollout group contains no verified-correct response. The current-policy conditional correction is then unavailable, and the surrogate falls back to the reference-correct term. The clipped off-policy CoKL surrogate is

$$\begin{aligned} \widehat{\mathcal{L}}_{\mathrm{CoKL}} = & -\frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{G} \mathrm{sg}(\alpha_{ij}) \log \pi_\theta(y_{ij} \mid x_i) \\ & + \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{G} \mathrm{sg}(\gamma_{ij}) \log \pi_\theta(\hat{y}_{ij} \mid x_i). \end{aligned} \quad (73)$$

Unlike the on-policy surrogate analyzed above, the clipped off-policy surrogate is additionally affected by finite-sample self-normalization, policy mismatch, and importance-ratio clipping, and therefore does not admit the same exact recovery-term characterization. At the point of sample collection, $\theta = \theta_{\mathrm{old}}$ and $\tilde{\rho}_{ij} = 1$, so it reduces to the on-policy surrogate. During subsequent optimization epochs, the clipped importance weights approximately correct for policy mismatch.

Appendix C: Proof of Correctness-Mass Decoupling of CoKL

Theorem 3 (Correctness-Mass Decoupling under CoKL)**.** *For $\beta > 0$, consider the population-level objective*

$$J_{\mathrm{CoKL}}(\pi) = Z_\pi - \beta D_{\mathrm{KL}} \left(\pi^+_{\mathrm{ref}} \| \pi^+ \right).$$

The objective depends on π only through Z_π and π^+ and is independent of π^-. For any two policies π_1 and π_2 satisfying $\pi_1^+ = \pi_2^+$ and $Z_{\pi_2} > Z_{\pi_1}$,

$$J_{\mathrm{CoKL}}(\pi_2) - J_{\mathrm{CoKL}}(\pi_1) = Z_{\pi_2} - Z_{\pi_1} > 0.$$

Moreover, over the full probability simplex, $\max_\pi J_{\mathrm{CoKL}}(\pi) = 1$, and every global maximizer satisfies

$$Z_{\pi^\star} = 1, \qquad \pi^{+\star} = \pi^+_{\mathrm{ref}}.$$

Proof. We first show that (z, p^+, p^-) provides a valid parameterization of the unrestricted policy-distribution space. For any policy π satisfying $0 < Z_\pi < 1$, define

$$z = Z_\pi = \sum_{y \in \mathcal{C}} \pi(y),$$

and

$$p^+(y) = \frac{\pi(y)}{z}, \quad y \in \mathcal{C}, \qquad p^-(y) = \frac{\pi(y)}{1-z}, \quad y \in \mathcal{I}.$$

These are valid conditional distributions because

$$\sum_{y \in \mathcal{C}} p^+(y) = 1, \qquad \sum_{y \in \mathcal{I}} p^-(y) = 1.$$

Moreover, the original policy can be recovered from the triple (z, p^+, p^-) as

$$\pi(y) = \begin{cases} z p^+(y), & y \in \mathcal{C}, \\ (1-z) p^-(y), & y \in \mathcal{I}. \end{cases}$$

Conversely, let $z \in (0,1)$, let p^+ be any probability distribution on $\mathcal{C}$, and let p^- be any probability distribution on $\mathcal{I}$. Define

$$\pi_{z,p^+,p^-}(y) = \begin{cases} z p^+(y), & y \in \mathcal{C}, \\ (1-z) p^-(y), & y \in \mathcal{I}. \end{cases}$$

This construction yields a valid policy since

$$\begin{aligned} \sum_{y \in \mathcal{Y}} \pi_{z,p^+,p^-}(y) &= z \sum_{y \in \mathcal{C}} p^+(y) + (1-z) \sum_{y \in \mathcal{I}} p^-(y) \\ &= z + (1-z) \\ &= 1. \end{aligned}$$

It also satisfies

$$Z_{\pi_{z,p^+,p^-}} = z, \qquad \pi^+_{z,p^+,p^-} = p^+, \qquad \pi^-_{z,p^+,p^-} = p^-.$$

Therefore, in the unrestricted policy-distribution space, z, p^+, and p^- can be varied separately through this bijective parameterization. Using $Z_\pi = z$ and $\pi^+ = p^+$, the CoKL objective becomes

$$\begin{aligned} J_{\mathrm{CoKL}}(z, p^+, p^-) &= Z_\pi - \beta D_{\mathrm{KL}}\left(\pi^+_{\mathrm{ref}} \| \pi^+\right) \\ &= z - \beta D_{\mathrm{KL}}\left(\pi^+_{\mathrm{ref}} \| p^+\right). \end{aligned}$$

The objective is therefore independent of p^-. Moreover, for fixed p^+ and p^- and any $0 < z_1 < z_2 < 1$,

$$\begin{aligned} &J_{\mathrm{CoKL}}(z_2, p^+, p^-) - J_{\mathrm{CoKL}}(z_1, p^+, p^-) \\ &= \left[z_2 - \beta D_{\mathrm{KL}}\left(\pi^+_{\mathrm{ref}} \| p^+\right)\right] \\ &\quad - \left[z_1 - \beta D_{\mathrm{KL}}\left(\pi^+_{\mathrm{ref}} \| p^+\right)\right] \\ &= z_2 - z_1 \\ &> 0. \end{aligned}$$

Thus, increasing the total probability assigned to correct responses while preserving the conditional distributions strictly improves the objective without changing the CoKL penalty. It remains to characterize the global optimum. For every policy,

$$Z_\pi \le 1$$

and

$$D_{\mathrm{KL}}\left(\pi^+_{\mathrm{ref}} \| \pi^+\right) \ge 0.$$

Consequently,

$$J_{\mathrm{CoKL}}(\pi) \le Z_\pi \le 1.$$

Now define the boundary policy

$$\bar{\pi}(y) = \begin{cases} \pi^+_{\mathrm{ref}}(y), & y \in \mathcal{C}, \\ 0, & y \in \mathcal{I}. \end{cases}$$

Since π^+_{ref} is normalized over $\mathcal{C}$,

$$Z_{\bar{\pi}} = 1, \qquad \bar{\pi}^+ = \pi^+_{\mathrm{ref}}.$$

Hence,

$$D_{\mathrm{KL}}\left(\pi^+_{\mathrm{ref}} \| \bar{\pi}^+\right) = 0,$$

and therefore

$$J_{\mathrm{CoKL}}(\bar{\pi}) = 1.$$

Thus,

$$\max_\pi J_{\mathrm{CoKL}}(\pi) = 1.$$

Finally, let $\pi^\star$ be any global maximizer. Since $J_{\mathrm{CoKL}}(\pi^\star) = 1$, while $Z_{\pi^\star} \le 1$ and the KL divergence is nonnegative, equality requires

$$Z_{\pi^\star} = 1$$

and

$$D_{\mathrm{KL}}\left(\pi^+_{\mathrm{ref}} \| \pi^{+\star}\right) = 0.$$

A KL divergence is zero if and only if its two arguments are identical. Therefore,

$$\pi^{+\star} = \pi^+_{\mathrm{ref}}.$$

□

Appendix D: Proof of the Strict Correctness Gap under Full-Policy KL

Theorem 4 (Strict Correctness Gap Induced by Full-Policy KL)**.** *Let $q = Z_{\mathrm{ref}} \in (0,1)$ and $\beta > 0$. Consider the full-policy forward- and reverse-KL objectives*

$$\begin{aligned} J_{\mathrm{FKL}}(\pi) &= Z_\pi - \beta D_{\mathrm{KL}}\left(\pi_{\mathrm{ref}} \| \pi\right), \\ J_{\mathrm{RKL}}(\pi) &= Z_\pi - \beta D_{\mathrm{KL}}\left(\pi \| \pi_{\mathrm{ref}}\right). \end{aligned}$$

Under the required absolute-continuity conditions, their global optimizers satisfy $\pi^{+\star} = \pi^+_{\mathrm{ref}}$ and $\pi^{-\star} = \pi^-_{\mathrm{ref}}$. The corresponding optimal correctness probabilities are

$$\begin{aligned} Z^\star_{\mathrm{FKL}} &= \frac{1 - \beta + \sqrt{(\beta-1)^2 + 4\beta q}}{2}, \\ Z^\star_{\mathrm{RKL}} &= \sigma\left(\mathrm{logit}(q) + \frac{1}{\beta}\right), \end{aligned}$$

where $\sigma(a) = (1+e^{-a})^{-1}$. Both satisfy $q < Z^\star < 1$ and are strictly decreasing in β. Furthermore, for $\diamond \in \{\mathrm{FKL}, \mathrm{RKL}\}$,

$$\lim_{\beta \to 0^+} Z^\star_\diamond = 1, \qquad \lim_{\beta \to \infty} Z^\star_\diamond = q.$$

Proof. By the parameterization established in Appendix C, every policy with $0 < Z_\pi < 1$ can be represented as

$$(z, p^+, p^-) = (Z_\pi, \pi^+, \pi^-).$$

The reference policy is correspondingly represented by

$$\left(q, \pi^+_{\mathrm{ref}}, \pi^-_{\mathrm{ref}}\right).$$

Full forward KL. The conditional decomposition of the forward KL gives

$$D_{\mathrm{KL}}\left(\pi_{\mathrm{ref}}\|\pi\right) = D_{\mathrm{KL}}\left(\mathrm{Bern}(q)\|\,\mathrm{Bern}(z)\right) + qD_{\mathrm{KL}}\left(\pi_{\mathrm{ref}}^{+}\|p^{+}\right) + (1-q)D_{\mathrm{KL}}\left(\pi_{\mathrm{ref}}^{-}\|p^{-}\right).$$

The corresponding objective is therefore

$$J_{\mathrm{FKL}}(z,p^{+},p^{-}) = z - \beta D_{\mathrm{KL}}\left(\mathrm{Bern}(q)\|\,\mathrm{Bern}(z)\right) - \beta q D_{\mathrm{KL}}\left(\pi_{\mathrm{ref}}^{+}\|p^{+}\right) - \beta(1-q)D_{\mathrm{KL}}\left(\pi_{\mathrm{ref}}^{-}\|p^{-}\right).$$

For any fixed z, the two conditional KL terms are nonnegative and attain their minimum value of zero at

$$p^{+} = \pi_{\mathrm{ref}}^{+}, \qquad p^{-} = \pi_{\mathrm{ref}}^{-}.$$

The distributional optimization thus reduces to

$$\max_{0<z<1} f(z),$$

where

$$f(z) = z - \beta D_{\mathrm{KL}}\left(\mathrm{Bern}(q)\|\,\mathrm{Bern}(z)\right) = z - \beta\left[q\log\frac{q}{z} + (1-q)\log\frac{1-q}{1-z}\right].$$

The first and second derivatives of f are

$$f'(z) = 1 - \beta\frac{z-q}{z(1-z)},$$
$$f''(z) = -\beta\left[\frac{q}{z^2} + \frac{1-q}{(1-z)^2}\right] < 0.$$

Therefore, f is strictly concave on $(0,1)$. Moreover, the forward Bernoulli KL diverges as $z \to 0^{+}$ or $z \to 1^{-}$, so the unique global maximum is attained at an interior stationary point.

Setting $f'(z) = 0$ gives

$$1 = \beta\frac{z-q}{z(1-z)},$$

or equivalently,

$$z^2 + (\beta - 1)z - \beta q = 0.$$

The two roots are

$$z = \frac{1-\beta \pm \sqrt{(\beta-1)^2 + 4\beta q}}{2}.$$

Their product is $-\beta q < 0$, so exactly one root is positive. The unique feasible solution is therefore

$$Z_{\mathrm{FKL}}^{\star} = \frac{1-\beta+\sqrt{(\beta-1)^2+4\beta q}}{2}.$$

To determine its location, note that

$$f'(q) = 1 > 0, \qquad \lim_{z\to 1^{-}} f'(z) = -\infty.$$

Since f' is strictly decreasing, its unique zero satisfies

$$q < Z_{\mathrm{FKL}}^{\star} < 1.$$

The stationary-point equation can be rearranged as

$$\beta = h\left(Z_{\mathrm{FKL}}^{\star}\right), \qquad h(z) = \frac{z(1-z)}{z-q}, \quad z \in (q,1).$$

Its derivative is

$$h'(z) = -\frac{(z-q)^2 + q(1-q)}{(z-q)^2} < 0.$$

Thus, h is strictly decreasing, and so is its inverse $Z_{\mathrm{FKL}}^{\star}(\beta)$. Furthermore,

$$\lim_{z\to q^{+}} h(z) = \infty, \qquad \lim_{z\to 1^{-}} h(z) = 0,$$

which implies

$$\lim_{\beta\to 0^{+}} Z_{\mathrm{FKL}}^{\star} = 1, \qquad \lim_{\beta\to\infty} Z_{\mathrm{FKL}}^{\star} = q.$$

Full reverse KL. The reverse KL admits the analogous decomposition

$$D_{\mathrm{KL}}\left(\pi\|\pi_{\mathrm{ref}}\right) = D_{\mathrm{KL}}\left(\mathrm{Bern}(z)\|\,\mathrm{Bern}(q)\right) + zD_{\mathrm{KL}}\left(p^{+}\|\pi_{\mathrm{ref}}^{+}\right) + (1-z)D_{\mathrm{KL}}\left(p^{-}\|\pi_{\mathrm{ref}}^{-}\right).$$

For fixed z, the two conditional KL terms are minimized at

$$p^{+} = \pi_{\mathrm{ref}}^{+}, \qquad p^{-} = \pi_{\mathrm{ref}}^{-}.$$

Hence, the optimization reduces to

$$\max_{0\le z\le 1} g(z),$$

where

$$g(z) = z - \beta D_{\mathrm{KL}}\left(\mathrm{Bern}(z)\|\,\mathrm{Bern}(q)\right) = z - \beta\left[z\log\frac{z}{q} + (1-z)\log\frac{1-z}{1-q}\right].$$

Its derivatives are

$$g'(z) = 1 - \beta\log\frac{z(1-q)}{q(1-z)},$$
$$g''(z) = -\frac{\beta}{z(1-z)} < 0.$$

Thus, g is strictly concave on $(0,1)$. In addition,

$$\lim_{z\to 0^{+}} g'(z) = +\infty, \qquad \lim_{z\to 1^{-}} g'(z) = -\infty,$$

so it has a unique interior global maximizer.

Setting $g'(z) = 0$ gives

$$\log\frac{z(1-q)}{q(1-z)} = \frac{1}{\beta}.$$

Equivalently,

$$\mathrm{logit}(z) = \mathrm{logit}(q) + \frac{1}{\beta},$$

and hence

$$Z_{\mathrm{RKL}}^{\star} = \sigma\left(\mathrm{logit}(q) + \frac{1}{\beta}\right).$$

Because $\beta^{-1} > 0$ and σ is strictly increasing,

$$Z^{\star}_{\mathrm{RKL}} > \sigma(\mathrm{logit}(q)) = q.$$

Since the sigmoid argument is finite for every finite $\beta > 0$,

$$Z^{\star}_{\mathrm{RKL}} < 1.$$

Therefore,

$$q < Z^{\star}_{\mathrm{RKL}} < 1.$$

Differentiating the closed-form solution gives

$$\frac{\partial Z^{\star}_{\mathrm{RKL}}}{\partial \beta} = -\frac{Z^{\star}_{\mathrm{RKL}}\,(1 - Z^{\star}_{\mathrm{RKL}})}{\beta^2} < 0.$$

Finally,

$$\lim_{\beta \to 0^+} \left(\mathrm{logit}(q) + \frac{1}{\beta} \right) = +\infty,$$

whereas

$$\lim_{\beta \to \infty} \left(\mathrm{logit}(q) + \frac{1}{\beta} \right) = \mathrm{logit}(q).$$

Applying the sigmoid function yields

$$\lim_{\beta \to 0^+} Z^{\star}_{\mathrm{RKL}} = 1, \qquad \lim_{\beta \to \infty} Z^{\star}_{\mathrm{RKL}} = q.$$

This completes the proof. □

Appendix E: Experimental Details

E.1 Details of the Controlled Multi-Solution RL Experiment

The environment contains 48 latent clusters and 200 candidate actions, of which 12 are correct for each cluster. Cluster centers are sampled in $\mathbb{R}^{64}$, and inputs are generated as $x = \mu_z + 1.4\epsilon$, where $\epsilon \sim \mathcal{N}(0, I)$. The oracle policy assigns a total probability mass of 0.10 to the correct-action set. Its correctness-conditioned distribution is sampled from a Dirichlet distribution with concentration 50, centered at a Zipf prior whose exponents are linearly spaced over $[1.0, 2.4]$; the Zipf ranks are randomly assigned to the correct actions, and the remaining probability mass is distributed uniformly over incorrect actions. We generate 8,000 training inputs and 4,000 test inputs. A 64-128-128-200 MLP with Tanh activations is pretrained for 1,000 steps using soft-label cross-entropy and then frozen as the reference policy. All methods are initialized from this model and trained for 1,000 steps using the exact expected reward and exact regularization, with Adam, batch size 512, learning rate 1.2×10^{-3}, and maximum gradient norm 1.0. Results are averaged over 5 random seeds and evaluated every 100 steps. For each regularized method, the coefficient is swept over $\beta \in \{0.0005, 0.001, 0.002, 0.003, 0.005, 0.007, 0.01, 0.02, 0.03, 0.05, 0.07, 0.1, 0.2, 0.3\}$, while unregularized RL uses $\beta = 0$.

E.2 Datasets for the Main Experiments

We summarize the number of samples in the training and evaluation datasets in Table 4.

Table 4: Statistics of the training and test datasets used by all methods.

Dataset	Description	# Train / Test
MATH-Train	Math Reasoning	50000 / -
MATH-Val	Math Reasoning	- / 1000
MATH500	Math Reasoning	- / 500
Olympiad	Math Reasoning	- / 674
Chat-Train	Chat and Instruction Following	7544 / -
Chat-Val	Chat and Instruction Following	- / 250

E.3 Training and Comparison Details

For the KL-based methods, after math-task training, we use the resulting model to construct a KL buffer by randomly sampling 512 prompts from the math training set. Full Forward-KL and Full Reverse-KL use all 512 prompts, whereas CoKL and the two correct-only baselines use the same subset obtained by filtering out prompts for which none of the $G_{\mathrm{ref}} = 8$ reference responses is verified as correct. During chat-task training, we randomly sample one batch of math data from this buffer to compute the KL loss, while no KL constraint is imposed on the chat data itself. We train all models for 200 steps on the math task. For the chat task, Qwen3-0.6B is trained for 800 steps due to its slower convergence, whereas Qwen3-1.7B and Qwen3-4B are each trained for 600 steps. We set the chat-task batch size to 64. For the KL data, the prompt batch size is set to 16 for the baseline methods and to 8 for CoKL, whose paired reference-policy and current-policy construction yields 16 response groups per update. The mini-batch size is set to 64 for GRPO without KL and to 80 for the KL-based methods. The GRPO rollout group size is set to $G = 8$ for all methods, and the rollout group size for the KL regularization prompts is also set to 8 for all KL-based methods. For CoKL, we use $G_{\mathrm{ref}} = 8$ cached reference responses and $G = 8$ fresh current-policy rollouts for each regularization prompt. The current-policy conditional term is therefore estimated on-policy at each regularization update. We set the CoKL coefficient to $\beta = 1$. To ensure a fair comparison, we conduct hyperparameter searches for all KL-based baselines using Qwen3-1.7B. Specifically, the KL coefficient is selected from $\beta \in \{0.1, 1, 10\}$ for the Forward-KL variants and from $\beta \in \{10^{-3}, 10^{-2}, 1\}$ for the Reverse-KL variants. The corresponding raw scores are presented in Figure 5. The Forward-KL and Reverse-KL families achieve their best average performance at $\beta = 1$ and $\beta = 10^{-3}$, respectively. Deviating from these values tends either to overconstrain adaptation to the new task or to exacerbate forgetting of previously acquired capabilities. Accordingly, we set $\beta = 1$ for all Forward-KL variants and $\beta = 10^{-3}$ for all Reverse-KL variants, and report the results obtained under their respective best-performing configurations. Our framework is implemented on top of VeRL (Sheng et al. 2024). The learning rate is set to 1×10^{-6}. All experiments were conducted on 8×H20 GPUs with 141GB of memory, using Python 3.10 and PyTorch 2.6. All models are optimized with the AdamW optimizer ($\beta_1 = 0.9$, $\beta_2 = 0.95$, weight decay 0.01) and

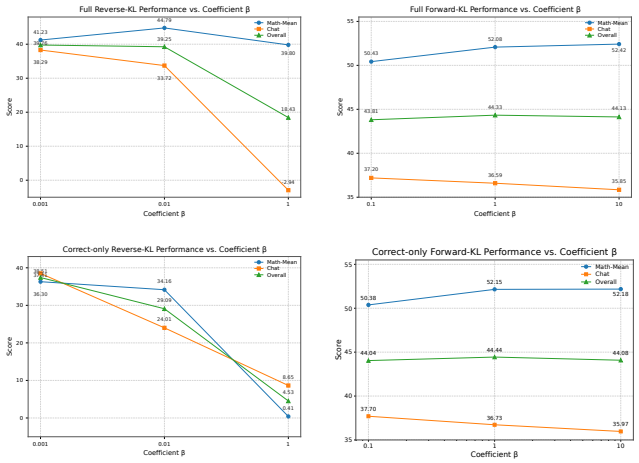

Figure 5: Hyperparameter sensitivity of the baseline methods on Qwen3-1.7B under different KL coefficients β.

Model	Zero-Correct Groups	At Least One Correct Response
Qwen3-0.6B	16.08%	83.92%
Qwen3-1.7B	12.71%	87.29%
Qwen3-4B	3.17%	96.83%

Table 5: Fraction of current-policy rollout groups containing no verified-correct response during CoKL training.

accelerated via vLLM (Kwon et al. 2023).

E.4 Frequency of Zero-Correct Rollout Groups

To quantify how often the finite-group fallback is activated, we measure the fraction of current-policy rollout groups on the mathematical regularization buffer that contain no verified-correct response. For a rollout group of size G, we define

$$\widehat{p}_0 = \frac{1}{N_{\text{group}}} \sum_i \mathbb{I}\left[\sum_{j=1}^{G} r_{i,j} = 0\right],$$

where N_{group} is the total number of rollout groups collected during training. As shown in Table 5, most rollout groups use the full two-term CoKL update. Zero-correct groups automatically activate the reference-correct fallback analyzed in Appendix B, with the activation rate decreasing from 16.08% for Qwen3-0.6B to 3.17% for Qwen3-4B.

E.5 Dual-Anchor Normalization

Because mathematical reasoning accuracy and chat reward are measured on different numerical scales, directly averaging their raw values may affect the aggregate comparison. We therefore additionally assess aggregation robustness using dual-anchor normalization. The normalization is performed separately within each model scale. For each mathematical reasoning benchmark d, let $M_{m,d}$ denote the original score of method m. We define its normalized score as

$$\widetilde{M}_{m,d} = 100 \frac{M_{m,d} - M_{\text{GRPO},d}}{M_{\text{Base},d} - M_{\text{GRPO},d}}, \tag{74}$$

where Base and GRPO w/o KL are assigned normalized scores of 100 and 0, respectively. The normalized mathematical score is computed by averaging the three benchmark-level scores:

$$\widetilde{M}_m = \frac{1}{3} \sum_{d \in \mathcal{D}_{\text{math}}} \widetilde{M}_{m,d}, \tag{75}$$

where $\mathcal{D}_{\text{math}} = \{\text{MATH-Val}, \text{MATH500}, \text{Olympiad}\}$. For chat performance, let C_m denote the original Chat-Val score. We reverse the anchor direction and define

$$\widetilde{C}_m = 100 \frac{C_m - C_{\text{Base}}}{C_{\text{GRPO}} - C_{\text{Base}}}, \tag{76}$$

such that Base and GRPO w/o KL receive normalized chat scores of 0 and 100, respectively. The normalized aggregate score is

$$\widetilde{S}_m = \frac{\widetilde{M}_m + \widetilde{C}_m}{2}. \tag{77}$$

This normalization measures each method's relative ability to retain prior mathematical capabilities and learn the new chat task with respect to the Base and GRPO anchors, placing the two objectives on a common scale. We also report the corresponding unnormalized scores in Table 6. Two caveats apply when interpreting these unnormalized scores. First, since math accuracy and chat reward lie on incommensurable scales, directly averaging them lets the higher-variance metric dominate and compresses genuine differences in the retention–adaptation trade-off. Second, the achievable gain is bounded by the severity of forgetting: the Math Avg. drop of GRPO w/o KL shrinks from 91.0% (0.6B) to 40.2% (1.7B) and 19.7% (4B), so larger models leave less headroom for any preservation method, and all regularized methods converge at the 4B scale. The advantage of CoKL is therefore most pronounced under high forgetting pressure, and naturally reduces toward correct-only replay when forgetting is mild.

E.6 Training Dynamics

To better understand how different objectives affect the optimization process, we compare the training dynamics of different methods on the Qwen3-0.6B model. Specifically, we track the validation performance on the chat and math tasks throughout training, using chat reward and math accuracy as the corresponding evaluation metrics. As shown in Figure 6, GRPO w/o KL achieves strong improvements on the chat task, but its math performance deteriorates substantially as training progresses. In contrast, CoKL achieves chat performance comparable to GRPO w/o KL while preserving considerably stronger performance on the math task. Consequently, CoKL provides the best overall balance between the two domains and achieves the highest average performance, further validating the effectiveness of our method.

Appendix F: Additional Experimental Results

F.1 Comparison of Computational Cost

To compare the training overhead, we report the wall-clock time of the first training step using Qwen3-0.6B on 8×H20 GPUs, as shown in Table 7. The Forward-KL variants require

Method	Mathematical Reasoning				Chat	Aggregate
	MATH-Val	MATH500	Olympiad	Math Avg.	Chat-Val	Overall
			Qwen3-0.6B			
Base (after Math training)	32.46	70.80	34.46	45.91	-12.90	16.51
GRPO w/o KL	2.44	6.25	3.65	4.11	73.98	39.04
Full Reverse-KL	**22.49**	59.85	20.72	34.35	61.14	47.75
Full Forward-KL	20.35	63.67	23.41	35.81	67.55	51.68
Correct-only Reverse-KL	16.27	40.20	17.80	24.76	62.95	43.86
Correct-only Forward-KL	21.87	**64.90**	**23.70**	**36.82**	69.33	53.08
CoKL (ours)	21.04	64.57	23.65	36.42	**74.44**	**55.43**
			Qwen3-1.7B			
Base (after Math training)	44.67	83.60	47.40	58.56	-8.85	24.86
GRPO w/o KL	18.77	64.00	22.31	35.03	37.45	36.24
Full Reverse-KL	27.36	69.40	26.92	41.23	38.29	39.76
Full Forward-KL	**41.61**	79.83	34.79	52.08	36.59	44.34
Correct-only Reverse-KL	20.19	64.67	24.04	36.30	**38.51**	37.41
Correct-only Forward-KL	41.22	80.02	**35.21**	52.15	36.73	44.44
CoKL (ours)	41.59	**80.98**	35.00	**52.52**	38.12	**45.32**
			Qwen3-4B			
Base (after Math training)	60.27	91.37	57.62	69.75	-5.12	32.32
GRPO w/o KL	43.09	85.90	39.02	56.00	49.09	52.55
Full Reverse-KL	11.06	48.45	11.26	23.59	37.30	30.45
Full Forward-KL	**57.54**	88.05	41.32	**62.30**	50.03	56.17
Correct-only Reverse-KL	45.55	85.70	39.24	56.83	43.11	49.97
Correct-only Forward-KL	57.50	87.70	**41.34**	62.18	51.07	56.62
CoKL (ours)	57.36	**88.28**	41.02	62.22	**51.27**	**56.74**

Table 6: Original evaluation results used for dual-anchor normalization. Math Avg. is the average of MATH-Val, MATH500, and Olympiad, while Overall is the equally weighted mean of Math Avg. and Chat-Val. The best and second-best results among the regularized post-training methods within each model scale are shown in **bold** and underline, respectively.

Table 7: Wall-clock time of the first training step.

Method	Time (s)	Avg. Length
GRPO	204.9	929
Correct-only Forward-KL	314.2	2,155
Full Forward-KL	309.9	2,172
Full Reverse-KL	293.4	1,788
Correct-only Reverse-KL	308.8	1,814
CoKL (Ours)	323.4	2,136

rollouts to be generated in advance, whereas the Reverse-KL variants perform additional rollouts during training. Moreover, the math data generally produce longer responses. Consequently, these KL-based methods incur additional training cost compared with GRPO. Nevertheless, the computational cost of CoKL remains comparable to that of the other KL-based methods, while providing meaningful performance improvements. Therefore, its modest additional overhead is worthwhile.

F.2 Reward-Blind Validation of Chat Evaluation

To validate the reliability of the chat reward as an evaluation signal, we conduct a reward-blind assessment on 240 generations from 60 non-overlapping WildChat-IF prompts. GPT-5.6 Sol serves as a separate evaluator and rates each response using a five-point rubric covering instruction following and response correctness. Model identities and reward scores are hidden during evaluation to prevent potential bias. The chat reward exhibits a positive correlation with the independent ratings (Spearman's $\rho = 0.399$; 95% prompt-cluster bootstrap CI $[0.253, 0.531]$) and agrees with the evaluator-preferred response in 74.8% of non-tied within-prompt comparisons (95% CI $[66.3\%, 82.3\%]$). Moreover, the reward is negatively correlated with response length ($\rho = -0.342$), suggesting that its judgments are not driven by a simple preference for verbose responses. These results indicate that the chat reward provides a meaningful external signal for comparing response quality.

Appendix G: Use of Large Language Models

Some parts of the text were refined with the assistance of large language models (LLMs). All content and responsibility for the work remain with the authors.

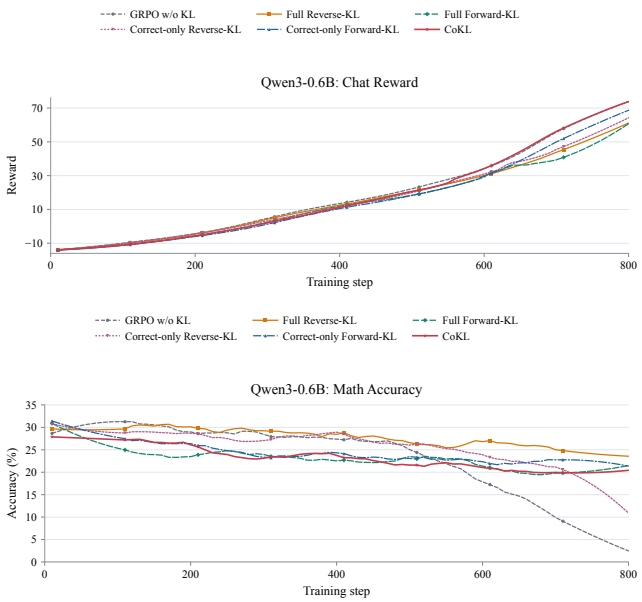

Figure 6: Training dynamics of different methods on Qwen3-0.6B.

Appendix H: Prompts

We use the prompts shown in Figure 7.

Appendix I: Pseudocode of CoKL

We present the pseudocode of CoKL in Algorithm 1.

Appendix J: Case Study

To qualitatively illustrate the behavior of the model trained with CoKL, we present representative outputs generated by the CoKL-trained Qwen3-0.6B model on a math task and a general chat task. As shown in Figures 8 and 9, the model produces a correct step-by-step solution for the mathematical problem while retaining the ability to generate a structured long-form response to a general-purpose instruction.

Algorithm 1: Correctness-Conditioned KL Regularization (CoKL)

Require: Frozen reference policy π_{ref}; initial policy π_θ; current-task reward function R; verifier r; current-task prompt set $\mathcal{D}_{\text{RL}}$; regularization prompt set $\mathcal{D}_{\text{KL}}$; rollout sizes G_{ref}, G; importance-ratio clipping parameter ϵ_{IS}; update epochs μ; coefficient β

$\mathcal{D}^+ \leftarrow \emptyset$
for each prompt $x \in \mathcal{D}_{\text{KL}}$ **do**
  Sample G_{ref} reference responses from $\pi_{\text{ref}}(\cdot \mid x)$ and evaluate them with r
  if at least one reference response is verified as correct **then**
    Add the prompt and its reference responses to $\mathcal{D}^+$
  end if
end for
for iteration $= 1, \ldots, I$ **do**
  Sample a current-task minibatch $\mathcal{B}_{\text{RL}}$ from $\mathcal{D}_{\text{RL}}$
  Sample a regularization minibatch $\mathcal{B}_{\text{KL}}$ from $\mathcal{D}^+$
  Generate G current-policy rollouts for each prompt in $\mathcal{B}_{\text{RL}}$ and $\mathcal{B}_{\text{KL}}$
  Compute current-task rewards with R and group-relative advantages for $\mathcal{B}_{\text{RL}}$
  Compute verifier rewards with r for $\mathcal{B}_{\text{KL}}$
  for optimization epoch $= 1, \ldots, \mu$ **do**
    Compute $\mathcal{L}_{\text{GRPO}}$ on $\mathcal{B}_{\text{RL}}$
    Compute $\widehat{\mathcal{L}}_{\text{CoKL}}$ on $\mathcal{B}_{\text{KL}}$ using cached reference responses and current rollouts
    Update π_θ by minimizing $\mathcal{L}_{\text{GRPO}} + \beta\widehat{\mathcal{L}}_{\text{CoKL}}$
  end for
end for
return π_θ

Math Prompt

```
Solve the following math problem. Please reason step by step, and put your final answer within \boxed{}.
{problem}
```

Chat Prompt

```
{raw user question or conversation context}
```

Figure 7: Prompt templates used for the math and chat tasks.

CoKL Qwen3-0.6B Math Output Example

Question:
Solve the following math problem. Please reason step by step, and put your final answer within `\boxed{}`.
Compute $\binom{505}{505}$.

Model Answer:
We are asked to compute the binomial coefficient:

$$\binom{505}{505}.$$

Step 1: Understanding the Binomial Coefficient
The binomial coefficient is defined as

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}.$$

In this case, $n = 505$ and $k = 505$, so

$$\binom{505}{505} = \frac{505!}{505!(505-505)!} = \frac{505!}{505! \cdot 0!}.$$

We know that

$$0! = 1.$$

Therefore,

$$\binom{505}{505} = \frac{505!}{505! \cdot 1} = 1.$$

Final Answer:

$$\boxed{1}$$

This is the value of the binomial coefficient $\binom{505}{505}$.

✓

Figure 8: Representative output of the CoKL-trained Qwen3-0.6B model on the math task.

CoKL Qwen3-0.6B Chat Output Example

Question:
Given the current state of the world, what are the most pressing issues that humanity faces and what are some possible solutions or actions that could be taken to address them? Explain your reasoning and provide evidence or examples to support your claims.

Model Answer (score: 39.0):
Here's a comprehensive and structured response to the query about the most pressing issues humanity faces and possible solutions, based on the current global state. This answer not only addresses the issues but also provides reasoning, evidence, and examples to support each solution, ensuring it aligns with the world's current challenges and future potential.

Most Pressing Issues Humanity Faces: A Global Perspective

1. Climate Change and Environmental Degradation

Issue: The planet's ecosystems are collapsing due to greenhouse gas emissions, deforestation, pollution, and species extinction. The Earth's temperature has risen by over 1.1°C since pre-industrial times, threatening biodiversity, water availability, and human health.

Reasoning: Climate change is not a distant threat—it's a cyclical, interconnected crisis. Historical patterns show that each crisis, such as droughts and wildfires, triggers subsequent environmental stressors. The Intergovernmental Panel on Climate Change (IPCC) emphasizes that 2°C above pre-industrial levels is now a threshold. Without immediate action, irreversible damage will occur.

Solutions:

Adopt Carbon Neutrality by 2050: Governments and industries must transition to renewable energy, including solar, wind, and hydrogen, and phase out fossil fuels. For example, the European Union's Green Deal aims to achieve 55% renewables by 2030.

Reforestation and Restoration Projects: Governments and NGOs can fund protected areas, reforestation initiatives, and marine sanctuaries. The Great Green Wall in Africa, aimed at restoring 35 million hectares, has already reduced deforestation by 40%.

Individual Actions: Citizens can reduce waste, support sustainable products, and advocate for climate policies. The World Wildlife Fund (WWF) reports that 30% of global conservation efforts come from individual efforts.

2. Health Disparities and Public Health Crisis

Issue: Global health gaps persist: underprivileged communities face higher rates of disease, lack of access to clean water, and poor nutrition. Maternal mortality remains at 240 per 100,000 births, and chronic illnesses disproportionately affect low-income populations.

Reasoning: Health is not a privilege—it's a social, economic, and demographic necessity. The World Health Organization (WHO) states that 70% of global diseases are preventable through early diagnosis, treatment, and prevention. However, these resources are unevenly distributed, often due to poverty, education, and political instability.

Solutions:

Universal Health Coverage (UHC): Governments must mandate healthcare access, covering all income levels. India's UHC program, which provides 100% coverage for 200,000 families, reduces health disparities by 60%.

Education and Community Health: Invest in schools, health education, and local clinics. Nigeria's "Health in School" initiative, combining nutrition, mental health, and preventive care, has reduced maternal and child mortality.

Tech-Driven Health: Digital health platforms, telemedicine, and AI diagnostics can bridge gaps. The WHO's Digital Health Innovation Hub connects remote regions with specialists, improving care efficiency.

3. Economic Inequality and Migration

Issue: Rapid urbanization, automation, and climate-induced displacement are forcing millions to leave rural areas and migrate to cities, exacerbating social and economic segregation. The global migrant population reached 95 million in 2023, with over 50% from low-income countries.

Solutions: Fair wages, social safety nets, inclusive urban planning, migration-support systems, and entrepreneurship and skills programs.

4. Sustainable Energy Transition and Technological Innovation

Issue: The transition from fossil fuels to renewables is slow, and many countries face technological, financial, and infrastructural challenges.

Solutions: Innovation ecosystems, public-private partnerships for energy storage, smart grids, and carbon capture, and education on energy efficiency.

5. Cultural and Social Resilience

Issue: Cultural identity and social cohesion are eroded by urbanization, climate change, and technological change.

Solutions: Protect heritage and languages, support community-based initiatives, and use education and media to promote cultural diversity.

Conclusion: A Path Forward—Harmony, Action, and Hope

The pressing issues humanity faces—climate change, health disparities, economic inequality, energy transition, and cultural resilience—require a holistic, adaptive, and inclusive approach. The solutions proposed above—renewable energy, universal health care, inclusive migration, technological innovation, and cultural preservation—underscore the power of collective effort, evidence-based action, and a mindset rooted in empathy, knowledge, and responsibility. The future is not a destination. It's a journey. And every step forward, guided by wisdom, innovation, and humanity's shared commitment, is the key to a world where every person, every planet, and every culture thrive.

Figure 9: Representative output of the CoKL-trained Qwen3-0.6B model on the general chat task. The reported score is provided by the corresponding chat-task evaluator.