CoReD: Council-weighted Reasoning Distillation

Anonymous authors
Paper under double-blind review

Abstract

Standard chain-of-thought distillation transfers reasoning ability to small models, but typically treats all reasoning steps uniformly despite substantial differences in their importance within a reasoning trajectory. Existing step-aware methods largely use importance signals for hard selection or pruning, leaving the retained steps weighted uniformly. We propose **Co**uncil-Weighted **Re**asoning **D**istillation (CoReD), which derives continuous step weights from a deliberately diversified council of LoRA experts. For each step, we combine the experts' average predictive uncertainty with their normalized disagreement, capturing whether a step is difficult for individual experts or controversial across the council. To make disagreement informative, the experts are explicitly diversified at uncertain reasoning steps before being consolidated in weight space. The merged adapter initializes the student and acts as a frozen distributional anchor during subsequent step-weighted optimization. After training, the learned adapter is merged into the backbone, introducing no additional inference overhead. Across mathematical reasoning benchmarks, CoReD consistently outperforms standard CoT distillation and recent step-aware reasoning methods.

1 Introduction

Chain-of-thought (CoT) reasoning improves language models on multi-step tasks such as mathematical problem solving (Wei et al., 2022; Kojima et al., 2022). Because this capability is most reliable in large models, teacher-generated reasoning trajectories are commonly distilled into smaller students through supervised fine-tuning (Hsieh et al., 2023; Li et al., 2023; Chen et al., 2024). The standard maximum-likelihood objective, however, weights every token equally despite substantial variation in the training value of different reasoning steps.

Predictive entropy in CoT trajectories is concentrated at a small number of decision points (Wang et al., 2025; Li et al., 2026), motivating step-aware distillation methods that discard, mask, or truncate reasoning according to estimated importance (Li et al., 2026; Fan et al., 2026; Wang et al., 2026; Liu et al., 2026). Although effective for compression, such binary decisions cannot distinguish the relative training value of retained steps. Continuous weighting has been explored using student-teacher divergence (Zhong et al., 2026), but this signal is derived from a single teacher and does not capture variation across plausible predictors.

We propose **Co**uncil-Weighted **Re**asoning **D**istillation (CoReD), which turns predictive uncertainty and cross-expert disagreement into continuous step-level training weights. CoReD constructs a council of LoRA experts (Hu et al., 2022; Wang et al., 2023a; Balabanov & Linander, 2025; Mühlematter et al., 2026) that share a frozen backbone but are explicitly encouraged to maintain diverse predictions at uncertain reasoning steps. For each reasoning step, we measure both the experts' mean predictive entropy and their disagreement, quantified by the mutual information between the predicted token and the expert identity (Houlsby et al., 2011; Depeweg et al., 2018; Malinin & Gales, 2018; Riaz et al., 2026). Unlike single-teacher entropy, which only captures how uncertain an individual predictor is, council disagreement reveals when otherwise confident predictors arrive at different conclusions, providing a signal of variation across alternative predictions.

CoReD uses two training phases:

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Phase 1 prepares the LoRA council to be both accurate and predictively diverse. We first adaptively truncate teacher reasoning trajectories by retaining the shortest prefix that allows the base model to recover the correct answer with sufficient reliability, while always preserving the final answer supervision. We then train multiple LoRA experts with three objectives: standard next-token prediction with step-level supervision dropout to expose experts to different subsets of the reasoning trajectory, dual orthogonality regularization (Wang et al., 2023b) to reduce structural overlap among their LoRA updates, and functional repulsion (D'Angelo & Fortuin, 2021; Steger et al., 2024) to separate their predictive distributions.

Phase 2 freezes the trained council, computes the continuous step weights, and consolidates the experts through weight-space merging using TIES (Yadav et al., 2023; Yu et al., 2024). The resulting merged adapter provides a single-model target distribution for the KL objective and also initializes the student LoRA adapter (Hu et al., 2022). During subsequent step-weighted optimization, the merged adapter remains frozen and acts as a distributional anchor, providing a stable reference while the student learns from the weighted reasoning trajectory. The council is therefore required only during training, unlike routing-based LoRA architectures that retain multiple experts at inference (Wu et al., 2024; Zhuang et al., 2026; Huang et al., 2026; Saliencro et al., 2026).

Our contributions are as follows.

- We propose **Co**uncil-Weighted **Re**asoning **D**istillation, a continuous step-weighting framework that combines predictive uncertainty and cross-expert disagreement from a council of LoRA experts, moving beyond binary step selection in reasoning distillation.
- We develop a targeted diversity-training strategy that concentrates cross-expert disagreement at high-entropy reasoning steps. This enables the council to provide step-level signals without relying on reward models or step-level supervision.
- We consolidate the expert council through weight-space merging and use the merged adapter to initialize the student and as a frozen distributional anchor during distillation.
- Experiments on four mathematical reasoning benchmarks and three student backbones show that CoReD attains the best average Pass@1 and Pass@3 over standard CoT distillation and recent step-aware reasoning baselines.

2 Related Work

Step-level supervision in CoT distillation. Fine-tuning small models on teacher-generated rationales is an effective way to transfer reasoning ability (Hsieh et al., 2023; Li et al., 2023; Magister et al., 2023; Muennighoff et al., 2025), although small students may struggle to learn from long traces produced by stronger reasoners (Li et al., 2025). Since standard supervised fine-tuning weights all tokens equally, recent work estimates the importance of individual steps and restricts supervision accordingly. Li et al. (2026) prune low-entropy steps; Fan et al. (2026) suppress steps whose gradients fall outside a low-rank consensus subspace; Wang et al. (2026) use integrated-gradient attribution to mask the loss on unimportant segments; and Liu et al. (2026) adaptively select teacher prefixes to guide student completions, with both providing supervision. In contrast, Zhong et al. (2026) continuously reweight steps using student-teacher divergence in on-policy distillation. CoReD likewise uses continuous weights, but derives them from the uncertainty and disagreement of a council of experts rather than from a single model or a single teacher.

Ensemble uncertainty and diversity in LoRA ensembles. The predictive uncertainty of an ensemble decomposes into the expected uncertainty of its members and the disagreement among them, a quantity used as the mutual-information acquisition score in Bayesian active learning (Houlsby et al., 2011). LoRA ensembles provide a parameter-efficient alternative to full deep ensembles (Lakshminarayanan et al., 2017; Wang et al., 2023a; Balabanov & Linander, 2025; Mühlematter et al., 2026), but sharing a frozen backbone can limit diversity among members. Prior work promotes diversity through particle-based repulsion (Liu & Wang, 2016; D'Angelo & Fortuin, 2021; Trinh et al., 2024), orthogonal subspace constraints (Wang et al., 2023b), bootstrap resampling (Breiman, 1996), dropout (Srivastava et al., 2014), or diversity-regularized reward ensembles (Zhai et al., 2024). CoReD instead uses council signals to construct reasoning-step weights and concentrates repulsion specifically on high-uncertainty steps.

Multi-LoRA composition and model merging. Multiple LoRA adapters (Hu et al., 2022) are commonly composed through learned routing, including mixtures of LoRA experts with static or dynamic gates (Wu et al., 2024; Zhuang et al., 2026), which retain multiple experts at inference. Model merging instead combines fine-tuned models directly in parameter space (Wortsman et al., 2022; Ilharco et al., 2023; Yadav et al., 2023; Yu et al., 2024). Unlike most model-merging settings, where the merged model is itself the deployed artifact, CoReD treats the merged adapter as an intermediate object: it defines the soft target of a KL objective (Hinton et al., 2015) and initializes the student, which is further optimized under step-weighted supervision.

3 Council-Weighted Reasoning Distillation

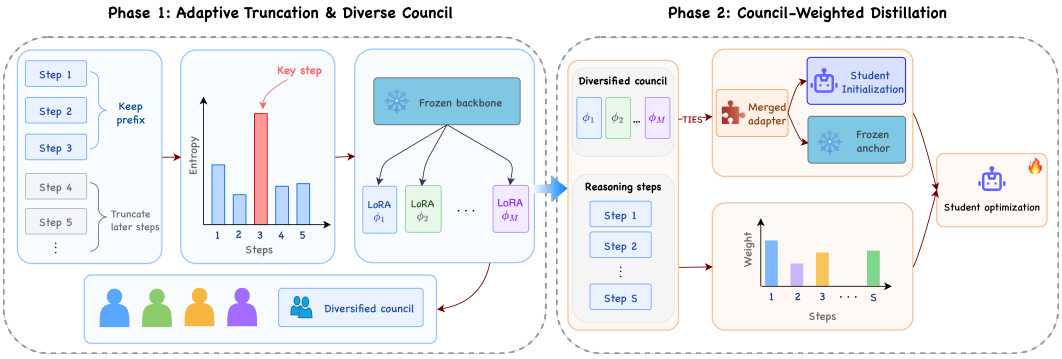

Figure 1: Overview of the CoReD framework.

As illustrated in Figure 1, CoReD assigns continuous weights to reasoning steps using two complementary council signals: within-expert predictive uncertainty and relative cross-expert disagreement. Phase 1 constructs a diversified LoRA council from adaptively truncated teacher trajectories. Phase 2 freezes the council, computes the step weights, merges the experts using TIES, and optimizes a merge-initialized student with weighted NLL and KL regularization.

3.1 Problem Formulation

Let θ_{base} denote the parameters of a frozen language model. Each training example consists of a problem x and a teacher-generated response R, which is segmented at paragraph boundaries into S reasoning steps $\mathcal{S} = \{s_1, \ldots, s_S\}$. Let $y_{i,1:T_i}$ denote the T_i tokens in step s_i. We attach a council of M trainable LoRA adapters $\Phi = \{\phi_1, \ldots, \phi_M\}$. For each adapted layer ℓ, expert m has $A_m^{(\ell)} \in \mathbb{R}^{r \times d_{\text{in}}^{(\ell)}}$ and $B_m^{(\ell)} \in \mathbb{R}^{d_{\text{out}}^{(\ell)} \times r}$, yielding the effective weight

$$W_m^{(\ell)} = W_{\text{base}}^{(\ell)} + \frac{\alpha}{r} B_m^{(\ell)} A_m^{(\ell)}. \tag{1}$$

We write $\theta_{\text{base}} \oplus \phi_m$ for the model whose adapted layers use the weights $W_m^{(\ell)}$ in Equation 1; its other backbone weights remain unchanged.

For any adapter ϕ, with i indexing reasoning steps and t indexing tokens within step s_i, the teacher-forced next-token distribution for $y_{i,t}$ and the length-normalized step loss are

$$\hat{p}_{i,t}^{\phi} = p_{\theta_{\text{base}} \oplus \phi}(\cdot \mid x, s_{<i}, y_{i,<t}), \qquad \mathcal{L}_i^{\phi} = -\frac{1}{T_i} \sum_{t=1}^{T_i} \log \hat{p}_{i,t}^{\phi}(y_{i,t}). \tag{2}$$

3.2 Phase 1: Selective Diversification of the LoRA Council

Phase 1 trains the LoRA council Φ to be simultaneously accurate and mutually diverse. Because LoRA experts sharing a frozen backbone naturally converge to similar predictive behaviors (Zhai et al., 2024), we first compress the training trajectories for efficiency, then apply a three-term objective enforcing diversity across data, parameter, and output spaces.

Adaptive CoT truncation. Teacher rationales are often overly verbose. To refine the data, we test reasoning prefixes at several ratios and retain the shortest tested prefix that leads the base model to a correct answer in at least half of N independent completions:

$$\mathcal{F} := \{0.15, 0.30, 0.45, 0.60, 0.75\}, \qquad k_f := \lceil fS \rceil \quad (f \in \mathcal{F}),$$
$$k^\star = \min\left\{ k_f : f \in \mathcal{F},\ \frac{1}{N}\sum_{n=1}^{N} \mathrm{Verify}(\hat{a}_{k_f,n}, a) \geq \frac{1}{2} \right\}. \tag{3}$$

Here, $\hat{a}_{k_f,n}$ is the final answer extracted from the n-th completion prompted with $(x, s_{\leq k_f})$ (see Appendix C.4 for the prompt template), and a is the ground-truth answer. The binary function $\mathrm{Verify}(\hat{a}, a)$ indicates whether the extracted answer is correct. If no prefix passes, we retain the full trajectory ($k^\star = S$). The final target is $R_{\mathrm{trunc}} = (s_1, \ldots, s_{k^\star}, a)$. As demonstrated in our ablations (Section D.2), this truncation reduces target tokens by 17% while improving accuracy over training on full trajectories.

Phase 1 objective. For notational simplicity, let S hereafter denote the total number of reasoning steps in the truncated trajectory R_{trunc}. Using this refined corpus, we optimize the council via three complementary mechanisms:

(i) Step-level supervision dropout (Data-space diversity). To expose experts to unique reasoning paths, we randomly mask the negative log-likelihood loss of entire steps rather than individual tokens. Using $d_{i,m} \sim \mathrm{Bernoulli}(1 - p_{\mathrm{drop}})$, dropping steps forces each expert to bridge different logical gaps:

$$\mathcal{L}_{\mathrm{NLL}}^{\phi_m} = \sum_{i=1}^{S} d_{i,m} \mathcal{L}_i^{\phi_m}. \tag{4}$$

(ii) Dual orthogonality in LoRA space (Parameter-space diversity). To prevent representational collapse, we penalize the alignment between the row spaces of $A_m^{(\ell)}$ and $A_{m'}^{(\ell)}$ and between the column spaces of $B_m^{(\ell)}$ and $B_{m'}^{(\ell)}$ for every expert pair (m, m'). This bounded penalty $\mathcal{L}_{\mathrm{ortho}}^{(\ell)}(m, m') \in [0, 2]$ (Lemma 1) encourages experts to encode distinct features:

$$\mathcal{L}_{\mathrm{ortho}}^{(\ell)}(m, m') = \frac{\left\| A_m^{(\ell)} (A_{m'}^{(\ell)})^\top \right\|_F^2}{\left\| A_m^{(\ell)} \right\|_F^2 \left\| A_{m'}^{(\ell)} \right\|_F^2} + \frac{\left\| (B_m^{(\ell)})^\top B_{m'}^{(\ell)} \right\|_F^2}{\left\| B_m^{(\ell)} \right\|_F^2 \left\| B_{m'}^{(\ell)} \right\|_F^2},$$
$$\mathcal{L}_{\mathrm{ortho}}(\Phi) = \sum_{m=1}^{M} \sum_{\substack{m'=1 \\ m' \neq m}}^{M} \sum_{\ell} \mathcal{L}_{\mathrm{ortho}}^{(\ell)}(m, m'). \tag{5}$$

In practice, each squared norm in the denominators is clamped from below, $\|\cdot\|_F^2 \mapsto \max(\|\cdot\|_F^2, \varepsilon_{\mathrm{o}})$ with $\varepsilon_{\mathrm{o}} = 10^{-8}$, which keeps the penalty well-defined at the standard LoRA initialization $B = 0$ and preserves the bound in Lemma 1.

(iii) Functional repulsion on key steps (Output-space diversity). To maintain diversity among experts sharing the same backbone, we penalize similarity between their next-token distributions. For a set of token positions $\mathcal{I}$, the pairwise distance between experts m and m' in output space is

$$d_{m,m'}(\mathcal{I}) = \sum_{(i,t) \in \mathcal{I}} \left\| \log \hat{\mathbf{p}}_{i,t}^{\phi_m} - \mathrm{sg}\left[\log \hat{\mathbf{p}}_{i,t}^{\phi_{m'}} \right] \right\|_2^2. \tag{6}$$

We define the functional repulsion loss as

$$\mathcal{L}_{\mathrm{rep}}(\Phi; x, R; \mathcal{I}) = \frac{1}{M-1} \sum_{m=1}^{M} \sum_{m' \neq m} \exp(-\gamma d_{m,m'}(\mathcal{I})). \tag{7}$$

Here, $\mathrm{sg}[\cdot]$ treats expert m''s predictions as fixed when computing the gradient for expert m, and $\gamma > 0$ controls the loss sensitivity to predictive distance.

Crucially, uniform repulsion would force artificial disagreement on deterministic steps (e.g., trivial arithmetic). We therefore restrict it to a set of *key steps* $\mathcal{K}$ ranked highest by base-model predictive entropy; our main experiments select the top 10% of steps. The key-step-ratio ablation (Section D.1) compares this setting with broader and narrower selection. The corresponding token positions are

$$\mathcal{I}_{\mathcal{K}} = \{(i,t) \,|\, i \in \mathcal{K},\ 1 \le t \le T_i\}\,. \tag{8}$$

We apply Equation 7 with $\mathcal{I} = \mathcal{I}_{\mathcal{K}}$.

The complete Phase 1 objective over a minibatch $\mathcal{B}$ is formulated as:

$$\mathcal{L}_{\text{phase1}}(\Phi;\mathcal{B}) = \frac{1}{|\mathcal{B}|}\sum_{(x,R)\in\mathcal{B}}\left[\sum_{m=1}^{M}\mathcal{L}_{\text{NLL}}^{\phi_m}(x,R) + \lambda_{\text{rep}}\,\mathcal{L}_{\text{rep}}\big(\Phi;x,R;\mathcal{I}_{\mathcal{K}}\big)\right] + \lambda_{\text{ortho}}\,\mathcal{L}_{\text{ortho}}(\Phi). \tag{9}$$

3.3 Phase 2: Council-Weighted Distillation

Phase 2 converts the frozen council into a single deployable student. We first use the individual experts to identify which reasoning steps warrant greater supervision. We then consolidate their parameter updates into a merged adapter, which initializes the student and serves as a frozen reference during weighted distillation.

Extracting Council Signals. For each reasoning step s_i, we characterize the council using two complementary signals. The **mean uncertainty** U_i measures the average predictive entropy within individual experts, whereas the **disagreement** V_i measures the additional entropy induced by variation across experts:

$$U_i = \frac{1}{T_i}\sum_{t=1}^{T_i}\frac{1}{M}\sum_{m=1}^{M}H\big(\hat{p}_{i,t}^{\phi_m}\big), \qquad V_i = \frac{1}{T_i}\sum_{t=1}^{T_i}\left[H(\bar{p}_{i,t}) - \frac{1}{M}\sum_{m=1}^{M}H(\hat{p}_{i,t}^{\phi_m})\right], \tag{10}$$

where $\bar{p}_{i,t} = \frac{1}{M}\sum_{m=1}^{M}\hat{p}_{i,t}^{\phi_m}$ is the council-mean next-token distribution at the same teacher-forced context, and H denotes Shannon entropy. For each teacher-forced token context, the bracketed term equals the generalized Jensen-Shannon divergence among the experts, or equivalently the mutual information between the predicted token and a uniformly sampled expert identity (Proposition 1). Thus, V_i provides a principled measure of average token-level council disagreement.

Because V_i captures absolute disagreement without accounting for the uncertainty within individual experts, we define the **normalized disagreement score**

$$\rho_i = \frac{V_i}{U_i + \epsilon}, \tag{11}$$

where $\epsilon > 0$ ensures numerical stability. A high U_i with a low ρ_i indicates shared predictive uncertainty, whereas a low U_i with a high ρ_i signifies confident contradiction: individual experts are highly certain of their own predictions, yet disagree with one another substantially relative to their own uncertainty. The two signals therefore capture complementary sources of training difficulty.

We standardize U_i and ρ_i over the training corpus to obtain $\hat{U}_i$ and $\hat{\rho}_i$, and use $\tanh$ to bound the contribution of extreme values. With $\lambda_U, \lambda_D \ge 0$ and $\lambda_U + \lambda_D \le 1$, the resulting step weight is

$$w_i = 1 + \lambda_U \tanh(\hat{U}_i) + \lambda_D \tanh(\hat{\rho}_i), \tag{12}$$

The weight in Equation 12 combines the magnitude of within-expert uncertainty with the prominence of cross-expert disagreement relative to it. The U_i term increases emphasis on steps that individual experts find uncertain, whereas $\rho_i = V_i/(U_i + \epsilon)$ highlights disagreement that remains substantial relative to that uncertainty. Its inverse dependence on U_i at fixed V_i is intentional: the same disagreement receives less relative emphasis when individual predictions are already diffuse. Thus, above-average values of both signals reinforce one another, while opposing deviations can temper the weight instead of letting either cue dominate. The subsequent normalization redistributes supervision across the trajectory while retaining every step.

Figure 2 illustrates the nonuniform weights assigned to one truncated reasoning trajectory. CoReD retains every step while varying its contribution to the Phase 2 objective; these weights reflect the council's predictive signals rather than verified step correctness.

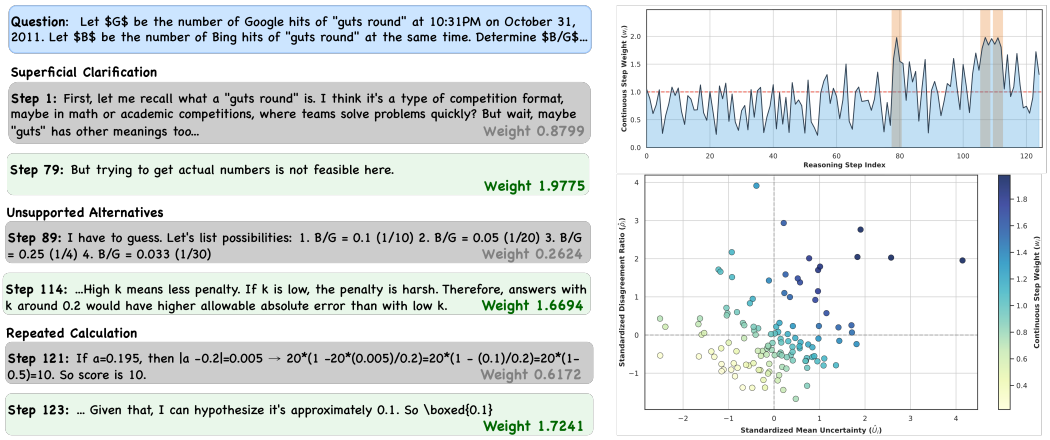

Figure 2: **Continuous emphasis across a reasoning trajectory.** CoReD retains every step of an adaptively truncated teacher response while varying its contribution to distillation according to the council's predictions (left: steps emphasized in green, $w_i > 1$, or downweighted in gray, $w_i < 1$; top right: weights across the trajectory; bottom right: relationship between step weights and the two standardized council signals).

Weight-space merging. After fixing the step weights, we consolidate the council using TIES (Yadav et al., 2023). Let $\Delta W_m^{(\ell)}$ denote the update to adapted layer ℓ induced by expert ϕ_m, as defined in Equation 1. TIES combines these updates at each adapted layer:

$$\Delta W_{\text{merge}}^{(\ell)} = \text{TIES}\left(\Delta W_1^{(\ell)}, \ldots, \Delta W_M^{(\ell)}\right). \tag{13}$$

The resulting layerwise updates collectively define the merged adapter ϕ_{merge}. We keep it frozen as the KL reference and use it to initialize the student, $\tilde{\theta} \leftarrow \phi_{\text{merge}}$. At initialization, their predictions coincide, so the KL term and its gradient are zero under the same forward-pass configuration (Proposition 2). As training updates $\tilde{\theta}$, the KL term constrains its deviation from the merged model.

Weighted distillation objective. With the council-derived step weights fixed, we optimize the student $\tilde{\theta}$ while keeping the merged adapter ϕ_{merge} frozen. The per-step KL loss compares their token distributions:

$$\mathcal{D}_i^{\text{KL}}(\tilde{\theta}) = \frac{1}{T_i} \sum_{t=1}^{T_i} \text{KL}\left(\hat{p}_{i,t}^{\phi_{\text{merge}}} \,\|\, \hat{p}_{i,t}^{\tilde{\theta}}\right). \tag{14}$$

The Phase 2 objective combines this loss with the teacher-forced NLL:

$$\mathcal{L}_{\text{distill}}(\tilde{\theta}) = \sum_{i=1}^{S} \frac{w_i}{\sum_{j=1}^{S} w_j} \left[\mathcal{L}_i^{\tilde{\theta}} + \lambda_{\text{KL}} \mathcal{D}_i^{\text{KL}}(\tilde{\theta})\right]. \tag{15}$$

The NLL trains the student on the target trajectories, while the KL term constrains deviations from the merged model. The same w_i emphasizes council-prioritized steps in both terms.

3.4 Inference

After Phase 2, the student update is merged into the backbone, so that inference uses only the fused weights, preserving the base model's parameter count and forward-pass structure regardless of council size M:

$$W_{\text{eff}}^{(\ell)} = W_{\text{base}}^{(\ell)} + \frac{\alpha}{r} \tilde{B}^{(\ell)} \tilde{A}^{(\ell)}. \tag{16}$$

4 Theoretical Analysis

This section states four properties of the quantities defined in Section 3: the entropy decomposition that identifies what the disagreement signal V_i measures (Proposition 1), the bounds that keep

the continuous step weights strictly positive (Proposition 3), the boundedness of the orthogonality penalty (Lemma 1), and the stationarity of the KL anchor at initialization (Proposition 2). These results characterize the council signals and objective scaling rather than accuracy, which we examine empirically in Section 5; full proofs appear in Appendix A, with additional statements in Appendix B.

First, we establish that our formulation of "disagreement" has a precise information-theoretic interpretation.

Proposition 1 (Entropy decomposition and non-negativity of disagreement)**.** *Fix a reasoning step s_i and token position t. Let $\hat{p}_{i,t}^{\phi_1}, \ldots, \hat{p}_{i,t}^{\phi_M}$ be the predictive probability distributions over the vocabulary from the M experts. Let $\bar{p}_{i,t} = \frac{1}{M}\sum_{m=1}^{M} \hat{p}_{i,t}^{\phi_m}$ be the mean predictive distribution, and $u_{i,t} = \frac{1}{M}\sum_{m=1}^{M} H(\hat{p}_{i,t}^{\phi_m})$ be the mean expected entropy. Then:*

(a) $H(\bar{p}_{i,t}) = u_{i,t} + v_{i,t}$, where $v_{i,t} := H(\bar{p}_{i,t}) - u_{i,t}$. Averaging over all positions $t \in [1, T_i]$ yields $\frac{1}{T_i}\sum_{t=1}^{T_i} H(\bar{p}_{i,t}) = U_i + V_i$.

(b) $v_{i,t} \geq 0$, with equality holding if and only if $\hat{p}_{i,t}^{\phi_1} = \cdots = \hat{p}_{i,t}^{\phi_M}$.

(c) Let $Z \sim$ Unif$\{1, \ldots, M\}$ represent the identity of an expert, and let $Y|(Z = m) \sim \hat{p}_{i,t}^{\phi_m}$ represent the generated token. Then $v_{i,t} = I(Y; Z)$, meaning the disagreement $v_{i,t}$ is exactly the mutual information between the token and the expert identity. Equivalently, $v_{i,t}$ is the uniformly-weighted generalized Jensen-Shannon divergence among the experts.

This licenses calling V_i "disagreement" in the precise sense of Bayesian active learning (Houlsby et al., 2011), quantifying how much a token reveals about which expert produced it.

Proposition 2 (Stationarity of the KL anchor at initialization)**.** *Let p be a fixed target distribution and $\theta \mapsto q_\theta$ a family of distributions satisfying Assumption 2. If $\theta_0 \in \Theta$ satisfies $q_{\theta_0} = p$, then $D_{\mathrm{KL}}(p\|q_{\theta_0}) = 0$ and $\nabla_\theta D_{\mathrm{KL}}(p\|q_\theta)\big|_{\theta_0} = 0$.*

Applying this to CoReD: because the student $\tilde{\theta}$ is initialized exactly with ϕ_{merge}, both the KL term $\mathcal{D}_i^{\mathrm{KL}}(\tilde{\theta})$ and its gradient vanish at the start of Phase 2. The KL term therefore exerts no pull at initialization, so early updates are driven entirely by the weighted NLL; as training proceeds, it acts as a soft anchor that limits drift from the merged model.

5 Experiments

5.1 Experimental Setup

Datasets & Models. We use the s1K-1.1 dataset (Muennighoff et al., 2025) for training, which utilizes DeepSeek-R1 (DeepSeek-AI et al., 2025) as the teacher. We evaluate on four mathematical reasoning benchmarks: MATH500 (Hendrycks et al., 2021; Lightman et al., 2024), AIME24 (Zhang & Math-AI Team, 2024), AIME25 (Zhang & Math-AI Team, 2025), and AMC12 (AI-MO, 2024). Our student models are Qwen2.5-7B-Instruct (Yang et al., 2024), Qwen3-8B (Yang et al., 2025), and DeepSeek-R1-Distill-Qwen-1.5B (DeepSeek-AI et al., 2025). For the 1.5B model, CoReD trains a council of LoRA adapters in Phase 1. At the start of Phase 2, we fold the merged adapter update into the base model's weights, and fine-tune all parameters of the student. The baselines on DeepSeek-R1-Distill-Qwen-1.5B also use full-parameter fine-tuning.

Baselines. We compare CoReD against Zero-shot, SFT (Long-CoT), and three recent step-aware distillation methods evaluated under the same conditions: Spectral Guided Learning (Fan et al., 2026), Segment Selective SFT (Wang et al., 2026), and P-ALIGN (Liu et al., 2026).

Implementation. CoReD uses a council of $M = 3$ LoRA experts. All models are trained for 3 epochs with a learning rate of 5×10^{-5}. Full training details, algorithm hyperparameters, and the evaluation sampling protocol are detailed in Appendix C.[1]

[1]The experimental code will be released publicly upon acceptance.

5.2 Main Results

Table 1: Overall performance on Qwen2.5-7B-Instruct, Qwen3-8B, and DeepSeek-R1-Distill-Qwen-1.5B. Results are reported as Pass@1 / Pass@3 (%), with **bold** and underlining indicating the best and second-best values, respectively, within each backbone.

	AIME25		AIME24		AMC12		MATH500		Avg.	
Method	Pass@1	Pass@3	Pass@1	Pass@3	Pass@1	Pass@3	Pass@1	Pass@3	Pass@1	Pass@3
Qwen2.5-7B-Instruct										
Zero-shot	7.78	16.67	10.00	16.67	43.37	56.63	76.33	85.40	34.37	43.84
SFT (Long-CoT)	11.11	16.67	8.89	13.33	34.94	48.19	65.27	80.40	30.05	39.65
Spectral Guided Learning	8.89	**20.00**	8.89	16.67	44.58	60.24	75.13	85.40	34.37	45.58
Segment Selective SFT	6.67	13.33	12.22	**23.33**	42.57	65.06	66.20	81.40	31.92	45.78
P-ALIGN	8.89	16.67	13.33	20.00	42.17	56.63	72.73	**86.40**	34.28	44.93
CoReD (ours)	**13.33**	**20.00**	**14.44**	**23.33**	**45.78**	**66.27**	**77.20**	**86.40**	**37.69**	**49.00**
Qwen3-8B										
Zero-shot	20.00	23.33	22.22	33.33	52.61	68.67	77.00	84.80	42.96	52.53
SFT (Long-CoT)	20.00	23.33	22.22	33.33	57.43	71.08	81.47	89.80	45.28	54.39
Spectral Guided Learning	17.78	26.67	23.33	30.00	53.82	68.67	80.33	90.00	43.81	53.84
Segment Selective SFT	22.22	33.33	22.22	33.33	55.42	68.67	80.33	89.40	45.05	56.18
P-ALIGN	21.11	30.00	30.00	40.00	60.64	73.49	85.20	91.80	49.24	58.82
CoReD (ours)	**28.89**	**36.67**	**34.44**	**43.33**	**63.86**	**75.90**	**87.67**	**93.20**	**53.72**	**62.28**
DeepSeek-R1-Distill-Qwen-1.5B										
Zero-shot	10.00	20.00	15.56	**33.33**	49.80	63.86	72.33	81.80	36.92	49.75
SFT (Long-CoT)	10.00	20.00	15.56	30.00	44.58	61.45	68.40	83.00	34.64	48.61
Spectral Guided Learning	16.67	23.33	17.78	23.33	48.59	59.04	74.13	86.00	39.29	47.93
Segment Selective SFT	**21.11**	**30.00**	16.67	23.33	50.60	**65.06**	74.87	84.60	40.81	50.75
P-ALIGN	16.67	26.67	17.78	30.00	50.20	63.86	**78.40**	**87.60**	40.76	52.03
CoReD (ours)	20.00	26.67	**18.89**	**33.33**	**51.41**	**65.06**	77.07	86.80	**41.84**	**52.97**

As shown in Table 1, CoReD achieves the highest average Pass@1 and Pass@3 on all three backbones. It leads on every benchmark metric for Qwen3-8B, with particularly clear gains on AIME, and leads or ties every metric for Qwen2.5-7B-Instruct. On DeepSeek-R1-Distill-Qwen-1.5B, CoReD achieves the highest average Pass@1 and Pass@3. It leads on AIME24 and AMC12 Pass@1, whereas Segment Selective SFT performs best on AIME25 and P-ALIGN performs best on MATH500. These results indicate that Phase 2 remains effective under full-parameter fine-tuning. Notably, CoReD improves on average over recent step-aware baselines such as P-ALIGN and Segment Selective SFT.

5.3 Ablation Study

Unless stated otherwise, all ablations use Qwen2.5-7B-Instruct trained on s1K-1.1 as in Section 5.1, and report Pass@1 averaged over the four benchmarks of Table 1. The Phase 1 objective ablation (Section 5.3.1) is the only exception and is evaluated on a separate held-out corpus.

5.3.1 Building a Diverse Yet Competent Council

We ablate the three diversity mechanisms in the Phase 1 objective: step-level gradient dropout, LoRA-space orthogonality $\mathcal{L}_{\text{ortho}}$, and functional repulsion loss $\mathcal{L}_{\text{rep}}$. We report council-averaged Pass@1 and pairwise Jensen-Shannon divergence on key and non-key steps (JS-Key and JS-NonKey) over 10,000 held-out verified mathematical reasoning trajectories; construction details are provided in Appendix C.2.

Table 2: Ablation of the Phase 1 objective.

Configuration	**JS-Key**	**JS-NonKey**	**Pass@1**
Full objective	0.04145	0.01707	**35.81**
w/o $\mathcal{L}_{\text{rep}}$	0.00033	0.00018	35.18
w/o $\mathcal{L}_{\text{ortho}}$	**0.04172**	**0.01709**	34.86
w/o dropout	0.04145	0.01704	35.41
$\mathcal{L}_{\text{rep}}$ only	0.04159	0.01704	34.44
$\mathcal{L}_{\text{ortho}}$ only	0.00030	0.00017	35.23
Dropout only	0.00032	0.00017	35.14
All three removed	0.00029	0.00016	34.83

Table 2 shows that the three Phase 1 mechanisms serve different roles. Functional repulsion is the main driver of diversity: configurations using it produce clear predictive disagreement, whereas dropout or orthogonality alone leave the experts nearly indistinguishable. Higher JS-Key than JS-NonKey indicates that entropy-gated repulsion focuses this diversity on the intended reasoning positions, although shared LoRA parameters allow some spillover to other steps. Disagreement alone,

however, does not predict reasoning accuracy. Loss $\mathcal{L}_{\text{rep}}$ by itself yields a diversity profile similar to the full objective but performs worse, while dropout and orthogonality improve expert performance despite creating little separation in isolation. The mechanisms are therefore complementary: Functional repulsion generates distinct predictive hypotheses, and the other terms help make those hypotheses reliable. The full objective performs best by balancing targeted diversity with expert competence rather than maximizing disagreement.

5.3.2 From Council to Student

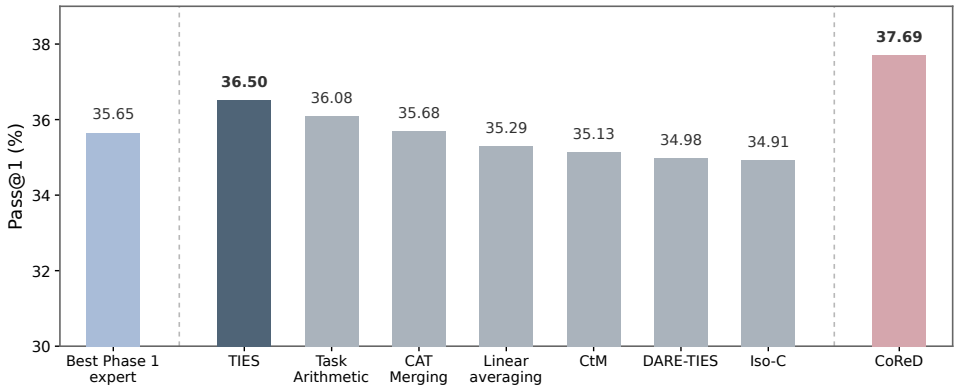

Figure 3: Contribution of merging and Phase 2 distillation, and comparison of merge operators. Merged adapters are evaluated without Phase 2 distillation.

Merging serves two purposes in CoReD: it removes the need to retain M adapters, and it produces a single starting point for Phase 2. We evaluate ϕ_{merge} directly, without Phase 2 distillation. As shown in Figure 3, merging is beneficial: the merged adapter reaches 36.50 Pass@1 against 35.65 for the strongest individual expert, indicating that council members encode complementary rather than redundant updates. Regarding the merging strategy, comparing TIES (Yadav et al., 2023) against Task Arithmetic (Ilharco et al., 2023), CAT Merging (Sun et al., 2025), linear averaging (Wortsman et al., 2022), CtM (He et al., 2026), DARE-TIES (Yu et al., 2024), and Iso-C (Marczak et al., 2025), all seven fall within 1.6 points of one another (36.50 to 34.91). We use TIES throughout for its explicit handling of sign conflicts across experts. Phase 2 adds a further 1.19 points ($36.50 \rightarrow 37.69$), and this gain comes from the council-derived weighting rather than from distillation as such: with uniform weights the same pipeline reaches only 35.84 (Section D.3), below the merged adapter it starts from.

Further ablation studies on adaptive truncation and key-step sensitivity are provided in Appendix D.

6 Conclusion

We introduced CoReD, a council-based framework for continuous step weighting in chain-of-thought distillation. A council of LoRA experts, diversified at high-entropy reasoning steps, provides uncertainty and disagreement signals that weight each step's contribution to distillation; the council is then consolidated into a single adapter that initializes the student and anchors it through a KL term. After training, the student update is folded into the backbone, eliminating council-dependent inference overhead. Across three student backbones and four mathematical reasoning benchmarks, CoReD achieves the highest average Pass@1 and Pass@3 among the compared methods.

AI use statement

We used generative AI tools only for auxiliary technical tasks: LaTeX formatting (table/figure layout, compilation errors, cross-references) and debugging software-environment issues. We did not use them for writing, literature search, research ideation and execution, or drafting. All AI-assisted outputs were reviewed by the authors, who take full responsibility for the content of this work.

Ethics statement

This work fine-tunes language models on publicly available, non-personal mathematical reasoning benchmarks and teacher-generated synthetic rationales; it does not involve human subjects, crowd-sourced annotation, or personally identifiable data. We are not aware of a direct dual-use or safety concern specific to improving distilled mathematical reasoning.

Reproducibility statement

The truncation procedure, council signals, and Phase 1 and Phase 2 objectives are defined in Section 3, with pseudocode for both training phases in Appendix C.1. Section 5.1 identifies the training source, teacher and student models, baselines, and evaluation benchmarks; Appendix C.3 reports the optimization and method settings. The held-out corpus used for the Phase 1 ablation is documented in Appendix C.2, and proofs are provided in Appendix A.

References

AI-MO. AIMO validation AMC. Hugging Face dataset, 2024.

Oleksandr Balabanov and Hampus Linander. Uncertainty quantification in fine-tuned LLMs using LoRA ensembles. *arXiv preprint arXiv:2402.12264*, 2025.

Leo Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, 1996.

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

Francesco D'Angelo and Vincent Fortuin. Repulsive deep ensembles are Bayesian. In *Advances in Neural Information Processing Systems*, 2021.

DeepSeek-AI et al. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Stefan Depeweg, José Miguel Hernández-Lobato, Finale Doshi-Velez, and Steffen Udluft. Decomposition of uncertainty in Bayesian deep learning for efficient and risk-sensitive learning. In *International Conference on Machine Learning*, pp. 1184–1193, 2018.

Sinan Fan, Xiaofeng Sun, Chen Shen, Chenxi Huang, Shaotian Yan, Bing Wang, Kaiyuan Liu, Xiaosong Yuan, Liang Xie, Wenxiao Wang, Jun Zhang, Hongyang Chen, and Jieping Ye. Uncovering the gradient geometry of long CoT: A spectral-guided approach to reasoning distillation. In *International Conference on Machine Learning*, 2026.

Zhengbao He, Ruiqi Ding, Zhehao Huang, Ruikai Yang, Tao Li, and Xiaolin Huang. Compress then merge: From multiple LoRAs into one low-rank adapter. In *Forty-third International Conference on Machine Learning*, 2026.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1*, 2021.

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

Neil Houlsby, Ferenc Huszár, Zoubin Ghahramani, and Máté Lengyel. Bayesian active learning for classification and preference learning. *arXiv preprint arXiv:1112.5745*, 2011.

Cheng-Yu Hsieh, Chun-Liang Li, Chih-kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alex Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 8003–8017, 2023.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Yitong Huang, Ziqi Yang, Zihui Wang, Jianzhong Qi, Rongshan Yu, Xiaoliang Fan, and Cheng Wang. Hybrid routing for a mixture of LoRA experts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 31211–31219, 2026. doi: 10.1609/aaai.v40i37.40383.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. In *International Conference on Learning Representations*, 2023.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*, 2022.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, 2017.

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

Yuetai Li, Xiang Yue, Zhangchen Xu, Fengqing Jiang, Luyao Niu, Bill Yuchen Lin, Bhaskar Ramasubramanian, and Radha Poovendran. Small models struggle to learn from strong reasoners. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 25366–25394, 2025.

Zeju Li, Jianyuan Zhong, Ziyang Zheng, Xiangyu Wen, Zhijian Xu, Yingying Cheng, Fan Zhang, and Qiang Xu. Making slow thinking faster: Compressing LLM chain-of-thought via step entropy. In *International Conference on Learning Representations*, 2026.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *International Conference on Learning Representations*, 2024.

Qiang Liu and Dilin Wang. Stein variational gradient descent: A general purpose Bayesian inference algorithm. In *Advances in Neural Information Processing Systems*, 2016.

Zhenghao Liu, Zhuoyang Wu, Xinze Li, Yukun Yan, Shuo Wang, Zulong Chen, Yu Gu, Ge Yu, and Maosong Sun. Long-chain reasoning distillation via adaptive prefix alignment. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 18041–18059, 2026.

Lucie Charlotte Magister, Jonathan Mallinson, Jakub Adamek, Eric Malmi, and Aliaksei Severyn. Teaching small language models to reason. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pp. 1773–1781, 2023.

Andrey Malinin and Mark Gales. Predictive uncertainty estimation via prior networks. In *Advances in Neural Information Processing Systems*, pp. 7047–7058, 2018.

Daniel Marczak, Simone Magistri, Sebastian Cygert, Bartłomiej Twardowski, Andrew D. Bagdanov, and Joost van de Weijer. No task left behind: Isotropic model merging with common and task-specific subspaces. In *International Conference on Machine Learning*, 2025.

Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. s1: Simple test-time scaling. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 20275–20321, 2025.

Dominik J. Mühlematter, Michelle Halbheer, Alexander Becker, Dominik Narnhofer, Helge Aasen, Konrad Schindler, and Mehmet Ozgur Turkoglu. LoRA-ensemble: Efficient uncertainty modelling for self-attention networks. *Transactions on Machine Learning Research*, 2026.

Kainat Riaz, Muhammad Ahmed Mohsin, Ahsan Bilal, Muhammad Umer, Ayesha Mohsin, Aqib Riaz, Ali Subhan, and John M. Cioffi. Epistemic uncertainty for test-time discovery. *arXiv preprint arXiv:2605.11328*, 2026.

Tom Saliencro, Rohan Desai, Priya Nair, Maya Lindqvist, and Daniel Whitmore. Spend experts where you are unsure: Confidence-adaptive routing for mixture-of-experts LoRA. *arXiv preprint arXiv:2607.26052*, 2026.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15:1929–1958, 2014.

Sophie Steger, Christian Knoll, Bernhard Klein, Holger Fröning, and Franz Pernkopf. Function space diversity for uncertainty prediction via repulsive last-layer ensembles. *arXiv preprint arXiv:2412.15758*, 2024.

Wenju Sun, Qingyong Li, Yangliao Geng, and Boyang Li. CAT merging: A training-free approach for resolving conflicts in model merging. In *Proceedings of the 42nd International Conference on Machine Learning*, 2025.

Trung Trinh, Markus Heinonen, Luigi Acerbi, and Samuel Kaski. Input-gradient space particle inference for neural network ensembles. In *International Conference on Learning Representations*, 2024.

Shenzhi Wang, Le Yu, Chang Gao, Chujie Zheng, Shixuan Liu, Rui Lu, Kai Dang, Xionghui Chen, Jianxin Yang, Zhenru Zhang, Yuqiong Liu, An Yang, Andrew Zhao, Yang Yue, Shiji Song, Bowen Yu, Gao Huang, and Junyang Lin. Beyond the 80/20 rule: High-entropy minority tokens drive effective reinforcement learning for LLM reasoning. In *Advances in Neural Information Processing Systems*, 2025.

Siyuan Wang, Yanchen Liu, and Xiang Ren. Segment-level attribution for selective learning of long reasoning traces. In *International Conference on Learning Representations*, 2026. URL `https://openreview.net/forum?id=C1UD4FLFPL`.

Xi Wang, Laurence Aitchison, and Maja Rudolph. LoRA ensembles for large language model fine-tuning. *arXiv preprint arXiv:2310.00035*, 2023a.

Xiao Wang, Tianze Chen, Qiming Ge, Han Xia, Rong Bao, Rui Zheng, Qi Zhang, Tao Gui, and Xuanjing Huang. Orthogonal subspace learning for language model continual learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 10658–10671, 2023b.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, 2022.

Mitchell Wortsman, Gabriel Ilharco, Samir Ya Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S. Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, and Ludwig Schmidt. Model soups: Averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *International Conference on Machine Learning*, 2022.

Xun Wu, Shaohan Huang, and Furu Wei. Mixture of LoRA experts. In *International Conference on Learning Representations*, 2024.

Prateek Yadav, Derek Tam, Leshem Choshen, Colin A. Raffel, and Mohit Bansal. TIES-merging: Resolving interference when merging models. In *Advances in Neural Information Processing Systems*, 2023.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

An Yang et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Le Yu, Bowen Yu, Haiyang Yu, Fei Huang, and Yongbin Li. Language models are super Mario: Absorbing abilities from homologous models as a free lunch. In *International Conference on Machine Learning*, 2024.

Yuanzhao Zhai, Han Zhang, Yu Lei, Yue Yu, Kele Xu, Dawei Feng, Bo Ding, and Huaimin Wang. Uncertainty-penalized reinforcement learning from human feedback with diverse reward LoRA ensembles. *arXiv preprint arXiv:2401.00243*, 2024.

Yifan Zhang and Math-AI Team. American invitational mathematics examination (AIME) 2024. Hugging Face dataset, 2024.

Yifan Zhang and Math-AI Team. American invitational mathematics examination (AIME) 2025. Hugging Face dataset, 2025.

Qiyong Zhong, Mao Zheng, Mingyang Song, Xin Lin, Jie Sun, Yiqi Zhao, Xiaodi Wang, Houcheng Jiang, Xiang Wang, and Junfeng Fang. SOD: Step-wise on-policy distillation for small language model agents. *arXiv preprint arXiv:2605.07725*, 2026.

Yuan Zhuang, Yi Shen, Yuexin Bian, Qing Su, Shihao Ji, Yuanyuan Shi, and Fei Miao. LD-MoLE: Learnable dynamic routing for mixture of LoRA experts. In *International Conference on Learning Representations*, 2026.

A Proofs of Theoretical Results

This appendix provides detailed, step-by-step mathematical proofs for the theoretical claims presented in Section 4.

A.1 Proof of Proposition 1 (Entropy Decomposition)

Part (a): By definition, we introduced the residual term $v_{i,t}$ such that $H(\bar{p}_{i,t}) = u_{i,t} + v_{i,t}$. Summing both sides over $t = 1, \ldots, T_i$ and dividing by T_i yields:

$$\frac{1}{T_i}\sum_{t=1}^{T_i} H(\bar{p}_{i,t}) = \frac{1}{T_i}\sum_{t=1}^{T_i} u_{i,t} + \frac{1}{T_i}\sum_{t=1}^{T_i} v_{i,t} \equiv U_i + V_i. \tag{17}$$

Part (b): The Shannon entropy $H(p) = -\sum_y p(y)\log p(y)$ is a strictly concave function on the probability simplex. According to Jensen's inequality for a concave function f, we have $f\left(\sum \lambda_m x_m\right) \geq \sum \lambda_m f(x_m)$. Applying this to the entropy function with uniform weights $\lambda_m = \frac{1}{M}$:

$$H\left(\frac{1}{M}\sum_{m=1}^{M} \hat{p}_{i,t}^{\phi_m}\right) \geq \frac{1}{M}\sum_{m=1}^{M} H\left(\hat{p}_{i,t}^{\phi_m}\right). \tag{18}$$

This translates directly to $H(\bar{p}_{i,t}) \geq u_{i,t}$, which implies $v_{i,t} \geq 0$. Because Shannon entropy is strictly concave on the interior of the simplex, equality holds if and only if all input distributions are identical: $\hat{p}_{i,t}^{\phi_1} = \cdots = \hat{p}_{i,t}^{\phi_M}$.

Part (c): Define $Z \sim \text{Unif}\{1, \ldots, M\}$ as the random variable representing the chosen expert. Conditional on $Z = m$, the distribution of the generated token Y is $P(Y \mid Z = m) = \hat{p}_{i,t}^{\phi_m}$.

First, we compute the conditional entropy of Y given Z:

$$H(Y|Z) = \sum_{m=1}^{M} P(Z = m)H(Y|Z = m) = \frac{1}{M}\sum_{m=1}^{M} H\left(\hat{p}_{i,t}^{\phi_m}\right) = u_{i,t}, \tag{19}$$

Next, we compute the marginal probability distribution of Y:

$$P(Y = y) = \sum_{m=1}^{M} P(Z = m)P(Y = y|Z = m) = \frac{1}{M}\sum_{m=1}^{M} \hat{p}_{i,t}^{\phi_m}(y) = \bar{p}_{i,t}(y), \tag{20}$$

Thus, the marginal entropy is $H(Y) = H(\bar{p}_{i,t})$.

By the definition of mutual information:

$$I(Y; Z) = H(Y) - H(Y|Z) = H(\bar{p}_{i,t}) - u_{i,t} = v_{i,t}. \tag{21}$$

Furthermore, the generalized Jensen-Shannon divergence (JSD) for a set of distributions $p_1, \ldots, p_M$ with weights π_m is defined as $JSD_\pi = H\left(\sum \pi_m p_m\right) - \sum \pi_m H(p_m)$. Setting $\pi_m = \frac{1}{M}$ perfectly recovers our definition of $v_{i,t}$.

A.2 Proof of Proposition 3 (Step-Weight Bounds)

The standard hyperbolic tangent function, $\tanh(x)$, has an image strictly bounded in the open interval $(-1, 1)$ for any real number $x \in \mathbb{R}$. Given the step weight $w_i = 1 + \lambda_U \tanh(\hat{U}_i) + \lambda_D \tanh(\hat{\rho}_i)$, we can establish the strict upper bound:

$$w_i < 1 + \lambda_U(1) + \lambda_D(1) = 1 + \lambda_U + \lambda_D, \tag{22}$$

And the strict lower bound:

$$w_i > 1 + \lambda_U(-1) + \lambda_D(-1) = 1 - (\lambda_U + \lambda_D). \tag{23}$$

Hence $1 - (\lambda_U + \lambda_D) < w_i < 1 + (\lambda_U + \lambda_D)$ for all finite $\hat{U}_i, \hat{\rho}_i$. Because $\lambda_U + \lambda_D \leq 1$, the lower bound is non-negative, so $w_i > 0$ for every step; in the extreme case $\lambda_U + \lambda_D = 1$, this gives $w_i \in (0, 2)$. Therefore, the lowest possible theoretical value for w_i strictly approaches 0 (but never reaches it for any finite standardized inputs). Consequently, it is mathematically guaranteed that $w_i \in (0, 2)$ strictly.

A.3 Proof of Lemma 1 (Boundedness of the Orthogonality Penalty)

Let $A \in \mathbb{R}^{d_1 \times k}$ and $B \in \mathbb{R}^{k \times d_2}$ be matrices. The (i, j)-th entry of the product AB is $(AB)_{ij} = \sum_{c=1}^{k} A_{ic} B_{cj}$. By applying the Cauchy-Schwarz inequality to the dot product of the i-th row of A and the j-th column of B, we obtain:

$$(AB)_{ij}^2 = \left(\sum_{c=1}^{k} A_{ic} B_{cj}\right)^2 \leq \left(\sum_{c=1}^{k} A_{ic}^2\right)\left(\sum_{c=1}^{k} B_{cj}^2\right), \tag{24}$$

To find the squared Frobenius norm of AB, we sum over all elements i and j:

$$\|AB\|_F^2 = \sum_{i=1}^{d_1}\sum_{j=1}^{d_2} (AB)_{ij}^2 \tag{25}$$

$$\leq \sum_{i=1}^{d_1}\sum_{j=1}^{d_2}\left[\left(\sum_{c=1}^{k} A_{ic}^2\right)\left(\sum_{c=1}^{k} B_{cj}^2\right)\right] \tag{26}$$

$$= \left(\sum_{i=1}^{d_1}\sum_{c=1}^{k} A_{ic}^2\right)\left(\sum_{j=1}^{d_2}\sum_{c=1}^{k} B_{cj}^2\right) \tag{27}$$

$$= \|A\|_F^2 \|B\|_F^2. \tag{28}$$

In the context of Equation 5, applying this inequality to $A_m^{(\ell)}$ and $(A_{m'}^{(\ell)})^\top$ implies that the ratio $\frac{\|A_m^{(\ell)}(A_{m'}^{(\ell)})^\top\|_F^2}{\|A_m^{(\ell)}\|_F^2 \|A_{m'}^{(\ell)}\|_F^2} \leq 1$. The same logic applies to the B matrices. Since the penalty $\mathcal{L}_{\text{ortho}}^{(\ell)}(m, m')$ is the sum of these two ratios (each bounded in $[0, 1]$), the total penalty lies in $[0, 2]$.

A.4 Proof of Proposition 2 (KL Stationarity at Initialization)

Gibbs' inequality states that for any two discrete probability distributions p and q_θ, the Kullback-Leibler divergence is non-negative: $D_{\text{KL}}(p \| q_\theta) \geq 0$, with equality holding if and only if p and q_θ are identical everywhere.

Consider the function $f(\theta) = D_{\mathrm{KL}}(p \| q_\theta)$. Because $q_{\theta_0} = p$, we have $f(\theta_0) = 0$. Since $f(\theta) \geq 0$ for all $\theta \in \Theta$, it is clear that θ_0 is a global minimizer of $f(\theta)$.

Under Assumption 2, the function $f(\theta)$ is differentiable on the open domain Θ. A standard result in continuous optimization dictates that if a differentiable function attains an interior global minimum at a point, its gradient evaluated at that point must be exactly the zero vector. Therefore, $\nabla_\theta f(\theta)\big|_{\theta_0} = \mathbf{0}$.

In CoReD Phase 2, because $\tilde{\theta}$ is copied identically from ϕ_{merge} at initialization, the student's output distribution perfectly matches the merged teacher's target distribution. Thus, the KL divergence $\mathcal{D}_i^{\mathrm{KL}}(\tilde{\theta})$ and its gradient are structurally forced to zero at step 0.

A.5 Extended Details on Phase 1 Population Risk and Gradient Coupling

Let $\mathcal{D}$ denote the distribution over post-truncation training pairs (x, R). Let $d_{i,m} \sim \mathrm{Bernoulli}(1 - p_{\mathrm{drop}})$ be the gradient dropout masks, drawn independently. The true population risk we aim to minimize is:

$$\mathcal{J}_{\mathrm{phase1}}(\Phi) = \mathbb{E}_{(x,R)\sim\mathcal{D}} \mathbb{E}_{d_{i,m}} \left[\sum_{m=1}^{M} \mathcal{L}_{\mathrm{NLL}}^{\phi_m}(x, R) + \lambda_{\mathrm{rep}} \mathcal{L}_{\mathrm{rep}}(\Phi; x, R; \mathcal{I}_{\mathcal{K}}) \right] + \lambda_{\mathrm{ortho}} \mathcal{L}_{\mathrm{ortho}}(\Phi). \tag{29}$$

When optimizing over a minibatch $\mathcal{B}$, we compute the empirical estimator $\hat{\mathcal{J}}_{\mathrm{phase1}}(\Phi; \mathcal{B})$ (which exactly corresponds to the empirical loss $\mathcal{L}_{\mathrm{phase1}}(\Phi; \mathcal{B})$ defined in Equation 9). Because the expectation operator is linear, and the minibatch $\mathcal{B}$ alongside masks $d_{i,m}$ are drawn i.i.d. from their respective underlying distributions, we naturally obtain:

$$\mathbb{E}_{\mathcal{B}, d_{i,m}} \left[\hat{\mathcal{J}}_{\mathrm{phase1}}(\Phi; \mathcal{B}) \right] = \mathcal{J}_{\mathrm{phase1}}(\Phi). \tag{30}$$

This confirms the claim in Remark 1 that the estimator is strictly unbiased.

To ensure that Stochastic Gradient Descent (SGD) converges correctly, we must make an additional explicit assumption regarding the interchange of the gradient and expectation operators:

Assumption 1 (Interchange of expectation and gradient)**.** *For every Φ in the region visited during training:*

$$\mathbb{E}_{\mathcal{B}, d_{i,m}} \left[\nabla_\Phi \hat{\mathcal{J}}_{\mathrm{phase1}}(\Phi; \mathcal{B}) \right] = \nabla_\Phi \mathbb{E}_{\mathcal{B}, d_{i,m}} \left[\hat{\mathcal{J}}_{\mathrm{phase1}}(\Phi; \mathcal{B}) \right]. \tag{31}$$

Under Assumption 1, $\mathbb{E}[\nabla_\Phi \hat{\mathcal{J}}_{\mathrm{phase1}}] = \nabla_\Phi \mathcal{J}_{\mathrm{phase1}}$, meaning our backward pass provides a formally unbiased stochastic gradient of the true population risk. We state this explicitly because LoRA parameters are unconstrained in $\mathbb{R}^d$, so standard dominated-convergence bounds required for this interchange are not automatically guaranteed a priori.

Note on Gradient Coupling: It is important to observe that minimizing $\mathcal{J}_{\mathrm{phase1}}$ is a *single joint optimization*, not M isolated problems. Taking the gradient with respect to a single expert m:

$$\nabla_{\phi_m} \hat{\mathcal{J}}_{\mathrm{phase1}} = \nabla_{\phi_m} \mathcal{L}_{\mathrm{NLL}}^{\phi_m} + \lambda_{\mathrm{rep}} \nabla_{\phi_m} \mathcal{L}_{\mathrm{rep}} + 2\lambda_{\mathrm{ortho}} \sum_{\substack{m'=1 \\ m' \neq m}}^{M} \sum_{\ell} \nabla_{\phi_m} \mathcal{L}_{\mathrm{ortho}}^{(\ell)}(m, m'). \tag{32}$$

(Note the factor of 2 arises from the symmetry $\mathcal{L}_{\mathrm{ortho}}^{(\ell)}(m, m') = \mathcal{L}_{\mathrm{ortho}}^{(\ell)}(m', m)$ in the full pairwise summation). Even though the functional repulsion term employs a 'stop-gradient' on other experts $m' \neq m$ during a specific pairwise evaluation to prevent recursive instability, expert m still explicitly receives gradients derived from its spatial relationship to every other expert m' through both the diversity and orthogonality terms. This ensures symmetric and mutual repulsion across the entire council during training.

B Additional Theoretical Statements

Section 4 informally introduced four key theoretical properties of the CoReD framework, detailing the entropy decomposition (Proposition 1) and KL stationarity (Proposition 2) in the main text. For

brevity, the formal statements regarding the numerical stability of step weights and the boundedness of the orthogonality penalty were deferred. We provide these statements below, alongside the regularity assumptions required for the optimization dynamics analyzed in Appendix A.

B.1 Step-Weight Bounds and Orthogonality Penalty

To guarantee that the continuous weighting scheme (Equation 12) is numerically stable and never assigns a zero or negative weight to a valid reasoning step, we establish strict algebraic bounds on the step weight w_i.

Proposition 3 (Step-weight bounds)**.** *For hyperparameters $\lambda_U, \lambda_D \geq 0$ such that $\lambda_U + \lambda_D \leq 1$, the continuous weight $w_i = 1 + \lambda_U \tanh(\hat{U}_i) + \lambda_D \tanh(\hat{\rho}_i)$ satisfies:*

$$1 - (\lambda_U + \lambda_D) < w_i < 1 + (\lambda_U + \lambda_D). \tag{33}$$

strictly for every finite $\hat{U}_i, \hat{\rho}_i \in \mathbb{R}$. Because $1 - (\lambda_U + \lambda_D) \geq 0$, the final weight w_i is strictly positive for every step, ensuring the student model never completely discards any valid reasoning step.

Similarly, to ensure the LoRA council remains diverse without causing parameter explosion, we demonstrate that the dual orthogonality penalty applied in Phase 1 is inherently bounded.

Lemma 1 (Boundedness of the orthogonality penalty)**.** *For any non-zero matrices A and B with conformable shapes, $\|AB\|_F^2 \leq \|A\|_F^2 \|B\|_F^2$. Consequently, the orthogonality penalty $\mathcal{L}_{\text{ortho}}^{(\ell)}(m, m')$ formulated in Equation 5 is strictly bounded within the interval $[0, 2]$.*

B.2 Regularity Assumptions for Optimization

The mathematical proofs in Appendix A rely on standard regularity assumptions for both training phases. To analyze the gradient behavior at the onset of Phase 2 (as utilized in Proposition 2), we require the student's predictive distribution to be differentiable.

Assumption 2. *The mapping $\tilde{\theta} \mapsto \hat{p}_{i,t}^{\tilde{\theta}}$ is differentiable on an open parameter domain Θ containing every parameter state $\tilde{\theta}_t$ visited during Phase 2 training.*

Finally, for Phase 1, the optimization process is structurally unbiased because the sampled stochastic elements exactly match the distribution of the population risk.

Remark 1 (Unbiasedness of the Phase-1 estimator)**.** *Because minibatches and gradient dropout masks are drawn i.i.d. from the exact distributions defining the Phase 1 population risk $\mathcal{J}_{phase1}$, the minibatch estimator $\hat{\mathcal{J}}_{phase1}$ is structurally unbiased.*

C Training and Reproducibility Details

C.1 Training Algorithms

Algorithms 1 and 2 summarize the two training stages. Phase 1 jointly trains the council with three diversity mechanisms. Phase 2 freezes the experts, derives step weights, merges their updates with TIES, and optimizes the student against the merged reference.

Algorithm 1 Phase 1 council training.

Input: frozen backbone θ_{base}; truncated corpus $\mathcal{D}$; key-step sets $\mathcal{K}$; council size M; Phase 1 coefficients
Initialize expert adapters $\Phi = \{\phi_1, \ldots, \phi_M\}$
for each training minibatch $\mathcal{B}$ **do**
  for each $(x, R) \in \mathcal{B}$ **do**
    Sample independent step dropout masks $d_{i,m}$ for each expert
    Compute expert predictions and masked NLL losses using Equation 4
    Compute functional repulsion on $\mathcal{I}_{\mathcal{K}}$ using Equations 8 and 7
  end for
  Compute the LoRA orthogonality penalty using Equation 5
  Form $\mathcal{L}_{\text{phase1}}(\Phi; \mathcal{B})$ using Equation 9
  Update all expert adapters jointly; keep θ_{base} frozen
end for
Output: trained council Φ

Algorithm 2 Phase 2 council-weighted distillation.

Input: frozen backbone θ_{base}; trained council Φ; truncated corpus $\mathcal{D}$; step-weight and KL coefficients
for each $(x, R) \in \mathcal{D}$ **do**
  Compute teacher-forced expert predictions without gradients
  Compute U_i, V_i, and ρ_i for every step using Equations 10 and 11
end for
Compute corpus-level standardization statistics for U_i and ρ_i
Compute and fix w_i for every training step using Equation 12
Merge the expert updates at each adapted layer using TIES (Equation 13); denote the result by ϕ_{merge}
Initialize $\tilde{\theta} \leftarrow \phi_{\text{merge}}$ and freeze ϕ_{merge}
for each Phase 2 training minibatch $\mathcal{B}$ **do**
  Obtain frozen reference predictions $\hat{p}_{i,t}^{\phi_{\text{merge}}}$, from a forward pass without gradients
  Compute student predictions $\hat{p}_{i,t}^{\tilde{\theta}}$
  Compute the weighted objective using Equations 2, 14, and 15
  Update only $\tilde{\theta}$
end for
Output: trained student $\tilde{\theta}$, folded into the backbone as in Equation 16

C.2 Held-Out Corpus for the Phase 1 Objective Ablation

The Phase 1 ablation reports JS-Key and JS-NonKey by comparing expert predictions at the same teacher-forced reasoning steps. This analysis requires long-CoT trajectories generated by DeepSeek-R1, the teacher used for our training data. Since the four benchmarks do not provide such trajectories, we evaluate these metrics on a separate held-out corpus of DeepSeek-R1 long-CoT trajectories.

We select 10,000 problems from the default split of `OpenR1-Math-220k`, whose reasoning traces were generated by DeepSeek-R1. We exclude records whose `problem` appears in `s1K-1.1`, the source of our training data. For each remaining problem, we select the first generation in source order marked as reasoning-complete and correct. Each released record contains `problem`, `solution`, `answer`, `uuid`, and `longcot`, with `longcot` storing the selected reasoning trajectory.

C.3 Hyperparameters and Evaluation Protocol

Table 3 summarizes the hyperparameters and evaluation settings used in our experiments.

Table 3: Training hyperparameters and evaluation settings for CoReD.

Setting	**Value**
Council size M	3
LoRA rank r	16
LoRA scale α/r	1
Key-step fraction	10%
Step dropout rate p_{drop}	0.1
Orthogonality coefficient λ_{ortho}	0.01
Functional repulsion coefficient λ_{rep}	1.0
Step-weight coefficients (λ_U, λ_D)	$(0.5, 0.5)$
Disagreement stabilizer ϵ	10^{-3}
KL coefficient λ_{KL}	1.0
Merge operator	TIES
Optimizer	AdamW, $\beta = (0.9, 0.999)$, weight decay 0
Learning rate and schedule	5×10^{-5}; cosine schedule with 0.1 warmup ratio
Training duration	3 epochs
Samples per problem n	3
Sampling temperature	0.6
Top-p	0.9
Repetition penalty	1.05

C.4 Prompt Template for Adaptive Truncation

During the adaptive chain-of-thought truncation described in Section 3.2, we evaluate whether a truncated reasoning prefix $s_{\leq k_f}$ is sufficient for the base model to arrive at the correct final answer. We format this completion task using the following prompt template:

```
You are an expert reasoner.  You are provided with a problem and a
partial reasoning process.  Your task is to continue the reasoning from
exactly where it left off to solve the problem.
Please format your final answer by enclosing it within \boxed{}.
Problem:
{x}
Partial Solution:
{s_{≤k_f}}
```

Here, $\{x\}$ is replaced by the original problem statement from the dataset, and $\{s_{\leq k_f}\}$ is the retained reasoning prefix of the teacher's rationale. The instruction, answer-formatting convention, and verification function $\text{Verify}(\cdot, \cdot)$ are adapted to the task domain; the mathematical instantiation shown here (with `\boxed{}` answers) matches our experimental setting in Section 5.1, while the truncation procedure of Equation 3 itself places no constraint on domain beyond requiring a verifiable final answer.

Handling of the final answer. In $R_{\text{trunc}} = (s_1, \ldots, s_{k^\star}, a)$, the final element is the teacher's final response that follows the reasoning trace. It is kept as a separate last segment rather than being merged into the last reasoning step. Truncation removes only reasoning steps and never alters or re-appends this segment, so each target contains it exactly once, including when the full trajectory is retained. For the training objectives, this answer segment is treated as an additional final step: the step sums in Equations 4 and 15 run over the reasoning steps and the answer segment. In Phase 1, the answer segment is subject to step-level supervision dropout and is a candidate for key-step selection in the same way as any other step. In Phase 2, the council signals U_i, V_i, and ρ_i and their corpus standardization statistics are computed over reasoning steps only; the answer segment instead receives a fixed weight of 1 and contributes to both the NLL and KL terms.

D Additional Ablation Studies

D.1 Sensitivity to the Key-Step Ratio

We vary the fraction of reasoning steps selected as key steps to examine how broadly functional repulsion should be applied. A ratio of 0% disables functional repulsion entirely, 100% applies repulsion at every reasoning step, and intermediate settings apply it only to the steps with the highest base-model predictive entropy.

Table 4: Sensitivity to the fraction of steps selected for functional repulsion.

Key-step ratio	Pass@1
0%	36.16
$\mathbf{10}\%$ (ours)	**37.69**
20%	34.96
100%	35.11

Repulsion helps only when it is narrowly targeted. The default 10% setting is the only one that improves on disabling repulsion (37.69 against 36.16 at 0%); widening the key-step set to 20% or applying repulsion everywhere both fall to roughly 35, below the no-repulsion baseline. Mean base-model entropy is 2.62 nats at the top-10% steps against 1.05 elsewhere, so a wider set admits increasingly deterministic positions, where enforced disagreement conflicts with the teacher-forcing objective rather than supplementing it. Accuracy is correspondingly sensitive to this ratio, and we treat 10% as a value to be set per domain rather than a universal constant. The teacher-forcing NLL is retained on the full trajectory in all settings, so the ratio governs only where the experts are encouraged to differ.

D.2 Effect of Adaptive Truncation

Table 5: Ablation of the trajectory-truncation strategy. Retained Tokens is the percentage of target tokens kept after truncation.

Strategy	Retained Tokens (%)	Pass@1	Pass@3
Fixed 45% prefix	57.3	35.45	46.84
Adaptive (ours)	82.6	**37.69**	**49.00**
No truncation	100	36.38	46.58

We compare adaptive truncation with retaining the full trajectories and with a fixed prefix ratio of 45%. Adaptive truncation achieves the best Pass@1 and Pass@3 while retaining only 82.6% of the target tokens, improving over training on full trajectories by 1.31 Pass@1 and 2.42 Pass@3. This suggests that the discarded suffixes, which often consist of redundant verification or re-derivation, contribute little useful supervision and may even dilute the training signal. Truncation is not beneficial per se, however: a fixed 45% prefix removes more tokens but falls 2.24 Pass@1 below adaptive truncation and also below the untruncated baseline, since a single ratio cuts into essential reasoning on harder problems while leaving easy ones longer than necessary. Selecting the prefix per example according to whether the base model can still recover the answer therefore matters more than the amount of truncation itself.

D.3 WHY TWO SIGNALS?

The step weight of Equation 12 combines the within-expert mean uncertainty $\hat{U}_i$ with the relative disagreement $\hat{\rho}_i$. We remove each signal in turn by setting $\lambda_D = 0$ (uncertainty only) or $\lambda_U = 0$ (disagreement only), and set $w_i \equiv 1$ for the uniform configuration. As shown in Figure 4, uncertainty-only weighting raises Pass@1 from 35.84 under uniform weighting to 36.83, whereas disagreement-only remains close to the uniform baseline at 35.91. When combined, however, the two signals further improve performance to 37.69. This interaction suggests that $\rho_i = V_i/(U_i + \epsilon)$ is most useful as a conditional modifier that distinguishes uncertain steps involving cross-expert conflict from those that are difficult for all experts, rather than as a standalone importance signal.

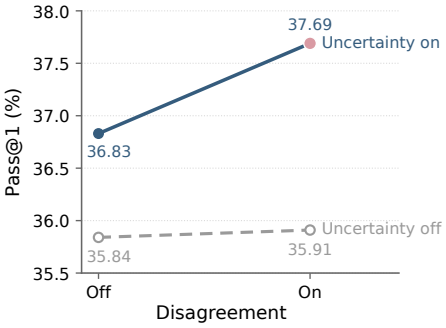

Figure 4: Interaction between uncertainty and disagreement in step weighting on Qwen2.5-7B-Instruct. Disagreement provides little improvement in isolation but yields a clear gain when uncertainty weighting is active.