CoRe: Collaborative Reasoning via Cross Teaching

Kshitij Mishra [1] **Mirat Aubakirov** [1] **Martin Takáč** [1] **Nils Lukas** [1] **Salem Lahlou** [1]

Abstract

Large language models exhibit complementary reasoning errors: on the same instance, one model may succeed with a particular decomposition while another fails. We propose Collaborative Reasoning (**CoRe**), a *training-time* collaboration framework that converts peer success into a learning signal via a **cross-teaching** protocol. Each problem is solved in two stages: a *cold* round of independent sampling, followed by a *contexted rescue* round in which failed models receive a hint extracted from a successful peer. **CoRe** optimizes a combined reward that balances (i) correctness, (ii) a lightweight DPP-inspired diversity term to reduce error overlap, and (iii) an explicit rescue bonus for successful recovery. We evaluate **CoRe** in low-data regimes across GSM8K, MATH, AIME, and GPQA. With only 1,000 training examples, a pair of small open-source models (3B+4B) reaches **99.54%** oracle Team Pass@2 on GSM8K and **92.08%** on MATH, compared to **82.50%** and **74.82%** for single-model training. On harder datasets, the same pair reaches **77.34%** oracle Team Pass@2 on GPQA (trained on 348 examples) and **79.65%** on AIME (trained on 792 examples), using a training-time budget of at most 1536 context tokens and 3072 generated tokens. A three-model non-oracle AIME evaluation further closes most of the oracle-selection gap (81.20% majority vote vs. 82.60% oracle Team Pass@2). Overall, these results show that training-time collaboration can reliably convert model complementarity into large gains without scaling model size.

[1]Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, United Arab Emirates. Correspondence to: Kshitij Mishra <kshitij.mishra@mbzuai.ac.ae>, Salem Lahlou <salem.lahlou@mbzuai.ac.ae>.

Proceedings of the 43rd International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026. Copyright 2026 by the author(s).

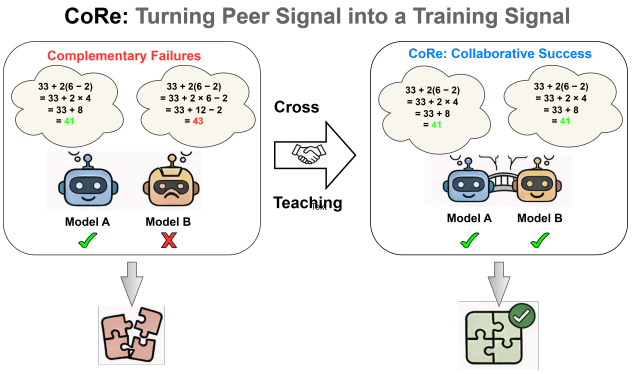

Figure 1. **CoRe**: Collaborative Reasoning via Cross Teaching Training.

1. Introduction

Large language models (LLMs) can solve many multi-step reasoning problems when prompted to produce intermediate steps (chain-of-thought; Wei et al. 2022). However, even strong models exhibit systematic and complementary failures: different architectures, pretraining mixes, and post-training pipelines lead to distinct error modes on the *same* dataset. In practice, ensembles and multi-agent prompting exploit this diversity at inference time (e.g., self-consistency Wang et al. 2023, ranking-and-fusion Jiang et al. 2023, and debate-style interactions Irving et al. 2018). A natural next question is *whether we can push this complementarity* into training *so that models become better collaborators*.

Most recent gains in LLM reasoning come from inference-time scaffolding (more samples, search, or multi-agent prompting), which increases compute but does not change the underlying policy (Wang et al., 2023; Yao et al., 2023a; Du et al., 2024). These methods often break under *correlated* failure: if a model misses a key insight, additional samples and even additional agents can converge to the same wrong plan; more generally, ensemble gains are limited when member errors are correlated (Dietterich, 2000; Kuncheva & Whitaker, 2003a). Training-time collaboration is challenging due to sparse correctness supervision (motivating process-level feedback), multi-model credit assignment, and the risk of diversity collapse through mutual imitation or agreement-style training (Lightman et al., 2024; Foerster et al., 2018; Hinton et al., 2015; Zhang et al., 2018).

This motivates a reinforcement learning (RL) view: we must jointly optimize (i) exploitation for correctness, (ii) exploration for complementary traces that reduce overlap, and (iii) rescue-specific reward on peer-disagreement cases.

We propose Collaborative Reasoning (**CoRe**) (Figure 1), an algorithm for training multiple LLMs to collaborate on reasoning tasks. **CoRe** is built around a simple micro-round protocol: (1) in a *cold* round, each model samples several independent solution traces; (2) if at least one model succeeds, we extract out a successful trace as hint; and (3) in a *contexted* round, each model resamples with hint, and models that previously failed are eligible for an explicit *rescue bonus* when they recover. The key intuition is that "peer success" provides a targeted supervision signal precisely on the instances where one model needs help.

Our goal is not only to improve single-model accuracy but also to reduce *error overlap* between collaborators, thereby increasing the probability that *at least one* model solves a problem. We operationalize this with an explore–exploit reward design. Exploitation rewards correctness (with optional partial credit), while exploration rewards encourage diverse reasoning traces using a greedy Determinantal Point Process (DPP) approximation. Our *key* contributions are three-fold:

1. We propose Collaborative Reasoning (**CoRe**), a novel algorithm to jointly train N LLM policies with a two-round protocol–a *cold* round of independent sampling followed by a *contexted rescue* round where failed models learn from hints extracted from successful peers, with *hint dropout* to improve robustness.

2. We propose a combined reward that simultaneously optimizes (i) correctness, (ii) exploration-exploitation of complementary traces (reducing correlated failures), and (iii) an explicit *rescue bonus* to incentivize successful recovery on peer-disagreement.

3. Extensive empirical analysis that show across GSM8K, MATH, AIME, GPQA, and MedMCQA, **CoRe** consistently improves collaborative performance, achieving **99.54% Oracle Team Pass@2** on GSM8K and strong gains on harder benchmarks (e.g., **92.08%** on MATH, **79.65%** on AIME, and **77.34%** on GPQA), while training on no more than 1,000 examples.

2. Related Work

Reasoning via prompting, search, and inference-time ensembling. Prompting methods elicit intermediate reasoning traces, most notably chain-of-thought (Wei et al., 2022), and can be strengthened with decompositional prompting such as least-to-most (Zhou et al., 2023). Sampling-based inference improvements like self-consistency (Wang et al., 2023) reduce variance by aggregating multiple traces, while search-based frameworks explicitly explore reasoning branches (e.g., tree-of-thought (Yao et al., 2023a)). Other lines interleave reasoning with actions or external tools (e.g., ReAct (Yao et al., 2023b)). Model combination methods further improve accuracy by selecting or fusing candidate generations (e.g., ranking-and-fusion in LLM-Blender (Jiang et al., 2023)). In contrast, **CoRe** moves collaboration *into training*: peer successes become an on-policy learning signal (via hints + rescue), rather than only an inference-time aggregation trick.

Reinforcement learning and preference optimization for LLM post-training. Alignment and post-training commonly rely on RLHF-style objectives (Ouyang et al., 2022) or preference optimization such as DPO (Rafailov et al., 2023). For reasoning-focused post-training, group-based policy optimization (e.g., GRPO) has shown strong empirical performance (Shao et al., 2024), and recent work proposes more stable variants such as sequence-level clipping (GSPO) (Zheng et al., 2025) and smooth, token-adaptive trust-region control (SAPO) (Gao et al., 2025). Prior work has also applied preference optimization directly to chain-of-thought traces to improve mathematical reasoning (Lahlou et al., 2025). **CoRe** is largely orthogonal to these optimizer choices: we contribute a *multi-model data collection protocol* at training time itself and reward shaping aimed at reducing correlated failures.

Diversity, exploration, and reducing correlated failures. Diversity is a classic lever for improving ensemble performance and reducing error correlation (Dietterich, 2000; Kuncheva & Whitaker, 2003b). In sequence generation, diverse decoding procedures such as Diverse Beam Search encourage exploration of multiple modes (Vijayakumar et al., 2016). Determinantal point processes (DPPs) provide a principled repulsive model for selecting diverse subsets (Kulesza & Taskar, 2012). The recent work SD-E^2 (Mishra et al., 2026) explores diverse reasoning traces within a single small model. Building on these ideas, we introduce a lightweight greedy approximation (*DPP-lite*) over hybrid semantic/structural distances between reasoning traces, and use it as an explicit exploration reward to reduce *overlap* between collaborators rather than merely diversifying outputs at inference time.

Multi-agent and multi-model collaboration. Multi-agent prompting and debate-style protocols can improve factuality and reasoning at inference time by having multiple instances interact (Irving et al., 2018; Du et al., 2023). Systems work has also proposed general multi-agent orchestration frameworks (e.g., CAMEL (Li et al., 2023) and AutoGen (Wu et al., 2023)) that coordinate role-based collaboration, tool use, and iterative refinement. Closer to our

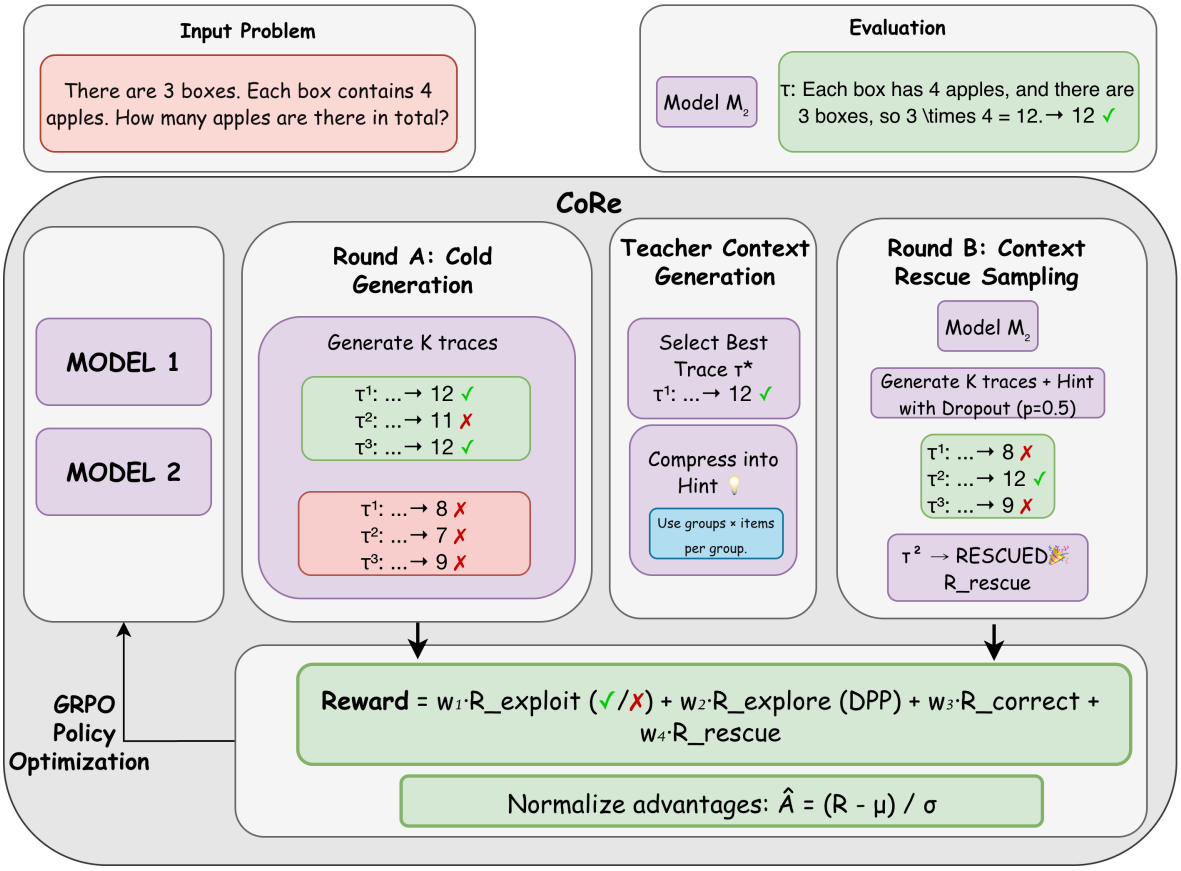

Figure 2. **CoRe training.** In Round A, each model samples K independent traces. If any trace succeeds, the best one is used as teacher context for Round B, where models resample with explore–exploit rewards and a rescue bonus for successful recoveries.

setting, MALT trains a sequential generator–verifier–refiner pipeline using trajectory expansion and value-style credit assignment (Motwani et al., 2025), while MAPoRL co-trains multiple LLMs with multi-agent RL and verifier rewards over multi-turn discussions (Park et al., 2025). **CoRe** differs in its training signal: peer success on the *same instance* is compressed into a hint for failed collaborators, and the rescue reward explicitly optimizes recovery on disagreement cases. Thus, **CoRe** is not an inference-time discussion protocol or a fixed role pipeline; it is an on-policy cross-teaching protocol that reshapes the collaborators' policies during post-training.

Collaborative learning, distillation, and self-training with rationales. Knowledge distillation and peer learning transfer information across models by matching outputs or training on teacher-generated data (Hinton et al., 2015; Zhang et al., 2018; Furlanello et al., 2018). For reasoning, prior work also leverages *rationales* as supervision via bootstrapping loops (STaR) (Zelikman et al., 2022) or distilling step-by-step solutions (Hsieh et al., 2023). **CoRe** differs in two key ways: (i) peer supervision is generated *on-policy* and *instance-conditioned* (only when a peer succeeds on the same example), and (ii) learning is explicitly targeted to the *disagreement set* via rescue bonuses.

3. Problem Formulation

Let $\mathcal{D} = \{(x, y^*)\}$ denote a dataset of reasoning problems. We train a set of N autoregressive policies $\{\pi_{\theta_i}\}_{i=1}^{N}$. A *reasoning trace* τ is a generated token sequence (including intermediate steps) and induces a final answer via an extraction function $\mathrm{Ans}(\tau)$. We define correctness as $\mathbb{I}[\mathrm{Ans}(\tau) = y^*]$.

Team success. A cooperative objective is to maximize the probability that *at least one* model solves a problem:

$$\max_{\theta_1,\ldots,\theta_N} \mathbb{E}_{(x,y^*)\sim\mathcal{D}}\Big[\mathbb{I}\big[\exists i \text{ s.t. } \mathrm{Ans}(\tau_i) = y^*\big]\Big], \qquad (1)$$

where $\tau_i \sim \pi_{\theta_i}(\cdot \mid x)$. This objective is sparse and does not indicate how to credit individual traces or encourage complementary reasoning. We therefore optimize a dense, per-trace surrogate reward.

For each problem x, models first sample a *cold* set of traces, then optionally resample in a *context-conditioned* rescue

round using a hint derived from a successful peer. Rescue traces are explicitly rewarded, steering learning toward instances where one model can help another. This turns collaboration from an inference-time heuristic into a training-time objective aligned with Equation (1).

4. Method

For each training problem x, **CoRe** collects traces in two micro-rounds.

Round A (cold). Each model i samples K traces independently: $\{\tau_{i,k}^{A}\}_{k=1}^{K} \sim \pi_{\theta_i}(\cdot \mid x)$. We compute rewards for all sampled traces and identify whether the model succeeded: $s_i = \max_k \mathbb{I}[\mathrm{Ans}(\tau_{i,k}^{A}) = y^*]$.

Teacher context construction. If any model succeeds in Round A, we select the highest-reward correct trace and build a teacher context c with a length budget L tokens.

Round B (contexted rescue). Each model samples K' additional traces conditioned on the context with *hint dropout*: with probability p_{hint} the model sees c, otherwise it sees the original prompt. A Round B trace is marked as *rescue-eligible* only when the model failed in Round A ($s_i = 0$) and the hint is provided.

4.1. Reward Design

Exploitation reward. For a trace τ we compute

$$R_{\text{et}}(\tau) = \mathbb{I}[\mathrm{Ans}(\tau) = y^*] + \alpha\, \mathrm{Pl}(\mathrm{Ans}(\tau), y^*), \quad (2)$$

where $\mathrm{Pl}(\cdot,\cdot) \in [0,1]$ is a lightweight overlap-based partial score and $\alpha \in [0,1]$ downweights partial credit (default $\alpha = 0.3$).

Exploration reward (DPP-lite). Given the set of traces for a single problem, we construct a diverse subset S with a greedy DPP approximation. For any trace $\tau \notin S$, we define

$$R_{\text{ee}}(\tau) = \max\Big(0,\ \min_{\tau' \in S} d(\tau, \tau') - \delta\Big), \quad (3)$$

where d is a hybrid semantic/structural distance. Traces in S receive zero exploration reward.

Rescue bonus. If a model fails in Round A but produces a correct trace in Round B *with the hint provided*, we add a fixed rescue bonus r_{teach}.

Combined reward. We define a binary rescue indicator $z(\tau) \in \{0,1\}$. The combined reward is

$$R(\tau) = w_1 R_{\text{et}}(\tau) + w_2 R_{\text{ee}}(\tau) + r_{\text{teach}} z(\tau) + R_{\text{aux}}(\tau). \quad (4)$$

Auxiliary rewards. We optionally include a trace-accuracy bonus and a cross-model complementarity term:

$$\begin{aligned} R_{\text{aux}}(\tau_{i,k}^{r}) = \lambda_{\text{ta}}\, \mathbb{I}[\mathrm{Ans}(\tau_{i,k}^{r}) = y^*]\,(1 - \rho_r(x)) \\ + \lambda_{\text{cm}}\, R_{\text{cross}}(\tau_{i,k}^{r}), \end{aligned} \quad (5)$$

where $r \in \{A, B\}$, $K_A = K$, $K_B = K'$, and $\rho_r(x) = \frac{1}{NK_r}\sum_{j,k'} \mathbb{I}[\mathrm{Ans}(\tau_{j,k'}^{r}) = y^*]$ is the fraction of correct traces for problem x in round r. The factor $(1 - \rho_r(x))$ upweights correct traces on harder or borderline instances where few traces succeed, acting as a soft curriculum around the disagreement boundary where cross-teaching is most useful. The complementarity term is secondary and is quality-gated below to avoid rewarding arbitrary diversity from poor traces.

Optional cross-model complementarity reward. We define a quality gate $q_j^r = \mathbb{I}\Big[\max_{k'} R_{\text{et}}(\tau_{j,k'}^{r}) \geq \eta\Big]$. For a trace $\tau_{i,k}^{r}$,

$$R_{\text{cross}}(\tau_{i,k}^{r}) = q_j^r \cdot \max\Big(0,\ \min_{k'} d(\tau_{i,k}^{r}, \tau_{j,k'}^{r}) - \delta_{\text{cm}}\Big). \quad (6)$$

In the main experiments we enable this term with a small weight (late-stage default $\lambda_{\text{cm}} = 0.1$) and ablate it in Appendix J; setting $\lambda_{\text{cm}} = 0$ recovers the simpler exploit–explore–rescue objective.

Policy optimization objective. Following group-based RL for LLMs (Shao et al., 2024), we normalize rewards across all traces sampled for the same problem instance:

$$A(\tau) = \frac{R(\tau) - \mu}{\sigma + \epsilon}, \quad \mu, \sigma \text{ computed over traces for } x. \quad (7)$$

This normalization stabilizes optimization and ensures that relative trace quality, rather than absolute reward scale, drives learning. Traces from Round B (contexted rescue) are included in the update with a configurable weight λ_B (default 0.8), allowing the policy to learn from rescue behavior while keeping the cold-round signal dominant and preventing over-reliance on hints. The **CoRe** training architecture is shown in Figure 2. The collaborative component of **CoRe** is *loss-agnostic*: collaboration enters exclusively through (i) micro-round and (ii) reward construction, and can be combined with any of the group based policy gradient optimization objectives.

4.2. Decomposition Theorem for Collaboration Gain

Our evaluation emphasizes *team success*—whether any collaborator solves a problem. For two models, team success admits a simple decomposition that clarifies what our collaboration objective optimizes and why reducing overlap matters.

Algorithm 1 CoRe

Require: Models M_1, M_2 with policies $\pi_{\theta_1}, \pi_{\theta_2}$; dataset $\mathcal{D}$

for epoch $e = 1$ to E **do**
  for problem $(x, y^*) \in \mathcal{D}$ **do**
    Round A: sample K traces per model; compute rewards; mark successes s_1, s_2
    if $s_1 \vee s_2$ **then**
      build teacher context c from best successful trace
    else
      $c \leftarrow \emptyset$
    end if
    Round B: sample K' traces per model with hint dropout; add rescue bonus if applicable
    compute group-normalized advantages and update θ_1, θ_2 with chosen policy loss
  end for
end for

Theorem 4.1 (Collaboration gain equals complementary correctness)**.** *Let $C_1, C_2 \in \{0, 1\}$ indicate whether models 1 and 2 solve an instance, and let $T = \mathbb{I}[C_1 \vee C_2]$ denote the team-success indicator. Define accuracies $p_i = \mathbb{P}(C_i = 1)$ and joint success $p_{12} = \mathbb{P}(C_1 = 1, C_2 = 1)$. Then*

$$\mathbb{P}(T = 1) = p_1 + p_2 - p_{12}. \tag{8}$$

Assume without loss of generality that $p_2 \geq p_1$. Then the collaboration gain over the best single model, $\Delta \triangleq \mathbb{P}(T = 1) - p_2$, equals

$$\Delta = \mathbb{P}(C_1 = 1, C_2 = 0). \tag{9}$$

Proof. Equation (8) follows from inclusion–exclusion:

$$\mathbb{P}(C_1 \vee C_2) = \mathbb{P}(C_1) + \mathbb{P}(C_2) - \mathbb{P}(C_1 \wedge C_2). \tag{10}$$

For Equation (9),

$$\begin{aligned} \Delta &= (p_1 + p_2 - p_{12}) - p_2 \\ &= p_1 - p_{12} \\ &= \mathbb{P}(C_1 = 1) \\ &\quad - \mathbb{P}(C_1 = 1, C_2 = 1) \\ &= \mathbb{P}(C_1 = 1, C_2 = 0). \end{aligned} \tag{11}$$

□

Remark 4.2 (Limits of collaboration gain). Since $\Delta = \mathbb{P}(C_1 = 1, C_2 = 0)$, it is bounded as

$$0 \leq \Delta \leq \min\{p_1,\ 1 - p_2\}. \tag{12}$$

Thus, even perfect collaboration cannot exceed the weaker model's marginal accuracy p_1, nor can it exceed the stronger model's error rate $1 - p_2$.

Implications for training. Theorem 4.1 implies that beating the best single model requires increasing *complementary correctness*–the weaker model must be right where the stronger one is wrong (i.e., lower success overlap p_{12} for fixed p_1). **CoRe** targets this by promoting diverse traces via DPP-lite exploration and focusing learning on disagreement cases through the rescue signal.

5. Experiments

We study two collaborating small language model pairs, constraining the combined parameter count to at most ∼7B, to test whether **CoRe**'s collaboration gains persist beyond weak baselines and extend to stronger reasoning-tuned models. Specifically, we consider: (1) a *Qwen instruct* pair, `Qwen2.5-3B-Instruct` (Team, 2024) and `Qwen3-4B-Instruct` (Yang et al., 2025), chosen to enable direct comparison with the SD-E^2 state-of-the-art baseline (Mishra et al., 2026); and (2) a *compact reasoning* pair, `Phi-4-mini-reasoning` (Xu et al., 2025) and `Ministral-3-3B-Reasoning` (Liu et al., 2026). We report *training-time* collaboration with *cold* test-time evaluation (no inference-time hints); all models use parameter-efficient LoRA (Hu et al., 2022). Unless otherwise stated, team metrics are oracle Team Pass@K and are used to measure error overlap; we also include non-oracle majority-vote and agreement-gated selection results for the 3-model AIME setting.

Each training step comprises two micro-rounds: Round A (cold), where each model samples K independent traces from the original prompt, and Round B (contexted), where each model samples K' traces with max_token budget of 3072 while optionally conditioning on a teacher context extracted from the best successful peer trace in Round A. For training, we use batch_size=4. To improve robustness, we apply hint dropout with probability p_{hint}, so Round B interleaves contexted and non-contexted rollouts. The teacher context is the peer trace's *correct strategy block* (i.e., the strategy giving outcome same as the ground truth) inserted *verbatim* as context; we cap it to a fixed budget (default 1536 tokens, via a lightweight estimate) and truncate any excess. We remove explicit answer lines/tags from the teacher context to prevent answer leakage; truncation-related failure modes are analyzed in Section 7 and Appendix A.1.

All experiments optimize the combined reward. Crucially, we **enable** the cross-model complementarity term R_{cross} with a small non-zero weight (default late-stage $w_{\text{cross}} = 0.1$), and we ablate it in Appendix J. We use $K = 2$, $K' = 1$, $p_{\text{hint}} = 0.75$ and include the rescue bonus for successful Round B recovery. Our exploration and cross-model rewards depend on a hybrid distance $d(\tau, \tau')$ between traces.

We implement:

$$d(\tau, \tau') = \lambda_{\text{emb}}\Big(1 - \cos\big(e(\tau), e(\tau')\big)\Big) + \lambda_{\text{struct}}\Big(1 - \text{Jaccard}\big(op(\tau), op(\tau')\big)\Big). \quad (13)$$

where $e(\tau)$ is a sentence embedding of the trace and $op(\tau)$ is a set of extracted operation signatures. We use the Sentence-Transformers embedding model `sentence-transformers/all-MiniLM-L6-v2` (Reimers & Gurevych, 2019; Reimers & community, 2020; Wang et al., 2020) for $e(\cdot)$, and extract operations via simple regex triggers for algebraic/arithmetic and reasoning operators (e.g., addition/multiplication, simplification, substitution, case analysis). Default weights are $\lambda_{\text{emb}} = 0.6$ and $\lambda_{\text{struct}} = 0.4$. Full extraction rules are provided in Appendix A.1.

5.1. Datasets and Baselines

Datasets. We evaluate on four reasoning benchmarks spanning arithmetic, formal mathematics, contest-style problem solving, and scientific QA: GSM8K (Cobbe et al., 2021), MATH (Hendrycks et al., 2021), AIME, and GPQA (Rein et al., 2024). Training follows a *low-data regime*: each run trains on at most 1,000 examples (and fewer when the dataset is smaller). For GSM8K, we use the standard train/test split and report results on the official test set. For MATH, AIME, and GPQA, we construct an 80:20 train–test split and report results on the resulting test sets (MATH: 2,500; AIME: 193; GPQA main: 90).

Baselines. We compare **CoRe** against three classes of baselines:

1. **Base (off-the-shelf):** We evaluate the pretrained models without **CoRe** training. To separate collaboration gains from merely using two models, we also report a *base-pair oracle*: run both base models independently under the same Pass@K budget and count an instance correct if either model succeeds.

2. **SD-E^2:** We compare against SD-E^2 (Mishra et al., 2026), a strong diverse-exploration baseline for small-model reasoning. For the Qwen setting, we match their setup using `Qwen2.5-3B-Instruct` and `Qwen3-4B-Instruct`. We also report *SD-E^2 + Oracle*, which combines the two SD-E^2 models (trained independently) and counts an instance correct if either model succeeds.

3. **SD-E^2 at 7B–8B scale:** To contextualize scale under our ∼7B total-parameter budget, Appendix D reports SD-E^2 results for `7B-8B` single models: `Qwen-2.5-7B-Instruct` (Team, 2024), `Phi-Small-8K-Instruct` (Abdin et al., 2024), and `Ministral-3-8B-Reasoning` (Liu et al., 2026).

5.2. Evaluation metrics

We perform all evaluation within a token budget of 4096 max_tokens. We evaluate both *individual* reasoning performance and the *collaborative* benefits of training-time cross-teaching. For each model, Pass@K measures whether at least one of K sampled solution traces is correct. In particular, Pass@1 uses a single trace (greedy or single-sample decoding), while Pass@2 samples two traces per model under stochastic decoding and counts an instance as solved if either trace is correct. We report **Pass@1** and **Pass@2** to capture both deterministic performance and robustness under limited sampling. **Team Pass@K** extends Pass@K to collaborators: an instance is counted as solved if *any* model produces a correct solution among its K traces. This is an *oracle* metric (it assumes perfect selection among candidates) and therefore should be read as an upper bound and a direct diagnostic of error overlap. To estimate deployable performance without oracle access, we also evaluate two simple non-oracle selectors in the 3-model AIME setting: **majority vote**, which returns the most frequent extracted answer, and **agreement-gated selection**, which returns the unanimous answer when all models agree and otherwise falls back to majority vote.

6. Results and Analysis

Table 1, Table 2 summarizes results for the `Qwen-2.5-3B-instruct`, `Qwen-3-4B-instruct`, `Phi-4-mini-reasoning` (Xu et al., 2025) and `Ministral-3-3B-Reasoning` (Liu et al., 2026) collaboration pair under three settings: Base (off-the-shelf), SD-E^2 training, and **CoRe** training. Across all four benchmarks, **CoRe** yields consistent improvements in both *individual* performance (Pass@1/Pass@2) and *team* success, indicating gains beyond simply using two models at test time.

In Table 1 it can be observed that On GSM8K, **CoRe** lifts the two models from ∼55% Base Pass@1 to 91.62% and 92.49% Pass@1, and from ∼56–57% Base Pass@2 to 92.95% and 93.33% Pass@2 (Table 1). Using inclusion–exclusion with the per-model Pass@2 values in Table 1, the estimated probability that *both are correct* is $p_{12} \approx 92.95 + 93.33 - 99.54 = 86.74\%$, leaving $\approx 12.80\%$ mass where exactly one model succeeds. This complementary-correctness region is precisely what enables near-ceiling team reliability without requiring either member to be near-perfect.

A similar pattern holds on MATH: Team Pass@2 is 80.80%,

Table 1. **Qwen pair results across datasets (Pass@1/Pass@2).** We report Base, SD-E^2 (Mishra et al., 2026), and **CoRe** for Qwen2.5-3B-Instruct and Qwen3-4B-Instruct.

Model	GSM8K		MATH		AIME		GPQA	
	Pass@1	Pass@2	Pass@1	Pass@2	Pass@1	Pass@2	Pass@1	Pass@2
Base								
Qwen2.5-3B-Instruct	54.66	56.30	32.40	38.10	6.74	7.50	4.20	5.90
Qwen3-4B-Instruct	55.32	57.10	33.55	39.40	7.04	7.80	4.55	6.25
Qwen2.5-3B + Qwen3-4B + Oracle	57.10	60.05	35.80	41.95	7.90	9.20	5.35	7.10
SD-E^2								
Qwen2.5-3B-Instruct	82.03	84.55	56.40	59.85	13.28	14.85	8.10	9.70
Qwen3-4B-Instruct	82.50	85.05	57.10	60.40	13.30	14.90	8.55	10.10
Qwen2.5-3B + Qwen3-4B + Oracle	85.02	87.20	59.30	63.10	17.14	17.90	10.20	11.95
CoRe								
Qwen2.5-3B-Instruct	91.62	92.95	70.56	71.96	15.70	16.82	11.90	13.05
Qwen3-4B-Instruct	92.49	93.33	71.65	72.52	16.10	17.04	10.75	12.10
Qwen2.5-3B + Qwen3-4B (Team)	**99.10**	**99.54**	**79.40**	**80.80**	**19.50**	**20.45**	**16.25**	**17.80**

Table 2. **Pass@1/Pass@2 across datasets (GSM8K/MATH/AIME/GPQA).** Values are Pass@1 (single-sample) and Pass@2 (two samples).

Model	GSM8K		MATH		AIME		GPQA	
	Pass@1	Pass@2	Pass@1	Pass@2	Pass@1	Pass@2	Pass@1	Pass@2
Base								
Phi-4-mini-reasoning	72.40	78.10	57.04	68.16	11.91	16.58	10.35	12.23
Ministral-3-3B-Reasoning	70.95	76.70	54.21	64.44	10.18	14.86	9.14	11.20
Phi-4-mini + Ministral-3-3B + Oracle	74.85	80.20	59.25	70.70	13.02	17.92	11.86	14.03
SD-E^2								
Phi-4-mini-reasoning	80.60	84.35	69.68	73.82	53.92	57.58	46.35	50.23
Ministral-3-3B-Reasoning	79.10	83.40	67.25	73.15	49.00	55.42	42.13	44.38
Phi-4-mini + Ministral-3-3B + Oracle	83.25	85.50	71.85	77.70	54.42	60.19	48.12	55.26
CoRe								
Phi-4-mini-reasoning	90.14	91.10	85.24	90.01	68.16	76.64	67.76	75.26
Ministral-3-3B-Reasoning	89.25	90.23	83.72	89.12	67.15	75.90	65.58	74.82
Phi-4-mini + Ministral-3-3B	**95.72**	**96.72**	**88.42**	**92.08**	**71.42**	**79.65**	**70.18**	**77.34**

compared to 71.96% and 72.52% for the individual models. Equivalently, only 19.20% of problems remain unsolved by *both* collaborators under a two-sample budget, which is a substantially lower both-wrong mass than the SD-E^2 oracle (36.90%) and the Base oracle (58.05%). On GPQA, Team Pass@2 improves from 11.95% (SD-E^2 oracle) and 7.10% (Base oracle) to 17.80% (**CoRe** team), indicating that cross-teaching still yields meaningful complementarity even in knowledge-heavy settings.

Although **CoRe** improves both individual and team performance on AIME/GPQA, these domains remain far from saturated for the Qwen pair, as also reflected in the comparison in Figure 3. On AIME, Team Pass@2 improves from 9.20% (Base oracle) and 17.90% (SD-E^2 oracle) to 20.45% (CoRe team), suggesting that many failures require longer-horizon insight than is consistently transferred through a truncated teacher context.

In Table 2 it can be observed that the pair of `Phi-4-mini-reasoning` (Xu et al., 2025) and `Ministral-3-3B-Reasoning` (Liu et al., 2026) starts stronger than the Qwen models and hence benefit substantially from post-training. The Base oracle pair provides only modest gains over the stronger Base model (e.g., GSM8K Pass@2: 80.20% vs. 78.10%; MATH Pass@2: 70.70% vs. 68.16%), indicating that off-the-shelf diversity is limited. SD-E^2 improves both competence and the oracle pair (e.g., MATH Pass@2 oracle: 77.70%), but **CoRe** yields substantially larger gains (MATH Team Pass@2: 92.08%). Thus, we argue that cross-teaching changes the learned policies in a way that increases the probability that *at least one*

Table 3. **Non-oracle selection on AIME.** The 3-model team uses Phi-4-mini-reasoning, Ministral-3-3B-Reasoning, and Ministral-3-8B-Reasoning.

Selection strategy	AIME Pass@2
Best single model	78.45
Agreement-gated team	80.40
Majority-vote team	81.20
Oracle Team Pass@2	82.60

model succeeds, rather than merely exploiting pre-existing diversity.

Across all datasets, **CoRe** improves Pass@2 for both members and the team (Table 2). For example, Team Pass@2 improves from 85.50% (SD-E^2 oracle) to 96.72% (CoRe team) on GSM8K, from 77.70% to 92.08% on MATH, and from 55.26% to 77.34% on GPQA, as also summarized visually in Figure 4. These improvements indicate that **CoRe** increases the probability mass on correct solution modes under stochastic decoding, not only the best trace under deterministic decoding.

6.1. Non-oracle selection

Because Team Pass@K is an oracle upper bound, we additionally evaluate deployable selection on a 3-model AIME team, where majority voting is meaningful. Table 3 shows that majority vote captures most of the oracle gain: the gap between oracle Team Pass@2 and majority vote is only 1.40 percentage points. Agreement-gated selection is slightly more conservative but remains above the best single model. These results suggest that **CoRe** not only increases the chance that a correct answer exists among candidates, but also makes agreement a more useful confidence signal by reducing correlated failure.

6.2. Collaboration Help via Cross Teaching

CoRe improves team success primarily by reducing *correlated failures* and increasing *complementary correctness* between collaborators under a fixed sampling budget. A direct diagnostic is the both-wrong probability $1 - p_{\text{team}}$ under a fixed Pass@K protocol. On the strong pair (Table 2, Pass@2), Team accuracy implies that only 3.28% (GSM8K), 7.92% (MATH), 20.35% (AIME), and 22.66% (GPQA) of instances remain unsolved by *both* models. These reductions in correlated failure relative to SD-E^2 are the most direct driver of oracle team improvements.

Training curves (Figure 5, 6, 7, 8) show steady reward improvement on GSM8K, MATH, AIME and GPQA, with faster stabilization on arithmetic and slower, noisier progress on scientific QA. This is consistent with cross-teaching concentrating learning signal on high-value disagreement cases where one model can provide a useful solution pattern or

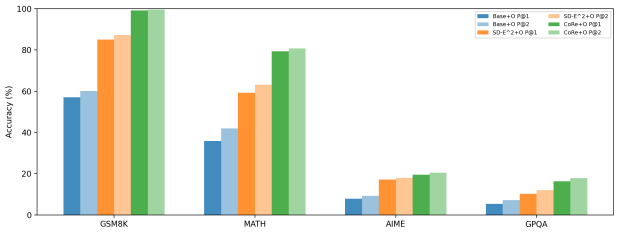

Figure 3. Base/SD-E^2 oracle baselines vs. the **CoRe** team for the Qwen2.5-3B + Qwen3-4B pair.

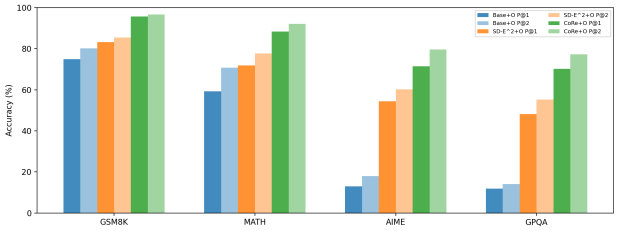

Figure 4. Base/SD-E^2 oracle baselines vs. the **CoRe** team for the Phi-4-mini + Ministral-3-3B pair.

missing step for its partner. Within-instance exploration (DPP-lite) mitigates mode collapse among a model's sampled traces, while the cross-model complementarity term R_{cross} explicitly discourages redundant high-quality traces across collaborators. Together, these incentives reshape the solution distribution so that success mass is spread across different reasoning modes, explaining why **CoRe** can outperform compute-matched oracle baselines formed from independently trained pairs (Tables 1 and 2) rather than merely benefiting from naive ensembling.

7. Error Analysis

We analyze remaining failures through two lenses: (1) **task-intrinsic reasoning difficulty** (where both models fail regardless of collaboration), and (2) **collaboration-specific failure modes** (where cross-teaching could help but fails due to hint selection, truncation, or brittleness).

A common failure pattern is *early-step correctness followed by drift*: both models choose the right plan but make a sign error, lose a factor, or mis-handle fractions/percentages midway. These errors are hard for a peer hint to repair if the hint does not explicitly highlight the fragile transformation. AIME amplifies this issue: many problems require a single "unlock" insight plus careful downstream algebra, so partial correctness is not enough. Another shared failure is missing edge cases (e.g., boundary values, domain restrictions, parity constraints). Even when models attempt case splits, they often prune too aggressively. This is particularly visible in contest-style problems where the entire solution hinges on enumerating all valid configurations. On GPQA, failures

are frequently driven by missing domain facts or overconfident elimination. Two models often converge to the same plausible distractor because it matches superficial patterns in the question.

Because teacher context is capped to a fixed budget, long peer traces can be truncated. If the truncated hint retains setup but drops the decisive step (e.g., a key identity or elimination rationale), the rescued model may confidently follow an incomplete plan. This motivates (i) selecting shorter teacher traces when multiple correct traces exist, or (ii) learned selection/truncation strategies.

A peer trace can be correct yet poorly justified or idiosyncratic; using it as teacher context may propagate brittle shortcuts or induce partial copying that fails to generalize. We mitigate this by selecting the highest exploitation-reward correct trace, stripping explicit answers, and applying hint dropout, but this risk remains. Our exploration and cross-model complementarity rewards further counteract such mode collapse by encouraging alternative, more robust reasoning paths.

Our hybrid distance is lightweight by design, but it can mis-score diversity: two traces may appear "diverse" under embeddings while sharing the same core mistaken assumption, or appear "similar" despite differing at the critical step. This suggests a future direction of verifier-informed distance or structured symbolic signatures for specific domains (e.g., algebraic transformations for MATH).

8. Conclusion

We proposed Collaborative Reasoning (**CoRe**), a training-time collaboration framework that converts *peer success* into an on-policy learning signal for reasoning LLMs. **CoRe** couples a two-round micro-protocol (cold sampling followed by contexted resampling with teacher context) with an explore–exploit reward that simultaneously improves correctness and discourages redundant solution modes via within-instance diversity and cross-model complementarity. Across GSM8K, MATH, AIME, and GPQA, **CoRe** improves both single-model Pass@K and oracle Team Pass@K, indicating meaningful reductions in correlated failures and increased complementary coverage at a fixed small-model parameter budget. Our extensive empirical analysis of our results support our proposed training-time collaboration as a practical route to more reliable reasoning without relying exclusively on scaling model size or introducing heavy inference-time search.

Impact Statement

CoRe may have positive impact by increasing the reliability of reasoning systems using multiple small, open models, which can improve robustness for educational tutoring, scientific assistance, and safety-critical decision support where single-model failures are costly. By explicitly reducing correlated failures, collaboration-trained systems may also be easier to monitor and verify using downstream selection mechanisms (e.g., verifiers or rankers), potentially improving trustworthiness in deployment. Potential risks include misuse for automated cheating or large-scale generation of convincing but incorrect solutions, as well as increased computational and environmental cost from running multiple models. Collaboration could also amplify shared biases if both collaborators inherit similar harmful behaviors from data or prompts, and oracle-style metrics may overstate real-world performance when selection is imperfect. Mitigations include restricting deployment in assessment settings, pairing **CoRe** with robust verification and safety filtering, reporting non-oracle selection results alongside oracle Team Pass@K, and using cost-aware serving policies (e.g., invoking the second model only when uncertainty is high).

References

Abdin, M., Aneja, J., Awadalla, H., Awadallah, A., Awan, A. A., Bach, N., Bahree, A., Bakhtiari, A., Bao, J., Behl, H., Benhaim, A., Bilenko, M., Bjorck, J., Bubeck, S., Cai, M., Cai, Q., Chaudhary, V., Chen, D., Chen, D., Chen, W., Chen, Y.-C., Chen, Y.-L., Cheng, H., Chopra, P., Dai, X., Dixon, M., Eldan, R., Fragoso, V., Gao, J., Gao, M., Gao, M., Garg, A., Giorno, A. D., Goswami, A., Gunasekar, S., Haider, E., Hao, J., Hewett, R. J., Hu, W., Huynh, J., Iter, D., Jacobs, S. A., Javaheripi, M., Jin, X., Karampatziakis, N., Kauffmann, P., Khademi, M., Kim, D., Kim, Y. J., Kurilenko, L., Lee, J. R., Lee, Y. T., Li, Y., Li, Y., Liang, C., Liden, L., Lin, X., Lin, Z., Liu, C., Liu, L., Liu, M., Liu, W., Liu, X., Luo, C., Madan, P., Mahmoudzadeh, A., Majercak, D., Mazzola, M., Mendes, C. C. T., Mitra, A., Modi, H., Nguyen, A., Norick, B., Patra, B., Perez-Becker, D., Portet, T., Pryzant, R., Qin, H., Radmilac, M., Ren, L., de Rosa, G., Rosset, C., Roy, S., Ruwase, O., Saarikivi, O., Saied, A., Salim, A., Santacroce, M., Shah, S., Shang, N., Sharma, H., Shen, Y., Shukla, S., Song, X., Tanaka, M., Tupini, A., Vaddamanu, P., Wang, C., Wang, G., Wang, L., Wang, S., Wang, X., Wang, Y., Ward, R., Wen, W., Witte, P., Wu, H., Wu, X., Wyatt, M., Xiao, B., Xu, C., Xu, J., Xu, W., Xue, J., Yadav, S., Yang, F., Yang, J., Yang, Y., Yang, Z., Yu, D., Yuan, L., Zhang, C., Zhang, C., Zhang, J., Zhang, L. L., Zhang, Y., Zhang, Y., Zhang, Y., and Zhou, X. Phi-3 technical report: A highly capable language model locally on your phone, 2024. URL https://arxiv.org/abs/2404.14219.

Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., et al. Training verifiers to solve math word problems.

arXiv preprint arXiv:2110.14168, 2021.

Dietterich, T. G. Ensemble methods in machine learning. In *International workshop on multiple classifier systems*, pp. 1–15. Springer, 2000.

Du, Y., Li, S., Torralba, A., Tenenbaum, J. B., and Mordatch, I. Improving factuality and reasoning in language models through multiagent debate. *arXiv preprint arXiv:2305.14325*, 2023. URL https://arxiv.org/abs/2305.14325.

Du, Y., Li, S., Torralba, A., Tenenbaum, J. B., and Mordatch, I. Improving factuality and reasoning in language models through multiagent debate. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 11733–11763. PMLR, 2024. URL https://proceedings.mlr.press/v235/du24e.html.

Foerster, J. N., Farquhar, G., Afouras, T., Nardelli, N., and Whiteson, S. Counterfactual multi-agent policy gradients. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 2974–2982, 2018. doi: 10.1609/AAAI.V32I1.11794.

Furlanello, T., Lipton, Z. C., Tschannen, M., Itti, L., and Anandkumar, A. Born again neural networks. *ICML*, 2018.

Gao, C., Zheng, C., Chen, X.-H., Dang, K., Liu, S., Yu, B., Yang, A., Bai, S., Zhou, J., and Lin, J. Soft adaptive policy optimization. *arXiv preprint arXiv:2511.20347*, 2025.

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

Hinton, G., Vinyals, O., and Dean, J. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Hsieh, C.-Y., Li, C.-L., Yeh, C.-k., Nakhost, H., Fujii, Y., Ratner, A., Krishna, R., Lee, C.-Y., and Pfister, T. 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, Toronto, Canada, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.507. URL https://aclanthology.org/2023.findings-acl.507/.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. Lora: Low-rank adaptation of large language models. *ICLR*, 2022.

Irving, G., Christiano, P., and Amodei, D. Ai safety via debate. *arXiv preprint arXiv:1805.00899*, 2018.

Jiang, D., Ren, X., and Lin, B. Y. Llm-blender: Ensembling large language models with pairwise ranking and generative fusion. *ACL*, 2023.

Kulesza, A. and Taskar, B. Determinantal point processes for machine learning. *Foundations and Trends in Machine Learning*, 5(2–3):123–286, 2012.

Kuncheva, L. I. and Whitaker, C. J. Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Machine Learning*, 51:181–207, 2003a. doi: 10.1023/A:1022859003006.

Kuncheva, L. I. and Whitaker, C. J. Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Machine Learning*, 51(2):181–207, 2003b. doi: 10.1023/A:1022859003006. URL https://doi.org/10.1023/A:1022859003006.

Lahlou, S., Abubaker, A., and Hacid, H. Port: Preference optimization on reasoning traces. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 10989–11005, 2025.

Li, G., Hammoud, H. A. A. K., Itani, H., Khizbullin, D., and Ghanem, B. Camel: Communicative agents for "mind" exploration of large language model society. *arXiv preprint arXiv:2303.17760*, 2023. URL https://arxiv.org/abs/2303.17760. Accepted at NeurIPS 2023 (per arXiv).

Lightman, H., Kosaraju, V., Burda, Y., Edwards, H., Baker, B., Lee, T., Leike, J., Schulman, J., Sutskever, I., and Cobbe, K. Let's verify step by step. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://arxiv.org/abs/2305.20050.

Liu, A. H., Khandelwal, K., Subramanian, S., Jouault, V., Rastogi, A., Sadé, A., Jeffares, A., Jiang, A., Cahill, A., Gavaudan, A., Sablayrolles, A., Héliou, A., You, A., Ehrenberg, A., Lo, A., Eliseev, A., Calvi, A., Sooriyarachchi, A., Bout, B., Rozière, B., Monicault, B. D., Lanfranchi, C., Barreau, C., Courtot, C., Grattarola, D., Dabert, D., de las Casas, D., Chane-Sane, E., Ahmed, F., Berrada, G., Ecrepont, G., Guinet, G., Novikov, G., Kunsch, G., Lample, G., Martin, G., Gupta, G., Ludziejewski, J., Rute, J., Studnia, J., Amar, J., Delas, J., Roberts, J. S., Yadav, K., Chandu, K., Jain, K., Aitchison, L., Fainsin, L., Blier, L., Zhao, L., Martin, L., Saulnier, L., Gao, L., Buyl, M., Jennings, M., Pellat, M., Prins, M., Poirée, M., Guillaumin, M., Dinot, M., Futeral, M., Darrin, M., Augustin, M., Chiquier, M., Schimpf, M., Grinsztajn, N.,

Gupta, N., Raghuraman, N., Bousquet, O., Duchenne, O., Wang, P., von Platen, P., Jacob, P., Wambergue, P., Kurylowicz, P., Muddireddy, P. R., Chagniot, P., Stock, P., Agrawal, P., Torroba, Q., Sauvestre, R., Soletskyi, R., Menneer, R., Vaze, S., Barry, S., Gandhi, S., Waghjale, S., Gandhi, S., Ghosh, S., Mishra, S., Aithal, S., Antoniak, S., Scao, T. L., Cachet, T., Sorg, T. S., Lavril, T., Saada, T. N., Chabal, T., Foubert, T., Robert, T., Wang, T., Lawson, T., Bewley, T., Bewley, T., Edwards, T., Jamil, U., Tomasini, U., Nemychnikova, V., Phung, V., Maladière, V., Richard, V., Bouaziz, W., Li, W.-D., Marshall, W., Li, X., Yang, X., Ouahidi, Y. E., Wang, Y., Tang, Y., and Ramzi, Z. Ministral 3, 2026. URL https://arxiv.org/abs/2601.08584.

Mishra, K., Lukas, N., and Lahlou, S. SD-e2: Semantic exploration for reasoning under token budgets. In Demberg, V., Inui, K., and Marquez, L. (eds.), *Findings of the Association for Computational Linguistics: EACL 2026*, pp. 6144–6157, Rabat, Morocco, March 2026. Association for Computational Linguistics. ISBN 979-8-89176-386-9. doi: 10.18653/v1/2026.findings-eacl.323. URL https://aclanthology.org/2026.findings-eacl.323/.

Motwani, S. R., Smith, C., Das, R. J., Rafailov, R., Laptev, I., Torr, P. H. S., Pizzati, F., Clark, R., and de Witt, C. S. MALT: Improving reasoning with multi-agent LLM training, 2025. URL https://arxiv.org/abs/2412.01928. Published at COLM 2025.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. Training language models to follow instructions with human feedback. *NeurIPS*, 2022.

Park, C., Han, S., Guo, X., Ozdaglar, A. E., Zhang, K., and Kim, J.-K. MAPoRL: Multi-agent post-co-training for collaborative large language models with reinforcement learning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 30215–30248, Vienna, Austria, July 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.acl-long.1459. URL https://aclanthology.org/2025.acl-long.1459/.

Rafailov, R., Sharma, A., Mitchell, E., Ermon, S., Manning, C. D., and Finn, C. Direct preference optimization: Your language model is secretly a reward model. *NeurIPS*, 2023.

Reimers, N. and community, S. sentence-transformers/all-minilm-l6-v2. https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2, 2020. Hugging Face model card (accessed 2026-01-29).

Reimers, N. and Gurevych, I. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3982–3992. Association for Computational Linguistics, 2019. doi: 10.18653/v1/D19-1410.

Rein, D., Hou, B. L., Stickland, A. C., Petty, J., Pang, R. Y., Dirani, J., Michael, J., and Bowman, S. R. GPQA: A Graduate-Level Google-Proof Q&A Benchmark. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=Ti67584b98. arXiv:2311.12022.

Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Zhang, M., Li, Y., Wu, Y., and Guo, D. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Team, Q. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Vijayakumar, A. K., Cogswell, M., Selvaraju, R. R., Sun, Q., Lee, S., Crandall, D., and Batra, D. Diverse beam search: Decoding diverse solutions from neural sequence models. *arXiv preprint arXiv:1610.02424*, 2016. URL https://arxiv.org/abs/1610.02424.

Wang, W., Wei, F., Dong, L., Bao, H., Yang, N., and Zhou, M. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *arXiv preprint arXiv:2002.10957*, 2020. doi: 10.48550/arXiv.2002.10957.

Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A., and Zhou, D. Self-consistency improves chain of thought reasoning in language models. *ICLR*, 2023.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., and Zhou, D. Chain-of-thought prompting elicits reasoning in large language models. *NeurIPS*, 2022.

Wu, Q., Bansal, G., Zhang, J., Wu, Y., Li, B., Zhu, E., Jiang, L., Zhang, X., Zhang, S., Liu, J., Awadallah, A. H., White, R. W., Burger, D., and Wang, C. Autogen: Enabling next-gen llm applications via multi-agent conversation. *arXiv preprint arXiv:2308.08155*, 2023. URL https://arxiv.org/abs/2308.08155.

Xu, H., Peng, B., Awadalla, H., Chen, D., Chen, Y.-C., Gao, M., Kim, Y. J., Li, Y., Ren, L., Shen, Y., Wang, S., Xu, W., Gao, J., and Chen, W. Phi-4-mini-reasoning: Exploring the limits of small reasoning language models in math, 2025. URL https://arxiv.org/abs/2504.21233.

Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., Zheng, C., Liu, D., Zhou, F., Huang, F., Hu, F., Ge, H., Wei, H., Lin, H., Tang, J., Yang, J., Tu, J., Zhang, J., Yang, J., Yang, J., Zhou, J., Zhou, J., Lin, J., Dang, K., Bao, K., Yang, K., Yu, L., Deng, L., Li, M., Li, M., Xue, M., Zhang, P., Wang, P., Zhu, Q., Men, R., Gao, R., Liu, S., Luo, S., Li, T., Tang, T., Yin, W., Ren, X., Wang, X., Zhang, X., Ren, X., Fan, Y., Su, Y., Zhang, Y., Zhang, Y., Wan, Y., Liu, Y., Wang, Z., Cui, Z., Zhang, Z., Zhou, Z., and Qiu, Z. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T., Cao, Y., and Narasimhan, K. Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems*, 36:11809–11822, 2023a.

Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., and Cao, Y. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*, 2023b. URL https://arxiv.org/abs/2210.03629. ICLR 2023 camera-ready version on arXiv.

Zelikman, E., Wu, Y., Mu, J., and Goodman, N. D. Star: Bootstrapping reasoning with reasoning. *arXiv preprint arXiv:2203.14465*, 2022. URL https://arxiv.org/abs/2203.14465. NeurIPS 2022 version exists; arXiv is the easiest stable citation.

Zhang, Y., Xiang, T., Hospedales, T. M., and Lu, H. Deep mutual learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4320–4328, 2018.

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

Zhou, D., Schärli, N., Hou, L., Wei, J., Scales, N., Wang, X., Schuurmans, D., Cui, C., Bousquet, O., Le, Q., and Chi, E. Least-to-most prompting enables complex reasoning in large language models. *arXiv preprint arXiv:2205.10625*, 2023. URL https://arxiv.org/abs/2205.10625. ICLR 2023 version on arXiv.

A. Reproducibility Details

A.1. Configuration summary

Table 4 lists key hyperparameters from the default configuration used in our main GSM8K runs.

The cold prompt is:

```
Question: <x>

Let's solve this step by step:
```

Table 4. Key hyperparameters.

Hyperparameter	Value
Cold traces (K)	2
Contexted traces (K')	1
Hint dropout (p_{hint})	0.5
Max teacher context tokens (L)	1536
Rescue bonus (r_{teach})	0.15
Explore margin (δ)	0.15
Explore set size cap	10
LoRA rank / alpha	16 / 32
LR (per model)	10^{-5}
Max grad norm	1.0

The contexted prompt injects the teacher context:

```
Question: <x>

Hint:
<c>

Let's solve this step by step:
```

We pass the teacher's reasoning *directly* as context and enforce a fixed length cap of 1536 tokens. Given ROUND A samples across models, we select the *highest exploitation-reward correct* trace as the teacher source. This reduces the probability of selecting spurious "lucky" correct outputs. We extract the `<strategy id="...">` ... `</strategy>` block whose strategy-level outcome matches the final answer. This yields a shorter, more focused teacher context than including all strategies. To prevent trivial leakage, we strip explicit answer markers (e.g., `Final answer:`, `\boxed{}`, `<final_answer>`) from the teacher context by default. This encourages the rescued model to learn *how* to solve the problem rather than copy the final token sequence. Truncation can remove critical steps for long proofs; we discuss this failure mode in Section 7.

Our evaluator extracts the final answer from common patterns such as "#### `<answer>`", "Final answer:", and bracketed formats. For numerical datasets like GSM8K, we normalize whitespace and punctuation before comparison.

A.2. Compute cost and Efficiency

A key practical advantage of **CoRe** is that all training runs operate in a low-data regime (at most 1,000 training examples per dataset), whereas SD-E^2 fine-tunes on the full training split (7,473 instances). (Mishra et al., 2026) Despite using substantially fewer training examples, **CoRe** attains large gains in both single-model Pass@K and Team Pass@K (Tables 1 and 2), indicating that cross-teaching provides a strong learning signal without requiring full-dataset RL runs.

Under our runs with batch size 4 and two traces per prompt, SD-E^2 training over the full ∼7.5k-example regime takes roughly 48–56 hours across backbones, while **CORE** completes in roughly 20–24 hours when trained on 1,000 examples (same batch size and trace budget). This reflects a favorable trade-off: **CORE** incurs additional per-step coordination (teacher selection, context construction, cross-model scoring), but the overall time-to-result is lower in our setting because training is intentionally sample-efficient.

CORE introduces analogous overhead sources: (i) selecting a teacher trace and forming the capped teacher context, (ii) computing diversity/complementarity terms under the same hybrid distance used for DPP-lite and R_{cross}, and (iii) lightweight structural extraction for the trace-distance. In practice these costs are modest compared to decoding, but they are paid twice per step (micro-rounds A/B) and across both collaborators.

Complexity analysis. We provide a coarse-grained compute model per training step, separating (a) *decoding*, which dominates wall-clock time, from (b) *scoring/selection* overheads. Let N be the number of collaborating models (here $N = 2$), K be the number of cold traces in ROUND A, and K' be the number of contexted traces in ROUND B. Let L denote the average generated trace length (tokens) and L_c the teacher-context budget (tokens). Let $C_{\text{gen}}(\cdot)$ be the per-token decoding cost for the backbone, and let $C_{\text{emb}}(\cdot)$ be the cost of computing a sentence embedding for a trace (used in $d(\tau, \tau')$).

(1) Decoding cost. Each step samples $N(K + K')$ traces, so the dominant term scales as

$$T_{\text{decode}} = \Theta\Big(N(K + K') \cdot L \cdot C_{\text{gen}}\Big) + \Theta\Big(NK' \cdot L_c \cdot C_{\text{attn}}\Big). \tag{14}$$

The second term captures the extra attention work from injecting the teacher context into ROUND B; in practice this is bounded by the fixed cap L_c (default 1536) and amortized by the fact that K' is small.

(2) Teacher selection and context construction. Selecting the best successful teacher among NK traces is linear:

$$T_{\text{select}} = \Theta(NK). \tag{15}$$

Constructing the teacher context is a single pass over the chosen trace and therefore scales with its length:

$$T_{\text{context}} = \Theta(L) \quad \text{(with truncation to } L_c). \tag{16}$$

Both are negligible relative to decoding.

(3) Distance-based rewards (DPP-lite and R_{cross}). Let $M = N(K + K')$ be the number of traces scored for the instance. Computing embeddings for all traces is

$$T_{\text{emb}} = \Theta(M \cdot C_{\text{emb}}). \tag{17}$$

Given embeddings and operation-signature sets, evaluating a single distance $d(\tau, \tau')$ is $O(1)$ for cosine distance plus $O(|op(\tau)| + |op(\tau')|)$ for Jaccard (small in practice due to coarse regex-triggered signatures). The greedy DPP-lite construction with an explore-set cap $s_{\max}$ evaluates distances from each candidate to the current set; with capped set size,

$$T_{\text{DPP-lite}} = \Theta(M \cdot s_{\max}). \tag{18}$$

The cross-model complementarity term compares each trace to its partner's traces (under a quality gate), yielding

$$T_{\text{cross}} = \Theta\big(NK_r^2\big) \ \text{ per round } r, \tag{19}$$

which reduces to $\Theta(K^2 + K'^2)$ in the two-model setting and is small for our defaults ($K = 2$, $K' = 1$).

(4) Total per-step complexity. Combining the above, the per-step runtime is

$$T_{\text{step}} = \Theta\Big(N(K + K') \cdot L \cdot C_{\text{gen}}\Big) + \Theta\Big(M \cdot C_{\text{emb}} + M \cdot s_{\max} + NK + L\Big), \tag{20}$$

where the first term (decoding) typically dominates, and the remaining terms are bounded by small constants under our capped budgets ($K, K', s_{\max}, L_c$). Thus, **CORE** adds modest *linear* overheads in the number of sampled traces, while the practical driver of wall-clock time remains the number/length of generated trajectories.

Overall, **CORE** trades a small amount of extra coordination per training step for substantially improved *team-level* reliability (reduced both-wrong mass) at a fixed small-model budget, and achieves competitive time-to-quality in a low-data regime. When compute or annotation budget is constrained, this "train less data, learn more from peers" behavior is a key part of **CORE**'s efficiency.

B. DPP-lite exploration reward properties

Proposition B.1 (Separation of the diverse set)**.** *Let S be the diverse set returned by the greedy construction with margin δ. Then every pair of distinct traces $\tau, \tau' \in S$ satisfies $d(\tau, \tau') \geq \delta$.*

Proof. A trace is added to S only if its minimum distance to the current set is at least δ. This maintains the invariant that each newly added trace is at distance $\geq \delta$ from all existing members. □

C. Policy optimization losses

For completeness, we summarize the loss functions implemented in `src/losses/`. Let $r_t = \exp(\log \pi_\theta(a_t \mid s_t) - \log \pi_{\theta_{\text{old}}}(a_t \mid s_t))$ be the token-level importance ratio, and let A denote a group-normalized advantage.

GRPO (token-level hard clipping). (Shao et al., 2024) GRPO uses PPO-style clipping at the token level:

$$\mathcal{L}_{\text{GRPO}} = -\mathbb{E}_t\Big[\min\Big(r_t A,\ \text{clip}\big(r_t, 1-\epsilon, 1+\epsilon_{\text{high}}\big)A\Big)\Big] + \beta\,\text{KL}\big(\pi_\theta \,\|\, \pi_{\text{ref}}\big). \tag{21}$$

GSPO (sequence-level ratio and clipping). (Zheng et al., 2025) GSPO computes a sequence-level ratio using the geometric mean of token ratios:

$$r_{\text{seq}} = \exp\Big(\tfrac{1}{|\tau|}\sum_t (\log \pi_\theta - \log \pi_{\theta_{\text{old}}})\Big), \tag{22}$$

then applies tight sequence-level clipping before multiplying the mean log-probability of the sequence.

SAPO (soft sigmoid gate). (Gao et al., 2025) SAPO replaces hard clipping with a smooth gate $g_t = \sigma(\tau\,(r_t - 1)) \cdot (4/\tau)$, with different temperatures for positive and negative advantages. The loss uses gated advantages:

$$\mathcal{L}_{SAPO} = -\mathbb{E}_t\big[g_t\, A\, \log \pi_\theta(a_t \mid s_t)\big] + \beta\,\text{KL}(\pi_\theta \| \pi_{\text{ref}}). \tag{23}$$

D. Additional benchmark results

We include a baseline comparison table for larger single models trained with SD-E^2 as well as an SD-E^2 oracle pair, evaluated on GSM8K, MATH, AIME, and GPQA. This "collaboration vs. scaling/compute" question: **CoRe** constrains collaboration to a combined budget of at most $\sim$7B parameters, so we contextualize the gains against SD-E^2-trained 7B–8B single models and a compute-matched SD-E^2 two-model oracle.

The 7B–8B SD-E^2 single-model baselines provide a reference for capacity and within-model diverse-exploration training. The AIME values in Table 5 are re-evaluated with the same 4096-token generation budget used in the main experiments to avoid truncating long contest-style solutions. Nevertheless, the **CoRe** team substantially exceeds these larger single models on all four benchmarks, despite operating under a smaller total parameter budget. This suggests that in the low-data post-training regime, many residual errors are not purely capacity-limited; they are often driven by systematic plan selection mistakes, brittle heuristics, or correlated blind spots that persist even when a model is encouraged to explore multiple traces.

SD-E^2 strengthens a *single* policy by promoting trace diversity, but correlated failure can still dominate when the model repeatedly chooses the same wrong decomposition or elimination strategy. In contrast, **CoRe** promotes *cross-model* complementarity: when one collaborator finds a viable path, that success becomes a training signal for the partner, and complementarity rewards discourage both models from collapsing onto the same dominant reasoning mode. As a result, the probability that *both* models fail decreases more aggressively than what is typically achieved by within-model diversity alone.

A strong compute-matched control is the SD-E^2 oracle pair (Phi-4-mini + Ministral-3-3B + Oracle), which already benefits from having two independently trained models and an oracle selector. That **CoRe** improves over this oracle across datasets indicates that gains cannot be explained by "just using two models" or by additional test-time compute. Instead, cross-teaching reshapes the learned policies: it improves individual competence while also shifting correct probability mass into complementary regions of the input space, thereby reducing overlap in errors and increasing oracle team success.

The largest gaps appear on the harder benchmarks (AIME/GPQA), where early wrong decisions often cascade and within-model sampling produces near-duplicate failures. Here, peer context provides a high-bandwidth corrective signal (e.g., a missing lemma, a successful elimination pattern, or a useful substitution), making collaboration training particularly effective. On GSM8K, although the task is easier, **CoRe** still closes much of the remaining reliability gap by reducing residual correlated arithmetic and plan-selection failures.

It can be concluded through Table 5 that **CoRe** delivers a qualitatively different improvement than modest parameter scaling or within-model diversity training: it explicitly targets correlated failures via training-time collaboration, yielding higher oracle reliability under tight data and compute budgets.

E. Comparison with MALT and MAPoRL

Table 6 summarizes the closest collaborative training baselines raised during review. The comparison is not intended as a perfectly controlled leaderboard because model families, training data, and metrics differ; rather, it clarifies the algorithmic distinction. MALT uses a sequential role-specialized pipeline (generator, verifier, refiner) with trajectory expansion and value-style credit assignment, whereas MAPoRL uses multi-agent PPO with a verifier reward over multi-turn discussions. **CoRe** instead uses instance-conditioned cross-teaching: a correct peer trace on the same problem becomes a hint for failed collaborators, and the rescue reward directly

Table 5. **Additional baseline comparison (Pass@1).** SD-E^2 baselines include 7B–8B single models and an SD-E^2 oracle pair; **CoRe** uses a $\leq$7B two-model collaboration budget.

Model	Training	GSM8K	MATH	AIME	GPQA
Qwen2.5-7B-Instruct	SD-E^2	86.21	44.00	26.00	30.00
Phi-3-small-8k-Instruct	SD-E^2	78.00	34.00	20.00	26.00
Ministral-3-8B-Reasoning	SD-E^2	84.50	47.00	51.50	33.00
Phi-4-mini + Ministral-3-3B + Oracle	SD-E^2	86.63	71.85	54.42	48.12
Phi-4-mini + Ministral-3-3B	**CoRe**	**97.04**	**88.42**	**71.42**	**70.18**

trains recovery on disagreement cases.

The main methodological distinction is that **CoRe** creates a targeted correction signal only when a peer succeeds on the same instance. This differs from fixed role specialization, generic discussion rewards, and inference-time debate: the learning target is the disagreement set identified online during training. These methods are therefore complementary; for example, a verifier from MAPoRL or a MALT-style refiner could be used as a downstream selector for **CoRe** candidates.

F. Hyperparameter and Distance Sensitivity

Reviewers asked whether the reward design is fragile due to the number of components. Table 7 sweeps the main task-specific parameters on MATH for the Phi-4-mini + Ministral-3B pair. Performance varies by at most about 0.9 percentage points across the tested ranges, suggesting that **CoRe** is not highly sensitive to exact tuning within reasonable ranges.

We also ablate the distance metric used by DPP-lite and R_{cross}. Table 8 shows that replacing the hybrid semantic/structural distance with cosine-only distance produces only a small drop on AIME. Thus, the core gain comes from cross-teaching, while the hybrid distance provides a modest benefit by capturing operation-level differences that embeddings alone may miss.

G. Scaling Beyond Two Collaborators

The micro-round protocol extends naturally to $N > 2$: Round A selects the best successful trace across all models, and Round B conditions failed models on the selected teacher context. Table 9 reports a preliminary 3-model AIME experiment. Adding Ministral-3-8B-Reasoning improves oracle Team Pass@2 by 2.95 percentage points over the 2-model team, suggesting that additional heterogeneous collaborators can contribute complementary correctness when their error sets are not redundant.

H. Cross-dataset Generalization

To test whether **CoRe** learns reusable reasoning strategies rather than only dataset-specific shortcuts, we evaluate AIME-trained **CoRe** models on MATH and GSM8K without additional fine-tuning. Table 10 shows favorable transfer: the AIME-trained team outperforms the base pair and the SD-E^2 oracle baseline, though per-dataset **CoRe** remains strongest.

I. AIME Split and Contamination Considerations

We use the AIME dataset from 1983–2025 and construct an 80:20 split (770 train / 193 test). Because many AIME problems are publicly available, contamination risk cannot be ruled out for any modern LLM. Our main conclusions rely on relative improvements under the same pretrained models and evaluation split: Base, SD-E^2, and **CoRe** are evaluated identically, so memorization risk affects all rows. We nevertheless report model-release cutoff considerations in Table 11 and treat absolute AIME accuracy with appropriate caution.

J. Ablations and Design Choices

We ablate the main components of **CoRe** to isolate which mechanisms drive gains. To keep ablations comparable to the main results, we report **oracle Team Pass@K** under the same Pass@1/Pass@2 protocol used in Tables 1 and 2. Unless otherwise noted, we use the default configuration (K=2, K'=1, hint dropout, DPP-lite, rescue bonus, and R_{cross} enabled).

Table 12 reports GSM8K team performance for the `Qwen2.5-3B-Instruct + Qwen3-4B-Instruct` pair, while Table 13 presents corresponding ablations for the stronger reasoning pair (`Phi-4-mini-reasoning` + `Ministral-3-3B-Reasoning`) on the harder benchmarks MATH, AIME, and GPQA.

Across both the easier GSM8K regime (Table 12) and the harder benchmarks (Table 13), removing Round B yields the largest degradation. This supports the core mechanism

Table 6. **Comparison to related collaborative/multi-agent post-training methods.** Reported numbers use each paper's own setting and are included to clarify evaluation context, not as a strictly controlled leaderboard.

Method	Collaboration stage	Learning signal	Representative setting
MALT (Motwani et al., 2025)	Sequential multi-agent post-training	Search tree with value-style credit assignment	Llama-3.1-8B, MV@3 on GSM8K/MATH/CSQA
MAPoRL (Park et al., 2025)	Multi-agent RL co-training	Verifier reward over multi-turn discussion	Phi-3-mini scale, GSM8K/ANLI
CoRe	On-policy cross-teaching	Peer success on the same instance plus rescue reward	3B–4B pairs, GSM8K/MATH/AIME/GPQA

Table 7. **Hyperparameter sensitivity on MATH.** Values are oracle Team Pass@2 for the Phi-4-mini + Ministral-3B pair.

Param.	Values tested	Team P@2
w_1	0.5/1.0/1.5	91.22/92.08/91.85
w_2	0.1/0.2/0.4	91.45/92.08/91.52
r_{teach}	0.25/0.50/0.75/1.0	92.08/92.15/91.90/91.62
p_{hint}	0.5/0.75/1.0	91.28–92.08

Table 8. **Distance metric sensitivity on AIME.** Values are oracle Team Pass@2 for the Phi-4-mini + Ministral-3B pair.

Distance metric	AIME Team P@2
Hybrid semantic/structural (default)	79.65
Cosine-only	78.70

Table 9. **Preliminary scaling to three models on AIME.**

Configuration	Oracle Team Pass@2
2-model **CoRe**	79.65
3-model **CoRe**	82.60

Table 10. **Zero-shot transfer from AIME-trained CoRe.** Values are Pass@1.

Setting	MATH	GSM8K
Base pair oracle	59.25	74.85
SD-E^2 oracle (per-dataset)	71.85	83.25
CoRe (AIME-trained $\rightarrow$ transfer)	75.90	86.35
CoRe Team (per-dataset)	88.42	95.72

of **CoRe**: when one model succeeds in Round A, its trace becomes a targeted correction signal for its partner on exactly the disagreement set. On the strong pair, this effect is especially pronounced (e.g., MATH P@2 drops from 92.08 to 88.90; GPQA P@2 drops from 77.34 to 72.20), reflecting that many remaining errors are *recoverable* given a correct peer plan or elimination pattern.

On GSM8K, removing R_{cross} yields a modest team drop because performance is already near-saturated and the both-wrong mass is tiny under Pass@2. However, on MATH/AIME/GPQA, ablating R_{cross} consistently harms both Pass@1 and Pass@2 (Table 13), indicating that explicit inter-model repulsion helps preserve complementary high-quality modes when the search space is larger and correlated failures are more common. Intuitively, DPP-lite prevents *within-model* trace collapse, while R_{cross} prevents *between-model* redundancy–the latter becomes more important as tasks demand multiple distinct solution templates.

Ablating DPP-lite ("w/o Explore Reward") hurts performance consistently, but the magnitude is smaller than removing rescue on hard datasets. This suggests that diversity is most valuable as a guardrail against mode collapse rather than as the primary source of gains: it increases the probability that at least one trace explores a different reasoning path, which in turn enables more frequent successful teaching events. Increasing K, K' provides only marginal improvements on GSM8K (Table 12), implying that once policies are strong, simply sampling more traces yields limited additional benefit. In contrast, the larger improvements from Round B and complementarity rewards indicate that the key bottleneck is not just candidate generation, but reshaping policies to reduce correlated failures and improve recovery dynamics.

Finally, Table 13 shows that the single-model baseline (M1 only) underperforms the full team, even though it shares the same backbone and training recipe. This supports the central claim: **CoRe** is not merely regularizing a single policy, but actively reallocating correctness mass across collaborators by teaching on disagreement instances and discouraging redundant high-quality reasoning modes.

K. Discussion

This section interprets the quantitative gains in Tables 1 and 2, the mechanism diagnostics in Section 6.2, and the component contributions in the ablations (Tables 12 and 13). We close with a qualitative analysis grounded in Table 14 and practical implications for deployment.

Team Pass@K is an *oracle* reliability measure: it reports whether *any* of the K samples from either collaborator is correct. Although it assumes an ideal selector, it is the cleanest diagnostic for **error overlap** and the **upper bound of selection**. In particular, Team Pass@K directly reflects the probability of *correlated failure* (the "both-wrong" mass),

Table 11. **AIME contamination considerations.** Cutoffs are approximate and intended only to contextualize absolute scores; relative comparisons use the same models and test split.

Model	Estimated training cutoff	AIME 2024 in data?	AIME 2025 in data?
Qwen2.5-3B-Instruct	late 2023	unlikely	no
Qwen3-4B-Instruct	around Mar. 2025	likely	possibly
Phi-4-mini-reasoning	around Jun. 2024	likely	no
Ministral-3-3B-Reasoning	late 2024	likely	unlikely

Figure 5. Train curves for AIME with `Phi-4-mini-reasoning` and `Ministral-3-3B-Reasoning`

Table 12. **Ablations on GSM8K for the Qwen team (oracle Team Pass@1/Pass@2).**

Variant	P@1	Δ	P@2	Δ
CoRe Team (Full)	99.10	–	99.54	–
w/o Round B (no rescue)	98.20	-0.90	98.90	-0.64
w/o Explore Reward	98.70	-0.40	99.20	-0.34
w/o Cross-Model Reward	98.50	-0.60	99.05	-0.49
K=2, K'=1 (default)	99.10	–	99.54	–
K=4, K'=2	99.20	+0.10	99.60	+0.06

which is the limiting factor for multi-model systems under small sampling budgets.

The results show that **CoRe** increases Team Pass@K for both pairs while also raising individual Pass@K (Tables 1, 2). This matters because oracle improvements can arise from two different sources: (i) *pure competence* (each model becomes more accurate), or (ii) *de-correlation* (models make different errors). **CoRe** targets both: Round B converts a peer's success into an instance-conditioned correction signal, and R_{cross} discourages redundant high-quality modes across collaborators.

A compute-matched baseline for collaboration is an oracle over two independently trained models (Tables 1, 2, Base/SD-E^2 oracle rows). These baselines already benefit from two-shot ensembling and diversity from different initializations/backbones. However, **CoRe** still yields substantially larger Team Pass@K gains, indicating that the improvement is not merely a test-time selection effect. The key difference is that **CoRe** couples the collaborators during training. When one model succeeds in Round A, its successful trace provides a high-bandwidth hint for the partner in Round B on the *same instance*. This converts sparse correctness signals into dense, targeted learning on exactly the disagreement set. Ablations confirm this: removing Round B causes the largest degradation for both the saturated GSM8K Qwen team (Table 12) and the stronger pair on harder datasets (Table 13).

Within-model exploration (DPP-lite) combats mode collapse among a model's own samples, but does not explicitly prevent two models from converging to the *same* strong-but-brittle strategy. The cross-model term R_{cross} addresses this by discouraging redundant high-quality traces across collaborators. This distinction explains why R_{cross} is modestly helpful on near-ceiling GSM8K but more consistently valuable on MATH/AIME/GPQA (Table 13), where correlated failure dominates oracle limits.

The Qwen pair begins weaker but reaches near-ceiling team performance on GSM8K under **CoRe** (Table 1). In contrast, on AIME/GPQA the Qwen team remains far from saturated, suggesting that some failures are driven by missing longer-horizon insights, factual knowledge gaps, or brittle transformations that are not reliably transferred by hints. The Phi+Ministral pair begins substantially stronger (Table 2) and benefits more sharply on the harder benchmarks: Team Pass@2 rises to 92.08 on MATH and 77.34 on GPQA, indicating that cross-teaching is especially effective when (i) each model has enough base competence to occasionally solve the instance, and (ii) remaining errors are often *recoverable* given a correct peer decomposition or elimination

Table 13. **Ablations on MATH/AIME/GPQA.** Values are Pass@1/Pass@2 for Phi-4-mini + Ministral-3B.

Variant	MATH		AIME		GPQA	
	P@1	P@2	P@1	P@2	P@1	P@2
CoRe Team (Full)	88.42	92.08	71.42	79.65	70.18	77.34
w/o Round B (no rescue)	84.50	88.90	66.20	74.80	64.80	72.20
w/o Explore Reward	86.70	90.60	69.40	78.10	68.00	75.90
w/o Cross-Model Reward	85.90	89.90	68.60	77.40	66.90	74.60
Single Model (M1 only)	85.24	90.01	68.16	76.64	67.76	75.26

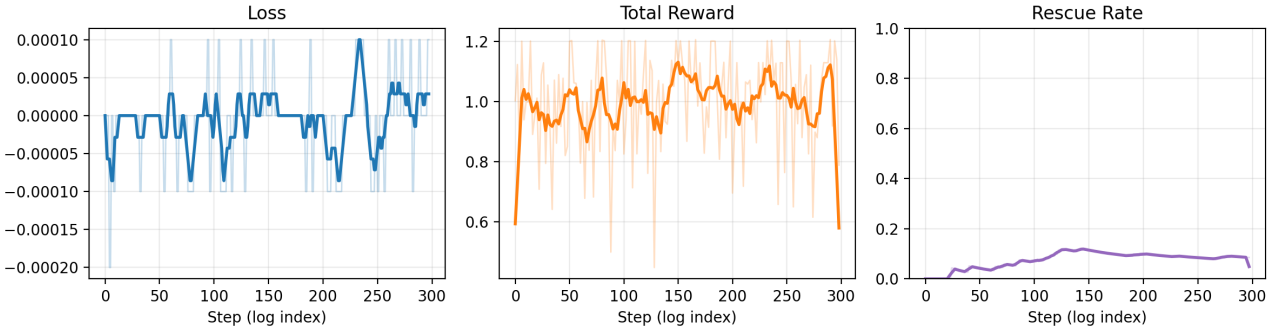

Figure 6. Train curves for GSM8K with `Phi-4-mini-reasoning` and `Ministral-3-3B-Reasoning`

rationale. This aligns with the "teacher-available" region: **CoRe** can only rescue when at least one collaborator succeeds in Round A; as models become more competent, this region expands, creating a positive feedback loop in training.

K.1. Deployment Implications

A practical system must select a single answer without oracle access. Our results suggest that **CoRe** is a particularly strong fit for settings where selection is imperfect but improving:

- **Verifier-based selection.** Since **CoRe** increases the probability that at least one candidate is correct, verifier capacity is spent distinguishing a smaller set of plausible solutions rather than searching for correctness from scratch. This is precisely the regime where lightweight verifiers tend to be most effective.
- **Disagreement-triggered routing.** Because collaboration gain concentrates on disagreement instances, a simple router can invoke the second model or additional sampling only when the first model's answer is uncertain (low log-prob margin) or when two collaborators disagree. This targets compute to the hard tail while keeping average serving cost low.
- **Agreement as a confidence heuristic.** Although agreement does not guarantee correctness, **CoRe** reduces redundant failure modes, so agreement becomes more informative: if both models converge, it is more likely to reflect a robust shared reasoning mode rather than a shared shortcut.

Importantly, these strategies are complementary to **CoRe**: better selection converts oracle gains into deployable gains, and **CoRe** makes selection easier by increasing complementary correctness.

K.2. Qualitative Analysis

We inspected pass@1 evaluation traces for GSM8K, MATH, AIME, and GPQA in `evaluation_results` and compared model outputs from M1 (Phi-4-mini) and M2 (Ministral-3B). Table 14 highlights representative cases where collaboration matters. Two consistent patterns emerge. First, **complementary rescue**: M2 often resolves geometry or domain-knowledge questions that M1 misses (AIME/GPQA), while M1 is more reliable on compact algebraic simplifications (MATH). Second, **format sensitivity**: several incorrect outcomes are due to malformed or missing final answers (e.g., not listing comma-separated roots), which hurts conditional accuracy even when a partial derivation is plausible. This suggests that lightweight answer-normalization and stricter format prompts can further improve deployable accuracy without changing the underlying reasoning model.

At a finer granularity, we observe dataset-specific tendencies: (i) **GSM8K**: errors often stem from scale and formatting slips (e.g., missing zeros) rather than algebraic reasoning;

Figure 7. Train curves for GPQA with `Phi-4-mini-reasoning` and `Ministral-3-3B-Reasoning`

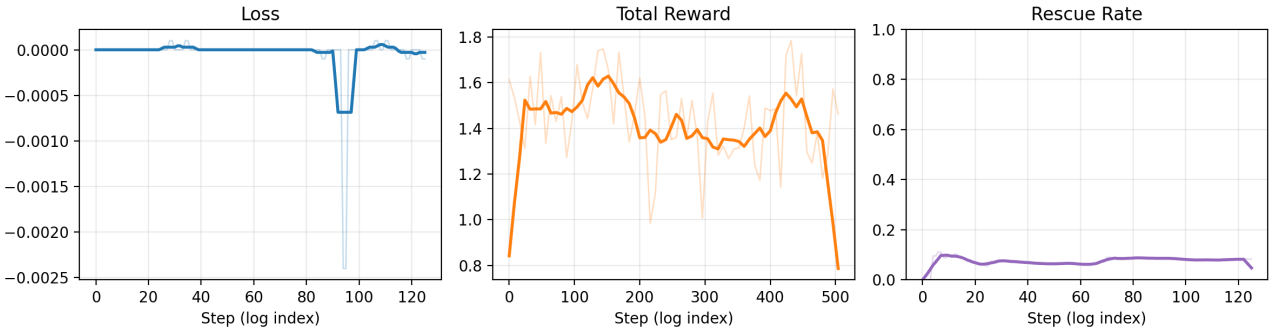

Figure 8. Train curves for MATH with `Phi-4-mini-reasoning` and `Ministral-3-3B-Reasoning`

(ii) **AIME**: M2 more consistently handles geometric constraint propagation and symmetry arguments, while M1 is more brittle when the solution depends on a specific configuration or auxiliary construction; (iii) **MATH**: M1 does better on compact algebraic products and substitutions, but both models are sensitive to answer-format conventions (intervals, ordered pairs, and multi-root lists); (iv) **GPQA**: chemistry and physics questions exhibit the most pronounced knowledge gaps, where both models can latch onto plausible distractors unless stoichiometry or mechanistic details are explicit. These trends mirror the quantitative gains in Table 2: collaboration helps most where the models' error sets are complementary rather than identical.

K.3. Limitations and Open Problems

Cross-teaching activates only when at least one collaborator succeeds in Round A. On extremely hard instances, both models fail and the method reduces to exploration/exploitation. Combining **CoRe** with retrieval, tool-use, or a verifier that can identify partially correct solutions could expand the teacher-available region. The reward structure extends naturally to $N>2$, but coordination raises questions: which teacher to trust, how to allocate sampling budgets, and how to prevent group collapse onto a single dominant style. Structured teacher selection and curriculum strategies (e.g., rotating teachers, difficulty-based routing) are promising directions.

Table 14. **Qualitative examples from pass@1 evaluation logs.** M1 is Phi-4-mini; M2 is Ministral-3B. Entries show final answers extracted from the `<final_answer>` tag when present. Row colors: blue = M2 correct, green = M1 correct, orange = both wrong.

Dataset	Problem (abbrev.)	GT	M1	M2	Notes
GSM8K	Jill earns \$20/hr teaching and \$30/hr coaching, 35+15 hrs/week for 50 weeks; annual salary?	57500	575	57500	M2 correct; M1 truncates zeros.
AIME	Two cubics share a complex root $m + \sqrt{n}\,i$; given real roots -20 and -21, find $m + n$.	330	310	330	M2 correct; M1 underestimates $m + n$.
AIME	Isosceles right triangle with interior point P s.t. $\angle PAB = \angle PBC = \angle PCA$ and $AP = 10$; find area.	250	200	250	M2 correct; M1 misses geometric constraint.
MATH	Let r be a root of $x^2 + 5x + 7 = 0$; compute $(r-1)(r+2)(r+6)(r+3)$.	13	13	001	M1 correct; M2 collapses to small value.
MATH	Integer roots of $2x^4 + 4x^3 - 5x^2 + 2x - 3 = 0$ (comma-separated).	1,-3	ANSWER	1, -3	M2 correct; M1 formatting failure.
GPQA	Enthalpy of neutralization for mixed $HCl/H_2SO_4/Ba(OH)_2$ (MCQ).	A	B	A	M2 correct; M1 picks distractor.
GPQA	$NaNH_2$ on 1-bromobenzene-2-d: number of organic products (MCQ).	A	C	B	Both wrong; likely benzyne-mechanism confusion.