

COLLABORATIVE LEARNING AND REASONING OVER MODEL SLICES

Yuheng Wang[†] Rundu Yang[†] Zhuohan Bao Qiran Zhang Huatao Li
Yiyang Zheng Lin Wu Jingru Fan Shu Yao Yihan Li Zhi Li Chen Qian[✉]

School of Artificial Intelligence, Shanghai Jiao Tong University

ABSTRACT

Large language models (LLMs) excel across diverse tasks, benefiting from knowledge priors, in-context reasoning, and test-time computation. Despite these strengths, single-model reasoning remains vulnerable to error propagation, as later steps depend on the model’s earlier outputs, leaving reasoning blind spots difficult to independently reassess. To address this limitation, we develop *collaborative learning*, encompassing the construction of heterogeneous collaborators and the learning of effective cross-model collaboration. Specifically, *model slicing* replaces a large model with smaller, heterogeneous models under a comparable parameter budget, enabling prior reasoning to be reassessed across distinct model parameterizations. *Pyramid supervision* combines protocol distillation with step-level supervision to refine reasoning and interaction decisions throughout collaboration. We evaluate our approach across four task categories: multi-hop question answering, mathematical reasoning, code generation, and open-ended constrained generation. The trained heterogeneous models achieve comparable or better task performance than the original large model and established multi-agent baselines, including reinforcement-learning-based methods, while also achieving higher inference efficiency than the multi-agent baselines and exhibiting differentiated collaboration patterns across tasks. These findings show that learned collaboration turns model heterogeneity into complementary reasoning capabilities. Code is available at: github.com/SJTUFreshman/collaborative-learning.

1 INTRODUCTION

Large language models (LLMs) have demonstrated capabilities across tasks (Wang et al., 2024b). Advances highlight the benefits of scaling test-time computation (Snell et al., 2025; Muennighoff et al., 2025; Wu et al., 2025). Many approaches seek to improve task performance by allowing a single model to structure its reasoning process (Besta et al., 2024) and generate, inspect, and revise intermediate steps (Zhou et al., 2024; Zhang et al., 2024; Kumar et al., 2025). However, when the same model both generates and evaluates a solution, it may exhibit self-bias, favoring its own outputs and limiting its ability to identify and correct reasoning errors (Huang et al., 2024; Xu et al., 2024). Consequently, errors in intermediate steps may remain undetected and influence subsequent reasoning (Zheng et al., 2025). This motivates *reasoning-state isolation* across models, allowing prior reasoning to be independently reassessed under distinct model parameterizations rather than repeatedly evaluated within the same model’s reasoning space (Du et al., 2024; Liang et al., 2024).

Accordingly, we develop a *collaborative learning* framework for reasoning and learning across heterogeneous models, as illustrated in Figure 1. Model slicing replaces a single large model with several smaller, heterogeneous models under a comparable total parameter budget (Liu et al., 2026b). Through explicit handoffs of task state between models (Chen et al., 2024c), prior reasoning can be independently reassessed under distinct model parameterizations, thereby enabling model-specific reasoning-state isolation while producing interaction trajectories that capture reasoning and interaction decisions. Together, this isolation and the resulting trajectories provide the structural basis for

[†] Equal contribution.

✉ Corresponding author.

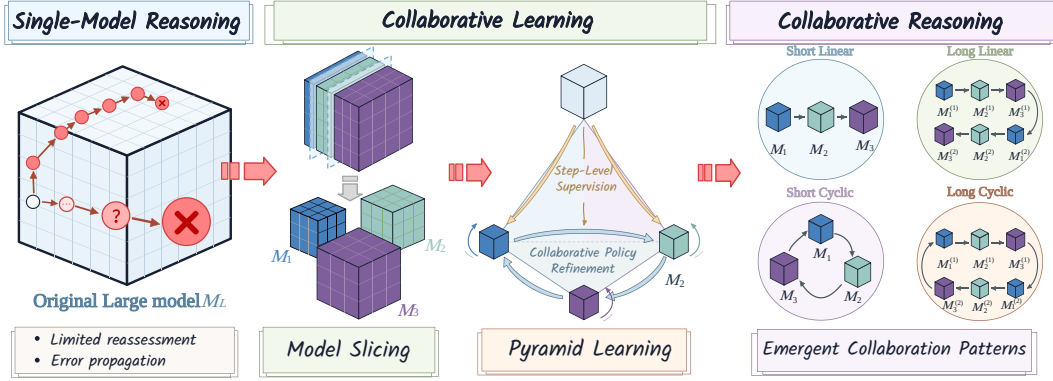


Figure 1: **Collaborative Learning and Reasoning over Model Slices.** Model slicing reallocates a single large model into multiple smaller, heterogeneous models under a comparable parameter budget, enabling reasoning to be reassessed across model parameterizations. Building on this, collaborative learning trains heterogeneous models through pyramid supervision, combining protocol distillation for interaction initialization with step-level supervision over collaboration trajectories².

training heterogeneous models to leverage complementary capabilities (Wang et al., 2025; Abdulaal et al., 2025; Chen et al., 2025) when questioning, extending, or revising prior conclusions.

Although this structure provides a foundation for cross-model collaboration, outcome-level supervision alone provides insufficient guidance for intermediate reasoning and interaction decisions throughout the collaborative process (Lightman et al., 2024; Wang et al., 2024a; Setlur et al., 2025). To address this limitation, we propose *pyramid supervision*, which uses supervisory signals from larger models to guide the heterogeneous models produced by model slicing toward effective cross-model collaboration. It combines interaction initialization with fine-grained supervision over reasoning and interaction decisions along model-generated collaboration trajectories, while integrating task-level rewards to jointly optimize intermediate decision quality and overall task performance. Through this training mechanism, heterogeneous models learn effective collaborative reasoning and develop differentiated interaction patterns across tasks.

We evaluate our approach across four task categories: multi-hop question answering, mathematical reasoning, code generation, and open-ended constrained generation. With a total parameter budget comparable to that of the original large model, the trained heterogeneous models collectively match or outperform the original model as well as established multi-agent system (MAS) and multi-agent reinforcement learning (MARL) baselines. Analysis of interaction trajectories further reveals reduced reasoning redundancy and differentiated collaboration patterns characterized by questioning, complementing, and revising prior reasoning. Together, these results suggest that collaborative reasoning across heterogeneous models can improve task performance while enabling complementary contributions and the reassessment of prior reasoning across distinct model parameterizations.

Our key contributions are summarized in three points:

- We propose a trainable paradigm for collaborative reasoning and learning across heterogeneous models, enabling independent reassessment across distinct model parameterizations and progressive learning of effective cross-model collaboration from interaction trajectories.
- Collaborative learning comprises model slicing and pyramid supervision. Model slicing constructs heterogeneous models under a comparable budget and enables reasoning-state isolation through task-state handoffs. Pyramid supervision combines interaction initialization with fine-grained supervision from the original large model over reasoning and interaction decisions.
- We validate collaborative learning across four task categories. The trained heterogeneous models achieve comparable or better performance than the original large model, established MAS baselines, and MARL methods, while demonstrating higher inference efficiency than MAS and MARL baselines and exhibiting emergent task-adaptive cross-model collaboration patterns.

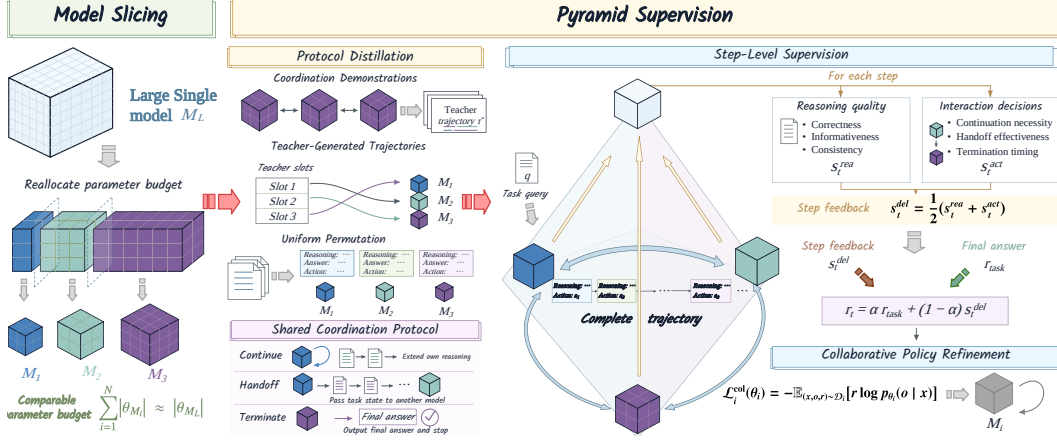


Figure 2: **Overview of collaborative learning.** (a) Model slicing constructs smaller, heterogeneous models under a comparable total parameter budget. (b) Pyramid supervision initializes collaboration through protocol distillation and refines it through step-level supervision. The original large model evaluates each step’s reasoning and interaction decisions within the complete trajectory. Step-level scores and final task rewards jointly guide collaborative policy refinement.

2 METHOD

Collaborative reasoning characterizes an iterative reasoning process among heterogeneous models, where explicit task-state handoffs under a unified interaction protocol allow prior reasoning to be continually reassessed by models with different parameters (▷ Section 2.1). To realize this process, collaborative learning constructs a heterogeneous model set through model slicing and learns effective cross-model collaboration through pyramid supervision (▷ Section 2.2).

2.1 COLLABORATIVE REASONING

Collaborative reasoning refers to a process in which multiple models jointly construct a solution through successive contributions (Yin et al., 2023; Chen et al., 2024a). We define a unified interaction protocol that specifies how heterogeneous models generate reasoning and retain, transfer, or terminate control. For a task input q , each interaction step t is represented by

$$z_t = (i_t, x_t, y_t, a_t) \quad (1)$$

where i_t denotes the active model M_{i_t} , x_t contains the task input and preceding interaction history, y_t is the current reasoning, and a_t is the interaction action. The action space is defined as

$$a_t \in \{\text{continue}, \text{terminate}\} \cup \{\text{handoff}(j) \mid j \in \{1, \dots, N\}, j \neq i_t\} \quad (2)$$

where `continue` keeps control with the model, `handoff(j)` transfers the task state and control to M_j , and `terminate` ends collaborative reasoning. A complete trajectory is represented as

$$\tau = (z_1, z_2, \dots, z_T) \quad (3)$$

where T denotes the total number of interaction steps recorded in the complete trajectory.

Through `handoff`, models use distinct parameters to extend, question, or revise prior conclusions, enabling cross-model reassessment across the evolving reasoning process. The resulting trajectories record reasoning and interaction decisions for subsequent collaborative learning.

2.2 COLLABORATIVE LEARNING

Collaborative learning encompasses both the construction of heterogeneous collaborators and the learning of cross-model collaboration, with the pipeline illustrated in Figure 2. Model slicing constructs heterogeneous collaborators under a comparable parameter budget, providing the structural

²Short linear, long linear, short cyclic, and long cyclic trajectories are defined in Section 3.5.

basis for cross-model reasoning. Pyramid supervision guides collaborative learning through protocol distillation and step-level supervision, progressively refining models’ reasoning and interaction decisions while integrating task-level rewards throughout the collaborative reasoning process.

2.2.1 MODEL SLICING

Model slicing replaces a large model with smaller, independently parameterized models under a comparable parameter budget, each operating as a standalone model and interacting through explicit task-state handoffs. Let M_L denote the large model with parameters θ_L . We have

$$\mathcal{M} = \{M_1, \dots, M_N\}, \quad \sum_{i=1}^N |\theta_i| \approx |\theta_L| \quad (4)$$

where θ_i denotes the parameters of M_i , and $|\theta|$ denotes the parameter count.

These models differ in size and parameters and are not obtained by directly partitioning the weights of M_L . Together with task-state handoffs, this heterogeneous model set enables model-specific *reasoning-state isolation* across distinct model parameterizations during collaboration.

2.2.2 PYRAMID SUPERVISION

We introduce *pyramid supervision*, which refines collaboration through supervision across model scales. Protocol distillation uses the largest model slice to initialize cross-model interaction, while step-level supervision uses the original large model to evaluate collaboration trajectories.

Protocol Distillation. Although the interaction protocol specifies how models communicate and transfer control, heterogeneous models do not naturally learn to use it effectively as reasoning evolves. In particular, they must learn when to retain control, hand off the task state, or terminate based on the evolving reasoning context (Zeng et al., 2024). We therefore use protocol distillation to initialize cross-model interaction from teacher-generated trajectories (Chen et al., 2024b).

We select the largest model in $\mathcal{M}$ by parameter count as the teacher M^* and instantiate N homogeneous copies to form the teacher system:

$$\mathcal{M}_{\text{teacher}} = \{M_1^*, \dots, M_N^*\}, \quad M_j^* \equiv M^*, \quad j = 1, \dots, N \quad (5)$$

Following the same interaction protocol as the target models, the teacher system generates multi-step trajectories for distillation inputs $\mathcal{Q}_{\text{distill}} = \{q^{(m)}\}_{m=1}^K$:

$$\tau^{(m)} \sim P_{\mathcal{M}_{\text{teacher}}}(\tau \mid q^{(m)}), \quad m = 1, \dots, K \quad (6)$$

We retain qualified trajectories as demonstrations for subsequent distillation.

To avoid a fixed correspondence between teacher slots and target models, we sample a uniform permutation for each trajectory:

$$\pi^{(m)} \sim \text{Uniform}(S_N), \quad \tilde{i}_t^{(m)} = \pi^{(m)}(i_t^{(m)}), \quad t = 1, \dots, T_m \quad (7)$$

where S_N denotes the set of permutations over $\{1, \dots, N\}$. Within each trajectory, the same permutation is applied consistently to all occurrences associated with each teacher slot, including model indices, handoff targets, and corresponding references in the interaction history. This operation changes only the correspondence between teacher slots and target models across trajectories, while preserving the reasoning content, step order, and handoff structure of the original trajectory.

Let $\tilde{x}_t^{(m)}$ and $\tilde{o}_t^{(m)}$ denote the reassigned context and response. Supervision covers the complete response, including reasoning and interaction decisions (Chen et al., 2023). For each model M_i , the reassigned context–response pairs satisfying $\tilde{i}_t^{(m)} = i$ form its training set $\mathcal{D}_i^{\text{pro}}$. Following sequence-level knowledge distillation (Kim & Rush, 2016), each model minimizes

$$\mathcal{L}_i^{\text{pro}} = -\mathbb{E}_{(x,o) \sim \mathcal{D}_i^{\text{pro}}} [\log p_{\theta_i}(o \mid x)] \quad (8)$$

This objective initializes cross-model interaction capabilities in the heterogeneous models, providing a stable starting point for trajectory-level refinement of reasoning and interaction decisions.

Step-Level Supervision. After protocol distillation, the heterogeneous models perform collaborative reasoning under the shared interaction protocol, generating trajectories for subsequent optimization. For a given training task q , a collaborative trajectory is sampled as

$$\tau = (z_1, \dots, z_T) \sim P_{\mathcal{M}}(\tau \mid q) \quad (9)$$

where $P_{\mathcal{M}}$ is induced by the current models and the shared interaction protocol.

Learning effective collaboration from these trajectories requires identifying the contribution of each reasoning and interaction decision to the overall process (Foerster et al., 2018). The value of an individual step depends not only on its local correctness but also on how it influences subsequent collaboration. We therefore retain the original large model M_L as an evaluator $\mathcal{J}$, which assesses each step in the context of the complete trajectory. Given the task q and trajectory τ , the evaluator produces two complementary scores:

$$\{(s_t^{\text{rea}}, s_t^{\text{act}})\}_{t=1}^T = \mathcal{J}(q, \tau), \quad s_t^{\text{rea}}, s_t^{\text{act}} \in [-1, 1] \quad (10)$$

The reasoning score s_t^{rea} evaluates the model’s reasoning contribution in terms of correctness, informativeness, and relevance to the task state and interaction history. The interaction score s_t^{act} evaluates whether its decision to continue, hand off the task state, or terminate appropriately advances the collaborative process. Separating these dimensions avoids conflating reasoning quality with interaction quality. We average the two scores to obtain a step-level score:

$$s_t^{\text{del}} = \frac{1}{2} (s_t^{\text{rea}} + s_t^{\text{act}}) \quad (11)$$

Reward integration. Step-level scores provide fine-grained supervision for intermediate reasoning and interaction decisions, but local decision quality does not necessarily translate into successful task completion. To provide decision-specific feedback while also accounting for the overall task objective, we integrate each step-level score with the final task reward. Let $m(q, \hat{y}) \in [0, 1]$ denote the task-specific evaluation metric for the final prediction $\hat{y}$. We map it to the same $[-1, 1]$ range as the step-level score and define

$$r_{\text{task}} = 2m(q, \hat{y}) - 1, \quad r_t = \alpha r_{\text{task}} + (1 - \alpha) s_t^{\text{del}} \quad (12)$$

where $\alpha \in [0, 1]$ controls the trade-off between the task-level reward and the step-level score. The resulting reward distinguishes the contributions of intermediate steps across the complete collaborative trajectory while aligning their optimization with the final task objective.

Collaborative Policy Refinement. Using the combined rewards, we extract the steps at which each model was active from the collaborative trajectories and construct model-specific training data. For model M_i , the corresponding training examples are collected as

$$\mathcal{D}_i = \left\{ (x_t^{(m)}, o_t^{(m)}, r_t^{(m)}) \mid i_t^{(m)} = i \right\} \quad (13)$$

where $o_t^{(m)}$ denotes the complete response at step t of trajectory m , serialized under the shared interaction protocol and including its reasoning, answer, and interaction fields. Following the score-function policy-gradient form, we optimize each model with the reward-weighted objective

$$\mathcal{L}_i^{\text{col}}(\theta_i) = -\mathbb{E}_{(x, o, r) \sim \mathcal{D}_i} [r \log p_{\theta_i}(o \mid x)] \quad (14)$$

This objective increases the likelihood of responses assigned positive rewards and decreases the likelihood of those assigned negative rewards. Because each reward combines step-level evaluation conditioned on the complete trajectory with the final task outcome, each model can adapt its policy according to how its own decisions contribute to the collaborative result, allowing differentiated collaboration patterns to progressively emerge through training (Cui et al., 2026).

3 EXPERIMENTS

3.1 EXPERIMENTAL SETUP

Datasets. We use five benchmarks across four task categories: multi-hop question answering (MuSiQue (Trivedi et al., 2022)), mathematical reasoning (GSM-Hard (Gao et al., 2023)) and

MATH (Hendrycks et al., 2021)), code generation (an eight-language MultiPL-E subset (Cassano et al., 2023))³, and open-ended constrained generation (Conifer (Sun et al., 2024)).

Baselines. (1) Single-agent baselines: Qwen3-14B (Yang et al., 2025a) and its self-evaluated RL variant (SE-RL)⁴. (2) A homogeneous MAS of three Qwen3-8B models (24B total) without collaborative training, serving as a higher-capacity multi-model baseline. (3) A homogeneous collaborative-learning system (Homo. CL) of three Qwen3-4B models (12B total), trained with the same procedure as our heterogeneous system to assess the effect of model heterogeneity in collaborative reasoning. (4) Established multi-agent frameworks: MAD (Du et al., 2024), AgentVerse (Chen et al., 2024c), AFlow (Zhang et al., 2025), and GPTSwarm (Zhuge et al., 2024). (5) MARL methods: MAGRPO (Liu et al., 2026b), MAPoRL (Park et al., 2025), and AT-GRPO (Zhao et al., 2026b).

Metrics. We report accuracy and F1 on MuSiQue (Trivedi et al., 2022), GSM-Hard (Gao et al., 2023), and MATH (Hendrycks et al., 2021). We use weighted pass@1 and macro-language pass@1 for MultiPL-E (Cassano et al., 2023), and Coverage and Explicit for Conifer (Sun et al., 2024). Beyond task performance, we analyze handoffs, anomaly rates, reasoning redundancy, and model-wise behavioral distributions to characterize training-induced changes in collaboration patterns.

Implementation Details. Model slicing replaces Qwen3-14B (M_L) (Yang et al., 2025a) with Qwen3-1.7B, Qwen3-4B, and Qwen3-8B (13.7B total). The three collaborating models share the same prompt and action space. Three Qwen3-8B instances generate protocol distillation demonstrations, while Qwen3-14B evaluates reasoning and interaction decisions within complete trajectories for step-level supervision. Experiments run on $8 \times$ NVIDIA A800-SXM4-80GB GPUs. Each model is trained with a rank-64 LoRA adapter (Hu et al., 2022) while its base parameters remain frozen.

3.2 DOES COLLABORATIVE LEARNING LEAD TO IMPROVED PERFORMANCE?

Table 1: **Performance comparison with baseline methods.** Scores are percentages (%); higher is better. Acc. denotes accuracy; Wtd. and Mac. denote weighted and macro-language pass@1, respectively; Cov. and Expl. denote Coverage and Explicit, respectively. CL. denotes collaborative learning. Best and second-best results are shown in **bold** and underline, respectively. $\uparrow$ and $\downarrow$ indicate scores above and below the corresponding baseline mean for each metric in the table. $\dagger$ indicates that the corresponding baseline differs significantly from CL. ($p \leq 0.05$).

Method	MuSiQue		GSM-Hard		MATH		MultiPL-E		Conifer	
	Acc.	F1	Acc.	F1	Acc.	F1	Wtd.	Mac.	Cov.	Expl.
Qwen3-14B	39.55 $\uparrow$ _{10.87}	49.70 $\uparrow$ _{10.47}	68.94 $\uparrow$ _{11.65}	<u>80.69</u> $\uparrow$ _{13.19}	74.60 $\uparrow$ _{14.40}	82.86 $\uparrow$ _{13.25}	55.18 $\uparrow$ _{11.08}	57.34 $\uparrow$ _{12.59}	88.00 $\uparrow$ _{10.17}	95.72 $\uparrow$ _{10.62}
SE-RL	41.75 $\uparrow$ _{13.07}	52.71 $\uparrow$ _{13.48}	68.18 $\uparrow$ _{10.89}	79.67 $\uparrow$ _{12.17}	<u>76.00</u> $\uparrow$ _{15.80}	80.70 $\uparrow$ _{11.09}	60.43 $\uparrow$ _{16.33}	62.55 $\uparrow$ _{17.80}	88.00 $\uparrow$ _{10.17}	95.81 $\uparrow$ _{10.71}
Homo. MAS	26.40 $\uparrow$ _{12.28}	32.82 $\uparrow$ _{11.64}	<u>71.21</u> $\uparrow$ _{13.92}	79.93 $\uparrow$ _{12.43}	68.20 $\uparrow$ _{12.00}	76.41 $\uparrow$ _{13.20}	56.36 $\uparrow$ _{12.26}	57.20 $\uparrow$ _{12.45}	88.02 $\uparrow$ _{10.19}	94.27 $\uparrow$ _{10.83}
Homo. CL.	36.00 $\uparrow$ _{12.68}	47.19 $\uparrow$ _{12.04}	60.61 $\uparrow$ _{16.68}	76.51 $\uparrow$ _{10.99}	68.60 $\uparrow$ _{11.60}	77.95 $\uparrow$ _{11.66}	53.33 $\uparrow$ _{10.77}	52.65 $\uparrow$ _{12.10}	88.15 $\uparrow$ _{10.32}	95.22 $\uparrow$ _{10.12}
MAD	39.51 $\uparrow$ _{10.83}	50.59 $\uparrow$ _{11.36}	66.67 $\uparrow$ _{10.62}	74.18 $\uparrow$ _{13.32}	72.40 $\uparrow$ _{12.20}	81.32 $\uparrow$ _{11.71}	55.03 $\uparrow$ _{10.93}	55.35 $\uparrow$ _{10.60}	88.55 $\uparrow$ _{10.72}	95.84 $\uparrow$ _{10.74}
AgentVerse	<u>41.99</u> $\uparrow$ _{13.31}	52.85 $\uparrow$ _{13.62}	71.97 $\uparrow$ _{14.68}	78.44 $\uparrow$ _{10.94}	68.20 $\uparrow$ _{12.00}	79.42 $\uparrow$ _{10.19}	59.25 $\uparrow$ _{15.15}	<u>59.83</u> $\uparrow$ _{15.08}	87.74 $\uparrow$ _{10.09}	95.86 $\uparrow$ _{10.76}
AFlow	41.83 $\uparrow$ _{13.15}	50.36 $\uparrow$ _{11.13}	68.18 $\uparrow$ _{10.89}	76.04 $\uparrow$ _{11.46}	65.00 $\uparrow$ _{15.20}	76.53 $\uparrow$ _{13.08}	42.83 $\uparrow$ _{11.27}	42.46 $\uparrow$ _{12.29}	87.92 $\uparrow$ _{10.09}	96.51 $\uparrow$ _{11.41}
GPTSwarm	36.20 $\uparrow$ _{12.48}	47.37 $\uparrow$ _{11.86}	70.45 $\uparrow$ _{13.16}	78.39 $\uparrow$ _{10.89}	69.20 $\uparrow$ _{11.00}	79.90 $\uparrow$ _{10.29}	51.63 $\uparrow$ _{12.47}	52.43 $\uparrow$ _{12.32}	88.17 $\uparrow$ _{10.34}	<u>95.96</u> $\uparrow$ _{10.86}
MAGRPO	40.13 $\uparrow$ _{11.45}	51.74 $\uparrow$ _{12.51}	65.91 $\uparrow$ _{11.38}	76.44 $\uparrow$ _{11.06}	65.00 $\uparrow$ _{15.20}	76.01 $\uparrow$ _{13.60}	49.93 $\uparrow$ _{14.17}	50.48 $\uparrow$ _{14.27}	<u>88.58</u> $\uparrow$ _{10.75}	95.82 $\uparrow$ _{10.72}
MAPoRL	40.59 $\uparrow$ _{11.91}	52.74 $\uparrow$ _{13.51}	63.64 $\uparrow$ _{13.65}	75.34 $\uparrow$ _{12.16}	70.80 $\uparrow$ _{10.60}	80.87 $\uparrow$ _{11.26}	53.25 $\uparrow$ _{10.85}	53.27 $\uparrow$ _{11.48}	86.37 $\uparrow$ _{11.46}	89.19 $\uparrow$ _{15.91}
AT-GRPO	41.54 $\uparrow$ _{12.86}	53.51 $\uparrow$ _{14.28}	64.39 $\uparrow$ _{12.90}	76.86 $\uparrow$ _{10.64}	74.20 $\uparrow$ _{14.00}	83.70 $\uparrow$ _{14.09}	57.91 $\uparrow$ _{13.81}	58.74 $\uparrow$ _{13.99}	86.60 $\uparrow$ _{11.23}	95.85 $\uparrow$ _{10.75}
CL.	42.37 $\uparrow$ _{13.69}	<u>53.43</u> $\uparrow$ _{14.20}	71.97 $\uparrow$ _{14.68}	81.15 $\uparrow$ _{13.65}	78.40 $\uparrow$ _{13.20}	85.25 $\uparrow$ _{15.64}	<u>59.39</u> $\uparrow$ _{15.29}	59.63 $\uparrow$ _{14.88}	88.90 $\uparrow$ _{11.07}	95.90 $\uparrow$ _{10.80}

As shown in Table 1, **collaborative learning achieves the highest average performance across five benchmarks, with best or joint-best results on most metrics.** With a total parameter budget of 13.7B, it outperforms the original Qwen3-14B on all ten metrics, with an average absolute gain of 2.38%, and surpasses SE-RL on eight. It also outperforms the 24B homogeneous MAS without collaborative training on all ten metrics, despite using fewer parameters. Compared with Homo. CL, which trains homogeneous models using the same procedure, the heterogeneous collaborative system achieves higher scores across all ten metrics, with an average absolute gain of 6.02%. Its

³Following the language selection in the Qwen3 technical report (Yang et al., 2025a), we restrict code-generation evaluation to Python, C++, Java, PHP, TypeScript, C#, Bash, and JavaScript.

⁴SE-RL uses Qwen3-14B self-feedback for comparison with cross-model feedback.

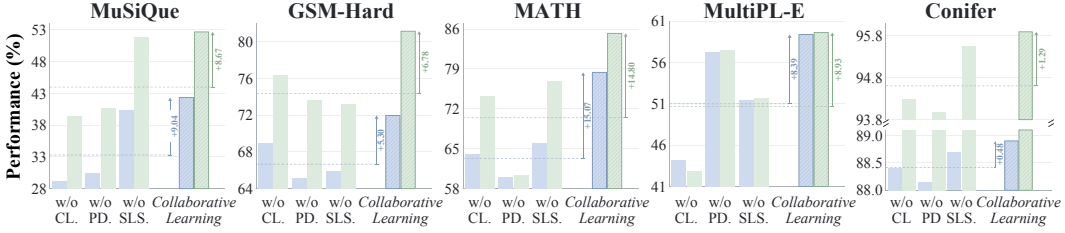


Figure 3: **Ablation study of the mechanism across five benchmarks.** We compare a zero-shot variant that receives no training (w/o CL.), a variant without Protocol Distillation (w/o PD.), and a variant without the step-level score (w/o SLS.). Arrows indicate the gain of *collaborative learning* over the arithmetic mean of the available variants. For each benchmark, ■ denotes Acc., ■ Wtd., and ■ Cov., while ■ denotes the corresponding second metric (F1, Mac., or Expl.).

average absolute gains over established MAS and MARL baselines range from 2.08% to 6.87% across the ten metrics, demonstrating competitive performance against both classes.

Trajectory analysis reveals task-specific characteristics of collaboration and how intermediate behaviors translate into final outcomes. On MuSiQue, subsequent interactions correct initial errors, while on MATH, the trained heterogeneous models better preserve correct intermediate answers than MA-PoRL. On MuSiQue and GSM-Hard, differences in partial-match scores on questions missed by both the collaborative system and the baseline help explain why accuracy and F1 rankings do not fully align. Code-generation performance varies across task subsets: the collaborative system solves more MBPP tasks but fewer HumanEval tasks than SE-RL. On most tasks solved only by SE-RL, it produces no passing intermediate program, suggesting that this gap primarily reflects failure to generate correct solutions rather than the loss of correct candidates during subsequent interactions. On Conifer, the collaborative system achieves the highest Coverage and almost always maintains compliance in the final output when an intermediate answer already satisfies all constraints.

3.3 HOW MUCH DOES EACH MECHANISM CONTRIBUTE TO PERFORMANCE?

As shown in Figure 3, we compare three ablation variants: without collaborative training (w/o CL.), without protocol distillation (w/o PD.), and without the step-level score (w/o SLS.). Averaged over the ten reported metrics, these variants incur absolute score drops of 9.37, 8.85, and 5.41 points, respectively, and each underperforms the full method on every metric. These results indicate that **heterogeneity alone is insufficient to achieve the full performance gains, and that both protocol distillation and step-level scoring contribute meaningfully to learned collaboration.**

Beyond aggregate performance, trajectory-level behavior provides a more direct view of how the two training components shape collaboration throughout collaborative reasoning. For example, without protocol distillation, first-answer EM on MuSiQue increases from 21.02% to 25.24%, yet the number of initially incorrect answers corrected by the final response falls from 594 to 171. Meanwhile, the proportion of Conifer trajectories that initially satisfy all constraints but fail at termination increases, and protocol-following failures on MATH also become more frequent. These observations suggest that protocol distillation primarily supports effective downstream interaction rather than simply improving the quality of the initial response. Similarly, removing the step-level score increases the average MuSiQue trajectory length from 3.4 to 3.8 steps while reducing the number of corrected answers by 18.5%, indicating that step-level feedback helps models use interaction steps more productively and achieve more corrections with fewer interactions on this task.

3.4 HOW EFFICIENT AND EFFECTIVE IS COLLABORATIVE REASONING?

Alongside improvements in task performance, collaborative learning leads to more efficient inference. As shown in Figure 4, the collaborative system requires fewer model invocations than MAD, AgentVerse, AFlow, and GPTSwarm across all five benchmarks, with reductions ranging from 39.1% to 84.8%. Specifically, it requires an average of 4.0, 4.1, 1.8, 3.0, and 3.3 invocations per question on MuSiQue, GSM-Hard, MultiPL-E, MATH, and Conifer, respectively. This efficiency advantage also extends to the evaluated MARL baselines, with invocation reductions ranging from 46.7% to 86.0%. In terms of parameter-weighted output tokens, **the collaborative system achieves**

the lowest count among the compared methods on four of five benchmarks. Compared with the MARL baselines, it further reduces parameter-weighted output tokens by 8.2%–89.4%.

We further compare reasoning-overlap rates before and after collaborative learning to examine redundancy across successive reasoning steps. On MuSiQue, the overlap rates of the three collaborating models decrease from 48.3%, 24.1%, and 55.2% to 0.2%, 0.3%, and 0.5%, respectively. On MultiPL-E, they decrease from 34.8%, 2.3%, and 17.6% to 7.4%, 1.8%, and 10.7%. These results suggest that **collaborative learning substantially reduces repetitive reasoning across model interactions, yielding less redundant subsequent contributions throughout the process.**

To further assess collaborative effectiveness, we measure correct-answer retention, defined as the proportion of trajectories whose final answer remains correct among those containing a correct intermediate candidate. On MuSiQue, GSM-Hard, MultiPL-E, and MATH, the trained heterogeneous models achieve 80.42%, 97.94%, 98.26%, and 98.00% retention, respectively, ranking first on the latter two tasks, second on GSM-Hard, and third on MuSiQue among comparison methods with recoverable intermediate candidates⁵. These results show that **the trained models generally preserve available correct intermediate answers throughout subsequent collaborative interactions.**

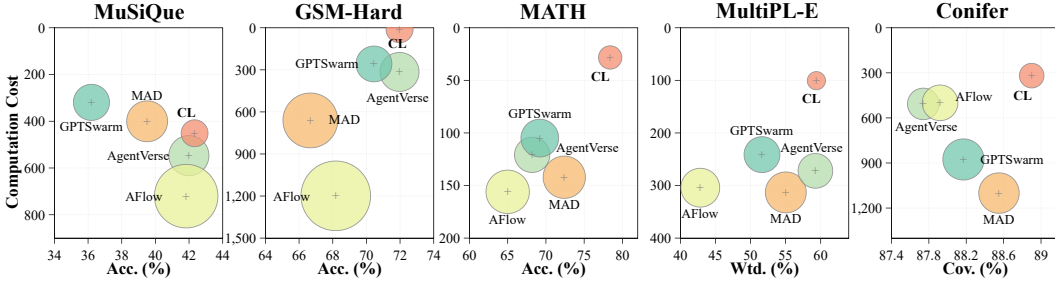


Figure 4: **Computation cost versus task performance.** The center of each circle indicates task performance and parameter-weighted output tokens for the collaborative system and four multi-agent baselines across five benchmarks; its area is proportional to the number of model invocations.

3.5 WHAT COLLABORATION PATTERNS EMERGE FROM COLLABORATIVE LEARNING?

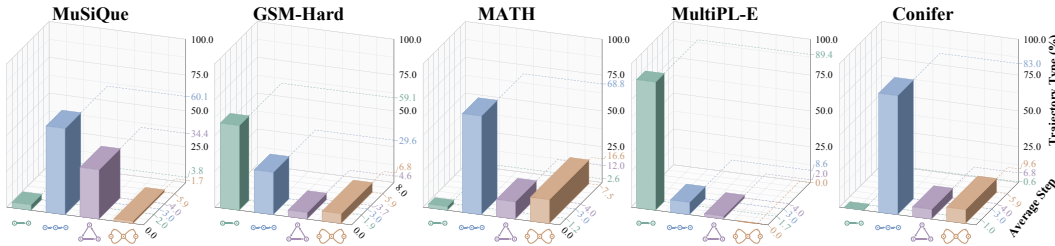


Figure 5: **Post-training interaction topologies across tasks.** $\square$, $\square$, $\square$, and $\square$ denote short linear, long linear, short cyclic, and long cyclic trajectories, respectively. Bar height indicates the proportion of each trajectory type, while depth indicates the average number of interaction steps.

To examine how collaborative learning reorganizes model interactions across tasks, we compare trajectory topologies and interaction lengths before and after training. We categorize trajectories as short linear, long linear, short cyclic, or long cyclic. Linear trajectories involve each model at most once, with at most two interaction steps considered short; cyclic trajectories revisit a previously participating model, with at most four steps considered short. On MuSiQue, the proportion of long linear trajectories increases from 9.18% to 60.12%, showing a shift toward longer linear interactions. In contrast, this proportion decreases from 10.95% to 8.58% on MultiPL-E. Short cyclic trajectories evolve differently across tasks: their average interaction length decreases from 4.00 to 3.67 steps on GSM-Hard, but increases from 3.33 to 4.00 on MATH. These results show that collaborative learning does not uniformly increase interaction complexity or length, but instead induces task-dependent collaboration topologies rather than converging to a single universal interaction pattern.

⁵Conifer is excluded because its open-ended outputs lack an intermediate-candidate correctness criterion.

Task-dependent functional specialization also emerges across models during collaborative learning. On MuSiQue, Qwen3-4B’s answer-revision rate increases from 13% to 69%, its hand-off rate from 48% to 96%, and its termination rate decreases from 50% to 4%, indicating a shift toward intermediate revision and handoff. Meanwhile, Qwen3-8B increasingly provides the final answer, with its share rising from 13.0% to 58.4%. This division of labor changes across tasks: on GSM-Hard, Qwen3-4B instead becomes the more frequent final-answer provider, with its share increasing from 34.09% to 64.39%. Thus, the same model can assume different functions across tasks, revealing task-dependent divisions of labor rather than a fixed collaboration structure.

4 RELATED WORK

Reasoning in LLMs. Existing work improves LLM reasoning by eliciting existing capabilities or strengthening them through training (Zhao et al., 2026a; Li et al., 2026c; Liu et al., 2026a). Inference-time methods use chain-of-thought prompting (Wei et al., 2022), systematic exploration (Hariri et al., 2026; Li et al., 2026a), evaluation (Wu & Raghunathan, 2026), and revision (Yang et al., 2026a); training-based methods learn from bootstrapped demonstrations (Zelikman et al., 2022; Dong et al., 2026), distill teacher reasoning (Yang et al., 2025b; Liu et al., 2026c; Fu et al., 2026), or use outcome- and process-level supervision. Yet these methods remain bounded by a single model’s knowledge and reasoning biases, leaving blind spots unaddressed and allowing intermediate errors to persist through subsequent reasoning (Xie et al., 2026).

From Single-Model Reasoning to Multi-Model Composition. These limitations motivate moving beyond single-model reasoning toward multi-model composition, which exploits complementary knowledge and task-specific strengths across models (Yang et al., 2026b; Zhang et al., 2026c; Huot et al., 2026). Existing approaches select models suited to an input (Ong et al., 2025; Cao et al., 2026; Feng et al., 2026b; Fisch et al., 2026), distribute tasks or subtasks (Agashe et al., 2025; Chen et al., 2026), organize sequential model invocations (Dekoninck et al., 2025; Li et al., 2026b), determine when additional models are needed (Rajib et al., 2026; Qin et al., 2026), and aggregate candidate outputs (Zhang et al., 2026b). However, system-level composition does not necessarily constitute collaborative reasoning in which models jointly develop, evaluate, and revise a shared solution.

Multi-Model Collaborative Reasoning. Collaborative reasoning directly addresses how models jointly develop solutions through interaction. Existing approaches range from hand-designed roles and protocols to optimized workflows (Hu et al., 2025; Zhou et al., 2026; Lee et al., 2026; Koh et al., 2026) and learned collaborative policies (Xue et al., 2026; Zhang et al., 2026a). Representative systems, however, often retain an explicitly prescribed division of labor (Zhao et al., 2026b; Feng et al., 2026a; Ding et al., 2026). In contrast, we train heterogeneous small models to collaborate without predefined roles under a total parameter budget comparable to that of a single large model.

5 CONCLUSION

We introduced a collaborative learning framework for reasoning across heterogeneous models. Model slicing replaces a large model with smaller heterogeneous models under a comparable parameter budget, enabling reasoning-state isolation through task-state handoffs. Pyramid supervision enables cross-model collaboration by combining protocol distillation for interaction initialization with step-level supervision over model-generated trajectories. Across four task categories, the trained heterogeneous models match or outperform the large model and MAS and MARL baselines, while requiring fewer model invocations than multi-agent frameworks. Analysis reveals reduced reasoning redundancy, stronger correct-answer retention, and task-dependent interaction topologies and functional specialization. These results suggest that collaborative learning can transform model heterogeneity into complementary reasoning capabilities across distinct model parameterizations.

Our study covers three model sizes and five benchmarks, leaving generalization to larger model pools, longer interaction horizons, and unseen task distributions unexplored. Pyramid supervision depends on inspector and evaluator quality, whose biases may affect step-level feedback. Comparable parameter counts do not imply equal computational cost, as collaborative inference adds latency and memory overhead beyond our parameter-weighted output-token metric. Future work will study generalization and robust supervision through multiple inspectors or uncertainty-aware aggregation.

REFERENCES

- Ahmed Abdulaal, Chen Jin, Nina Montaña Brown, Aryo Pradipta Gema, Daniel Castro, Daniel Alexander, Philip Teare, Tom Diethe, Dino Oglic, and Amrutha Saseendran. Balancing act: Diversity and consistency in large language model ensembles. In *International Conference on Learning Representations (ICLR)*, 2025.
https://proceedings.iclr.cc/paper_files/paper/2025/hash/a5554e55d7a21e62d0d7a028ec0ealc7-Abstract-Conference.html.
- Saaket Agashe, Kyle Wong, Vincent Tu, Jiachen Yang, Ang Li, and Xin Eric Wang. Agent S2: A compositional generalist-specialist framework for computer use agents. In *Conference on Language Modeling (COLM)*, 2025.
<https://openreview.net/forum?id=zg5is4GJ3R>.
- Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michal Podstawski, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Hubert Niewiadomski, Piotr Nyczyk, and Torsten Hoefler. Graph of thoughts: Solving elaborate problems with large language models. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2024.
<https://ojs.aaai.org/index.php/AAAI/article/view/29720>.
- Qi Cao, Shuhao Zhang, Ruizhe Zhou, Ruiyi Zhang, Peijia Qin, and Pengtao Xie. Models under SCOPE: Scalable and controllable routing via pre-hoc reasoning. In *International Conference on Machine Learning (ICML)*, 2026. <https://arxiv.org/abs/2601.22323>.
- Federico Cassano, John Gouwar, Daniel Nguyen, Sydney Nguyen, Luna Phipps-Costin, Donald Pinckney, Ming-Ho Yee, Yangtian Zi, Carolyn Jane Anderson, Molly Q Feldman, Arjun Guha, Michael Greenberg, and Abhinav Jangda. MultiPL-E: A scalable and polyglot approach to benchmarking neural code generation. In *IEEE Transactions on Software Engineering (TSE)*, 2023. <https://doi.org/10.1109/TSE.2023.3267446>.
- Baian Chen, Chang Shu, Ehsan Shareghi, Nigel Collier, Karthik Narasimhan, and Shunyu Yao. FireAct: Toward language agent fine-tuning. In *arXiv preprint arXiv:2310.05915*, 2023.
<https://arxiv.org/abs/2310.05915>.
- Guanzhong Chen, Shaoxiong Yang, Chao Li, Wei Liu, Jian Luan, and Zenglin Xu. End-to-end optimization of LLM-driven multi-agent search systems via heterogeneous-group-based reinforcement learning. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2026. <https://aclanthology.org/2026.acl-long.1399/>.
- Justin Chen, Swarnadeep Saha, and Mohit Bansal. ReConcile: Round-table conference improves reasoning via consensus among diverse LLMs. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024a.
<https://aclanthology.org/2024.acl-long.381/>.
- Justin Chen, Swarnadeep Saha, Elias Stengel-Eskin, and Mohit Bansal. MAGDi: Structured distillation of multi-agent interaction graphs improves reasoning in smaller language models. In *International Conference on Machine Learning (ICML)*, 2024b.
<https://proceedings.mlr.press/v235/chen24ah.html>.
- Weize Chen, Yusheng Su, Jingwei Zuo, Cheng Yang, Chenfei Yuan, Chi-Min Chan, Heyang Yu, Yaxi Lu, Yi-Hsin Hung, Chen Qian, Yujia Qin, Xin Cong, Ruobing Xie, Zhiyuan Liu, Maosong Sun, and Jie Zhou. AgentVerse: Facilitating multi-agent collaboration and exploring emergent behaviors. In *International Conference on Learning Representations (ICLR)*, 2024c.
https://proceedings.iclr.cc/paper_files/paper/2024/hash/578e65cdee35d00c708d4c64bce32971-Abstract-Conference.html.
- Weize Chen, Ziming You, Ran Li, Yitong Guan, Chen Qian, Chenyang Zhao, Cheng Yang, Ruobing Xie, Zhiyuan Liu, and Maosong Sun. Internet of agents: Weaving a web of heterogeneous agents for collaborative intelligence. In *International Conference on Learning Representations (ICLR)*, 2025. <https://openreview.net/forum?id=olEt3MogPw>.

- Ganqu Cui, Lifan Yuan, Zefan Wang, Hanbin Wang, Yuchen Zhang, Jiacheng Chen, Wendi Li, Bingxiang He, Yuchen Fan, Tianyu Yu, Qixin Xu, Weize Chen, Jiarui Yuan, Huayu Chen, Kaiyan Zhang, Xingtai Lv, Shuo Wang, Yuan Yao, Xu Han, Hao Peng, Yu Cheng, Zhiyuan Liu, Maosong Sun, Bowen Zhou, and Ning Ding. Process reinforcement through implicit rewards. In *Transactions on Machine Learning Research (TMLR)*, 2026. <https://openreview.net/forum?id=9SkkifLopZ>.
- Jasper Dekoninck, Maximilian Baader, and Martin Vechev. A unified approach to routing and cascading for LLMs. In *International Conference on Machine Learning (ICML)*, 2025. <https://proceedings.mlr.press/v267/dekoninck25a.html>.
- Jie Ding, Rui Sun, Xinyuan Zhang, Zeyu Zhang, and Xin Liu. APEx: Distillation of agent procedural experience for adaptive deep research question answering. In *arXiv preprint arXiv:2609.02253*, 2026. <https://arxiv.org/abs/2609.02253>.
- Leichao Dong, Dongxu Zhang, Yiding Sun, Qirui Wang, Yuhan Wang, Lin Chen, and Jihua Zhu. Better starts, better ends: Bootstrapped iterative self-reasoning distillation for compressed reasoning. In *arXiv preprint arXiv:2607.15736*, 2026. <https://arxiv.org/abs/2607.15736>.
- Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. In *International Conference on Machine Learning (ICML)*, 2024. <https://proceedings.mlr.press/v235/du24e.html>.
- Lang Feng, Longtao Zheng, Shuo He, Fuxiang Zhang, and Bo An. Dr. MAS: Stable reinforcement learning for multi-agent LLM systems. In *arXiv preprint arXiv:2602.08847*, 2026a. <https://arxiv.org/abs/2602.08847>.
- Tao Feng, Fangxu Yu, Haozhen Zhang, Zhongjie Dai, Liangqi Yuan, Zijie Lei, Weizhi Zhang, Kunlun Zhu, Haodong Yue, Keyang Xuan, Ge Liu, and Jiaxuan You. LLMRouter: Unified infrastructure for developing, evaluating, and deploying LLM routers. In *arXiv preprint arXiv:2608.06867*, 2026b. <https://arxiv.org/abs/2608.06867>.
- Adam Fisch, Shubhendu Trivedi, Fantine Huot, William W. Cohen, Michael Kaisers, Mirella Lapata, Kate Larson, and Jacob Eisenstein. Pandora’s AI model routing box: Efficient allocation with costly value estimation. In *arXiv preprint arXiv:2608.20316*, 2026. <https://arxiv.org/abs/2608.20316>.
- Jakob Foerster, Gregory Farquhar, Triantafyllos Afouras, Nantas Nardelli, and Shimon Whiteson. Counterfactual multi-agent policy gradients. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2018. <https://ojs.aaai.org/index.php/AAAI/article/view/11794>.
- Zixuan Fu, Bingxiang He, Yuxin Zuo, Haohuan Huang, Jinqian Zhang, Ruhang Xiao, Cheng Qian, Qinyu Luo, Huan-ang Gao, Yudong Wang, Zhiyuan Liu, Ning Ding, and Chaojun Xiao. Rethinking on-policy distillation of large language models II: One training example. In *arXiv preprint arXiv:2609.04172*, 2026. <https://arxiv.org/abs/2609.04172>.
- Luyu Gao, Aman Madaan, Shuyan Zhou, Uri Alon, Pengfei Liu, Yiming Yang, Jamie Callan, and Graham Neubig. PAL: Program-aided language models. In *International Conference on Machine Learning (ICML)*, 2023. <https://proceedings.mlr.press/v202/gao23f.html>.
- Mohsen Hariri, Weicong Chen, Nahal Shahini, Vikash Singh, Kai Ye, Amirhossein Samandar, Debargha Ganguly, Sreehari Sankar, Yanyan Zhang, Shouren Wang, Jerry Peng, Biyao Zhang, Michael Hinczewski, and Vipin Chaudhary. Test-time scaling in reasoning LLMs: Inference regimes, evaluation, and reproducibility. In *arXiv preprint arXiv:2608.04001*, 2026. <https://arxiv.org/abs/2608.04001>.
- 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 *Conference on Neural Information Processing Systems, Datasets and Benchmarks Track*

- (*NeurIPS*), 2021. https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/hash/be83ab3ecd0db773eb2dc1b0a17836a1-Abstract-round2.html.
- Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022. <https://openreview.net/forum?id=nZeVKeeFYf9>.
- Shengran Hu, Cong Lu, and Jeff Clune. Automated design of agentic systems. In *International Conference on Learning Representations (ICLR)*, 2025. https://proceedings.iclr.cc/paper_files/paper/2025/hash/36b7acf6f6010652b3f2a433774a66fe-Abstract-Conference.html.
- Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Yu, Xinying Song, and Denny Zhou. Large language models cannot self-correct reasoning yet. In *International Conference on Learning Representations (ICLR)*, 2024. https://proceedings.iclr.cc/paper_files/paper/2024/hash/8b4add8b0aa8749d80a34ca5d941c355-Abstract-Conference.html.
- Fantine Huot, Michael Kaisers, and Mirella Lapata. When is routing meaningful? diversity and robustness in language model societies. In *arXiv preprint arXiv:2607.09197*, 2026. <https://arxiv.org/abs/2607.09197>.
- Yoon Kim and Alexander M. Rush. Sequence-level knowledge distillation. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2016. <https://aclanthology.org/D16-1139/>.
- Donghyeok Koh, Gyuwan Kim, Jinyeong Bak, Seung-Hoon Na, Tao Yang, Haneol Jang, and Cheoneum Park. Global optimization and inference-time region grafting for agentic workflows. In *arXiv preprint arXiv:2608.02353*, 2026. <https://arxiv.org/abs/2608.02353>.
- Aviral Kumar, Vincent Zhuang, Rishabh Agarwal, Yi Su, John D. Co-Reyes, Avi Singh, Kate Baumli, Shariq Iqbal, Colton Bishop, Rebecca Roelofs, Lei Zhang, Kay McKinney, Disha Shrivastava, Cosmin Paduraru, George Tucker, Doina Precup, Feryal Behbahani, and Aleksandra Faust. Training language models to self-correct via reinforcement learning. In *International Conference on Learning Representations (ICLR)*, 2025. https://proceedings.iclr.cc/paper_files/paper/2025/hash/871ac99fdc5282d0301934d23945ebaa-Abstract-Conference.html.
- Jung Hwan Lee, Kyu Ho Lee, and Gwang Hoon Yoo. MEGA: Self-evolving agent optimization infrastructure via wisdom graph. In *arXiv preprint arXiv:2608.10504*, 2026. <https://arxiv.org/abs/2608.10504>.
- Muheng Li, Jian Qian, and Wenlong Mou. Predicting and improving test-time scaling laws via reward tail-guided search. In *arXiv preprint arXiv:2602.01485*, 2026a. <https://arxiv.org/abs/2602.01485>.
- Songyuan Li, Ahmed M. Abdelmoniem, and Shiqiang Wang. ProgRouter: Online progress-guided orchestration for multi-agent LLM workflows under quality-cost tradeoffs. In *Findings of the Association for Computational Linguistics: Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2026b. <https://arxiv.org/abs/2608.25992>.
- Yang Li, Semih Yavuz, and Shafiq Joty. RISE: Recursive improvement via self-extrapolating policy distillation. In *arXiv preprint arXiv:2609.05295*, 2026c. <https://arxiv.org/abs/2609.05295>.
- Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. Encouraging divergent thinking in large language models through multi-agent debate. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2024. <https://aclanthology.org/2024.emnlp-main.992/>.

- Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison 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 (ICLR)*, 2024. https://proceedings.iclr.cc/paper_files/paper/2024/hash/aca97732e30bcf1303bc22ac3924fd16-Abstract-Conference.html.
- Bo Liu, Simon Yu, Yiding Jiang, Ao Qu, Andrew Zhao, Zichen Liu, Junsu Kim, Zijian Zhou, Seungone Kim, Tongzheng Ren, Mickel Liu, Hanfei Yu, Zhaorun Chen, Weiyan Shi, Paul Pu Liang, Luke Zettlemoyer, Yejin Choi, and Natasha Jaques. SPADE: Self-play in adaptive synthetic executable environments. In *arXiv preprint arXiv:2608.19197*, 2026a. <https://arxiv.org/abs/2608.19197>.
- Shuo Liu, Zeyu Liang, Xueguang Lyu, and Christopher Amato. LLM collaboration with multi-agent reinforcement learning. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2026b. <https://ojs.aaai.org/index.php/AAAI/article/view/40487>.
- Zhishuai Liu, Xingzi Xu, Mehmet Saygin Seyfioglu, Pan Xu, and Karim Bouyarmane. Extremely sparse supervision incentivizes reasoning ability. In *arXiv preprint arXiv:2609.04565*, 2026c. <https://arxiv.org/abs/2609.04565>.
- 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 *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2025. <https://aclanthology.org/2025.emnlp-main.1025/>.
- Isaac Ong, Amjad Almahairi, Vincent Wu, Wei-Lin Chiang, Tianhao Wu, Joseph E. Gonzalez, M. Kadous, and Ion Stoica. RouteLLM: Learning to route LLMs from preference data. In *International Conference on Learning Representations (ICLR)*, 2025. https://proceedings.iclr.cc/paper_files/paper/2025/hash/5503a7c69d48a2f86fc00b3dc09de686-Abstract-Conference.html.
- Chanwoo Park, Seungju Han, Xingzhi Guo, Asuman E. Ozdaglar, Kaiqing Zhang, and Joo-Kyung Kim. MAPoRL: Multi-agent post-co-training for collaborative large language models with reinforcement learning. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2025. <https://aclanthology.org/2025.acl-long.1459/>.
- Peijia Qin, Qi Cao, and Pengtao Xie. ATLAS: Agentic test-time learning-to-allocate scaling. In *arXiv preprint arXiv:2606.01667*, 2026. <https://arxiv.org/abs/2606.01667>.
- Rakibul Hasan Rajib, Mengxin Zheng, and Qian Lou. Learning what to retain: Gated-memory routing for efficient collaboration in multi-agent LLM systems. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2026. <https://arxiv.org/abs/2609.00237>.
- Amrith Setlur, Chirag Nagpal, Adam Fisch, Xinyang Geng, Jacob Eisenstein, Rishabh Agarwal, Alekh Agarwal, Jonathan Berant, and Aviral Kumar. Rewarding progress: Scaling automated process verifiers for LLM reasoning. In *International Conference on Learning Representations (ICLR)*, 2025. https://proceedings.iclr.cc/paper_files/paper/2025/hash/98711dea460bdefe0e651ca23ec98ba2-Abstract-Conference.html.
- Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling LLM test-time compute optimally can be more effective than scaling parameters for reasoning. In *International Conference on Learning Representations (ICLR)*, 2025. https://proceedings.iclr.cc/paper_files/paper/2025/hash/1b623663fd9b874366f3ce019fdffd44-Abstract-Conference.html.
- Haoran Sun, Lixin Liu, Junjie Li, Fengyu Wang, Baohua Dong, Ran Lin, and Ruohui Huang. Conifer: Improving complex constrained instruction-following ability of large language models. In *arXiv preprint arXiv:2404.02823*, 2024. <https://arxiv.org/abs/2404.02823>.
- Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. MuSiQue: Multihop questions via single-hop question composition. In *Transactions of the Association for Computational Linguistics (TACL)*, 2022. <https://aclanthology.org/2022.tacl-1.31/>.

- Junlin Wang, Jue Wang, Ben Athiwaratkun, Ce Zhang, and James Y. Zou. Mixture-of-agents enhances large language model capabilities. In *International Conference on Learning Representations (ICLR)*, 2025.
https://proceedings.iclr.cc/paper_files/paper/2025/hash/5434be94e82c54327bb9dcdf7fca52b6-Abstract-Conference.html.
- Peiyi Wang, Lei Li, Zhihong Shao, Runxin Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. Math-Shepherd: Verify and reinforce LLMs step-by-step without human annotations. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024a.
<https://aclanthology.org/2024.acl-long.510/>.
- Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyang Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhui Chen. MMLU-Pro: A more robust and challenging multi-task language understanding benchmark. In *Conference on Neural Information Processing Systems, Datasets and Benchmarks Track (NeurIPS)*, 2024b.
https://proceedings.neurips.cc/paper_files/paper/2024/hash/ad236edc564f3e3156e1b2feafb99a24-Abstract-Datasets_and_Benchmarks_Track.html.
- 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 *Conference on Neural Information Processing Systems (NeurIPS)*, 2022.
https://proceedings.neurips.cc/paper_files/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html.
- Chen Henry Wu and Aditi Raghunathan. Self-trained verification for training- and test-time self-improvement. In *arXiv preprint arXiv:2605.30290*, 2026.
<https://arxiv.org/abs/2605.30290>.
- Yangzhen Wu, Zhiqing Sun, Shanda Li, Sean Welleck, and Yiming Yang. Inference scaling laws: An empirical analysis of compute-optimal inference for LLM problem-solving. In *International Conference on Learning Representations (ICLR)*, 2025.
https://proceedings.iclr.cc/paper_files/paper/2025/hash/8c3caae2f725c8e2a55ecd600563d172-Abstract-Conference.html.
- Si'an Xie, Jiaxun Liu, Biao Yang, Wei Yuan, Fan Yang, Tingting Gao, and Ming Wu. From reasoning depth to reasoning breadth: Evaluating multi-point associative reasoning in large language models. In *arXiv preprint arXiv:2608.10444*, 2026.
<https://arxiv.org/abs/2608.10444>.
- Wenda Xu, Guanglei Zhu, Xuandong Zhao, Liangming Pan, Lei Li, and William Wang. Pride and prejudice: LLM amplifies self-bias in self-refinement. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024.
<https://aclanthology.org/2024.acl-long.826/>.
- Xiangyuan Xue, Yifan Zhou, Guibin Zhang, Zaibin Zhang, Yijiang Li, Chen Zhang, Zhenfei Yin, Philip Torr, Wanli Ouyang, and Lei Bai. CoMAS: Co-evolving multi-agent systems via interaction rewards. In *International Conference on Learning Representations (ICLR)*, 2026.
<https://openreview.net/forum?id=ihwAzktmWc>.
- Aijun Yang, Qianxue Guo, Ziyi Huang, Yuxuan Chen, Shiyu Qian, and Jian Cao. Self-correcting long-horizon search agents via tree-structured memory. In *arXiv preprint arXiv:2608.10676*, 2026a. <https://arxiv.org/abs/2608.10676>.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger

- Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report. In *arXiv preprint arXiv:2505.09388*, 2025a. <https://arxiv.org/abs/2505.09388>.
- Ling Yang, Zhaochen Yu, Tianjun Zhang, Minkai Xu, Joseph E. Gonzalez, Bin Cui, and Shuicheng Yan. SuperCorrect: Advancing small LLM reasoning with thought template distillation and self-correction. In *International Conference on Learning Representations (ICLR)*, 2025b. https://proceedings.iclr.cc/paper_files/paper/2025/hash/0967d7c8b171dd81b77c43067c02bebf-Abstract-Conference.html.
- Yunhao Yang, Yuexin Bian, Yunjie Tian, Di Fu, Tianjin Huang, Yuanyuan Shi, Ziang Xiao, Nuno Vasconcelos, and Yijiang Li. Co-RL: Unsupervised reasoning emerges from diverse cohort in multi-agent RL. In *arXiv preprint arXiv:2608.17253*, 2026b. <https://arxiv.org/abs/2608.17253>.
- Zhangyue Yin, Qiushi Sun, Cheng Chang, Qipeng Guo, Junqi Dai, Xuanjing Huang, and Xipeng Qiu. Exchange-of-thought: Enhancing large language model capabilities through cross-model communication. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023. <https://aclanthology.org/2023.emnlp-main.936/>.
- Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah Goodman. STaR: Bootstrapping reasoning with reasoning. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2022. https://proceedings.neurips.cc/paper_files/paper/2022/hash/639a9a172c044fbb64175b5fad42e9a5-Abstract-Conference.html.
- Aohan Zeng, Mingdao Liu, Rui Lu, Bowen Wang, Xiao Liu, Yuxiao Dong, and Jie Tang. AgentTuning: Enabling generalized agent abilities for LLMs. In *Findings of the Association for Computational Linguistics: Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024. <https://aclanthology.org/2024.findings-acl.181/>.
- Dan Zhang, Sining Zhoubian, Ziniu Hu, Yisong Yue, Yuxiao Dong, and Jie Tang. ReST-MCTS*: LLM self-training via process reward guided tree search. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2024. https://proceedings.neurips.cc/paper_files/paper/2024/hash/76ec4dc30e9faaf0e4b6093eaa377218-Abstract-Conference.html.
- Jiayi Zhang, Jinyu Xiang, Zhaoyang Yu, Fengwei Teng, Xionghui Chen, Jiaqi Chen, Mingchen Zhuge, Xin Cheng, Sirui Hong, Jinlin Wang, Bingnan Zheng, Bang Liu, Yuyu Luo, and Chenglin Wu. AFlow: Automating agentic workflow generation. In *International Conference on Learning Representations (ICLR)*, 2025. https://proceedings.iclr.cc/paper_files/paper/2025/hash/5492ecbce4439401798dcd2c90be94cd-Abstract-Conference.html.
- Yaolun Zhang, Yujie Zhao, Nan Wang, Yiran Wu, Jiayu Chang, Yizhao Chen, Qingyun Wu, Jishen Zhao, and Huazheng Wang. MetaAgent-X: Breaking the ceiling of automatic multi-agent systems via end-to-end reinforcement learning. In *arXiv preprint arXiv:2605.14212*, 2026a. <https://arxiv.org/abs/2605.14212>.
- Zheyuan Zhang, Kaiwen Shi, Zhengqing Yuan, Zehong Wang, Tianyi Ma, Keerthiram Murugesan, Vincent Galassi, Chuxu Zhang, and Yanfang Ye. AgentRouter: A knowledge-graph-guided LLM router for collaborative multi-agent question answering. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2026b. <https://aclanthology.org/2026.acl-long.33/>.
- Zhixia Zhang, Zixuan Huang, Gongxun Li, Huaiyang Wang, Chengyi Yuan, Xin Xia, Deqing Wang, Fuzhen Zhuang, Shuai Ma, Ning Ding, Yaodong Yang, Jianxin Li, and Yikun Ban. Heterogeneous agent collaborative reinforcement learning. In *arXiv preprint arXiv:2603.02604*, 2026c. <https://arxiv.org/abs/2603.02604>.
- Siyan Zhao, Zhihui Xie, Mengchen Liu, Jing Huang, Guan Pang, Feiyu Chen, and Aditya Grover. Self-distilled reasoner: On-policy self-distillation for large language models. In *International Conference on Machine Learning (ICML)*, 2026a. <https://openreview.net/forum?id=Jpxfof0EaS>.

- Yujie Zhao, Lanxiang Hu, Yang Wang, Minmin Hou, Hao Zhang, Ke Ding, and Jishen Zhao. Stronger-MAS: Multi-agent reinforcement learning for collaborative LLMs. In *International Conference on Learning Representations (ICLR)*, 2026b.
<https://openreview.net/forum?id=IdF6JqXWzx>.
- Chujie Zheng, Zhenru Zhang, Beichen Zhang, Runji Lin, Keming Lu, Bowen Yu, Dayiheng Liu, Jingren Zhou, and Junyang Lin. ProcessBench: Identifying process errors in mathematical reasoning. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2025.
<https://aclanthology.org/2025.acl-long.50/>.
- Han Zhou, Xingchen Wan, Ruoxi Sun, Hamid Palangi, Shariq Iqbal, Ivan Vulić, Anna Korhonen, and Sercan Arik. Multi-agent design: Optimizing agents with better prompts and topologies. In *International Conference on Learning Representations (ICLR)*, 2026.
<https://iclr.cc/virtual/2026/poster/10010348>.
- Pei Zhou, Jay Pujara, Xiang Ren, Xinyun Chen, Heng-Tze Cheng, Quoc V. Le, Ed H. Chi, Denny Zhou, Swaroop Mishra, and Huaixiu Steven Zheng. SELF-DISCOVER: Large language models self-compose reasoning structures. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2024.
https://proceedings.neurips.cc/paper_files/paper/2024/hash/e41efb03e20ca3c231940a3c6917ef6f-Abstract-Conference.html.
- Mingchen Zhuge, Wenyi Wang, Louis Kirsch, Francesco Faccio, Dmitrii Khizbullin, and Jürgen Schmidhuber. GPTSwarm: Language agents as optimizable graphs. In *International Conference on Machine Learning (ICML)*, 2024.
<https://proceedings.mlr.press/v235/zhuge24a.html>.