NEITHER SYCOPHANTIC NOR STUBBORN: PRESERVING FEEDBACK DISCERNMENT IN MERGED REASONING MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Model merging offers a practical approach to reducing reasoning length while preserving answer accuracy. However, accuracy and output length alone do not capture how merged models respond to external feedback. Across diverse mathematical reasoning benchmarks, we find that some evaluated merging configurations improve accuracy and reduce output length yet degrade feedback discernment, the ability to retain correct answers under misleading advice and correct erroneous answers under valid feedback. This degradation is not inevitable, as other configurations preserve this capability. Motivated by these findings, we formulate discernment-aware merging as a constrained optimization problem that maximizes feedback discernment subject to accuracy and token-cost constraints. We propose **Dis**cernment-guided **C**onditional Fl**o**w Merge (DISCO-Merge), an active optimization framework that couples conditional flow matching with Bayesian optimization. Conditioned on a parent configuration and its behavioral profile, a residual flow learns a conditional distribution of updates to module-wise merging coefficients from evaluated neighboring model pairs. Gaussian-process surrogates estimate candidate performance and uncertainty, and an acquisition function selects candidates for expensive model evaluations. We select the model with the highest discernment among confirmed candidates that meet the token-cost and accuracy constraints. Experiments on eight mathematical benchmarks with Qwen2.5, Qwen3, and Phi-4-mini show that DISCO-Merge surpasses all evaluated merging baselines in macro-averaged discernment and accuracy–efficiency–discernment hypervolume (HV).

1 INTRODUCTION

Model merging combines the parameters of compatible models into a single deployable checkpoint, reusing specialized capabilities without retaining multiple experts or jointly retraining on their original data. Weight averaging, task-vector composition, and interference-aware methods enable such composition while preserving single-model inference (Wortsman et al., 2022; Ilharco et al., 2023; Yadav et al., 2023; Yu et al., 2024).

Long-to-short (L2S) reasoning is an emerging application. Long reasoning traces can improve problem solving, but repeated solutions and verification may add cost with limited benefit (Chen et al., 2025). Prompting and training-based approaches seek to retain accuracy with shorter outputs (Xu et al., 2025; Xia et al., 2025; Luo et al., 2026; Hou et al., 2026). L2S merging offers a complementary post-hoc route by combining short- and long-reasoning checkpoints without additional language-model fine-tuning (Wu et al., 2025; Lan et al., 2025; Wu et al., 2026). Existing approaches seek merging configurations that preserve answer accuracy while generating fewer tokens (Yao et al., 2025; Zhu et al., 2026; Iacobelli et al., 2026).

Are single-turn accuracy (Acc) and output length (Tok) sufficient to assess a merged reasoning model? In multi-step mathematical reasoning, a user may correctly identify a sign error in an intermediate equation or mistakenly challenge a valid transformation. Accepting the mistaken feedback can derail a correct derivation, while rejecting the valid correction can leave an error that propagates to the final answer. This calls for both *Retention* (Ret), preserving correct answers under misleading

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

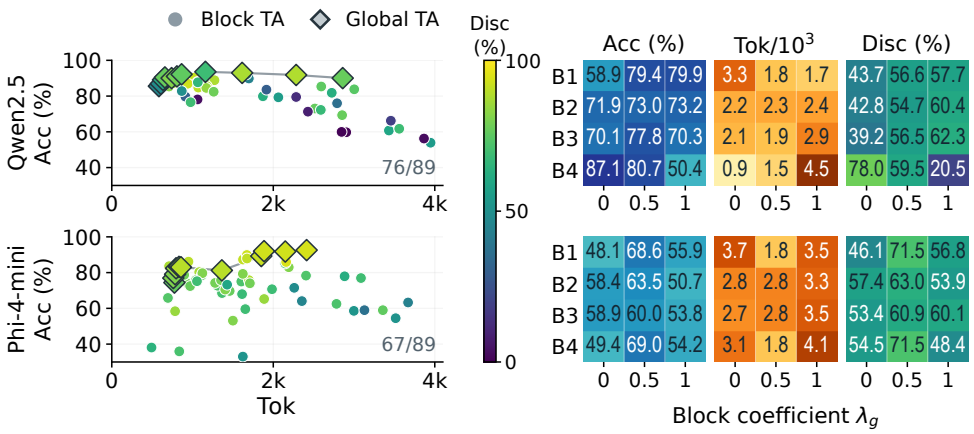

Figure 1: **Illustration of motivation.** We grid-scan global TA ($\lambda = 0, 0.1, \ldots, 1$) and four-block TA ($\lambda_g \in \{0, 0.5, 1\}$), evaluating 89 distinct configurations per family on 512 MATH/GSM8K diagnostic questions. **Left:** Acc–Tok distributions in the common condition (Tok ≤ 4000), colored by Disc. **Right:** Marginal effects of λ_g, averaging 27 settings of the other coefficients per cell. *Accuracy and token cost alone do not fully characterize feedback behavior; merged models must also be evaluated for discernment.*

feedback, and *Correction* (Cor), correcting errors under valid feedback (Sharma et al., 2024; Laban et al., 2023; Wang et al., 2025; Sinha, 2026). We summarize these complementary abilities by their balanced mean, *Discernment* (Disc).

Figure 1 illustrates the resulting mismatch: configurations with similar accuracy and response length in the common condition can exhibit different feedback outcomes, and block-wise coefficient changes do not affect the three objectives uniformly. This motivates *discernment-aware merging*: explicitly considering feedback behavior when evaluating and selecting an accurate, efficient merged checkpoint. Appendix B provides the scan setup and per-configuration results.

Searching for such merges requires costly model evaluations, motivating us to reuse previous evaluations to guide candidate generation. A single parent model may admit several useful update directions, each involving coordinated changes across module-wise coefficients. We therefore use conditional flow matching (Lipman et al., 2023) to learn flexible distributions of local coefficient updates from evaluated configurations, conditioned on the parent's configuration and behavioral profile. This provides diverse proposals, while Bayesian selection determines which candidates to evaluate under a fixed budget.

We propose **Dis**cernment-guided **C**onditional Fl**o**w Merge (DISCO-Merge), which couples these conditional proposals with Bayesian evaluation. A persistent archive supplies residuals between evaluated neighbors for velocity-regression training, conditioned on each parent's configuration and behavioral profile. Gaussian-process surrogates fitted to the same archive estimate candidate performance and uncertainty, and an acquisition function allocates evaluations under a fixed budget.

Our main contributions are summarized as follows:

- We show that configurations with similar accuracy and output length can differ substantially in feedback discernment, motivating a constrained merging objective that maximizes discernment while meeting accuracy and token-cost requirements.
- We introduce DISCO-Merge, combining conditional residual flows trained on archived behavioral comparisons with cost-aware Bayesian evaluation to generate diverse coefficient updates and allocate costly model evaluations.
- Experiments on Qwen2.5, Qwen3, and Phi-4-mini across eight mathematical benchmarks demonstrate higher macro-averaged discernment and hypervolume than all evaluated merging baselines. Studies across intermediate models and three-seed ablations further examine the framework's effectiveness and component contributions.

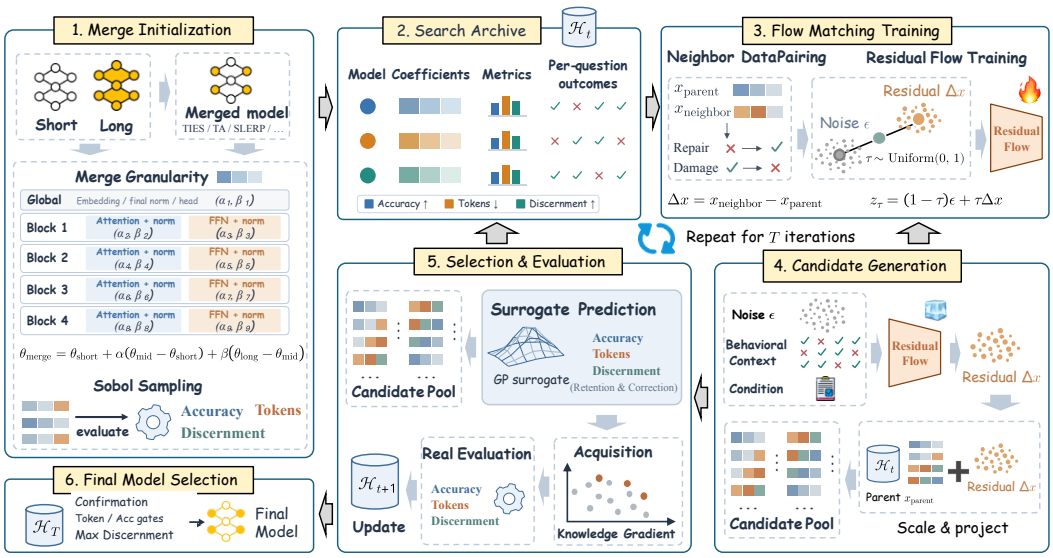

Figure 2: *Overview of DISCO-Merge.* Initial evaluations form a persistent archive. Conditional Flow learns coefficient updates from archived behavioral contrasts; Gaussian processes and cost-aware selection allocate evaluations among proposals. Measured outcomes update the archive. After search, confirmation selects one merged model for high discernment under accuracy and token constraints.

2 Related Work

Model Merging for Long-to-Short Reasoning. Beyond uniform weight combination, model merging uses entropy-based coefficient adaptation, activation statistics, and parameter sensitivity to preserve specialized capabilities (Yang et al., 2024b; Heyrani Nobari et al., 2025; Liu et al., 2025). Long-to-short (L2S) merging applies parameter composition to concise and deliberative models to balance answer accuracy and generation cost (Wu et al., 2025). The Thinking Spectrum and Revisiting Model Interpolation characterize how merging strength changes reasoning behavior and accuracy–efficiency trade-offs (Lan et al., 2025; Wu et al., 2026). Layer-adaptive approaches use activation information in ACM (Yao et al., 2025), outlier statistics and input-dependent adjustment in ORCA (Zhu et al., 2026), or random-token Fisher estimates in FIM-Merging (Xia, 2026). QA-Merging calibrates layers with divergent reasoning patterns for query-adaptive reasoning (Zhong et al., 2026), whereas Ada-R1 follows long/short model merging with preference training (Luo et al., 2025). Evo-L2S directly optimizes the accuracy–length trade-off (Iacobelli et al., 2026). Prior L2S work also examines self-critique and self-correction (Wu et al., 2025).

Feedback Discernment in Language Models. Iterative self-feedback is explored in Self-Refine (Madaan et al., 2023), while intrinsic self-correction studies identify limitations without external evidence (Huang et al., 2024). Sycophancy research documents agreement over correctness (Sharma et al., 2024), and the FlipFlop experiment shows that a challenge without corrective evidence can induce harmful answer changes (Laban et al., 2023). SycEval distinguishes beneficial and harmful changes (Fanous et al., 2025), and SYCON Bench examines sustained conversational pressure (Hong et al., 2025). Feedback Indices disentangle sycophancy and stubbornness through controlled histories (Dunlap et al., 2026), while SycoBench-600 pairs misleading pressure with correct suggestions (Sinha, 2026). The Critic-Discernment Game trains both behaviors through self-play (Wang et al., 2025).

Discussion. Existing research has advanced accurate and efficient model merging and clarified how language models respond to feedback, yet accuracy and token cost alone do not capture whether merged models resist misleading advice while benefiting from valid corrections. This motivates studying discernment as an additional criterion for evaluating and selecting merged models.

3 METHODOLOGY

3.1 PROBLEM FORMULATION

Merge space and granularity. We optimize combination coefficients for compatible, fixed Short and Long models, θ_{short} and θ_{long}, optionally using an intermediate merge θ_{mid}.

We divide Transformer layers into depth blocks and assign separate coefficients to attention and feed-forward parameters within each block, including their associated normalization parameters. Embeddings, the final normalization, and the output head form a global group. For each group g, coefficients $\alpha_g, \beta_g \in [0, 1]$ define

$$\theta_{\text{merge},g}(x) = \theta_{\text{short},g} + \alpha_g(\theta_{\text{mid},g} - \theta_{\text{short},g}) + \beta_g(\theta_{\text{long},g} - \theta_{\text{mid},g}). \tag{1}$$

With G groups, $x = (\alpha_1, \beta_1, \ldots, \alpha_G, \beta_G) \in \mathcal{X} = [0, 1]^{2G}$ specifies the full merged model, denoted $\theta(x)$. Uniform pairs $(0, 0)$, $(1, 0)$, and $(1, 1)$ recover Short, the intermediate model, and Long, respectively.

An intermediate model is optional. Without one, we directly combine Short and Long with one coefficient per group, yielding a G-dimensional search space. Appendix A.5 gives this formulation and the grouping details.

Evaluation objectives. We measure *accuracy* $\text{Acc}(x)$ and mean output *tokens* $\text{Tok}(x)$ in the common condition. To evaluate feedback behavior, each configuration receives the same fixed histories: a correct history followed by misleading advice, or an incorrect history followed by valid advice. Correctness in these two conditions defines *retention* $\text{Ret}(x)$ and *correction* $\text{Cor}(x)$, respectively. We define feedback discernment as their balanced mean:

$$\text{Disc}(x) = \tfrac{1}{2}\big(\text{Ret}(x) + \text{Cor}(x)\big). \tag{2}$$

Benchmark comparisons use equal-weight dataset means; correctness scores are percentages. High accuracy, low token cost, and high discernment define the desired trade-off. Here we study preference-constrained selection of one model. Before search, we fix a reference x_{ref}, defaulting to the intermediate model when available, a token multiplier $\kappa \geq 1$, and an accuracy tolerance ε in percentage points.

Let $\mathcal{F}(\mathcal{S}) = \{x \in \mathcal{S} : \text{Tok}(x) \leq \kappa \text{Tok}(x_{\text{ref}})\}$ denote the token-feasible subset of an evaluated set $\mathcal{S}$. We select

$$x^\star \in \arg\max_{x \in \mathcal{F}(\mathcal{S})} \text{Disc}(x) \quad \text{s.t.} \quad \text{Acc}(x) \geq \max_{z \in \mathcal{F}(\mathcal{S})} \text{Acc}(z) - \varepsilon. \tag{3}$$

This rule prioritizes discernment among candidates meeting the token allowance and approaching the best feasible accuracy. Retention and correction remain separate measurements throughout evaluation.

3.2 DISCO-MERGE

DISCO-Merge alternates proposal generation, batch selection and evaluation (Figure 2; Algorithm 1 in Appendix A.1). We denote the initial design size, proposal-pool size, evaluation-batch size and number of rounds by N_0, P, q, T.

Initialization and archive. We construct N_0 distinct configurations, including the source endpoints, the reference, the intermediate model when available, and Sobol-based space-filling designs. Each merged model is evaluated using common questions, feedback histories, decoding settings, and scorers. The resulting archive is

$$\mathcal{H}_t = \{(x_i, y_i, B_i)\}_{i=1}^{n_t}, \qquad y_i = (\text{Acc}_i, \text{Tok}_i, \text{Ret}_i, \text{Cor}_i), \tag{4}$$

where B_i stores aligned question-level correctness and output lengths, and $\mathcal{H}_0$ contains the initial evaluations. We retain all measured configurations, including dominated ones, to preserve local behavioral contrasts.

Flow matching training. We learn local coefficient updates conditioned on a parent's configuration and behavioral profile.

(i) Construct update examples. For each archived x_i, we retrieve nearby measured configurations x_j in coefficient space and form directed residuals $\Delta x_{ij} = x_j - x_i$. Comparing aligned responses identifies repairs ($0 \to 1$) and damage ($1 \to 0$) on the same questions. The parent context c_i contains its coefficients, dataset-and-condition correctness rates, and rates of losing answers that the reference solves. These examples need not be actual parent–child transitions; Appendix A.3 specifies the pairing and context construction.

(ii) Learn conditional updates. We use a shared velocity multilayer perceptron (MLP) $v_\phi(z, \tau, c_i, h)$, where $h \in \{\mathrm{Acc}, \mathrm{Ret}, \mathrm{Cor}\}$ indexes a proposal mode, z is the interpolated residual state, and τ is Flow time. Following flow matching (Lipman et al., 2023), sample $\epsilon_0 \sim \mathcal{N}(0, \sigma_0^2 I)$ and $\tau \sim \mathcal{U}[0,1]$. The path $z_\tau = (1-\tau)\epsilon_0 + \tau \Delta x_{ij}$ has target velocity $\Delta x_{ij} - \epsilon_0$. We train the network by minimizing

$$\mathcal{L}_{\mathrm{FM}}(\phi) = \mathbb{E}_{(i,j),h,\epsilon_0,\tau}\left[\omega_{ij,h} \frac{\|v_\phi(z_\tau, \tau, c_i, h) - (\Delta x_{ij} - \epsilon_0)\|_2^2}{d}\right], \qquad d = \dim(\mathcal{X}). \tag{5}$$

The weights $\omega_{ij,h}$ are obtained by permuting gain-and-damage weight vectors across pairs once per round, preserving contexts and residual targets. This supplies random reweighting rather than pair-specific gain supervision. After an initial exploration period, each eligible round trains a fresh network on pairs reconstructed from the cumulative archive. Insufficient contrasts defer training and retain ordinary exploration within the same budget. Appendix A.3 gives the architecture, weights, and training criteria.

Candidate generation. *(i) Choose parents.* We refresh a parent pool from the measured archive by interleaving high-accuracy, high-discernment, and low-token nondominated candidates within a token allowance. Each parent may generate multiple proposals.

(ii) Sample coefficient updates. For parent x_p, fix its context c_p and a mode h, then integrate the learned velocity from a Gaussian draw:

$$\frac{\mathrm{d}z_\tau}{\mathrm{d}\tau} = v_\phi(z_\tau, \tau, c_p, h), \qquad z_0 \sim \mathcal{N}(0, \sigma_0^2 I), \qquad \widehat{\Delta x} = z_1. \tag{6}$$

Different noise draws and modes yield multiple update samples. We normalize the sampled direction and match its pre-projection step norm to an ordinary mutation η:

$$x' = \Pi_{\mathcal{X}}\left(x_p + \|\eta\|_2 \frac{\widehat{\Delta x}}{\|\widehat{\Delta x}\|_2}\right). \tag{7}$$

Projection clips coefficients to their bounds. After discarding negligible directions, we combine Flow updates with ordinary local mutations and global samples into P distinct, unevaluated configurations $\mathcal{C}_t$. Ordinary exploration supplies the pool before Flow becomes eligible. Only selected configurations are materialized as models.

Selection and evaluation. Performance surrogates allocate costly model evaluations among inexpensive coefficient proposals.

(i) Predict performance. At each round, four independent Gaussian processes (GPs) fit Acc, $\log \mathrm{Tok}$, Ret, and Cor from all coefficient–performance observations in $\mathcal{H}_t$. Question-level records estimate observation noise. The posteriors predict performance and uncertainty for each proposal, with discernment derived from the two feedback outputs.

(ii) Allocate evaluations. Posterior utility is expected discernment, set to zero outside the token allowance or accuracy floor. We shortlist candidates with high utility or discernment uncertainty. A cost-aware finite-pool knowledge gradient (KG) (Frazier et al., 2009) estimates improvement in the best posterior utility per evaluation cost, conditioning the GP on possible outcomes with hyperparameters, thresholds and decision set fixed. Appendix A.4 gives the equations.

Acquisition ranking and a random exploration quota select q candidates. We construct and evaluate these merged models, append their real outcomes to $\mathcal{H}_{t+1}$, and repeat. Predictions and hypothetical outcomes never become archive entries.

Final model selection. After T rounds, we promote representative candidates, including the reference, to larger development evaluations and recompute scores at a common level. Equation 3 selects $x^\star$, which is locked before full evaluation. The returned model $\theta(x^\star)$ requires neither Flow nor GPs at deployment. Appendix A.2 analyzes computational costs and evaluation budgets.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Models. We merge three Short/Long pairs: Qwen2.5-Math-7B / DeepSeek-R1-Distill-Qwen-7B (7B), Qwen3-4B-Instruct-2507 / Qwen3-4B-Thinking-2507 (4B), and Phi-4-mini-instruct / Phi-4-mini-reasoning (3.8B). The first and second checkpoints in each pair are Short and Long, respectively. Merging and the fixed evaluation runtime are defined within each family.

Baselines. We compare static merging methods Average (Wortsman et al., 2022), TA (Ilharco et al., 2023), TIES (Yadav et al., 2023), SLERP (Goddard et al., 2024), KARCHER (Wang et al., 2026), ACM (Yao et al., 2025) and Sens-Merging (Liu et al., 2025), together with evolutionary search methods Evo-L2S (Iacobelli et al., 2026) and EvoGM (Jiang et al., 2026). Main results compare complete methods.

Benchmarks and Tasks. We assess arithmetic word problems on GSM8K (Cobbe et al., 2021), SVAMP (Patel et al., 2021) and ASDiv (Miao et al., 2020); advanced mathematical reasoning on MATH500 (Hugging Face H4, 2024), Minerva Math (Lewkowycz et al., 2022) and OlympiadBench (He et al., 2024); and competition problems on AMC23 and AIME24 (Mathematical Association of America, n.d.).

4.2 MAIN RESULTS

Qwen2.5-7B Models. We evaluate merges of Qwen2.5-Math-7B and DeepSeek-R1-Distill-Qwen-7B across eight mathematical benchmarks, reporting one selected model per method. As shown in Table 1, DISCO achieves the highest macro Acc and Disc, reaching 68.03% and 74.99%. Compared with Long, it improves Acc by 3.17 points with 62.59% fewer tokens. DISCO also exceeds TIES by 1.92 Acc and 9.96 Disc points with 9.71% fewer tokens, and EvoGM by 2.17 Acc and 9.75 Disc points with 8.35% fewer tokens. Notably, TA improves clean accuracy over Long but reduces Disc, showing that accuracy gains alone do not guarantee better feedback behavior.

Table 1: Qwen2.5-7B performance on eight mathematical benchmarks.

Model / Method	Acc↑	Tok↓	Disc↑	Ret↑	Cor↑	HV↑
Qwen2.5-Math (Short)	47.38	767.86	41.91	24.58	59.24	0.18370
R1-Distill (Long)	64.86	3,693.01	61.23	54.37	68.10	0.25391
Average (22)	65.58	1,699.17	61.88	56.41	67.34	0.33846
TA (23)	<u>67.00</u>	2,425.54	52.79	49.53	56.05	0.26989
TIES (23)	66.11	1,530.35	65.03	61.56	68.50	0.36566
SLERP (24)	65.93	1,524.17	62.31	56.94	67.68	0.34967
Sens-Merging (25)	65.80	2,369.13	52.98	49.68	56.28	0.26795
ACM-TA (25)	66.04	3,535.21	<u>68.02</u>	<u>64.33</u>	71.71	0.29411
ACM-TIES (25)	63.33	<u>1,411.57</u>	66.88	57.96	<u>75.80</u>	0.36519
KARCHER (26)	65.10	1,622.03	64.40	57.62	71.18	0.35281
Evo-L2S (26)	66.08	1,762.05	54.75	51.85	57.66	0.29954
EvoGM (26)	65.86	1,507.65	65.24	59.44	71.04	<u>0.36640</u>
DISCO-Merge	**68.03**	**1,381.72**	**74.99**	**72.43**	**77.54**	**0.44130**

Qwen3-4B Models. To examine whether these benefits extend to another source pair, we evaluate merges of Qwen3-4B-Instruct-2507 and Qwen3-4B-Thinking-2507 under the common evaluation runtime. Table 2 shows that DISCO achieves the highest Acc on six of eight benchmarks, including one tie, and the highest Disc on all eight. Its macro Acc and Disc reach 74.76% and 85.97%.

Compared with ACM-TIES, the strongest static baseline in macro Acc, DISCO gains 9.52 Acc and 11.51 Disc points with 19.34% fewer tokens. EvoGM produces shorter answers but trails DISCO by 6.06 Acc and 34.37 Disc points. These results show that shorter answers need not preserve reasoning quality or feedback behavior.

Table 2: Qwen3-4B performance on eight mathematical benchmarks.

Model / Method	Acc↑	Tok↓	Disc↑	Ret↑	Cor↑	HV↑
Qwen3-Instruct (Short)	58.94	3,591.14	45.62	38.43	52.82	0.17461
Qwen3-Thinking (Long)	51.09	4,986.74	47.19	48.01	46.38	0.12369
Average (22)	58.41	4,288.36	74.63	74.27	74.99	0.25338
TA (23)	48.42	5,065.95	61.98	62.95	61.02	0.15166
TIES (23)	58.16	4,278.79	73.88	72.71	75.05	0.25013
SLERP (24)	58.52	4,276.58	74.50	71.16	77.84	0.25390
Sens-Merging (25)	46.23	5,028.02	56.55	56.99	56.12	0.13308
ACM-TA (25)	53.29	4,846.48	46.03	46.17	45.89	0.12919
ACM-TIES (25)	65.23	3,870.71	74.46	70.69	78.22	0.30211
KARCHER (26)	57.26	4,281.44	74.40	71.12	77.68	0.24789
Evo-L2S (26)	63.06	3,957.98	76.01	72.89	79.14	0.29406
EvoGM (26)	68.70	**2,612.27**	51.60	38.34	64.85	0.26404
DISCO-Merge	**74.76**	3,121.97	**85.97**	**85.49**	**86.44**	**0.44674**

Phi-4-mini-3.8B Models. We further assess the framework on Phi-4-mini-instruct and Phi-4-mini-reasoning using native Phi chat formatting and the same eight benchmarks. As shown in Table 3, DISCO achieves 69.44% Acc and 77.92% Disc, with the highest Disc on seven benchmarks, including one tie. Compared with Long, it uses 18.75% fewer tokens and gains 4.01 Disc points at a modest 0.39-point Acc cost. ACM-TIES uses 1,203 tokens but reaches only 61.63% Acc. Compared with EvoGM, DISCO gains 2.85 Disc points with 12.31% fewer tokens at a 0.13-point Acc cost. Its improvements over Long involve both Ret and Cor.

Table 3: Phi-4-mini-3.8B performance on eight mathematical benchmarks.

Model / Method	Acc↑	Tok↓	Disc↑	Ret↑	Cor↑	HV↑
Phi4m-Instruct (Short)	55.79	1,228.02	60.70	58.67	62.73	0.29806
Phi4m-Reasoning (Long)	**69.83**	3,495.50	73.91	70.59	77.23	0.33995
Average (22)	60.13	1,438.33	70.92	68.74	73.10	0.36656
TA (23)	65.99	2,812.91	75.02	73.37	76.66	0.35905
TIES (23)	61.06	1,538.98	69.72	69.66	69.77	0.36168
SLERP (24)	60.59	1,401.36	70.89	70.53	71.25	0.37073
KARCHER (26)	62.36	1,353.28	69.32	67.65	70.98	0.37512
ACM-TA (25)	67.34	3,316.14	71.70	67.77	75.63	0.32647
ACM-TIES (25)	61.63	**1,203.36**	68.04	66.46	69.61	0.37005
Sens-Merging (25)	64.39	2,843.01	72.22	69.55	74.90	0.33592
Evo-L2S (26)	69.21	3,435.54	74.31	70.65	77.97	0.34176
EvoGM (26)	69.57	3,239.01	75.07	71.73	78.42	0.35710
DISCO-Merge	69.44	2,840.14	**77.92**	**76.83**	**79.02**	**0.39103**

4.3 Effectiveness Across Intermediate Models

To examine whether DISCO improves different intermediate models, we use TIES, Average, SLERP and KARCHER merges of the same Qwen2.5-7B source pair. We keep parameter grouping, initial candidate models, search budget, evaluation settings and final-selection protocol fixed. DISCO-Module directly merges the two source models under the same search framework, providing a control for the role of the intermediate model. Table 4 reports each intermediate model and its corresponding DISCO output. All four midpoints gain 2.17–3.38 Acc points, 7.65–12.41 Disc points and 19.15–27.69% HV; both Ret and Cor improve. Three variants reduce Tok. SLERP uses 8.78%

Table 4: Qwen2.5 results; red parentheses show absolute changes from each midpoint.

Model / Variant	Acc↑	Tok↓	Disc↑	Ret↑	Cor↑	HV↑
TIES (23)	66.11	1,530.35	65.03	61.56	68.50	0.36566
DISCO-TIES	68.81 (+2.70)	1,524.71 (-5.65)	**75.76** (+10.73)	71.58 (+10.03)	**79.94** (+11.44)	**0.44368** (+0.07802)
Average (22)	65.58	1,699.17	61.88	56.41	67.34	0.33846
DISCO-Average	67.75 (+2.17)	1,410.85 (-288.32)	73.98 (+12.10)	70.73 (+14.31)	77.23 (+9.89)	0.43217 (+0.09371)
SLERP (24)	65.93	1,524.17	62.31	56.94	67.68	0.34967
DISCO-SLERP	**69.31** (+3.38)	1,658.06 (+133.89)	74.72 (+12.41)	**72.65** (+15.71)	76.79 (+9.10)	0.43402 (+0.08435)
KARCHER (26)	65.10	1,622.03	64.40	57.62	71.18	0.35281
DISCO-KARCHER	67.51 (+2.41)	**1,389.77** (-232.25)	72.05 (+7.65)	66.79 (+9.17)	77.31 (+6.14)	0.42038 (+0.06757)
DISCO-Module	67.51	1,491.91	73.74	70.57	76.91	0.42526

more tokens for the highest Acc, while TIES leads Disc/HV and KARCHER minimizes Tok. The framework therefore improves several intermediate merges with distinct quality–cost trade-offs.

4.4 Feedback Behavior Analysis

We test whether DISCO retains correct answers under misleading advice and corrects errors after valid advice. Comparing with Long assesses whether merging preserves these abilities; comparing with the highest-Disc merging baseline tests improvement over existing methods. Per-benchmark Ret and Cor changes, computed on identical fixed histories, reveal whether gains are consistent across tasks.

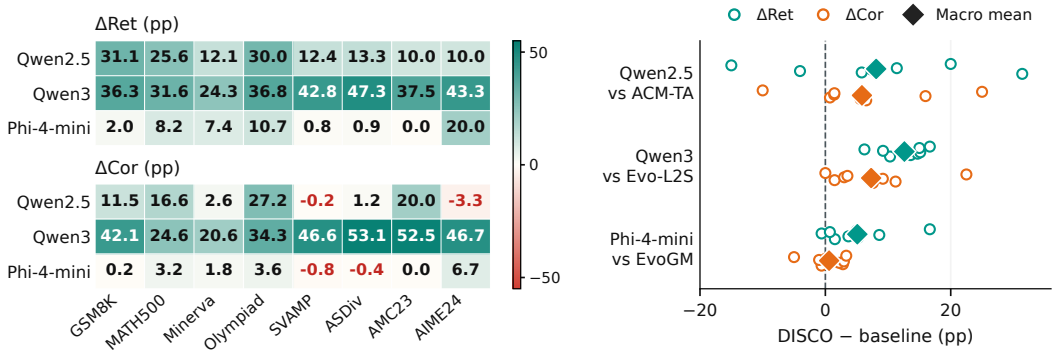

Figure 3: **Feedback changes across benchmarks (percentage points). Left:** Ret and Cor changes relative to Long. **Right:** Changes relative to each family's highest-Disc merging baseline; circles denote datasets and diamonds denote macro means.

Against Long, Ret improves in 23 of 24 family–benchmark combinations and Cor in 19, with four Cor declines. Against the highest-Disc merging baselines, both macro Ret and Cor improve in all three families, while individual datasets can regress. These comparisons show gains over both source and merged models, with Ret improving more consistently than Cor. Additional mathematical feedback results are reported in Appendix G.

4.5 Component and Search Analysis

Ablation Study. We evaluate the contributions of DISCO's core components on eight Qwen2.5-7B mathematical benchmarks. We compare the full framework with five variants: *w/ Gaussian* replaces Flow with a simpler conditional Gaussian proposal model; *w/o KG* allocates evaluations uniformly; *w/o Disc. in Acq.* and *w/o Disc. in Selection* replace discernment with accuracy at the respective decision stages; and *w/o Reward Shuffle* uses unshuffled training rewards. Table 5 reports means and sample standard deviations over three random search seeds.

Table 5: Qwen2.5 ablation study: mean ± sample SD over three random seeds.

Method	Acc↑	Tok↓	Disc↑	Ret↑	Cor↑	HV↑
w/ Gaussian	**68.97** ± 1.36	1519.94 ± 149.35	73.49 ± 1.91	69.42 ± 0.61	77.57 ± 3.92	0.43151 ± 0.00925
w/o KG	67.36 ± 0.23	1504.13 ± 98.62	71.00 ± 3.83	64.97 ± 3.03	77.04 ± 4.78	0.40810 ± 0.02493
w/o Disc. in Acq.	68.15 ± 0.44	1535.32 ± 299.75	72.93 ± 0.44	69.37 ± 1.17	76.48 ± 0.33	0.42257 ± 0.01810
w/o Reward Shuffle	67.04 ± 0.90	1478.63 ± 84.15	73.88 ± 3.19	69.86 ± 4.53	**77.90** ± 2.76	0.42384 ± 0.02126
w/o Disc. in Selection	68.77 ± 0.97	1424.89 ± 66.73	72.47 ± 3.10	69.90 ± 4.26	75.03 ± 1.98	0.42919 ± 0.02484
DISCO-Merge	**68.97** ± 1.20	**1378.84** ± 120.37	**74.18** ± 4.13	**71.10** ± 4.69	77.25 ± 3.59	**0.44314** ± 0.03609

DISCO achieves the highest mean HV and the lowest mean Tok. Every ablated variant has lower mean HV, supporting the contribution of each tested component to joint performance. The *w/o KG* variant loses 1.61 Acc and 3.17 Disc points, highlighting the value of informed allocation. Replacing Disc in acquisition and selection lowers mean Disc by 1.25 and 1.71 points, respectively, supporting feedback-aware decisions at both stages. Without reward shuffling, Acc falls by 1.93 points despite higher Cor. Compared with Gaussian proposals, Flow maintains similar Acc while reducing Tok by 9.28% and improving Disc by 0.68 points, supporting its ability to find more efficient merging configurations.

Search dynamics. To test whether search improves on initialization across different random seeds, we track the model selected by the frozen output rule at each iteration. Figure 4 reports search-set Acc, Tok and Disc for three seeds; iteration 0 denotes initialization, followed by ten search iterations.

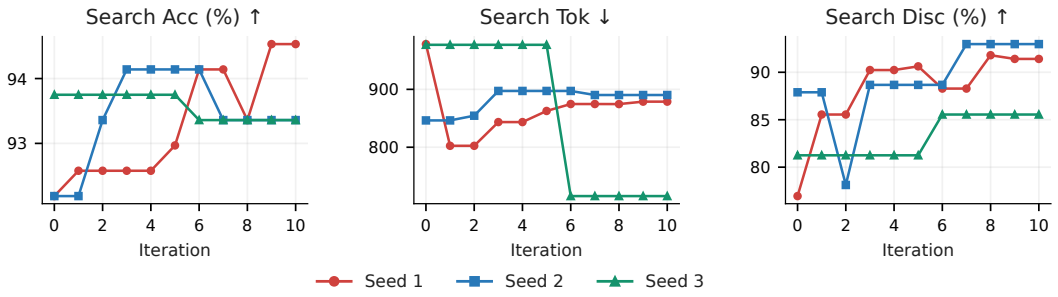

Figure 4: Search-set performance across iterations for three random seeds.

All three final search selections improve Disc by 4.30–14.45 points over initialization. Two reduce Tok by 10.16–26.75%; the third uses 5.18% more tokens. All three confirmation-selected models originate after initialization, supporting search beyond the initial pool. Seed identities and control trajectories appear in Appendix K.

5 Conclusion and Limitations

We showed that accuracy and output length alone do not fully characterize feedback behavior in merged reasoning models. DISCO-Merge combines behavior-conditioned coefficient proposals with Bayesian evaluation to maximize discernment under accuracy and token-cost constraints. Experiments across three model families and eight mathematical benchmarks demonstrate improved joint performance, supported by analyses of feedback behavior, intermediate models, and search components. Our evaluation primarily covers mathematical reasoning with fixed feedback histories, so it does not establish reliability in open-ended, multi-turn interactions. The search also requires repeated candidate evaluations, adding offline computational cost.

AI USE STATEMENT

We used generative AI tools, including ChatGPT and Codex, to assist with refining the conceptual framing and research hypotheses, discussing methodological choices and experimental designs, implementing and debugging research code, and interpreting experimental results. We also used these tools for manuscript drafting, translation and language editing, the preparation and revision of figures and tables, and literature discovery. We reviewed the AI-assisted material and made the final decisions regarding the methodology, experimental protocol, analysis, and presentation. Verification included checking relevant references against their original sources, inspecting and testing research code, and checking numerical statements and interpretations against experimental records. We distinguish AI assistance in the research workflow from the language models evaluated as experimental subjects, whose roles are described in the methodology and experimental setup. We retain full responsibility for the accuracy, originality, and integrity of the submitted work.

REFERENCES

Xingyu Chen, Jiahao Xu, Tian Liang, Zhiwei He, Jianhui Pang, Dian Yu, Linfeng Song, Qiuzhi Liu, Mengfei Zhou, Zhuosheng Zhang, Rui Wang, Zhaopeng Tu, Haitao Mi, and Dong Yu. Do NOT Think That Much for 2+3=? On the Overthinking of Long Reasoning Models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 9487–9499, 2025. URL https://proceedings.mlr.press/v267/chen25bx.html.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021. URL https://arxiv.org/abs/2110.14168.

Samuel Daulton, Maximilian Balandat, and Eytan Bakshy. Parallel Bayesian Optimization of Multiple Noisy Objectives with Expected Hypervolume Improvement. In *Advances in Neural Information Processing Systems*, volume 34, pp. 2187–2200, 2021. URL https://papers.nips.cc/paper/2021/hash/11704817e347269b7254e744b5e22dac-Abstract.html.

DeepSeek-AI. DeepSeek-R1-Distill-Qwen-7B model card. Hugging Face, 2025. URL https://huggingface.co/deepseek-ai/DeepSeek-R1-Distill-Qwen-7B.

Justin C. Dunlap, Anne-Simone Parent, and Ralf Widenhorn. Feedback indices to evaluate LLM responses to rebuttals for multiple choice type questions. *Discover Artificial Intelligence*, 6:855, 2026. doi: 10.1007/s44163-026-01521-8.

Aaron Fanous, Jacob Goldberg, Ank Agarwal, Joanna Lin, Anson Zhou, Sonnet Xu, Vasiliki Bikia, Roxana Daneshjou, and Sanmi Koyejo. SycEval: Evaluating LLM Sycophancy. In *AAAI/ACM Conference on AI, Ethics, and Society*, volume 8, pp. 893–900, 2025. doi: 10.1609/aies.v8i1.36598.

Peter Frazier, Warren Powell, and Savas Dayanik. The knowledge-gradient policy for correlated normal beliefs. *INFORMS Journal on Computing*, 21(4):599–613, 2009. doi: 10.1287/ijoc.1080.0314.

Charles Goddard, Shamane Siriwardhana, Malikeh Ehghaghi, Luke Meyers, Vladimir Karpukhin, Brian Benedict, Mark McQuade, and Jacob Solawetz. Arcee's MergeKit: A Toolkit for Merging Large Language Models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pp. 477–485. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-industry.36.

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, Jie Liu, Lei Qi, Zhiyuan Liu, and Maosong Sun. OlympiadBench: A Challenging Benchmark for Promoting AGI with Olympiad-Level Bilingual Multimodal Scientific Problems. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3828–3850. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.211.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring Mathematical Problem Solving With the MATH Dataset. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/hash/be83ab3ecd0db773eb2dc1b0a17836a1-Abstract-round2.html.

Amin Heyrani Nobari, Kaveh Alimohammadi, Ali ArjomandBigdeli, Akash Srivastava, Faez Ahmed, and Navid Azizan. Activation-Informed Merging of Large Language Models. In *Advances in Neural Information Processing Systems*, volume 38, pp. 72975–72998, 2025. doi: 10.52202/085713-2446.

Jiseung Hong, Grace Byun, Seungone Kim, and Kai Shu. Measuring sycophancy of language models in multi-turn dialogues. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 2239–2259. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-emnlp.121.

Bairu Hou, Yang Zhang, Jiabao Ji, Yujian Liu, Kaizhi Qian, Jacob Andreas, and Shiyu Chang. ThinkPrune: Pruning Long Chain-of-Thought of LLMs via Reinforcement Learning. *Transactions on Machine Learning Research*, 2026. ISSN 2835-8856. URL https://openreview.net/forum?id=V51gPu1uQD.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. Large Language Models Cannot Self-Correct Reasoning Yet. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=IkmD3fKBPQ.

Hugging Face H4. MATH-500 dataset. Hugging Face, 2024. URL https://huggingface.co/datasets/HuggingFaceH4/MATH-500.

Mario Iacobelli, Adrian Robert Minut, Tommaso Mencattini, Donato Crisostomi, Andrea Santilli, Iacopo Masi, and Emanuele Rodolà. Multi-objective Evolutionary Merging Enables Efficient Reasoning Models. In *Conference on Language Modeling*, 2026. URL https://arxiv.org/abs/2604.06465.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing Models with Task Arithmetic. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=6t0Kwf8-jrj.

Tao Jiang, Xinmeng Yu, Chenhao Yi, Yiling Wu, Yan Li, Ran Cheng, Dongmei Jiang, and Jianguo Zhang. EvoGM: Learning to Merge LLMs via Evolutionary Generative Optimization. In *International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=R2CoYV0Gf6.

Philippe Laban, Lidiya Murakhovs'ka, Caiming Xiong, and Chien-Sheng Wu. Are You Sure? Challenging LLMs Leads to Performance Drops in The FlipFlop Experiment. *arXiv preprint arXiv:2311.08596*, 2023. URL https://arxiv.org/abs/2311.08596.

Xiaochong Lan, Yu Zheng, Shiteng Cao, and Yong Li. The Thinking Spectrum: An Empirical Study of Tunable Reasoning in LLMs through Model Merging. *arXiv preprint arXiv:2509.22034*, 2025. URL https://arxiv.org/abs/2509.22034.

Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, Yuhuai Wu, Behnam Neyshabur, Guy Gur-Ari, and Vedant Misra. Solving Quantitative Reasoning Problems with Language Models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 3843–3857. Curran Associates, Inc., 2022. doi: 10.52202/068431-0278.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=PqvMRDCJT9t.

Shuqi Liu, Han Wu, Bowei He, Xiongwei Han, Mingxuan Yuan, and Linqi Song. Sens-Merging: Sensitivity-Guided Parameter Balancing for Merging Large Language Models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 19243–19255. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-acl.984.

Haotian Luo, Haiying He, Yibo Wang, Jinluan Yang, Rui Liu, Naiqiang Tan, Xiaochun Cao, Dacheng Tao, and Li Shen. Ada-R1: Hybrid-CoT via Bi-Level Adaptive Reasoning Optimization. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://papers.neurips.cc/paper_files/paper/2025/hash/558da25ef295121e2d79209a95e45436-Abstract-Conference.html.

Haotian Luo, Haiying He, Yibo Wang, Shiwei Liu, Wei Li, Xiaochun Cao, Dacheng Tao, Naiqiang Tan, and Li Shen. O1-Pruner: Length-Harmonizing Fine-Tuning for O1-Like Reasoning Pruning. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 14242–14257. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.findings-acl.697.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. Self-Refine: Iterative Refinement with Self-Feedback. In *Advances in Neural Information Processing Systems*, volume 36, pp. 46534–46594, 2023. doi: 10.52202/075280-2019.

Mathematical Association of America. American mathematics competitions, n.d. URL https://maa.org/student-programs/amc/. Official competition resource; accessed September 26, 2026. Evaluated AMC23 and AIME24 subsets are specified in the experimental protocol.

Shen-yun Miao, Chao-Chun Liang, and Keh-Yih Su. A Diverse Corpus for Evaluating and Developing English Math Word Problem Solvers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 975–984. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.92.

Microsoft. Phi-4-mini-instruct model card. Hugging Face, 2025a. URL https://huggingface.co/microsoft/Phi-4-mini-instruct.

Microsoft. Phi-4-mini-reasoning model card. Hugging Face, 2025b. URL https://huggingface.co/microsoft/Phi-4-mini-reasoning.

Arkil Patel, Satwik Bhattamishra, and Navin Goyal. Are NLP Models really able to Solve Simple Math Word Problems? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 2080–2094. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.naacl-main.168.

Qwen Team. Qwen3-4B-Instruct-2507 model card. Hugging Face, 2025a. URL https://huggingface.co/Qwen/Qwen3-4B-Instruct-2507.

Qwen Team. Qwen3-4B-Thinking-2507 model card. Hugging Face, 2025b. URL https://huggingface.co/Qwen/Qwen3-4B-Thinking-2507.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, Shauna Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. Towards Understanding Sycophancy in Language Models. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/0105f7972202c1d4fb817da9f21a9663-Abstract-Conference.html.

Debu Sinha. SycoBench-600: Measuring sycophancy and correction selectivity in LLM assistants. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 35278–35284. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.findings-acl.1759.

Jiayu Wang, Zuojun Ye, and Wenpeng Yin. Functionality-oriented LLM merging on the Fisher–Rao manifold. *arXiv preprint arXiv:2603.04972*, 2026. URL https://arxiv.org/abs/2603.04972.

Pinzheng Wang, Juntao Li, Zecheng Tang, Haijia Gui, and Min Zhang. Improving Rationality in the Reasoning Process of Language Models through Self-playing Game. In *International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 63417–63438, 2025. URL https://proceedings.mlr.press/v267/wang25bb.html.

Mitchell Wortsman, Gabriel Ilharco, Samir Ya Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S. Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, and Ludwig Schmidt. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 23965–23998, 2022. URL https://proceedings.mlr.press/v162/wortsman22a.html.

Han Wu, Yuxuan Yao, Shuqi Liu, Zehua Liu, Xiaojin Fu, Xiongwei Han, Xing Li, Hui-Ling Zhen, Tao Zhong, and Mingxuan Yuan. Unlocking Efficient Long-to-Short LLM Reasoning with Model Merging. *arXiv preprint arXiv:2503.20641*, 2025. URL https://arxiv.org/abs/2503.20641.

Taiqiang Wu, Runming Yang, Tao Liu, Jiahao Wang, and Ngai Wong. Revisiting model interpolation for efficient reasoning. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8624–8638. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.389.

Heming Xia, Chak Tou Leong, Wenjie Wang, Yongqi Li, and Wenjie Li. TokenSkip: Controllable Chain-of-Thought Compression in LLMs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 3351–3363. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-main.165.

Tian Xia. Data-Free Layer-Adaptive Merging via Fisher Information for Long-to-Short Reasoning LLMs. *arXiv preprint arXiv:2603.21705*, 2026. URL https://arxiv.org/abs/2603.21705.

Silei Xu, Wenhao Xie, Lingxiao Zhao, and Pengcheng He. Chain of Draft: Thinking Faster by Writing Less. *arXiv preprint arXiv:2502.18600*, 2025. URL https://arxiv.org/abs/2502.18600.

Prateek Yadav, Derek Tam, Leshem Choshen, Colin Raffel, and Mohit Bansal. TIES-Merging: Resolving Interference When Merging Models. In *Advances in Neural Information Processing Systems*, volume 36, pp. 7093–7115, 2023. doi: 10.52202/075280-0310.

An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu, Jingren Zhou, Junyang Lin, Keming Lu, Mingfeng Xue, Runji Lin, Tianyu Liu, Xingzhang Ren, and Zhenru Zhang. Qwen2.5-Math Technical Report: Toward Mathematical Expert Model via Self-Improvement. *arXiv preprint arXiv:2409.12122*, 2024a. URL https://arxiv.org/abs/2409.12122.

Enneng Yang, Zhenyi Wang, Li Shen, Shiwei Liu, Guibing Guo, Xingwei Wang, and Dacheng Tao. AdaMerging: Adaptive Model Merging for Multi-Task Learning. In *International Conference on Learning Representations*, 2024b. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/62868cc2fc1eb5cdf321d05b4b88510c-Abstract-Conference.html.

Yuxuan Yao, Shuqi Liu, Zehua Liu, Qintong Li, Mingyang Liu, Xiongwei Han, Zhijiang Guo, Han Wu, and Linqi Song. Activation-Guided Consensus Merging for Large Language Models. In *Advances in Neural Information Processing Systems*, volume 38, pp. 124041–124069, 2025. doi: 10.52202/085713-4133.

Le Yu, Bowen Yu, Haiyang Yu, Fei Huang, and Yongbin Li. Language Models are Super Mario: Absorbing Abilities from Homologous Models as a Free Lunch. In *International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 57755–57775, 2024. URL https://proceedings.mlr.press/v235/yu24p.html.

Zhaofeng Zhong, Wei Yuan, Tong Chen, Liang Qu, Xiangyu Zhao, Quoc Viet Hung Nguyen, and Hongzhi Yin. QA-Merging: Query-Adaptive Reasoning via Layer Selective Model Merging. In *ACM International Conference on Information and Knowledge Management*, 2026. URL https://arxiv.org/abs/2601.03506. Accepted for publication.

Qiyuan Zhu, Dezhi Li, Lujun Li, Xiaoyu Qin, Wei Li, Hao Gu, Hua Xu, Sirui Han, and Yike Guo. Outlier Matters: Efficient Long-to-Short Reasoning via Outlier-Guided Model Merging. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 35213–35221, 2026. doi: 10.1609/aaai.v40i41.40828.

APPENDIX

Appendix contents

Section	Content	Page
A	Additional Method Details	15
B	Motivation Scan: Setup and Detailed Results	18
C	Experimental Configuration and Evaluation Protocol	22
D	Complete Qwen2.5-7B Results	24
E	Complete Qwen3-4B Results	27
F	Complete Phi-4-mini-3.8B Results	28
G	Supplementary Mathematical Feedback Evaluation	31
H	Component Comparisons	31
I	Framework Performance Across Intermediate Models	32
J	Search Strategy and Merge Parameterization	34
K	Search Trajectories and Computational Accounting	35

A ADDITIONAL METHOD DETAILS

A.1 COMPLETE PROCEDURE

Algorithm 1 DISCO-Merge: iterative proposal and evaluation

Input: Source models $\theta_{\text{short}}, \theta_{\text{long}}$, optional intermediate model θ_{mid}, reference x_{ref}; search and confirmation data $\mathcal{Q}_s$, $\mathcal{Q}_c$
Settings: Initial size N_0, pool size P, batch size q, rounds T; preference (κ, ε)
Output: Final coefficients $x^\star$ and merged model $\theta(x^\star)$

Define $\theta(x)$ by Equation 1, or Equation 15 without an intermediate model
Choose N_0 initial configurations, including available input models, the reference, and Sobol samples
Evaluate these models on $\mathcal{Q}_s$; store coefficients, scores, and records in $\mathcal{H}_0$
for $t = 0, \ldots, T-1$ **do**
    Choose parent configurations from the measured archive $\mathcal{H}_t$
    Pair nearby archived configurations to obtain residuals and behavioral contexts
    if initial exploration is complete and enough training pairs are available **then**
        Train a fresh conditional Flow on the cumulative pairs using Equation 5
        Generate coefficient proposals from the parents using Equations 6–7
    else
        Begin with an empty proposal pool
    end if
    Add ordinary local and global samples to obtain P distinct, unevaluated candidates
    Fit four GPs to measured coefficients and performance in $\mathcal{H}_t$
    Predict candidate performance; shortlist high-utility or high-uncertainty candidates
    Select q candidates using cost-aware KG and a random exploration quota
    Construct the selected merged models and evaluate them on $\mathcal{Q}_s$
    Append these real observations to obtain $\mathcal{H}_{t+1}$
end for
Confirm promising archived candidates and the reference using expanded data $\mathcal{Q}_c$
Select $x^\star$ from the confirmed set by Equation 3; lock the choice
return $x^\star$, $\theta(x^\star)$

A.2 COMPUTATIONAL COMPLEXITY AND EVALUATION BUDGET

Search computation. For $n = n_t$ archived d-dimensional configurations, let k be the neighbor limit and Q the aligned records per model. Flow uses I training steps, batch size b, S integration steps, and per-example forward cost C_v. Let $\widetilde{P}$ count proposal attempts (including rejections), L the KG shortlist size, F fantasies per action, and $Z = n + P$ the terminal pool size.

Operation	Per-round arithmetic cost
Pairing and behavior statistics	$O(n^2(d+\log n)+knQ)$
Flow optimization	$O(IbC_v)$
Proposal generation and filtering	$O(\widetilde{P}(SC_v+d+Q))$
GP fitting	$O(n^3+n^2d)$
Candidate marginal predictions	$O(Pn^2+Pnd)$
Joint posterior and KG	$O(Zn^2+Z^2(n+d)+LFZ)$

Bounds assume aligned records, fixed numbers of GP outputs and kernel trials, and hash-based deduplication; pairing includes sorting and outcome scans. Flow samples I minibatches; KG reuses one posterior across fantasies. Bounded retries give $\widetilde{P}=O(P)$.

Evaluation budget. Only q new configurations per round undergo real evaluation:

$$n_t = N_0 + tq, \qquad N_{\text{search}} = N_0 + Tq. \tag{8}$$

Constructing a selected model requires $O(N_\theta)$ arithmetic for N_θ parameters. Let Q_s count logical search requests per model across all conditions. With n_j promoted models at confirmation stage j, ΔQ_j additional requests beyond the previous stage, and Q_{test} final-evaluation requests, the nominal generation-request budget is

$$N_{\text{gen}} = (N_0+Tq)Q_s + \sum_{j=1}^{J_c} n_j \Delta Q_j + Q_{\text{test}}. \tag{9}$$

A.3 Flow Training and Candidate Generation

For the reference 18-dimensional implementation, we set $T=10$ search rounds.

Training pairs and context. Let $B_{i,k,h,n} \in \{0,1\}$ denote correctness for model i, dataset k, condition h, and question n, with $n_{k,h}$ records in that stratum. For D datasets and three conditions, the $d+6D$ context entries are

$$c_i = \left[x_i, \left\{ \frac{\sum_n B_{i,k,h,n}}{n_{k,h}} \right\}_{k,h}, \left\{ \frac{\sum_n B_{\text{ref},k,h,n}(1-B_{i,k,h,n})}{n_{k,h}} \right\}_{k,h} \right]. \tag{10}$$

Rates are fractions over aligned questions and fixed histories. Each x_i has up to eight Euclidean-nearest measured neighbors within $\|x_j - x_i\|_\infty \le 0.4$, excluding negligible differences.

Sample weights. For pair $e=(i,j)$, let $g_{e,h}$ be the change in macro-averaged correctness under condition h, in percentage points, and set $g_{e,\text{Disc}} = (g_{e,\text{Ret}} + g_{e,\text{Cor}})/2$. Let $\bar{g}^{(2)}_{e,\text{Acc}}$ average the two lowest dataset-level accuracy changes. With $[u]_+ = \max(u,0)$, set $w_{e,h} = [g_{e,h}]_+ p_e$, where

$$p_e = \exp\left[-\min\left(20, \frac{[-g_{e,\text{Acc}}]_+}{2} + \frac{[-g_{e,\text{Disc}}]_+}{2} + \frac{[-\bar{g}^{(2)}_{e,\text{Acc}}]_+}{4} \right) \right]. \tag{11}$$

A seeded permutation π_t reassigns whole three-component weight vectors across the m pairs. The loss uses

$$\omega_{e,h} = \frac{w_{\pi_t(e),h}}{\max(10^{-8}, m^{-1}\sum_{e'} w_{e',h})}. \tag{12}$$

The permutation is fixed within each round; residual–context correspondence is unchanged.

Training and proposals. The velocity MLP takes $[c_i, h, z_\tau, \tau]$ with a three-way encoding of h, giving $2d+6D+4$ inputs and d outputs. Samples, modes, noise, and interpolation times are redrawn during training.

Setting	Value
Network	Two width-128 hidden layers; SiLU; linear output
Optimization	Adam, 180 steps, batch 128; learning rate 0.0015; gradient clipping 2
Noise and integration	$\sigma_0 = 0.1$; 12 Euler steps
Training eligibility	After two exploration rounds: at least 24 pairs, eight source parents, and four weights $> 10^{-5}$ per mode
Parent pool	Up to 12; token allowance $1.30\mathrm{Tok}(x_{\mathrm{ref}})$; four consecutive attempts per parent
Proposal mixture	Three Flow attempts per ordinary attempt; ordinary global-sampling probability 0.2
Local mutation scale	Standard deviation sampled from $\{0.035, 0.07, 0.12\}$; its sampled norm also scales Flow directions

If the parent token filter is empty, all archived models are eligible. Duplicate proposals and negligible directions are replaced.

A.4 Surrogate Modeling and Evaluation

Surrogate and acquisition. The GPs use a Matérn-$5/2$ kernel on coefficient RMS distance, with a shared length scale minimizing mean leave-one-out negative log predictive density. Noise uses stratified binomial estimates for correctness and a delta-method estimate for log Tokens. Noise correlations are omitted; latent correlations across configurations are retained.

Let $c_{\max} = \kappa \mathrm{Tok}(x_{\mathrm{ref}})$ and $a_t = \max_{x \in \mathcal{F}(\mathcal{S}_t)} \mathrm{Acc}(x) - \varepsilon$, where $\mathcal{S}_t = \{x_i\}_{i=1}^{n_t}$ is the evaluated configuration set and $\mathcal{F}$ is defined in Section 3.1. These thresholds use observed search-stage scores, not GP posterior means; the reference token cost is measured at the same stage. They remain fixed during fantasy conditioning. Confirmation instead recomputes gates from cumulative observed scores at each promotion stage. The posterior utility is

$$U_t(z; \mathcal{H}) = \mathbb{E}[\mathrm{Disc}(z)\mathbf{1}\{\mathrm{Acc}(z) \geq a_t,\ \mathrm{Tok}(z) \leq c_{\max}\} \mid \mathcal{H}]\,. \tag{13}$$

Let $V_t(\mathcal{H}) = \max(0, \max_{z \in \mathcal{Z}_t} U_t(z; \mathcal{H}))$, where $\mathcal{Z}_t$ contains archived and proposed configurations. The acquisition is

$$\mathcal{A}_t(x) = \frac{\mathbb{E}_{Y_x \mid \mathcal{H}_t}[V_t(\mathcal{H}_t^{x, Y_x})] - V_t(\mathcal{H}_t)}{\widehat{c}_t(x)}. \tag{14}$$

Y_x observes all four outputs with noise. Fantasies fix hyperparameters, thresholds, $\mathcal{Z}_t$, and Flow; actions are ranked individually. The cost proxy is $\widehat{c}_t(x) = 8 \times 180 + 512 \exp(\mu_{\log \mathrm{Tok}}(x))/300$.

Setting	Value
Length-scale candidates	$\{0.08, 0.16, 0.32, 0.64, 1.28\}$
Standardization floors	$(2, 0.1, 2, 2)$ for $(\mathrm{Acc}, \log \mathrm{Tok}, \mathrm{Ret}, \mathrm{Cor})$
Noise variance floor	0.05 after standardization
Pool and evaluation batch	$P = 128$, $q = 8$: six KG/cost selections and two uniform random selections from the remainder
KG shortlist and fantasies	Eight highest-utility candidates plus eight additional high-uncertainty candidates; eight antithetic fantasies
Selection preference	$\kappa = 1.15$; $\varepsilon = 1$ percentage point

Development confirmation. Search uses $(256, 128, 128)$ common/retention/correction requests per model. Confirmation adds $(256, 128, 128)$ requests for 24 candidates, then $(512, 256, 256)$ for six, reaching 512 and 1024 common-condition questions. Promotion retains the reference and sources alongside high-accuracy, high-discernment, or low-token candidates. Thresholds use stage-specific scores; ties favor accuracy, fewer tokens, then a stable identifier. Confirmation and test outcomes are excluded from search training.

A.5 Merge Parameterization and Variants

Grouping and initialization. For the 28-layer reference model, zero-indexed layers 0–6, 7–13, 14–20, and 21–27 form four depth blocks. Separating attention/input normalization from MLP/post-

attention normalization within each block, plus one global group for embeddings/final normalization/output head, gives $G = 9$ and $d = 18$. Figure 2 indexes the global group by $g = 1$, followed by attention/MLP groups $g = 2, 3$ through $g = 8, 9$ in block-depth order.

The $N_0 = 24$ initial configurations comprise Short, the intermediate model, Long, five uniform pairs $(0.25, 0.25)$, $(0.5, 0.5)$, $(0.75, 0.75)$, $(1, 0.5)$, and $(0.5, 1)$, and 16 scrambled Sobol points using the experiment seed. The reference setting uses TIES (Yadav et al., 2023) as both intermediate model and reference. Merges accumulate in FP32 and are stored in BF16.

Two-source and tied-block variants. Without an intermediate model, the same search loop uses one coefficient per group:

$$\theta_{\mathrm{merge},g}(x) = (1 - \lambda_g)\theta_{\mathrm{short},g} + \lambda_g \theta_{\mathrm{long},g}, \qquad x = (\lambda_1, \ldots, \lambda_G) \in [0, 1]^G. \tag{15}$$

An evaluated arithmetic merge (Ilharco et al., 2023) can serve as reference. The block variant ties attention/MLP coefficients, with embeddings in the first block and final normalization/output head in the last. Controlled intermediate-model comparisons fix the Average reference and compute $\alpha_g = \beta_g$ configurations directly to preserve identical initial checkpoints.

B Motivation Scan: Setup and Detailed Results

Coefficient grids. Using Equation 15, we evaluate global TA at $\lambda \in \{0, 0.1, \ldots, 1\}$ and all four-block combinations in $\{0, 0.5, 1\}^4$. The three uniform vectors at 0, 0.5, and 1 are shared, yielding $11 + 81 - 3 = 89$ distinct configurations per family (178 total). This grid is fixed before evaluation.

Diagnostic evaluation. The 512 questions comprise 234 MATH training, 214 MATH test-remainder, and 64 GSM8K training examples, with 64 per stratum: basic algebra, advanced algebra, intermediate algebra, prealgebra, number theory, counting/probability, geometry/precalculus, and word arithmetic. This previously observed subset is diagnostic, not an untouched test set. Each configuration receives 512 common, 256 retention, and 256 correction evaluations, with fixed feedback histories. Metric definitions follow Section 3.1; this diagnostic averages over questions within each condition, rather than the eight benchmarks used in the main evaluation.

Figure and pair selection. The left panel of Figure 1 retains 76 Qwen2.5 and 67 Phi-4-mini configurations with $\mathrm{Tok} \leq 4000$; shared grid points appear once as diamonds. The right panel uses all 81 block configurations, averaging the 27 combinations of the other blocks per cell. These marginal averages do not resolve block interactions.

For Table 6, we select the largest discernment gap per family among pairs with $\mathrm{Acc} \geq 80$, $\mathrm{Tok} \leq 4000$, an accuracy gap ≤ 1 percentage point, and a token gap $\leq 10\%$ of the smaller value. Qwen2.5 differs by 0.39 accuracy points and 31.0 tokens but 58.40 discernment points; Phi-4-mini differs by 0.78 accuracy points and 11.4 tokens but 24.41 discernment points.

Table conventions. All tables report correctness percentages and mean common-condition tokens, with coefficients in block-depth order. Inv. and Trunc. count invalid and truncated outputs among 1024 evaluations; these counts overlap. G, B, and G/B denote global TA, block TA, and shared configurations. Tables 7–8 retain all 178 configurations, rounded only for display.

Table 6: Illustrative pairs with similar accuracy and token cost but different discernment.

$(\lambda_1, \lambda_2, \lambda_3, \lambda_4)$	Acc	Tok	Ret	Cor	Disc	Inv.	Trunc.
Qwen2.5							
$(0, 0, 0.5, 0.5)$	86.52	909.3	21.48	46.48	33.98	326	326
$(1, 0.5, 1, 0)$	86.91	940.3	93.36	91.41	92.38	6	6
Phi-4-mini							
$(1, 1, 0, 1)$	85.16	2147.7	64.84	73.83	69.34	92	92
$(1, 1, 1, 0.5)$	85.94	2159.1	92.19	95.31	93.75	54	54

B.1 QWEN2.5 CONFIGURATION RESULTS

Table 7: Qwen2.5 motivation scan: all 89 distinct configurations.

Grid	$(\lambda_1, \lambda_2, \lambda_3, \lambda_4)$	Acc	Tok	Ret	Cor	Disc	Inv.	Trunc.
G/B	$(0, 0, 0, 0)$	85.55	583.3	33.20	59.77	46.48	115	115
G	$(0.1, 0.1, 0.1, 0.1)$	86.91	607.8	40.23	57.03	48.63	123	122
G	$(0.2, 0.2, 0.2, 0.2)$	89.06	634.0	51.56	71.09	61.33	89	89
G	$(0.3, 0.3, 0.3, 0.3)$	90.43	657.8	67.58	85.55	76.56	46	46
G	$(0.4, 0.4, 0.4, 0.4)$	90.04	740.6	72.66	90.23	81.45	51	51
G/B	$(0.5, 0.5, 0.5, 0.5)$	91.02	806.6	73.05	84.77	78.91	77	77
G	$(0.6, 0.6, 0.6, 0.6)$	92.19	864.5	72.66	78.52	75.59	121	121
G	$(0.7, 0.7, 0.7, 0.7)$	93.55	1159.3	67.97	69.14	68.55	173	173
G	$(0.8, 0.8, 0.8, 0.8)$	92.97	1614.5	85.16	89.06	87.11	87	87
G	$(0.9, 0.9, 0.9, 0.9)$	91.80	2283.2	82.81	93.36	88.09	92	92
G/B	$(1, 1, 1, 1)$	90.04	2859.7	68.36	85.16	76.76	161	161
B	$(0, 0, 0, 0.5)$	79.69	911.8	17.97	40.62	29.30	318	317
B	$(0, 0, 0, 1)$	18.55	6444.6	0.39	0.39	0.39	902	902
B	$(0, 0, 0.5, 0)$	88.67	706.5	62.50	76.95	69.73	76	76
B	$(0, 0, 0.5, 0.5)$	86.52	909.3	21.48	46.48	33.98	326	326
B	$(0, 0, 0.5, 1)$	2.34	8002.3	0.39	0.39	0.39	1008	1008
B	$(0, 0, 1, 0)$	88.67	1168.5	73.83	82.03	77.93	80	80
B	$(0, 0, 1, 0.5)$	50.00	4198.1	14.84	35.55	25.20	604	604
B	$(0, 0, 1, 1)$	1.17	8130.7	1.56	3.12	2.34	1006	1006
B	$(0, 0.5, 0, 0)$	89.65	618.8	48.83	77.73	63.28	55	55
B	$(0, 0.5, 0, 0.5)$	86.72	735.6	39.84	70.70	55.27	176	176
B	$(0, 0.5, 0, 1)$	8.59	7491.0	0.00	0.00	0.00	977	977
B	$(0, 0.5, 0.5, 0)$	90.82	709.9	79.30	89.06	84.18	19	19
B	$(0, 0.5, 0.5, 0.5)$	87.70	981.8	57.42	76.95	67.19	166	166
B	$(0, 0.5, 0.5, 1)$	7.23	7689.8	2.34	7.42	4.88	960	960
B	$(0, 0.5, 1, 0)$	88.87	1271.8	84.38	90.62	87.50	43	43
B	$(0, 0.5, 1, 0.5)$	53.91	3947.8	46.88	61.72	54.30	449	449
B	$(0, 0.5, 1, 1)$	9.18	7753.6	7.03	9.38	8.20	931	931
B	$(0, 1, 0, 0)$	89.26	669.0	64.84	81.25	73.05	28	28
B	$(0, 1, 0, 0.5)$	89.06	701.1	60.16	69.92	65.04	135	135
B	$(0, 1, 0, 1)$	15.43	6913.9	2.73	5.47	4.10	915	915
B	$(0, 1, 0.5, 0)$	91.02	798.5	86.33	91.41	88.87	18	18
B	$(0, 1, 0.5, 0.5)$	89.45	1005.7	79.30	72.27	75.78	131	131
B	$(0, 1, 0.5, 1)$	23.44	6560.0	6.64	10.94	8.79	857	857
B	$(0, 1, 1, 0)$	85.55	1178.2	82.03	93.75	87.89	40	40
B	$(0, 1, 1, 0.5)$	60.74	3431.9	59.38	58.98	59.18	394	394
B	$(0, 1, 1, 1)$	21.29	6916.3	7.03	6.25	6.64	881	881
B	$(0.5, 0, 0, 0)$	87.50	594.0	33.59	77.73	55.66	42	42
B	$(0.5, 0, 0, 0.5)$	88.09	630.4	27.73	68.75	48.24	176	176
B	$(0.5, 0, 0, 1)$	59.77	2903.5	1.17	0.78	0.98	664	664
B	$(0.5, 0, 0.5, 0)$	89.65	713.5	71.88	84.77	78.32	17	17
B	$(0.5, 0, 0.5, 0.5)$	90.43	820.5	48.05	74.61	61.33	149	149
B	$(0.5, 0, 0.5, 1)$	56.25	3866.4	1.17	4.30	2.73	703	703
B	$(0.5, 0, 1, 0)$	87.70	1077.8	79.69	85.94	82.81	33	33
B	$(0.5, 0, 1, 0.5)$	79.30	2063.0	42.19	67.19	54.69	284	284
B	$(0.5, 0, 1, 1)$	66.21	3455.8	8.20	28.52	18.36	582	582
B	$(0.5, 0.5, 0, 0)$	86.91	606.5	42.58	81.64	62.11	44	43
B	$(0.5, 0.5, 0, 0.5)$	89.45	660.1	31.64	62.50	47.07	186	186
B	$(0.5, 0.5, 0, 1)$	35.74	5324.5	0.78	0.78	0.78	827	827
B	$(0.5, 0.5, 0.5, 0)$	86.91	738.6	79.30	89.06	84.18	7	7
B	$(0.5, 0.5, 0.5, 1)$	79.49	2283.0	8.20	17.19	12.70	536	536
B	$(0.5, 0.5, 1, 0)$	90.23	1014.4	91.02	91.41	91.21	15	15
B	$(0.5, 0.5, 1, 0.5)$	73.05	2514.4	73.05	79.69	76.37	221	221
B	$(0.5, 0.5, 1, 1)$	81.84	2724.1	51.95	73.05	62.50	278	278
B	$(0.5, 1, 0, 0)$	88.09	644.5	65.62	83.98	74.80	12	12
B	$(0.5, 1, 0, 0.5)$	91.02	694.5	64.84	66.80	65.82	142	142
B	$(0.5, 1, 0, 1)$	37.11	5196.2	5.08	10.16	7.62	778	778

Continued on next page

Table 7 (continued)

Grid	$(\lambda_1, \lambda_2, \lambda_3, \lambda_4)$	Acc	Tok	Ret	Cor	Disc	Inv.	Trunc.
B	$(0.5, 1, 0.5, 0)$	90.82	788.0	88.28	93.75	91.02	13	13
B	$(0.5, 1, 0.5, 0.5)$	90.23	856.6	83.59	86.33	84.96	64	64
B	$(0.5, 1, 0.5, 1)$	89.84	1697.2	34.77	53.12	43.95	318	318
B	$(0.5, 1, 1, 0)$	84.77	1071.4	86.72	93.36	90.04	22	22
B	$(0.5, 1, 1, 0.5)$	69.34	2852.5	76.56	76.95	76.76	255	255
B	$(0.5, 1, 1, 1)$	83.79	3002.6	69.92	78.12	74.02	211	211
B	$(1, 0, 0, 0)$	84.96	562.4	50.39	84.77	67.58	13	13
B	$(1, 0, 0, 0.5)$	84.77	603.9	37.50	66.80	52.15	154	154
B	$(1, 0, 0, 1)$	78.12	1062.2	1.17	6.25	3.71	522	522
B	$(1, 0, 0.5, 0)$	85.35	708.4	77.34	86.72	82.03	12	12
B	$(1, 0, 0.5, 0.5)$	88.28	849.7	54.30	76.95	65.62	106	106
B	$(1, 0, 0.5, 1)$	71.29	2428.9	3.12	13.28	8.20	582	582
B	$(1, 0, 1, 0)$	87.89	1062.3	88.67	89.84	89.26	22	22
B	$(1, 0, 1, 0.5)$	79.88	1871.4	53.52	71.88	62.70	236	236
B	$(1, 0, 1, 1)$	75.98	2790.1	25.39	46.88	36.13	437	437
B	$(1, 0.5, 0, 0)$	84.77	602.4	58.20	78.12	68.16	32	32
B	$(1, 0.5, 0, 0.5)$	85.55	626.1	39.45	47.27	43.36	220	220
B	$(1, 0.5, 0, 1)$	59.96	2845.7	0.00	0.78	0.39	659	659
B	$(1, 0.5, 0.5, 0)$	87.11	756.6	86.72	86.33	86.52	3	3
B	$(1, 0.5, 0.5, 0.5)$	89.06	729.2	70.31	76.95	73.63	105	104
B	$(1, 0.5, 0.5, 1)$	83.59	1915.3	27.73	32.03	29.88	420	420
B	$(1, 0.5, 1, 0)$	86.91	940.3	93.36	91.41	92.38	6	6
B	$(1, 0.5, 1, 0.5)$	72.27	2588.6	67.19	67.58	67.38	281	281
B	$(1, 0.5, 1, 1)$	85.35	2594.1	71.09	78.91	75.00	194	194
B	$(1, 1, 0, 0)$	76.56	975.6	64.84	73.05	68.95	62	62
B	$(1, 1, 0, 0.5)$	82.81	866.6	46.88	51.17	49.02	221	221
B	$(1, 1, 0, 1)$	27.93	5787.4	3.91	5.08	4.49	832	832
B	$(1, 1, 0.5, 0)$	84.57	1170.4	81.64	85.55	83.59	57	57
B	$(1, 1, 0.5, 0.5)$	87.70	1065.7	62.89	68.36	65.62	179	179
B	$(1, 1, 0.5, 1)$	92.19	1668.4	51.17	66.02	58.59	234	234
B	$(1, 1, 1, 0)$	82.42	1261.8	72.66	83.20	77.93	95	95
B	$(1, 1, 1, 0.5)$	61.72	3559.8	68.75	68.36	68.55	336	336

B.2 PHI-4-MINI CONFIGURATION RESULTS

Table 8: Phi-4-mini motivation scan: all 89 distinct configurations.

Grid	$(\lambda_1, \lambda_2, \lambda_3, \lambda_4)$	Acc	Tok	Ret	Cor	Disc	Inv.	Trunc.
G/B	$(0, 0, 0, 0)$	74.41	775.7	72.66	79.69	76.17	38	38
G	$(0.1, 0.1, 0.1, 0.1)$	76.95	762.0	79.30	82.42	80.86	32	32
G	$(0.2, 0.2, 0.2, 0.2)$	79.10	784.3	79.30	82.81	81.05	33	33
G	$(0.3, 0.3, 0.3, 0.3)$	82.42	796.4	82.03	86.72	84.38	32	32
G	$(0.4, 0.4, 0.4, 0.4)$	83.20	832.4	87.11	86.72	86.91	31	31
G/B	$(0.5, 0.5, 0.5, 0.5)$	83.01	856.5	87.50	91.02	89.26	32	32
G	$(0.6, 0.6, 0.6, 0.6)$	81.25	1363.1	87.11	91.02	89.06	53	53
G	$(0.7, 0.7, 0.7, 0.7)$	89.26	1851.7	87.89	94.92	91.41	63	63
G	$(0.8, 0.8, 0.8, 0.8)$	91.99	1885.7	89.84	95.31	92.58	50	50
G	$(0.9, 0.9, 0.9, 0.9)$	91.80	2148.5	88.67	96.48	92.58	65	65
G/B	$(1, 1, 1, 1)$	92.58	2412.9	91.02	96.09	93.55	56	56
B	$(0, 0, 0, 0.5)$	68.55	1364.9	66.80	74.22	70.51	148	148
B	$(0, 0, 0, 1)$	31.25	5444.4	11.33	9.77	10.55	787	787
B	$(0, 0, 0.5, 0)$	77.73	705.6	78.52	83.98	81.25	30	30
B	$(0, 0, 0.5, 0.5)$	67.97	1613.5	65.62	69.92	67.77	209	209
B	$(0, 0, 0.5, 1)$	20.90	6440.5	3.91	8.20	6.05	879	879
B	$(0, 0, 1, 0)$	73.63	945.1	77.73	78.52	78.12	75	75
B	$(0, 0, 1, 0.5)$	58.59	2998.1	43.75	60.55	52.15	411	411
B	$(0, 0, 1, 1)$	9.38	7404.3	3.91	11.72	7.81	933	933
B	$(0, 0.5, 0, 0)$	78.32	799.0	81.64	84.77	83.20	35	35

Continued on next page

Table 8 (continued)

Grid	$(\lambda_1, \lambda_2, \lambda_3, \lambda_4)$	Acc	Tok	Ret	Cor	Disc	Inv.	Trunc.
B	$(0, 0.5, 0, 0.5)$	75.00	1603.9	64.84	78.52	71.68	191	191
B	$(0, 0.5, 0, 1)$	19.73	6605.6	0.78	3.52	2.15	910	910
B	$(0, 0.5, 0.5, 0)$	81.64	837.7	84.77	85.94	85.35	31	31
B	$(0, 0.5, 0.5, 0.5)$	70.70	1984.3	61.33	75.39	68.36	256	256
B	$(0, 0.5, 0.5, 1)$	17.77	6784.2	1.95	2.73	2.34	921	921
B	$(0, 0.5, 1, 0)$	80.27	1082.6	87.11	86.72	86.91	59	59
B	$(0, 0.5, 1, 0.5)$	40.43	4842.3	32.42	60.16	46.29	567	567
B	$(0, 0.5, 1, 1)$	5.47	7756.0	5.08	17.58	11.33	938	938
B	$(0, 1, 0, 0)$	75.78	1041.8	73.05	76.17	74.61	93	93
B	$(0, 1, 0, 0.5)$	58.98	3128.9	25.39	49.22	37.30	495	495
B	$(0, 1, 0, 1)$	9.38	7432.0	0.00	0.00	0.00	975	975
B	$(0, 1, 0.5, 0)$	72.46	970.1	75.39	77.73	76.56	74	74
B	$(0, 1, 0.5, 0.5)$	47.66	4279.7	17.58	44.53	31.05	603	603
B	$(0, 1, 0.5, 1)$	7.23	7629.1	0.78	3.91	2.34	974	974
B	$(0, 1, 1, 0)$	53.12	1499.9	80.08	81.64	80.86	113	113
B	$(0, 1, 1, 0.5)$	15.04	6436.9	16.80	35.55	26.17	773	773
B	$(0, 1, 1, 1)$	8.20	7546.6	3.52	14.06	8.79	927	927
B	$(0.5, 0, 0, 0)$	72.66	1310.5	68.36	76.56	72.46	90	90
B	$(0.5, 0, 0, 0.5)$	78.12	835.0	75.00	82.42	78.71	36	36
B	$(0.5, 0, 0, 1)$	64.06	2353.9	36.33	54.30	45.31	369	369
B	$(0.5, 0, 0.5, 0)$	74.22	1279.3	76.95	80.47	78.71	81	81
B	$(0.5, 0, 0.5, 0.5)$	78.71	892.6	80.47	82.81	81.64	43	43
B	$(0.5, 0, 0.5, 1)$	54.49	3515.6	41.41	60.94	51.17	442	442
B	$(0.5, 0, 1, 0)$	69.73	1459.8	76.56	76.95	76.76	106	106
B	$(0.5, 0, 1, 0.5)$	75.78	1352.6	75.78	81.64	78.71	119	119
B	$(0.5, 0, 1, 1)$	43.95	4557.8	39.84	66.41	53.12	501	501
B	$(0.5, 0.5, 0, 0)$	78.52	1114.8	69.92	82.81	76.37	74	74
B	$(0.5, 0.5, 0, 0.5)$	83.59	709.2	86.72	87.89	87.30	23	23
B	$(0.5, 0.5, 0, 1)$	71.48	2259.3	33.59	62.89	48.24	376	376
B	$(0.5, 0.5, 0.5, 0)$	79.69	1121.9	74.61	87.11	80.86	62	62
B	$(0.5, 0.5, 0.5, 1)$	78.32	2405.3	60.55	84.38	72.46	232	232
B	$(0.5, 0.5, 1, 0)$	70.51	1404.5	76.95	84.77	80.86	95	95
B	$(0.5, 0.5, 1, 0.5)$	75.78	1697.5	87.50	88.28	87.89	100	100
B	$(0.5, 0.5, 1, 1)$	83.01	2191.6	75.00	87.11	81.05	163	163
B	$(0.5, 1, 0, 0)$	59.57	1653.5	66.41	71.48	68.95	151	151
B	$(0.5, 1, 0, 0.5)$	65.82	695.4	65.23	81.25	73.24	60	60
B	$(0.5, 1, 0, 1)$	63.28	3669.4	20.31	66.80	43.55	467	467
B	$(0.5, 1, 0.5, 0)$	35.94	835.7	65.23	76.17	70.70	97	97
B	$(0.5, 1, 0.5, 0.5)$	58.40	783.9	80.47	88.28	84.38	44	43
B	$(0.5, 1, 0.5, 1)$	77.93	2883.1	46.48	77.73	62.11	296	296
B	$(0.5, 1, 1, 0)$	33.01	1623.1	46.09	57.42	51.76	234	234
B	$(0.5, 1, 1, 0.5)$	65.23	1886.9	78.12	90.62	84.38	101	101
B	$(0.5, 1, 1, 1)$	76.95	3086.0	57.81	80.86	69.34	267	267
B	$(1, 0, 0, 0)$	20.51	6428.2	8.98	12.50	10.74	831	831
B	$(1, 0, 0, 0.5)$	75.20	1054.9	67.97	69.92	68.95	46	46
B	$(1, 0, 0, 1)$	73.24	1428.7	53.12	64.84	58.98	195	195
B	$(1, 0, 0.5, 0)$	25.78	5997.0	15.62	33.20	24.41	720	720
B	$(1, 0, 0.5, 0.5)$	78.71	896.5	74.61	81.25	77.93	36	36
B	$(1, 0, 0.5, 1)$	62.50	2782.7	65.23	77.73	71.48	219	219
B	$(1, 0, 1, 0)$	14.06	6988.4	12.50	19.14	15.82	850	850
B	$(1, 0, 1, 0.5)$	74.02	1710.0	75.39	81.64	78.52	82	82
B	$(1, 0, 1, 1)$	58.98	3356.8	72.27	78.91	75.59	261	261
B	$(1, 0.5, 0, 0)$	8.59	7424.3	9.77	5.08	7.42	795	795
B	$(1, 0.5, 0, 0.5)$	80.47	749.8	57.03	59.77	58.40	27	27
B	$(1, 0.5, 0, 1)$	79.30	1622.8	76.95	85.16	81.05	107	107
B	$(1, 0.5, 0.5, 0)$	14.45	6895.8	14.84	10.55	12.70	756	756
B	$(1, 0.5, 0.5, 0.5)$	86.13	949.4	88.28	90.62	89.45	20	20
B	$(1, 0.5, 0.5, 1)$	87.30	1624.8	88.28	93.36	90.82	60	60
B	$(1, 0.5, 1, 0)$	5.47	7505.6	10.94	16.80	13.87	895	895
B	$(1, 0.5, 1, 0.5)$	89.84	1672.9	91.02	93.36	92.19	31	31
B	$(1, 0.5, 1, 1)$	91.02	1863.7	93.36	95.70	94.53	48	48

Continued on next page

Table 8 (continued)

Grid	$(\lambda_1, \lambda_2, \lambda_3, \lambda_4)$	Acc	Tok	Ret	Cor	Disc	Inv.	Trunc.
B	$(1, 1, 0, 0)$	0.20	7717.1	1.17	1.95	1.56	960	960
B	$(1, 1, 0, 0.5)$	38.09	492.1	57.81	70.31	64.06	39	39
B	$(1, 1, 0, 1)$	85.16	2147.7	64.84	73.83	69.34	92	92
B	$(1, 1, 0.5, 0)$	1.95	6911.0	0.78	3.52	2.15	904	904
B	$(1, 1, 0.5, 0.5)$	87.89	1674.5	89.45	91.41	90.43	25	25
B	$(1, 1, 0.5, 1)$	90.43	2175.4	88.67	96.48	92.58	63	63
B	$(1, 1, 1, 0)$	1.76	7436.8	2.34	2.34	2.34	949	949
B	$(1, 1, 1, 0.5)$	85.94	2159.1	92.19	95.31	93.75	54	54

C Experimental Configuration and Evaluation Protocol

C.1 Source Models and Evaluation Coverage

Table 9: Source models and checkpoint identifiers.

Family	Role	Public checkpoint identifier
Qwen2.5	Short	Qwen/Qwen2.5-Math-7B[1] (Yang et al., 2024a)
	Long	deepseek-ai/DeepSeek-R1-Distill-Qwen-7B[2] (DeepSeek-AI, 2025)
Qwen3	Short	Qwen/Qwen3-4B-Instruct-2507[3] (Qwen Team, 2025a)
	Long	Qwen/Qwen3-4B-Thinking-2507[4] (Qwen Team, 2025b)
Phi-4-mini	Short	microsoft/Phi-4-mini-instruct[5] (Microsoft, 2025a)
	Long	microsoft/Phi-4-mini-reasoning[6] (Microsoft, 2025b)

Short and Long denote paired roles. Qwen2.5 Short is a base mathematical model. Qwen2.5 Short uses its native context; the other mathematical rows share a common output cap. Qwen3 uses the common Thinking tokenizer/runtime with raw prompts; Phi uses its native chat template and the common Reasoning runtime. These are fixed family-specific evaluation settings.

Table 10: Mathematical evaluation subsets and feedback denominators.

Benchmark	Clean	Retention records	Correction records
GSM8K (Cobbe et al., 2021)	1,319	659	660
MATH500 (Hendrycks et al., 2021; Hugging Face H4, 2024)	500	500	500
Minerva Math (Lewkowycz et al., 2022)	272	272	272
OlympiadBench (He et al., 2024)	675	337	338
SVAMP (Patel et al., 2021)	1,000	500	500
ASDiv (Miao et al., 2020)	2,215	1,107	1,108
AMC23 (Mathematical Association of America, n.d.)	40	40	40
AIME24 (Mathematical Association of America, n.d.)	30	30	30
Total	6,051	3,445	3,448

Feedback conditions supply fixed histories of known correctness, so their denominators do not depend on each model's clean-answer accuracy. Counts describe the frozen evaluation subsets, not

[1] https://huggingface.co/Qwen/Qwen2.5-Math-7B

[2] https://huggingface.co/deepseek-ai/DeepSeek-R1-Distill-Qwen-7B

[3] https://huggingface.co/Qwen/Qwen3-4B-Instruct-2507

[4] https://huggingface.co/Qwen/Qwen3-4B-Thinking-2507

[5] https://huggingface.co/microsoft/Phi-4-mini-instruct

[6] https://huggingface.co/microsoft/Phi-4-mini-reasoning

complete upstream datasets. Every dataset has macro weight $1/8$, including the small competition subsets.

C.2 Search, Confirmation and Final Evaluation

Table 11: DISCO stages and incremental logical evaluation records.

Stage	Models	Incremental logical records/model	Decision supported
Search	104	512	Candidate proposals and allocation
Confirm to development512	24	512	Promote six models
Confirm to development1024	6	1,024	Choose one model
Full mathematical evaluation	1	12,944	Report frozen output

Stage counts are incremental logical records, not cumulative set sizes or necessarily newly generated outputs. Cached outputs are counted logically without regeneration; the measured reuse is separated in Appendix K. Search256 contains 256 clean, 128 misleading-feedback and 128 valid-feedback records. Confirmation adds disjoint chunks and recomputes cumulative scores; only search observations fit the GP and generator. Selection requires $\mathrm{Tok} \leq 1.15\mathrm{Tok}_{\mathrm{TIES}}$ and $\mathrm{Acc} \geq \max_{\mathrm{feasible}} \mathrm{Acc} - 1$, then maximizes Disc (ties: higher Acc, lower Tok, identifier). Full evaluation follows the identity lock; these benchmarks were previously observed.

C.3 Baseline Recipes, Budgets and Reporting Rules

Method years refer to publications. SLERP cites MergeKit (2024); KARCHER cites its method paper (Wang et al., 2026). DISCO and its controls are this work.

Table 12: Baseline merging recipes and adaptations.

Method	Configuration or adaptation
Average (22)	Equal Short–Long weights; identical to the verified Model Soup checkpoint in the source table.
TA (23)	Qwen2.5 retains the historical public-Base-referenced configuration. Qwen3 and Phi use Short as the task-vector reference: $S + 0.25(S - S) + 0.70(L - S)$.
TIES (23)	Qwen2.5 uses the exact historical TIES midpoint employed by DISCO. Qwen3/Phi displayed baselines use scale 0.5 and drop rate 0.2; their internal DISCO references use scale 1.0 and drop rate 0.8.
SLERP (24)	Short–Long midpoint, $t = 0.5$.
KARCHER (26)	Iteration limit 10, tolerance 10^{-5}; the frozen kernel matches the author-linked source revision `09bbb0ae` (Wang et al., 2026).
ACM (25)	Consensus threshold 1.2; TA and TIES parent variants shown separately. Calibration follows the fixed source configuration.
Sens-Merging (25)	$\alpha = 0.7$, $\tau = 3$, multiplier 2; 100 calibration examples, length 2,048. The reported results use the documented BF16 calibration adaptation.
Evo-L2S (26)	NSGA-II, population 20, ten generations, seed 42; maximize Acc and minimize Tok, with a separate 512-clean-question confirmation set.
EvoGM (26)	Qwen: Short-base scalar space, population 14, two rounds, four generator iterations per round, seed 0; select maximum search singleton HV. Phi settings are in Appendix F.

Table 13: Qwen2.5 search budgets and output rules.

Search	Candidate-evaluation calls	Space	Confirmation and output
DISCO	104	18D	$24 \to 6 \to 1$; token gate, accuracy floor, then Disc.
EvoGM (26)	140	1D	No extra confirmation; maximum search singleton HV.
Evo-L2S (26)	189 (200 nominal)	TA	Seven final-population models confirmed; highest Acc and lowest Tok within 2 accuracy points retained.
Direct BO	22	4D	23 candidates including external TIES, then six, then one under DISCO's rule.

Evo-L2S makes 189 optimizer evaluations ($9 + 9 \times 20$), confirms seven models and selects two; the original representative is retained. Qwen2.5 EvoGM measures 138 distinct candidates for 140 charged requests. Round two propagates the previous winner's direction with Short fixed: $\alpha = z\alpha_{\text{prev}}$, where z is the new scalar and α_{prev} is the previous winner's coefficient. The selected coefficient is 0.48806446791. All figures and tables use the name EvoGM. Qwen2.5 and Qwen3 use HV selection; Phi is the Accuracy-selection exception described in Appendix F. The separate Accuracy-selected Qwen2.5 branch is excluded.

C.4 Metric and Statistical Conventions

We use Accuracy (Acc), Tokens (Tok), Discernment (Disc), Retention (Ret) and Correction (Cor). Best comparable values are bold on light gray; second-best values are underlined. Tok is minimized; other performance metrics are maximized. Ties at displayed precision are retained, and highlights do not imply significance. Stacked cells follow caption order; HV uses macro scores. $\mathrm{Acc}/\mathrm{Tok}$ use clean questions; $\mathrm{Ret}/\mathrm{Cor}$ use the two fixed feedback conditions, with $\mathrm{Disc} = (\mathrm{Ret} + \mathrm{Cor})/2$. Macro metrics weight datasets equally. With reference $(0, 0, 0)$, singleton hypervolume is

$$H(\mathrm{Acc}, \mathrm{Tok}, \mathrm{Disc}) = \frac{\mathrm{Acc}}{100}\left(1 - \frac{\mathrm{Tok}}{10240}\right)\frac{\mathrm{Disc}}{100}.$$

Numerical comparisons and reported HV use unrounded metrics; HV uses token reference 10,240.

D Complete Qwen2.5-7B Results

Tables 14 and 15 retain all 13 configurations across eight benchmarks. Qwen2.5-Math and R1-Distill abbreviate Qwen2.5-Math-7B and DeepSeek-R1-Distill-Qwen-7B. The native-context Short row is a reference and excluded from common-cap ranking.

Table 14: Qwen2.5-7B performance (Acc/Tok/Disc) by benchmark, with Macro and HV.

Model / Method	GSM8K	MATH500	Minerva	Olympiad	SVAMP	ASDiv	AMC23	AIME24	Macro equal weight	HV↑
Qwen2.5-Math (Short)	74.22	71.40	26.10	33.93	47.10	62.12	47.50	16.67	47.38	0.18370
	334.8	652.4	1142.5	948.4	131.4	188.5	1068.8	1676.1	767.9	
	37.21	32.80	37.32	37.59	57.20	67.36	32.50	33.33	41.91	
R1-Distill (Long)	83.24	81.00	36.76	41.93	80.00	82.57	70.00	**43.33**	64.86	0.25391
	1594.9	3236.2	4590.6	5573.6	1710.0	1398.8	4701.8	6738.4	3693.0	
	71.33	66.00	49.82	39.25	84.80	86.18	57.50	35.00	61.23	
Average (22)	89.92	82.60	29.78	47.11	93.50	95.08	70.00	16.67	65.58	0.33846
	413.6	1191.3	2285.5	2294.4	296.7	**251.4**	**1789.8**	5070.7	1699.2	
	78.85	71.90	50.55	51.84	79.40	80.81	55.00	26.67	61.88	
TA (23)	**91.89**	84.80	31.62	44.30	93.60	94.76	**75.00**	20.00	67.00	0.26989
	594.4	1810.3	3240.2	3827.2	448.7	412.6	2958.7	6112.2	2425.5	
	65.12	68.60	41.36	51.11	52.30	51.74	58.75	33.33	52.79	
TIES (23)	89.54	83.60	30.15	44.59	92.90	94.76	70.00	23.33	66.11	0.36566
	378.9	1185.9	1686.9	2343.2	345.9	284.1	2008.8	4009.1	1530.4	
	83.55	77.90	53.31	57.18	74.20	74.94	67.50	31.67	65.03	
SLERP (24)	90.14	82.80	34.19	46.96	**94.10**	95.08	67.50	16.67	65.93	0.34967
	401.6	1140.5	2164.6	2309.2	**244.5**	256.9	1890.2	3785.7	1524.2	
	78.62	72.90	50.37	51.69	80.40	81.17	60.00	23.33	62.31	
Sens-Merging (25)	91.05	86.60	29.41	43.41	92.50	**95.12**	65.00	23.33	65.80	0.26795
	601.7	1667.2	3164.8	3870.5	482.7	370.7	3164.1	5631.3	2369.1	
	63.91	71.80	50.37	52.59	44.80	40.77	66.25	33.33	52.98	
ACM-TA (25)	90.30	84.80	31.25	46.37	89.40	92.87	60.00	33.33	66.04	0.29411
	1496.0	2949.6	4474.3	5140.9	1477.5	1180.3	4569.4	6993.6	3535.2	
	73.68	81.40	**58.82**	56.14	84.50	88.76	67.50	33.33	68.02	
ACM-TIES (25)	89.39	79.20	29.04	43.26	94.00	94.27	57.50	20.00	63.33	0.36519
	370.7	1136.9	2225.9	**1846.6**	266.7	269.3	2552.2	**2624.3**	1411.6	
	82.10	75.30	55.70	57.02	85.30	88.80	57.50	33.33	66.88	
KARCHER (26)	90.14	83.00	30.88	45.04	93.50	94.90	60.00	23.33	65.10	0.35281
	391.9	1160.8	2192.3	2155.8	263.4	262.1	1939.2	4610.6	1622.0	
	78.32	73.70	51.29	52.13	79.50	81.49	58.75	**40.00**	64.40	
Evo-L2S (26)	90.98	83.80	31.62	47.56	**94.10**	94.72	72.50	13.33	66.08	0.29954
	419.2	1317.9	2402.2	2599.0	270.3	283.4	2079.3	4725.0	1762.0	
	75.36	67.70	39.15	49.32	64.30	66.36	52.50	23.33	54.75	
EvoGM (26)	89.99	82.80	32.72	45.48	93.50	94.90	67.50	20.00	65.86	0.36640
	387.3	1193.3	1936.7	2063.0	269.4	261.4	2159.1	3791.0	1507.6	
	79.00	74.10	54.23	54.51	79.10	82.21	63.75	35.00	65.24	
DISCO-Merge	91.81	**86.80**	**37.87**	**49.04**	93.40	94.49	67.50	23.33	**68.03**	**0.44130**
	364.8	**1123.5**	**1446.8**	2157.2	257.9	266.8	1821.2	3615.5	**1381.7**	
	92.64	**87.10**	57.17	**67.84**	**90.90**	**93.41**	**72.50**	38.33	**74.99**	

Table 15: Qwen2.5-7B feedback (Ret/Cor) by benchmark, with Macro and HV.

Model / Method	GSM8K	MATH500	Minerva	Olympiad	SVAMP	ASDiv	AMC23	AIME24	Macro equal weight	HV↑
Qwen2.5-Math (Short)	21.85 52.58	20.80 44.80	22.79 51.84	11.28 63.91	49.60 64.80	60.34 74.37	10.00 55.00	0.00 66.67	24.58 59.24	0.18370
R1-Distill (Long)	60.24 82.42	60.00 72.00	43.01 56.62	30.56 47.93	77.40 92.20	79.58 92.78	47.50 67.50	36.67 33.33	54.37 68.10	0.25391
Average (22)	75.42 82.27	65.40 78.40	46.69 54.41	41.54 62.13	76.40 82.40	76.69 84.93	42.50 67.50	26.67 26.67	56.41 67.34	0.33846
TA (23)	59.94 70.30	67.20 70.00	39.71 43.01	48.96 53.25	44.60 60.00	47.52 55.96	55.00 62.50	33.33 33.33	49.53 56.05	0.26989
TIES (23)	80.27 86.82	74.60 81.20	52.94 53.68	52.82 61.54	64.80 83.60	67.03 82.85	70.00 65.00	30.00 33.33	61.56 68.50	0.36566
SLERP (24)	76.93 80.30	69.00 76.80	43.38 57.35	41.25 62.13	78.80 82.00	76.96 85.38	52.50 67.50	16.67 30.00	56.94 67.68	0.34967
Sens-Merging (25)	59.33 68.48	70.60 73.00	53.68 47.06	48.07 57.10	41.80 47.80	38.93 42.60	65.00 67.50	20.00 46.67	49.68 56.28	0.26795
ACM-TA (25)	59.94 87.42	79.80 83.00	**59.19** 58.46	53.12 59.17	78.40 90.60	85.00 92.51	**72.50** 62.50	26.67 40.00	64.33 71.71	0.29411
ACM-TIES (25)	75.87 88.33	66.60 84.00	43.38 **68.01**	45.40 68.64	77.40 **93.20**	84.19 93.41	57.50 57.50	13.33 **53.33**	57.96 75.80	0.36519
KARCHER (26)	76.78 79.85	66.60 80.80	44.12 58.46	42.73 61.54	75.40 83.60	76.96 86.01	45.00 72.50	33.33 46.67	57.62 71.18	0.35281
Evo-L2S (26)	73.90 76.82	64.20 71.20	38.97 39.34	41.84 56.80	63.80 64.80	64.59 68.14	47.50 57.50	20.00 26.67	51.85 57.66	0.29954
EvoGM (26)	75.72 82.27	68.20 80.00	47.06 61.40	46.59 62.43	74.60 83.60	78.32 86.10	65.00 62.50	20.00 50.00	59.44 71.04	0.36640
DISCO-Merge	**91.35** **93.94**	**85.60** **88.60**	55.15 59.19	**60.53** **75.15**	**89.80** 92.00	**92.86** **93.95**	57.50 **87.50**	**46.67** 30.00	**72.43** **77.54**	**0.44130**

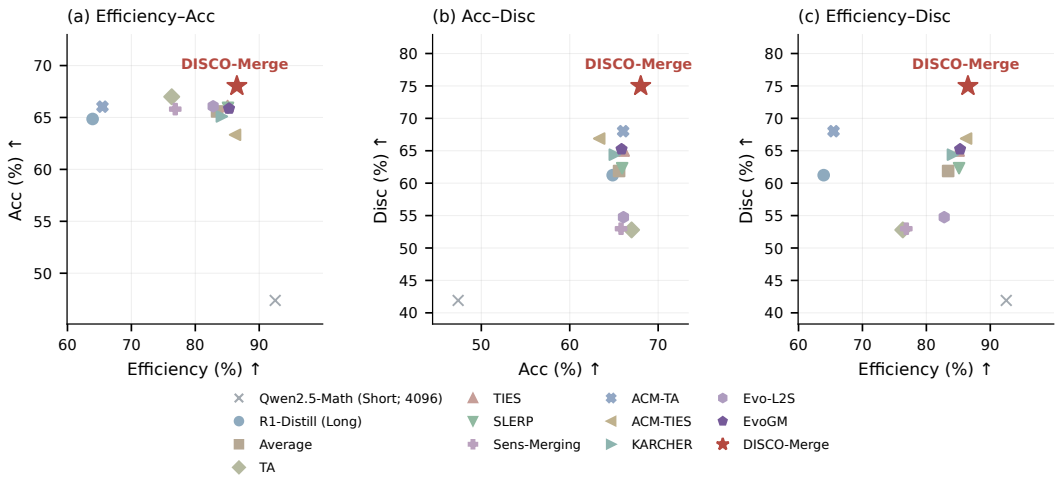

Figure 5: Qwen2.5-7B objective-pair comparisons. Muted markers denote baselines; the red star denotes DISCO. All axes favor the upper right. Efficiency is $100(1 - \text{Tok}/10240)$. Short is a native-context reference.

DISCO improves Accuracy over Long on six benchmarks but loses 2.50 points on AMC23 and 20.00 points on AIME24 (six of 30 questions). Macro improvements therefore do not imply per-benchmark dominance.

E COMPLETE QWEN3-4B RESULTS

Qwen3-Instruct and Qwen3-Thinking denote the 2507 releases. All 13 reported configurations use the common Thinking runtime, greedy BF16 decoding and a common output cap. EvoGM measures 138 distinct candidates in 140 requests; its effective coefficient is -0.26060366172.

Table 16: Qwen3-4B performance (Acc/Tok/Disc) by benchmark, with Macro and HV.

Model / Method	GSM8K	MATH500	Minerva	Olympiad	SVAMP	ASDiv	AMC23	AIME24	Macro equal weight	HV↑
Qwen3-Instruct (Short)	72.86	76.20	16.91	47.41	80.50	82.66	65.00	30.00	58.94	0.17461
	2225.2	2694.4	5584.2	4712.9	1577.9	1370.7	4032.3	6531.5	3591.1	
	50.02	47.70	42.65	34.81	61.90	63.74	42.50	21.67	45.62	
Qwen3-Thinking (Long)	71.72	67.80	24.63	32.59	81.20	79.91	37.50	13.33	51.09	0.12369
	3116.5	4442.8	6237.1	6783.8	2581.0	2471.8	6590.9	7670.0	4986.7	
	59.06	66.30	47.61	40.00	53.30	47.54	43.75	20.00	47.19	
Average (22)	69.45	73.60	31.99	41.04	80.50	81.58	62.50	26.67	58.41	0.25338
	3097.0	3565.6	5331.3	5908.4	2142.6	1915.9	5047.4	7298.7	4288.4	
	87.34	81.20	59.74	61.63	91.80	90.34	75.00	50.00	74.63	
TA (23)	73.84	58.60	23.90	22.81	77.20	73.54	47.50	10.00	48.42	0.15166
	3003.3	4643.3	6075.9	6942.5	2745.0	2812.1	6304.7	8000.8	5065.9	
	68.77	71.30	46.69	54.66	71.40	67.23	72.50	43.33	61.98	
TIES (23)	70.74	76.00	28.31	41.04	80.60	81.08	47.50	40.00	58.16	0.25013
	2962.3	3468.9	5576.9	5883.2	2055.4	1931.0	5430.8	6921.8	4278.8	
	89.46	80.80	56.43	60.59	91.40	91.11	71.25	50.00	73.88	
SLERP (24)	72.33	76.60	27.94	38.22	78.50	81.26	70.00	23.33	58.52	0.25390
	2887.4	3408.8	5330.3	6021.2	2252.0	1984.0	4792.1	7536.8	4276.6	
	87.94	81.80	58.27	61.48	91.90	91.24	75.00	48.33	74.50	
Sens-Merging (25)	75.06	53.80	31.25	21.78	78.80	77.52	25.00	6.67	46.23	0.13308
	2868.2	4879.9	5537.5	7031.6	2521.9	2424.9	6858.3	8101.9	5028.0	
	56.94	66.70	42.65	56.44	58.50	56.21	70.00	45.00	56.55	
ACM-TA (25)	80.21	62.80	26.84	33.04	78.30	78.47	50.00	16.67	53.29	0.12919
	2491.3	4678.1	6129.5	6737.6	2733.3	2527.8	5870.7	7603.6	4846.5	
	46.17	66.20	50.00	47.26	33.20	30.79	61.25	33.33	46.03	
ACM-TIES (25)	72.18	82.80	26.84	52.74	76.20	79.46	85.00	**46.67**	65.23	0.30211
	2565.3	3061.4	5310.0	5170.9	2290.1	1929.1	4077.0	6561.8	3870.7	
	86.27	79.00	59.19	59.99	88.90	90.20	78.75	53.33	74.46	
KARCHER (26)	70.74	75.80	28.68	38.67	77.30	80.23	60.00	26.67	57.26	0.24789
	2983.5	3460.0	5337.0	5991.6	2314.5	2072.4	4995.6	7096.9	4281.4	
	87.56	80.40	57.90	61.63	90.00	90.20	82.50	45.00	74.40	
Evo-L2S (26)	69.14	81.80	26.10	50.81	76.00	80.63	80.00	40.00	63.06	0.29406
	2828.9	3104.1	5356.6	5344.1	2292.2	1862.4	4070.2	6805.5	3958.0	
	89.31	82.00	61.95	63.40	91.90	92.87	70.00	56.67	76.01	
EvoGM (26)	90.30	86.00	22.43	51.85	**93.50**	**93.86**	75.00	36.67	68.70	0.26404
	685.5	**1825.9**	4372.7	**4084.3**	**497.1**	**377.3**	**3169.8**	**5885.5**	**2612.3**	
	64.96	58.10	37.87	34.80	74.40	78.05	41.25	23.33	51.60	
DISCO-Merge	**91.81**	**90.00**	**40.07**	**57.63**	90.10	91.78	**90.00**	**46.67**	**74.76**	**0.44674**
	1052.9	2352.5	**4142.6**	4953.7	1002.2	888.9	3992.6	6590.3	3122.0	
	98.26	**94.40**	**70.04**	**75.55**	**98.00**	**97.74**	**88.75**	**65.00**	**85.97**	

Table 17: Qwen3-4B feedback (Ret/Cor) by benchmark, with Macro and HV.

Model / Method	GSM8K	MATH500	Minerva	Olympiad	SVAMP	ASDiv	AMC23	AIME24	Macro equal weight	HV↑
Qwen3-Instruct (Short)	29.59 70.45	41.20 54.20	34.93 50.37	31.75 37.87	49.00 74.80	50.14 77.35	47.50 37.50	23.33 20.00	38.43 52.82	0.17461
Qwen3-Thinking (Long)	61.00 57.12	62.80 69.80	45.96 49.26	37.09 42.90	54.00 52.60	49.86 45.22	50.00 37.50	23.33 16.67	48.01 46.38	0.12369
Average (22)	84.98 89.70	81.20 81.20	54.04 65.44	58.75 64.50	88.20 95.40	89.52 91.16	77.50 72.50	60.00 40.00	74.27 74.99	0.25338
TA (23)	69.80 67.73	71.60 71.00	46.32 47.06	51.93 57.40	72.80 70.00	72.81 61.64	75.00 70.00	43.33 43.33	62.95 61.02	0.15166
TIES (23)	86.80 92.12	79.80 81.80	53.68 59.19	59.35 61.83	88.20 94.60	89.70 92.51	77.50 65.00	46.67 53.33	72.71 75.05	0.25013
SLERP (24)	84.22 91.67	81.00 82.60	53.31 63.24	60.24 62.72	89.40 94.40	90.24 92.24	67.50 82.50	43.33 53.33	71.16 77.84	0.25390
Sens-Merging (25)	58.12 55.76	67.40 66.00	40.81 44.49	51.93 60.95	60.20 56.80	58.27 54.15	72.50 67.50	46.67 43.33	56.99 56.12	0.13308
ACM-TA (25)	49.77 42.58	65.00 67.40	46.69 53.31	45.40 49.11	33.20 33.20	33.42 28.16	62.50 60.00	33.33 33.33	46.17 45.89	0.12919
ACM-TIES (25)	82.09 90.45	74.80 83.20	52.94 65.44	55.79 64.20	83.60 94.20	86.27 94.13	70.00 87.50	60.00 46.67	70.69 78.22	0.30211
KARCHER (26)	84.22 90.91	77.60 83.20	51.84 63.97	59.05 64.20	85.20 94.80	88.53 91.88	82.50 82.50	40.00 50.00	71.12 77.68	0.24789
Evo-L2S (26)	86.95 91.67	80.80 83.20	55.51 68.38	58.75 68.05	87.60 96.20	90.97 94.77	72.50 67.50	50.00 **63.33**	72.89 79.14	0.29406
EvoGM (26)	41.12 88.79	44.00 72.20	30.88 44.85	26.41 43.20	57.20 91.60	64.59 91.52	32.50 50.00	10.00 36.67	38.34 64.85	0.26404
DISCO-Merge	**97.27** **99.24**	**94.40** **94.40**	**70.22** **69.85**	**73.89** **77.22**	**96.80** **99.20**	**97.20** **98.29**	**87.50** **90.00**	**66.67** **63.33**	**85.49** **86.44**	**0.44674**

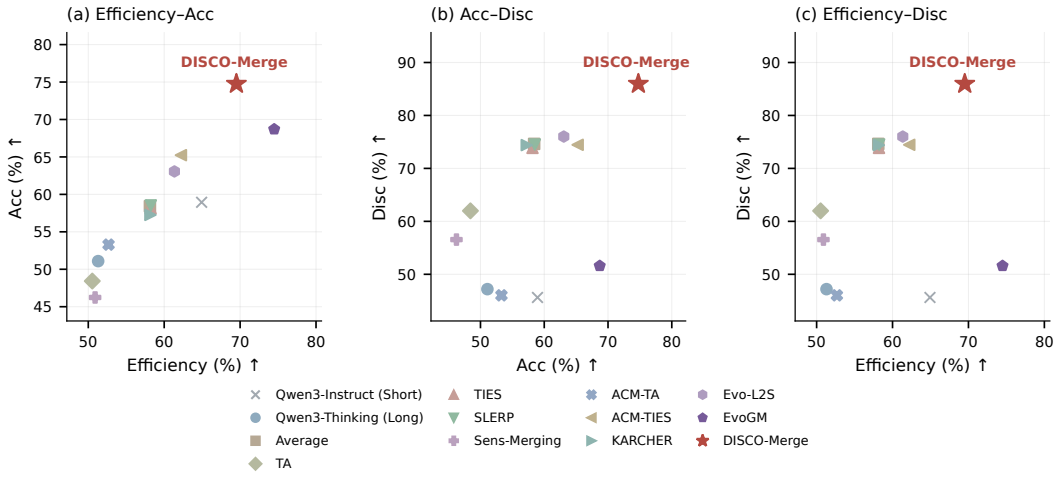

Figure 6: Qwen3-4B objective-pair comparisons. Styles and the Efficiency normalization match Figure 5.

F Complete Phi-4-mini-3.8B Results

Phi4m-Instruct and Phi4m-Reasoning abbreviate the 3.8B source models `microsoft/Phi-4-mini-instruct` and `microsoft/Phi-4-mini-reasoning` (Microsoft, 2025a;b).

Table 18: Phi-4-mini-3.8B performance (Acc/Tok/Disc) by benchmark, with Macro and HV.

Model / Method	GSM8K	MATH500	Minerva	Olympiad	SVAMP	ASDiv	AMC23	AIME24	Macro equal weight	HV↑
Phi4m-Instruct (Short)	87.79	63.80	31.99	28.59	92.10	92.91	42.50	6.67	55.79	0.29806
	267.6	1006.0	757.6	**1781.3**	182.0	**173.1**	**1675.3**	3981.3	1228.0	
	88.63	69.90	49.63	44.43	91.00	93.27	38.75	10.00	60.70	
Phi4m-Reasoning (Long)	93.93	83.20	**40.07**	44.74	95.00	95.03	**80.00**	26.67	**69.83**	0.33995
	1170.5	3141.9	4562.2	5454.7	905.2	827.6	4485.2	7416.8	3495.5	
	94.46	84.30	65.99	54.07	96.20	97.52	**73.75**	25.00	73.91	
Average (22)	91.43	72.40	31.25	37.33	92.70	94.27	45.00	16.67	60.13	0.36656
	259.7	1098.5	793.2	2361.5	199.2	194.2	1727.7	4872.7	1438.3	
	92.87	83.30	58.09	55.99	94.90	96.39	62.50	23.33	70.92	
TA (23)	93.78	81.60	36.76	41.04	**95.40**	95.17	67.50	16.67	65.99	0.35905
	664.8	2524.5	2383.8	4732.0	551.4	513.7	3877.8	7255.3	2812.9	
	94.62	86.10	65.44	55.10	95.70	97.34	72.50	33.33	75.02	
TIES (23)	91.43	74.20	31.62	38.37	93.20	94.63	55.00	10.00	61.06	0.36168
	246.3	1093.7	747.2	2293.3	**177.1**	192.4	2422.6	5139.2	1539.0	
	93.10	82.60	56.62	56.14	94.50	96.84	56.25	21.67	69.72	
SLERP (24)	91.66	72.20	31.25	38.07	92.60	94.76	47.50	16.67	60.59	0.37073
	256.9	1082.9	825.1	2260.0	195.7	200.3	2088.8	4301.2	1401.4	
	93.10	82.30	56.80	55.99	94.10	96.48	60.00	28.33	70.89	
KARCHER (26)	90.37	73.40	31.99	38.96	92.90	94.58	60.00	16.67	62.36	0.37512
	242.9	**1004.5**	888.4	2209.1	178.0	205.8	1922.3	4175.3	1353.3	
	93.25	82.90	56.62	56.44	94.50	97.07	53.75	20.00	69.32	
ACM-TA (25)	93.48	80.00	37.13	42.37	94.60	95.30	62.50	33.33	67.34	0.32647
	930.9	3047.8	3961.4	5310.8	731.8	686.0	4783.7	7076.8	3316.1	
	92.79	83.80	64.89	51.69	96.70	97.06	65.00	21.67	71.70	
ACM-TIES (25)	90.52	74.00	37.50	38.96	93.70	95.03	50.00	13.33	61.63	0.37005
	246.6	1064.5	**705.4**	2432.4	220.9	178.1	1975.8	**2803.3**	**1203.4**	
	92.42	81.10	54.04	53.18	93.70	96.52	50.00	23.33	68.04	
Sens-Merging (25)	93.10	78.40	37.50	41.19	94.70	95.21	55.00	20.00	64.39	0.33592
	689.6	2570.9	2520.7	4612.3	542.8	540.8	4360.8	6906.3	2843.0	
	92.19	86.40	64.89	52.74	95.60	96.39	66.25	23.33	72.22	
Evo-L2S (26)	93.86	82.80	38.60	45.78	94.90	95.26	72.50	30.00	69.21	0.34176
	1149.7	3159.0	4524.5	5410.9	896.4	814.3	4614.4	6915.1	3435.5	
	94.69	85.50	62.32	54.81	95.90	97.52	68.75	35.00	74.31	
EvoGM (26)	94.16	83.20	37.50	44.74	94.70	**95.62**	70.00	**36.67**	69.57	0.35710
	969.9	2802.1	3980.6	5124.8	758.5	703.9	4670.0	6902.1	3239.0	
	95.07	86.00	67.65	55.40	**97.00**	97.38	**73.75**	28.33	75.07	
DISCO-Merge	**94.47**	**84.60**	39.71	**47.11**	95.20	95.30	72.50	26.67	69.44	**0.39103**
	731.3	2419.0	2643.7	4646.2	582.9	580.6	4110.1	7007.3	2840.1	
	95.53	**90.00**	**70.59**	**61.18**	96.20	**97.79**	**73.75**	**38.33**	**77.92**	

Table 19: Phi-4-mini evaluation and selection settings.

Item	Setting
Source pair	Phi-4-mini-instruct / Phi-4-mini-reasoning
Prompt and runtime	Native Phi chat template; pinned Long tokenizer/runtime
Decoding	Greedy; BF16; common output cap
Coverage per completed model	6,051 clean + 3,445 Retention + 3,448 Correction records
DISCO	Original 18D policy; 104 candidates; 24-to-6-to-1 confirmation; seed 20260920
Evo-L2S	189 search candidates; 20 confirmation candidates; highest-confirmation-Accuracy output
EvoGM	Instruct-base adaptation; 198 distinct candidates / 200 requests; selection details below

Table 20: Phi-4-mini-3.8B feedback (Ret/Cor) by benchmark, with Macro and HV.

Model / Method	GSM8K	MATH500	Minerva	Olympiad	SVAMP	ASDiv	AMC23	AIME24	Macro equal weight	HV↑
Phi4m-Instruct (Short)	87.71	67.40	47.06	37.39	89.00	92.50	35.00	13.33	58.67	0.29806
	89.55	72.40	52.21	51.48	93.00	94.04	42.50	6.67	62.73	
Phi4m-Reasoning (Long)	92.26	81.00	62.50	48.66	93.80	96.48	70.00	20.00	70.59	0.33995
	96.67	87.60	69.49	59.47	98.60	**98.56**	77.50	30.00	77.23	
Average (22)	92.26	83.20	57.35	50.74	93.40	96.30	60.00	16.67	68.74	0.36656
	93.48	83.40	58.82	61.24	96.40	96.48	65.00	30.00	73.10	
TA (23)	92.72	84.40	62.87	50.45	93.80	96.93	**72.50**	33.33	73.37	0.35905
	96.52	87.80	68.01	59.76	97.60	97.74	72.50	33.33	76.66	
TIES (23)	93.32	82.20	58.09	54.01	92.80	96.84	60.00	20.00	69.66	0.36168
	92.88	83.00	55.15	58.28	96.20	96.84	52.50	23.33	69.77	
SLERP (24)	92.87	80.40	56.99	52.52	92.40	96.57	62.50	30.00	70.53	0.37073
	93.33	84.20	56.62	59.47	95.80	96.39	57.50	26.67	71.25	
KARCHER (26)	92.72	83.80	58.09	51.93	93.60	96.93	47.50	16.67	67.65	0.37512
	93.79	82.00	55.15	60.95	95.40	97.20	60.00	23.33	70.98	
ACM-TA (25)	89.07	79.00	62.13	43.92	94.60	95.93	57.50	20.00	67.77	0.32647
	96.52	88.60	67.65	59.47	**98.80**	98.19	72.50	23.33	75.63	
ACM-TIES (25)	92.11	80.80	52.57	48.37	92.00	95.84	50.00	20.00	66.46	0.37005
	92.73	81.40	55.51	57.99	95.40	97.20	50.00	26.67	69.61	
Sens-Merging (25)	89.23	83.20	63.97	51.34	93.00	95.66	60.00	20.00	69.55	0.33592
	95.15	89.60	65.81	54.14	98.20	97.11	72.50	26.67	74.90	
Evo-L2S (26)	92.11	83.00	58.09	49.55	94.00	96.75	65.00	26.67	70.65	0.34176
	97.27	88.00	66.54	60.06	97.80	98.29	72.50	**43.33**	77.97	
EvoGM (26)	92.72	84.00	66.18	50.74	**95.20**	96.66	65.00	23.33	71.73	0.35710
	97.42	88.00	69.12	60.06	**98.80**	98.10	**82.50**	33.33	78.42	
DISCO-Merge	**94.23**	**89.20**	**69.85**	**59.35**	94.60	**97.38**	70.00	**40.00**	**76.83**	**0.39103**
	96.82	**90.80**	**71.32**	**63.02**	97.80	98.19	77.50	36.67	**79.02**	

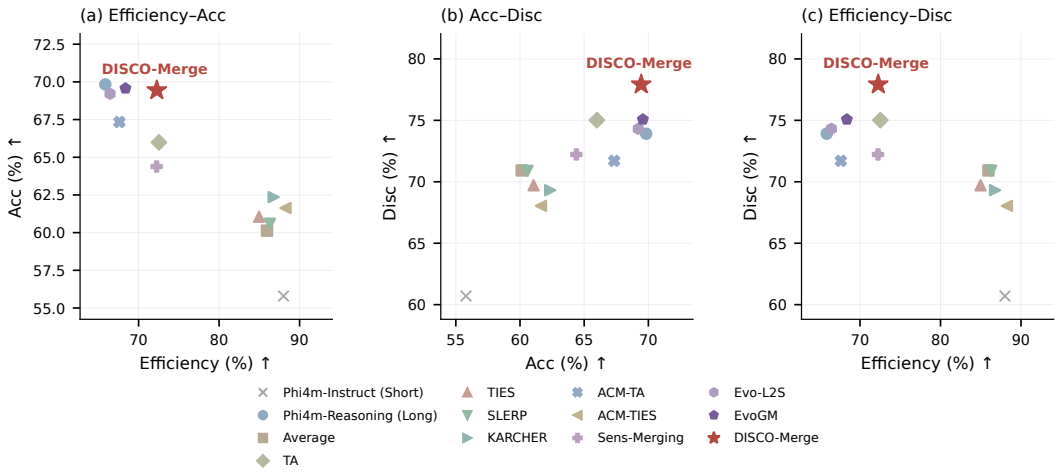

Figure 7: Phi-4-mini-3.8B objective-pair comparisons. Styles and the Efficiency normalization match Figure 5. All 13 completed models are plotted.

The two Evo-L2S outputs were selected on confirmation data; this comparison retains the highest-Accuracy representative and excludes the separate shorter-output representative. DISCO's output was locked before the full benchmark. All 13 completed rows cover 12,944 records each, with the same per-dataset denominators as Table 10.

Phi's EvoGM is the Instruct-base adaptation: $B = S$ in $B + c_0(S - B) + c_1(L - B)$, so the first initial task vector vanishes. Its completed search uses population 20, two rounds and four generator iterations per round: 200 logical requests and 198 distinct candidates. Unlike the Qwen experiments, its frozen objective and final rule maximize clean search Accuracy, with earliest-candidate tie-breaking. The selected identity was locked before full evaluation and covers all 12,944 records.

G Supplementary Mathematical Feedback Evaluation

G.1 Aggregate Mathematical Feedback Profiles

The right panel of the main feedback figure uses the highest-Disc non-DISCO merged model within each family: ACM-TA for Qwen2.5, Evo-L2S for Qwen3 and EvoGM for Phi. Sources are excluded from this comparator set. The reference is chosen once per family by macro Disc, rather than separately for Ret, Cor or individual benchmarks. DISCO's macro Ret/Cor changes are $+8.11/+5.83$, $+12.61/+7.30$ and $+5.10/+0.60$ points, respectively. The small circles show all eight dataset differences, including negative ones; the diamonds are equal-weight means, not uncertainty estimates.

This complementary view compares the absolute Ret and Cor of all reported models on the mathematical fixed-history evaluation. It visualizes the macro statistics in the main tables; Figure 3 resolves benchmark-level changes relative to both Long and each family's highest-Disc merging baseline.

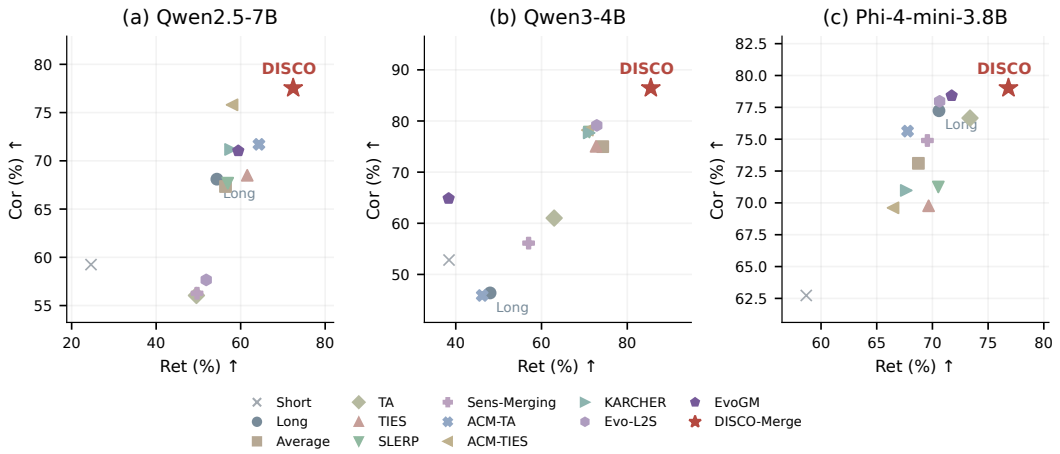

Figure 8: Mathematical feedback behavior across all reported models. Each point is an eight-benchmark macro mean; upper right is better. Qwen2.5 Short remains a native-context reference.

DISCO has the highest Ret and Cor point estimates in each family. Relative to Long, its macro Ret/Cor changes are $+18.06/+9.44$ points for Qwen2.5, $+37.49/+40.06$ for Qwen3 and $+6.24/+1.78$ for Phi.

H Component Comparisons

The component study examines the roles of proposal generation, evaluation allocation, reward assignment and feedback-aware selection under the same nominal search and confirmation budgets.

Table 5 reports the five selected controls and DISCO-Merge. Means and sample standard deviations use three search seeds: 20260917, 20260920 and 20260921.

Table 21: DISCO results with the independent eight-card seed 20260920.

Seed	Acc↑	Tok↓	Disc↑	Ret↑	Cor↑	HV↑
20260917	68.0303	1381.7190	74.9871	72.4327	77.5414	0.44130
20260920	70.3135	1257.0660	77.8358	74.9828	80.6887	0.48010
20260921	68.5598	1497.7445	69.7056	65.8906	73.5206	0.40800
Mean ± SD	68.9679 ± 1.1951	1378.8432 ± 120.3650	74.1762 ± 4.1253	71.1021 ± 4.6899	77.2502 ± 3.5929	0.44314 ± 0.03609

Table 22: Definitions of the five reported component controls.

Method	Intervention
w/ Gaussian	Replace neural Flow with a conditional diagonal shrinkage Gaussian; retain behavioral inputs and mixture rules.
w/o KG	Choose eight candidates uniformly from the same proposal pool; fit the same GP but do not use it for allocation.
w/o Disc. in Acq.	Replace terminal acquisition utility's Disc with Acc; retain the gates and other Disc-dependent components.
w/o Reward Shuffle	Use true reward assignment instead of shuffled assignments when training the generator.
w/o Disc. in Selection	Choose highest feasible Acc from the same six confirmed models in the original control runs.

DISCO uses reward-shuffled conditional Flow, the ordinary-proposal mixture and finite-pool KG/cost. Each search has 24 initial models, ten batches of eight, the same 18D space and $24 \to 6 \to 1$ confirmation. Applicable generator settings retain 180 Adam steps, 12 Euler steps and 128 proposals. For conditions that retain KG, allocation selects six candidates by KG and two randomly. The first two rounds are warmup.

In the original control study, initialization is matched within seeds; generator controls share the pre-intervention warmup and acquisition controls share initialization. The replacement DISCO run is independent. Exact-output reuse remains logically charged, and differing reuse fractions prevent a direct comparison of physical generation costs.

Table 23: All three seeds for the six reported component conditions.

Method	Seed	Acc↑	Tok↓	Disc↑	Ret↑	Cor↑	HV↑
w/ Gaussian	20260917	67.55	1348.10	**75.35**	68.99	81.71	**0.44196**
	20260920	69.11	1618.40	73.59	70.11	77.08	0.42820
	20260921	**70.26**	1593.32	71.53	69.15	73.92	0.42436
w/o KG	20260917	67.62	1413.47	74.45	67.08	**81.82**	0.43392
	20260920	67.22	1489.79	66.88	61.50	72.26	0.38417
	20260921	67.23	1609.14	71.69	66.34	77.03	0.40622
w/o Disc. in Acq.	20260917	**68.63**	**1277.78**	73.11	70.05	76.17	0.43912
	20260920	67.75	1463.82	73.25	70.05	76.45	0.42533
	20260921	68.07	1864.35	72.43	68.02	76.83	0.40325
w/o Reward Shuffle	20260917	67.30	1436.82	75.33	**73.56**	77.10	0.43584
	20260920	66.04	1423.58	70.22	64.81	75.63	0.39929
	20260921	67.78	1575.50	**76.09**	**71.19**	**80.98**	**0.43639**
w/o Disc. in Selection	20260917	67.93	1366.74	71.87	69.45	74.29	0.42304
	20260920	69.83	1410.18	75.82	74.37	77.28	0.45653
	20260921	68.56	**1497.74**	69.71	65.89	73.52	0.40800
DISCO-Merge	20260917	68.03	1381.72	74.99	72.43	77.54	0.44130
	20260920	**70.31**	**1257.07**	**77.84**	**74.98**	**80.69**	**0.48010**
	20260921	68.56	**1497.74**	69.71	65.89	73.52	0.40800

I Framework Performance Across Intermediate Models

This experiment tests whether improvements depend on the chosen intermediate merge while controlling initialization and evaluation budget; the direct two-source Module variant additionally probes parameterization.

The completed Qwen2.5 study uses TIES, Average, SLERP and KARCHER as intermediate models in $W = S + a(M - S) + b(L - M)$, with nine coefficient pairs. The Module control uses nine direct Short–Long coefficients. All five start from the same 24 byte-identical models and measured initial scores (seed 20260917), then run 104 candidates with the same generator, GP, acquisition and $24 \to 6 \to 1$ confirmation. The common constraint reference is Average.

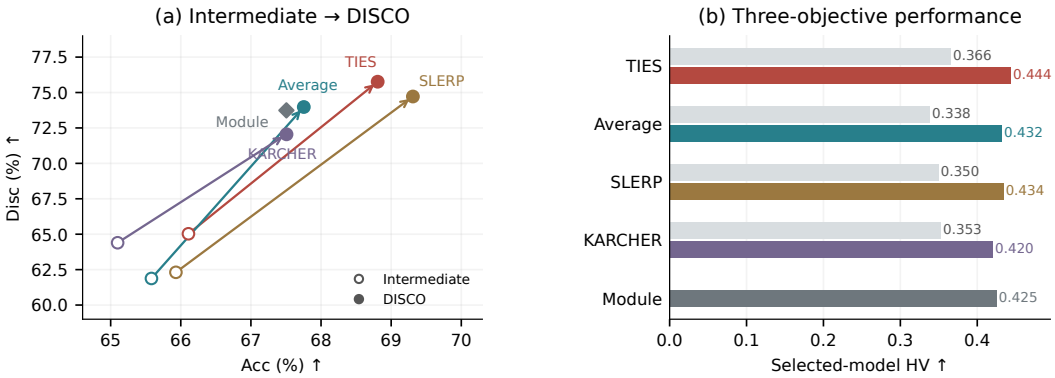

Figure 9: Four intermediate models and their selected DISCO outputs. Arrows connect each reference to its output; gray and colored bars show their respective HV. Module is a direct two-source control without an intermediate reference.

Table 24: Changes from each intermediate model to its selected DISCO output.

Intermediate	ΔAcc	ΔTok (%)	ΔDisc	ΔRet	ΔCor	ΔHV
TIES	+2.70	-0.37	+10.73	+10.03	**+11.44**	+0.07802
Average	+2.17	**-16.97**	+12.10	+14.31	+9.89	**+0.09371**
SLERP	**+3.38**	+8.78	**+12.41**	**+15.71**	+9.10	+0.08435
KARCHER	+2.41	-14.32	+7.65	+9.17	+6.14	+0.06757

Changes are output minus reference. Acc, Disc, Ret and Cor changes are percentage points; Tok changes are relative percentages. HV uses the reporting reference 10,240. TIES, Average and KARCHER improve all three objectives; SLERP exchanges longer outputs for higher accuracy and discernment. All four improve both feedback components.

The references are the exact historical TIES midpoint, the arithmetic Short–Long mean, SLERP at $t = 0.5$, and KARCHER with ten iterations and tolerance 10^{-5}. They retain their identities and full-evaluation scores from Table 1. The four midpoint variants have 18 coordinates; Module has nine. Different parameterizations change reachable weights and proposal steps, so this study evaluates complete framework configurations rather than isolating one component. Average's 18D form has effective rank nine in exact arithmetic, with BF16 residual effects in the implementation.

Each selected identity was locked before its 12,944-record full evaluation. All five completed independent full evaluations and the shared-initialization audit.

I.1 Complete Per-Benchmark Results

Table 25: Matched variants: Acc/Tok/Disc by benchmark, with Macro and HV.

Variant	GSM8K	MATH500	Minerva	Olympiad	SVAMP	ASDiv	AMC23	AIME24	Macro equal weight	HV↑
DISCO-TIES	**92.04**	**87.00**	33.82	**50.52**	94.50	**95.08**	**77.50**	20.00	68.81	**0.44368**
	378.6	1256.7	1341.2	2376.8	**263.9**	287.6	2124.4	4168.4	1524.7	
	91.51	87.70	**61.58**	66.80	90.40	92.28	**77.50**	38.33	**75.76**	
DISCO-Average	91.58	86.20	35.29	47.56	94.00	94.90	72.50	20.00	67.75	0.43217
	355.3	1225.4	1588.8	2284.7	313.9	269.2	**1627.9**	**3621.6**	1410.8	
	90.45	87.20	57.54	66.36	90.50	92.28	67.50	40.00	73.98	
DISCO-SLERP	91.36	86.00	35.66	47.70	**94.60**	94.99	**77.50**	**26.67**	**69.31**	0.43402
	373.6	1351.6	1697.4	2561.5	296.4	290.1	1963.2	4730.8	1658.1	
	92.04	86.30	60.66	**67.99**	90.40	89.52	67.50	**43.33**	74.72	
DISCO-KARCHER	91.05	84.60	**37.50**	47.11	94.20	94.76	67.50	23.33	67.51	0.42038
	353.4	1130.8	**1260.9**	**1982.4**	277.3	**263.4**	2137.8	3712.2	**1389.8**	
	91.43	84.80	58.09	64.13	90.40	92.55	65.00	30.00	72.05	
DISCO-Module	91.13	85.80	37.13	47.26	94.10	94.63	70.00	20.00	67.51	0.42526
	358.8	**1094.2**	1370.4	2148.6	300.3	267.4	2089.6	4306.0	1491.9	
	91.89	**88.20**	55.51	63.84	**91.10**	**93.54**	62.50	**43.33**	73.74	

Table 26: Matched variants: Ret/Cor by benchmark, with Macro and HV.

Variant	GSM8K	MATH500	Minerva	Olympiad	SVAMP	ASDiv	AMC23	AIME24	Macro equal weight	HV↑
DISCO-TIES	88.77	84.40	**60.29**	57.27	**86.40**	89.70	**72.50**	33.33	71.58	**0.44368**
	94.24	91.00	**62.87**	**76.33**	94.40	94.86	**82.50**	43.33	**79.94**	
DISCO-Average	89.68	84.80	55.88	56.97	**86.40**	90.42	65.00	36.67	70.73	0.43217
	91.21	89.60	59.19	75.74	94.60	94.13	70.00	43.33	77.23	
DISCO-SLERP	89.83	84.60	59.56	**62.61**	86.20	85.91	**72.50**	**40.00**	**72.65**	0.43402
	94.24	88.00	61.76	73.37	94.60	93.14	62.50	**46.67**	76.79	
DISCO-KARCHER	88.77	82.40	55.15	53.71	85.40	89.70	52.50	26.67	66.79	0.42038
	94.09	87.20	61.03	74.56	95.40	95.40	77.50	33.33	77.31	
DISCO-Module	**89.98**	**85.00**	51.84	57.27	**86.40**	**91.60**	62.50	**40.00**	70.57	0.42526
	93.79	**91.40**	59.19	70.41	**95.80**	**95.49**	62.50	**46.67**	76.91	

Each macro metric is an equal-weight mean over the eight benchmarks. Performance cells stack Acc/Tok/Disc; feedback cells stack Ret/Cor. Module has no intermediate-model reference, so no reference-to-output gain is assigned to it. Full evaluation covers 6,051 clean questions, 3,445 retention records and 3,448 correction records per selected model.

J Search Strategy and Merge Parameterization

To compare complete search strategies, we evaluate Direct BO with its native four-dimensional parameterization and 22 evaluations against DISCO with 18 dimensions and 104 evaluations. Both use the same development and confirmation question files. The experiment compares the delivered models under their respective workflows.

Table 27: Search strategies compared at their native evaluation budgets.

Strategy	Space / evaluations	Acc↑	Tok↓	Disc↑	Ret↑	Cor↑	HV↑
Direct BO	4D / 22	67.63	**1351.39**	74.42	71.98	76.86	0.43686
DISCO-Merge	18D / 104	**68.03**	1381.72	**74.99**	**72.43**	**77.54**	**0.44130**

Direct BO uses 12 initial points and 10 sequential noisy expected hypervolume improvement (NEHVI) queries (Daulton et al., 2021), then $23 \rightarrow 6 \rightarrow 1$ confirmation including a fixed TIES outside option. This comparison does not isolate the generator, acquisition function or evaluation efficiency.

Relative to Direct BO, DISCO gains 0.40 Accuracy and 0.57 Discernment points but uses 30.33 more tokens. Different search spaces and budgets prevent isolating Flow's effect.

The complete matched-initialization midpoint and Module comparison is reported in Section 4.3 and Appendix I.

J.1 Descriptive Acc/Tok Neighborhoods

To examine whether similar clean accuracy and length imply similar feedback behavior, we compare all common-cap model pairs within 0.6 Acc points and 5% symmetric Tok difference, then inspect their Disc, Ret and Cor gaps.

Table 28: Qwen2.5-7B feedback differences between nearby Acc/Tok configurations.

First / second model	ΔAcc	ΔTok **(%)**	ΔDisc	ΔRet	ΔCor
Average / KARCHER	-0.48	**-4.65**	+2.52	+1.21	**+3.84**
Average / Evo-L2S	**+0.50**	+3.63	-7.13	-4.56	-9.68
TIES / SLERP	-0.18	-0.40	-2.72	-4.62	-0.82
TIES / EvoGM	-0.25	-1.49	+0.21	-2.12	+2.54
SLERP / EvoGM	-0.07	-1.09	**+2.93**	**+2.50**	+3.36

Second minus first. Filter: $|\Delta\mathrm{Acc}| \leq 0.6$ pp and symmetric token difference $100|\mathrm{Tok}_2 - \mathrm{Tok}_1|/[(\mathrm{Tok}_1 + \mathrm{Tok}_2)/2] \leq 5\%$, using the displayed aggregate precision. These are observed configurations, not randomized or causally matched pairs.

Average and Evo-L2S differ by 0.50 Accuracy points and 3.63% in symmetric token difference, yet their Discernment differs by 7.13 points. The complete filtered set is shown, including pairs with small behavioral differences.

K Search Trajectories and Computational Accounting

K.1 Three-Seed Search-Stage Selections

We track the frozen output rule across search iterations to assess improvement beyond initialization and sensitivity to the random seed. Seeds 1, 2 and 3 in the main text correspond to 20260917, 20260920 and 20260921. Iteration 0 selects from the 24 initial models; each of iterations 1–10 adds eight evaluated models. All curves use search-set scores, without confirmation or full-test results.

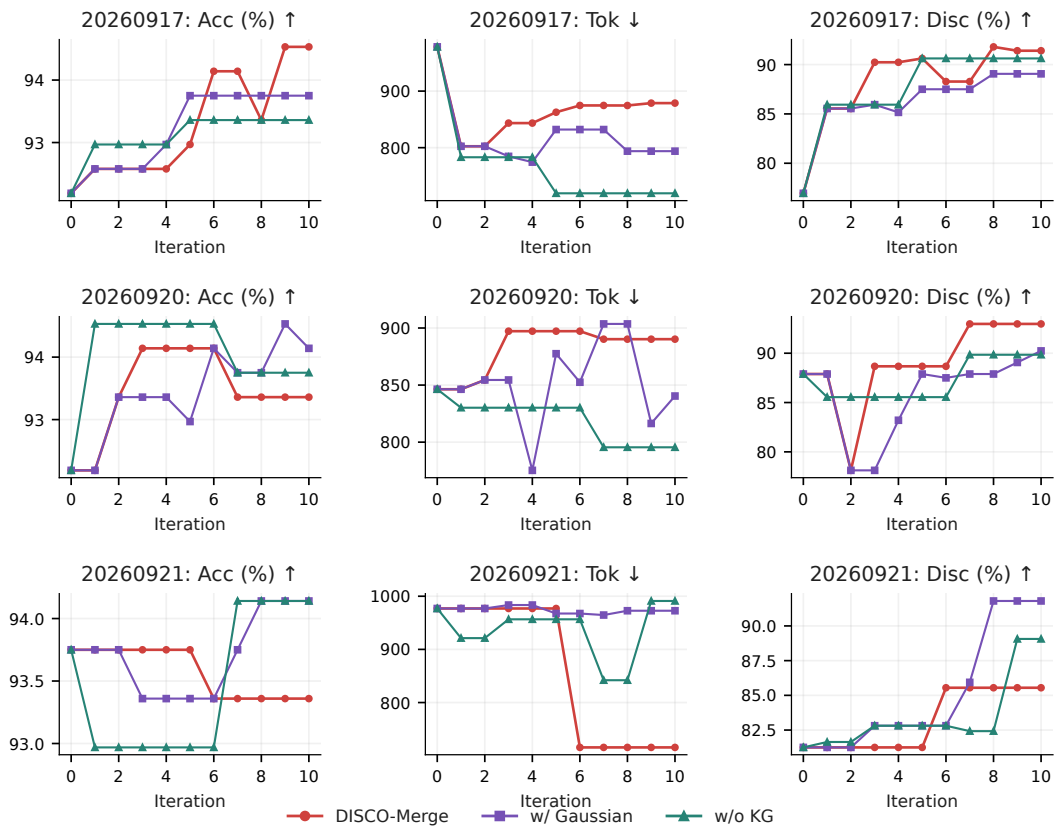

Figure 10: Search trajectories for three seeds.

K.2 Independent Eight-Card Execution and Cost

Table 29: Detailed stage costs of the eight-card run. Record counts distinguish logical, reused and newly generated retained records; time columns are worker-summed hours.

Stage	Logical	Reused	Newly generated	Generation h	Engine init. h	Merge h
Search	53,248	4,096	49,152	57.81	5.96	1.363
Confirmation to 512	12,288	0	12,288	11.42	1.49	0.295
Confirmation to 1,024	6,144	0	6,144	6.63	0.37	0.059
Final evaluation	12,944	0	12,944	5.40	0.52	0.012

Logical records count evaluations charged to the protocol. Reused records are cached outputs; newly generated counts include only outputs retained in completed bundles. Their sum equals the logical count. Generation times cover completed bundles and are not stage wall times.

Table 30: End-to-end cost accounting.

Quantity	Value	Interpretation
Elapsed completion	17.54 h	Includes recovery and gaps between attempts
Occupied capacity estimate	138.59 NPU-h	Eight-device lease intervals; includes initialization and idle waits
Final-evaluation wall time	50.89 min	One frozen model, 12,944 records
Generation duration	81.26 worker-h	Sum over completed generation bundles
New completed output tokens	97,538,190	Excludes output not retained in completed bundles
Recorded engine initializations	134	Excludes eight reused initial candidates and failed attempts
Logical records	84,624	Search + incremental confirmation + full evaluation
Reused records	4,096	Eight initial candidates, each with 512 cached records
Newly generated records	80,528	Retained new records; excludes reused records and unretained attempts