LLMs Evolutionary Merging for L2S Math Reasoning

August 2025

Mario Iacobelli - 1841427

Abstract

Model merging has emerged as an effective alternative to full fine-tuning, allowing for the combination and transfer of knowledge and behavioral traits across LLMs. In this work, inspired by *Activation-Guided Consensus Merging for LLMs*, we merge a verbose yet highly accurate Chain-of-Thought (CoT) model with a concise direct-answer model. The resulting hybrids aim to retain or enhance accuracy while significantly reducing response length. Leveraging **Mergenetic**, an evolutionary merging framework, we perform multi-objective optimization targeting both accuracy and output length for the **Long-to-Short** (L2S) math reasoning task. The merged models lie along the Pareto-optimal frontier of the accuracy–length trade-off, outperforming both base models and the ACM baseline in at least one dimension and, in several cases, surpassing them in both accuracy and compression.

1. Introduction

Mathematical reasoning tasks often reveal a fundamental trade-off between accuracy and efficiency. *Slow-thinking* (**System 2**) models, such as DeepSeek-R1 (DeepSeek-AI, 2025), rely on Chain-of-Thought (CoT) reasoning to produce highly accurate yet verbose outputs. In contrast, *fast-thinking* (**System 1**) models, like Qwen2.5-Math (Yang et al., 2024), generate concise and efficient responses, but are typically less robust and more prone to reasoning errors.

In this work, we investigate *model merging* as a strategy to combine these complementary behaviors without requiring additional fine-tuning. The objective is to compress the reasoning trace while preserving the accuracy of the original Slow-thinking model.

To achieve this, we leverage `Mergenetic` (Minut et al., 2025), an evolutionary merging framework built on top of `MERGE³` (Mencattini et al., 2025). The merging process is framed as a *multi-objective optimization* problem over accuracy and output length, enabling the discovery of a population of hybrid models that outperform not only the original components but also the Activation-Guided Consensus Merging (ACM) baseline (Yao et al., 2025).

For efficiency and reproducibility, our main experiments focus on the lightweight 1.5B-parameter variants, which can be fully evaluated on a single consumer-grade GPU. For completeness, evaluation tables and plots for the larger 7B-parameter variants are provided in the appendix.

2. Related Work

Model merging has recently emerged as an effective alternative to traditional approaches for Long-to-Short reasoning, such as token-level compression and reinforcement learning-based adaptive computation.

Unlocking L2S for model merging (Wu et al., 2025) applies standard merging techniques to merge fast- and slow-thinking models for L2S tasks. However, they emphasize that "it is highly desirable to develop a framework capable of automatically determining the optimal parameters," highlighting the limitations of manual coefficient tuning. Activation-Guided Consensus Merging addresses this challenge by deriving layer-wise merging coefficients from mutual information between base and target model activations. These coefficients are applied within established methods such as Task Arithmetic (Ilharco et al., 2023) and TIES (Yadav et al., 2023), yielding substantial response compression with minimal, or sometimes even positive, impact on accuracy.

In contrast, our work leverages multi-objective optimization to automatically discover effective merging configurations, jointly optimizing for accuracy and response length.

3. Method

We adopt `Mergenetic` to perform evolutionary model merging with two objectives: **accuracy**, measured as Exact Match on numerical answers (extracted via regex), and **response length**, quantified as number of generated tokens.

Email: Iacobelli <iacobelli.1841427@studenti.uniroma1.it>.

Deep Learning and Applied AI 2025, Sapienza University of Rome, 2nd semester a.y. 2024/2025.

`Mergenetic` integrates the following components:

- **MergeKit:** a flexible toolkit supporting merging strategies such as weight averaging, SLERP, Task Arithmetic (TA), TIES, and DARE. In our setup, MergeKit defines and applies parameter-space merging rules, encoded as a *genotype recipe*.
- **pymoo:** an optimization library offering evolutionary algorithms such as the Genetic Algorithm (GA) and NSGA-II, enabling both single- and multi-objective searches over merging coefficients.
- **lm-eval-harness:** a unified evaluation framework for generative LLMs, used to assess model accuracy during optimization, serving as a proxy fitness function.

Models. We consider two models that exhibit opposite reasoning behaviors:

- **DeepSeek-R1-DistillQwen-1.5B**[1]: a 1.5B-parameter model obtained by distilling DeepSeek-R1 (a 671B MoE model) into the Qwen2.5-Math architecture. It retains strong Chain-of-Thought reasoning abilities and supports long context windows (up to 32K tokens).
- **Qwen2.5-Math-1.5B**[2]: a compact 1.5B-parameter model fine-tuned from Qwen2.5 for mathematical reasoning, optimized for concise, direct-answer generation with shorter context lengths (up to 4K tokens).

Non-Canonical Merging. Unlike conventional model merging approaches, which assume a shared pretraining backbone, we merge two models with entirely different pretraining histories. Here, **DeepSeek-R1-DistillQwen** serves as the base model, while **Qwen2.5-Math** acts as the behavioral target.

Given this fixed-base setup, we apply the following merging strategies in simplified form:

- **Task Arithmetic (TA):** a scalar-weighted interpolation of the two models' parameters:

  $$\theta_{\text{merged}} = (1-w)\,\theta_{\text{DeepSeek}} + w\,\theta_{\text{Qwen}}$$

  where $w \in [0,1]$ controls the contribution of the target model.
- **TIES:** reduced to its initial *trimming* phase. We compute the task vector $\Delta\theta = \theta_{\text{Qwen}} - \theta_{\text{DeepSeek}}$, retain only the top-$d\%$ of parameters by absolute magnitude, and interpolate them as in TA, leaving the remaining parameters unchanged.

[1] deepseek-ai/DeepSeek-R1-Distill-Qwen-1.5B
[2] Qwen/Qwen2.5-Math-1.5B

Optimization Setup. We conduct both **multi-objective** optimization (jointly maximizing accuracy and minimizing response length) and **single-objective** optimization (targeting each metric individually). The search procedure uses the `Mergenetic` framework with the following configuration:

- **Evaluation subset: 30 anchors**, randomly sampled from GSM8K (out of 1319) and equally weighted, are used for the fitness proxy. This choice enables quick yet reliable evaluation and is intentionally smaller than the 100-sample calibration set used in ACM.
- **Algorithms:** NSGA-II for multi-objective optimization, and a standard Genetic Algorithm (GA) for single-objective runs.
- **Configuration:** a population size of **20 models** per generation, evolved over **10 generations**. The initial population is randomly initialized, and diversity is encouraged via polynomial mutation.

4. Results

Evaluation Setup. To ensure reproducibility, we adopted the official evaluation framework provided by Qwen[3], and fixed the random seed to 0 for all runs. Following best practices for both DeepSeek and Qwen models, we applied two distinct prompting strategies:

- The **fast-thinking** model was evaluated in a **few-shot** setting (4 examples per dataset) using a direct-question prompt; the maximum sequence length was set to 8k tokens.
- The **slow-thinking** model and all **merged** variants were evaluated in **zero-shot** mode using a chat-style `User/Assistant` prompt with the instruction "Please reason step by step ..."; the maximum sequence length was set to 10k tokens.

Accuracy deltas and length compression for ACM models are recomputed with respect to our own evaluation of DeepSeek-R1, as the original code and checkpoints are not available.

Quantitative Results. We evaluate all models on **six** math reasoning benchmarks — *AIME24, CollegeMath, GSM8K, MATH500, Minerva-MATH*, and *OlympiadBench* — reporting average accuracy (Exact Match on numerical answers) and mean output length in tokens.

Figure 1 illustrates the performance of both **Task Arithmetic** and **TIES**-based merged models, including Pareto

[3] QwenLM/Qwen2.5-Math

frontiers, single-objective optima, baselines, and ACM references. For comparison, we also report configurations from (Wu et al., 2025), which adopt a different setup by using Qwen2.5-Math as the base model and DeepSeek-R1 as the target. Specifically, we include results for **Average Merging**, **Task Arithmetic** with weight $w = 0.7$, and **TIES** with trim ratio $k = 0.8$ (i.e., density $= 0.2$) and weight $w = 1.0$.

Mergenetic-TA. The multi-objective search with Task Arithmetic merging strategy yields three clear groups of solutions:

1. **Length-oriented:** `TA-01`, `TA-09`, `TA-10`, and `TA-11` achieve 66–73% output length reduction while maintaining 40–42% accuracy, outperforming ACM-TA's 57.6% compression at comparable accuracy (41.4%).

2. **Balanced:** `TA-04`, `TA-06`, `TA-07`, and `TA-08` improve on both DeepSeek (45.5%) and ACM-TA, reaching 43–48% accuracy with 55–63% compression.

3. **Accuracy-oriented:** `TA-02`, `TA-03`, and `TA-05` achieve 49–50% accuracy while retaining 45–50% compression. The top-performing configuration, `TA-12`, reaches **51.2%** (+5.7 over DeepSeek) with a **55%** length reduction.

Mergenetic-TIES. Although none of the TIES-based models surpass the DeepSeek baseline in accuracy, several configurations provide strong compression with limited degradation. For instance, `TIES-02` achieves 42.5% accuracy with 62.7% compression, `TIES-10` reaches 43.4% with 60.0% compression, and `TIES-12` attains **44.9%** accuracy with approximately **55%** length reduction.

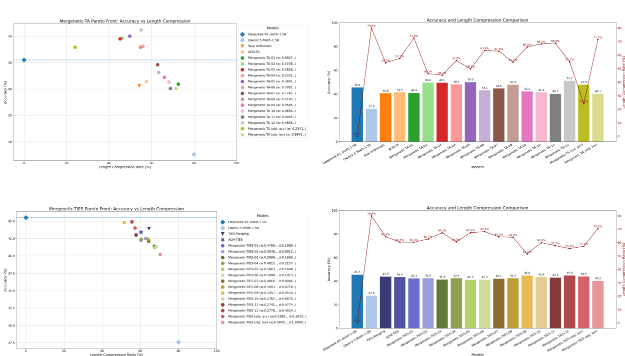

Figure 1. Evaluation plots for Mergenetic-TA (top) and Mergenetic-TIES (bottom) models. Each row shows Pareto Front (left) and a bar chart of accuracy and length (right). ACM and baseline results are included for reference.

Finally, in Fig. 2, we compare our top accuracy models (`Mergenetic-TA-12` and `Mergenetic-TIES-12`) with the ACM baselines and the original slow- and fast-reasoning models across the six math benchmarks: both of them achieve competitive accuracy across all tasks, while providing substantial length compression compared to the DeepSeek-R1 baseline. In particular, `Mergenetic-TA-12` consistently outperforms ACM-TA in average accuracy while maintaining similar or better compression, confirming the benefits of evolutionary search to tune merge coefficients. Conversely, `Mergenetic-TIES-12` achieves competitive accuracy with ACM-TIES while preserving its strength in output shortening.

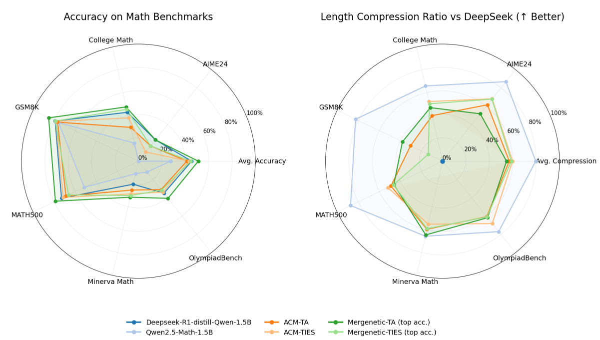

Figure 2. Comparison across six benchmarks for top Mergenetic models vs ACM and base models.

5. Conclusion

These experiments demonstrate that evolutionary merging can produce L2S hybrids that equal or exceed ACM in the accuracy-length trade-off. Single-objective optimization confirms that optimizing for accuracy or length in isolation produces inferior results, emphasizing the value of multi-objective optimization.

Although our focus was on lightweight 1.5B models to ensure efficiency and reproducibility, the results obtained on 7B variants indicate that the method is suitable for larger models. Future work should explore other model pairs and architectures or higher parameters versions, directly integrating the Qwen2.5-Math evaluation framework into mergenetic to unify evaluation in both search and test step.

See the appendix for the complete 1.5B evaluation table and the 7B plots and result tables. The full implementation is available at https://github.com/Marioiacobelli/Evolutionary-Merging-for-L2S-Math-Reasoning.git.

References

DeepSeek-AI. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025. URL https://arxiv.org/abs/2501.12948.

Ilharco, G., Wortsman, M., Gadre, S., Gontijo Lopes, R., Roelofs, R., Namkoong, H., Farhadi, A., and Schmidt, L. Editing models with task arithmetic. *arXiv preprint arXiv:2212.04089*, 2023. URL https://arxiv.org/abs/2212.04089.

Mencattini, T., Minut, A. R., Crisostomi, D., Santilli, A., and Rodolà, E. Merge3: Efficient evolutionary merging on consumer-grade gpus. *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, 2025. URL https://arxiv.org/abs/2502.10436.

Minut, A. R., Mencattini, T., Santilli, A., Crisostomi, D., and Rodolà, E. Mergenetic: a simple evolutionary model merging library, 2025. URL https://arxiv.org/abs/2505.11427. arXiv:2505.11427 [cs.LG].

Wu, H., Yao, Y., Liu, S., Liu, Z., Fu, X., Han, X., Li, X., Zhen, H.-L., Zhong, T., and Yuan, M. Unlocking efficient long-to-short llm reasoning with model merging. 2025. URL https://arxiv.org/abs/2503.20641.

Yadav, P., Vu, T., Lai, J., Chronopoulou, A., Faruqui, M., Bansal, M., and Munkhdalai, T. Ties-merging: Resolving interference when merging models. *arXiv preprint arXiv:2306.01708*, 2023. doi: 10.48550/arXiv.2306.01708. URL https://doi.org/10.48550/arXiv.2306.01708.

Yang, A., Zhang, B., Hui, B., Gao, B., Yu, B., Li, C., Liu, D., Tu, J., Zhou, J., Lin, J., Lu, K., Xue, M., Lin, R., Liu, T., Ren, X., and Zhang, Z. Qwen2.5-math technical report: Toward mathematical expert model via self-improvement. *arXiv preprint arXiv:2409.12122*, 2024.

Yao, Y., Liu, S., Liu, Z., Li, Q., Liu, M., Han, X., Guo, Z., Wu, H., and Song, L. Activation-guided consensus merging for large language models, 2025. URL https://arxiv.org/abs/2505.14009. arXiv:2505.14009 [cs.CL].

Appendix

Additional Experimental Attempts

In this section, we briefly report on several experimental configurations that were tested during development but ultimately discarded due to poor performance, methodological issues, or architectural incompatibilities.

- **Qwen2.5 as common pre-trained checkpoint.** The standard TA and TIES merging strategies assume two fine-tuned models derived from a shared pre-trained checkpoint. In our setup, however, we merged `DeepSeek-R1-Distill-Qwen` (distilled from DeepSeek-R1 with Qwen2.5-Math as the student) and `Qwen2.5-Math` (fine-tuned from Qwen2.5 on mathematical reasoning data). Since these models do not share an identical base checkpoint, the canonical assumption for task-vector approaches does not hold. We tested `Qwen2.5` as a shared base model, considering it the common origin for both DeepSeek-R1-Distill-Qwen and Qwen2.5-Math. However, this setup produced poor results: evolutionary merging was extremely slow, and convergence to low-accuracy solutions.

- **Qwen2.5-Math as common pre-trained checkpoint.** Given that `Qwen2.5-Math` shares its architecture with the student of DeepSeek-R1-Distill-Qwen, we tested it as a base pre-trained checkpoint. This, however, created a degenerate case in which one task vector was null, effectively reducing TA/TIES to a masked linear interpolation between parameters. As a result, this approach is equivalent of having a base model and a single target model.

- **Qwen2.5-Math as Base, DeepSeek as Target.** Another variant used `Qwen2.5-Math` as the base and `DeepSeek-R1-Distill-Qwen` as the target, similar to (Wu et al., 2025) work. This configuration was limited by Qwen2.5-Math's maximum positional embedding of 4096 tokens, which constrained the available context length, preventing the exploration of high accuracy solutions requiring longer contexts, so this setup was also discarded.

- **Different prompt templates.** In the ACM work, the `qwen25-math-cot` template is used, structured as a chat-style prompt with explicit `system`, `user`, and `assistant` roles. However, this template relies on special ChatLM tokens (<|im_start|>, <|im_end|>), which are not supported by DeepSeek's tokenizer. When we attempted to adopt this template for evaluating both baseline and merged models, we observed a deterioration in performance due to this incompatibility. Therefore, in our final evaluation setup we employ the `deepseek-math` prompt template, which is consistent with the model's recommended usage and was also used throughout the optimization process.

Negative Transfer Analysis

While the main analysis focused on optimizing accuracy and response length, it is equally important to verify that the merging process does not discard useful knowledge from the original models. To this end, we evaluate **Negative Transfer Rate (NTR)** for both TA and TIES-based merging.

We define negative transfer as cases where the merged model fails on an example that was correctly solved by at least one of the original models. For a given input i, let:

- $Y_{i,m_j} \in \{0, 1\}$ be the correctness of base model m_j.
- $Y_{i,\tilde{m}} \in \{0, 1\}$ be the correctness of merged model $\tilde{m}$.

We define a binary indicator $n_i = 1$ if there exists a j such that $Y_{i,m_j} = 1$ but $Y_{i,\tilde{m}} = 0$.

The Negative Transfer Rate (NTR) is then computed as:

$$\text{NTR} = \frac{\sum_i n_i}{\sum_i \mathbf{1}\big[\, \exists j \;:\; Y_{i,m_j} = 1 \,\big]}$$

This metric quantifies how often the merged model loses knowledge that was present in at least one of the base models. Lower NTR values indicate better preservation of original capabilities.

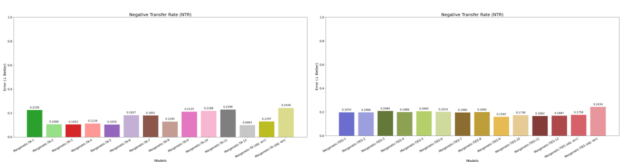

Figure 3. Negative Transfer Rate (NTR) for TA-based (left) and TIES-based (right) models

The results show that models obtained through multi-objective and accuracy-driven merging maintain a consistently low Negative Transfer Rate (NTR), confirming their ability to preserve knowledge from the original models. On the other hand, models optimized solely for output length tend to suffer higher negative transfer. This effect is also visible in some points along the Pareto front, particularly those representing non-dominated solutions on the length objective. These cases highlight the trade-off between preserving model capabilities and minimizing response verbosity.

Response Comparison

Question (GSM8k test set, idx = 383): Millie decides to open a lemonade stand. She spends $18 to buy enough supplies to make 3 pitchers of lemonade. Each pitcher holds 12 cups of lemonade. She sells each cup of lemonade for $1. She sells an average of 4 cups per hour that her lemonade stand is open. If Millie sells all of the lemonade, how much profit will she make per hour that she spends running the lemonade stand?

Gold Answer: 2

Qwen2.5-Math:

> Millie sells 3 pitchers of lemonade, and each pitcher holds 12 cups. So she sells 3 * 12 = 36 cups of lemonade. She sells an average of 4 cups per hour. So she sells 36 / 4 = 9 hours worth of lemonade. She spends $18 to buy supplies, so her profit is 9 * $1 - $18 = $9. The answer is 9.

Deepseek-Math:

> To determine Millie's profit per hour, we need to calculate both her total revenue and her total costs.
>
> First, let's find out how much she spends on supplies. She spends $18 to buy enough supplies for 3 pitchers of lemonade.
>
> Next, we'll calculate the total number of cups of lemonade she can make. Each pitcher holds 12 cups, so 3 pitchers equal 36 cups.
>
> She sells each cup for $1, so the total revenue from selling all the lemonade is $36.
>
> Now, we'll determine how many hours she needs to run her lemonade stand to sell all the lemonade. She sells an average of 4 cups per hour, so dividing the total cups by the hourly sales rate gives us 9 hours.
>
> To find her total profit, we subtract her costs from her total revenue: $36 - $18 = $18.
>
> Finally, to find the profit per hour, we divide the total profit by the number of hours she spends running the lemonade stand: $18 / 9 hours = $2 per hour.
>
> </think>
>
> To determine Millie's profit per hour, let's break down the costs and revenues step by step.
>
> 1. **Total Cost**
> Supplies: Millie spends $18 to buy enough supplies for 3 pitchers of lemonade.
>
> 2. **Total Revenue**
> Each pitcher holds 12 cups, so 3 pitchers = 36 cups.
> Selling Price = $1 per cup.
> Total Revenue = 36 cups × $1 = $36.
>
> 3. **Hours to Sell All Lemonade**
> 36 cups ÷ 4 cups/hour = 9 hours.
>
> 4. **Total Profit**
> Total Profit = $36 - $18 = $18.
>
> 5. **Profit per Hour**
> $18 ÷ 9 hours = $2 per hour.
>
> $$\boxed{2}$$

Mergenetic-TA:

> To determine Millie's profit per hour, we need to follow these steps:
>
> 1. **Calculate the total number of cups of lemonade she can make:**
> 3 pitchers × 12 cups = 36 cups.
>
> 2. **Determine the total revenue from selling all the lemonade:**
> 36 cups × $1 = $36.
>
> 3. **Calculate the total time her lemonade stand is open:**
> 36 cups ÷ 4 cups/hour = 9 hours.
>
> 4. **Determine the total cost of running the lemonade stand:**
> Total cost = $18.
>
> 5. **Calculate the total profit:**
> $36 - $18 = $18.
>
> 6. **Find the profit per hour:**
> $18 ÷ 9 hours = $2.
>
> Thus, Millie's profit per hour is $\boxed{2}$.

Evaluation Tables

Accuracy (%) and average response length (tokens, in parentheses) on six mathematical reasoning benchmarks. The last two columns report the mean accuracy and mean response length across all benchmarks, including the change in accuracy (Δ) and percentage length reduction relative to the DeepSeek-R1 baseline. Best results for each benchmark and overall are in **bold**.

Model/Method	AIME24	College Math	GSM8k	MATH500	Minerva Math	OlympiadBench	Avg. Accuracy (Δ)	Avg. Length (Δ%)
Deepseek-R1-Distill-Qwen-1.5B	**23.3** (8432.93)	42.8 (1776.60)	79.0 (628.74)	73.0 (2570.01)	20.2 (3014.85)	35.0 (5814.21)	45.5	3756.90
Qwen2.5-Math-1.5B	0.0 (1107.5)	15.8 (606.30)	75.7 (110.00)	51.2 (327.20)	11.0 (1032.20)	11.9 (1337.70)	27.6	753.50
Average Merging	13.3 (2428.60)	43.4 (852.77)	77.4 (418.52)	71.8 (1025.02)	29.0 (1035.47)	35.0 (1654.97)	45.0 (-0.5)	1235.73 (-67.1%)
Task Arithmetic	3.3 (3415.40)	38.2 (1082.90)	75.1 (684.02)	65.2 (1459.65)	29.0 (1343.04)	33.0 (2329.62)	40.6 (-4.9)	1719.10 (-54.2%)
TIES Merging	6.7 (2624.73)	44.5 (903.49)	78.5 (412.79)	71.4 (1111.39)	29.8 (1046.53)	32.4 (1918.40)	43.9 (-1.6)	1336.21 (-64.5%)
ACM-TA	16.7 (3240.90)	29.6 (1067.30)	76.8 (438.00)	68.8 (1309.60)	25.3 (1214.70)	31.1 (2291.60)	41.4 (-4.1)	1593.68 (-57.6%)
ACM-TIES	10.0 (2688.50)	37.9 (848.50)	78.4 (962.10)	71.4 (1235.60)	28.7 (1350.50)	33.8 (1849.60)	43.4 (-2.1)	1489.13 (-60.4%)
Mergenetic-TA-01	10.0 (2190.07)	42.0 (712.49)	72.6 (388.21)	63.6 (800.33)	29.0 (850.11)	28.4 (1235.35)	40.9 (-4.6)	1029.76 (**-72.6%**)
Mergenetic-TA-02	13.3 (4910.73)	46.8 (996.79)	85.5 (449.16)	77.2 (1472.97)	33.1 (1360.07)	41.8 (2947.60)	49.6 (+4.1)	2022.55 (-46.2%)
Mergenetic-TA-03	13.3 (5366.27)	47.3 (984.17)	**86.0** (449.82)	78.0 (1366.72)	30.5 (1326.47)	42.1 (2868.28)	49.5 (+4.0)	2059.29 (-45.2%)
Mergenetic-TA-04	10.0 (4109.77)	**47.8** (874.50)	84.3 (392.29)	75.6 (1208.28)	33.5 (1071.92)	37.5 (2319.09)	48.1 (+2.6)	1655.31 (-55.9%)
Mergenetic-TA-05	13.3 (4491.00)	46.7 (994.72)	85.5 (442.36)	78.6 (1365.52)	**33.8** (1271.67)	**42.2** (2757.72)	50.0 (+4.5)	1887.50 (-49.8%)
Mergenetic-TA-06	10.0 (3895.20)	45.0 (812.10)	76.0 (408.27)	67.6 (1026.99)	27.6 (1070.97)	32.6 (1576.86)	43.1 (-2.4)	1375.07 (-63.4%)
Mergenetic-TA-07	16.7 (3677.10)	45.1 (792.77)	75.7 (394.22)	68.2 (1144.98)	30.1 (932.77)	32.1 (1643.13)	44.6 (-0.9)	1397.16. (-62.8%)
Mergenetic-TA-08	6.7 (4148.77)	47.2 (885.37)	84.7 (408.53)	76.6 (1255.91)	32.4 (1033.71)	39.6 (2270.06)	47.9 (+2.4)	1700.06 (-54.8%)
Mergenetic-TA-09	16.7 (3088.83)	42.8 (738.95)	74.3 (424.81)	63.6 (879.30)	26.1 (787.86)	29.9 (1347.95)	42.2 (-3.3)	1277.12 (-66.0%)
Mergenetic-TA-10	10.0 (3188.70)	42.4 (722.52)	73.9 (405.63)	62.4 (829.29)	29.4 (814.54)	29.8 (1208.89)	41.3 (-4.2)	1194.93 (-68.2%)
Mergenetic-TA-11	6.7 (2466.63)	42.1 (719.15)	72.9 (392.17)	61.6 (817.31)	28.3 (841.39)	29.2 (1193.05)	40.1 (-5.4)	1171.12 (-68.8%)
Mergenetic-TA-12	**23.3** (4071.43)	47.5 (943.80)	85.2 (389.46)	**78.8** (1383.28)	31.6 (1067.35)	40.6 (2223.76)	**51.2** (+5.7)	1686.68 (-55.1%)
Mergenetic-TIES-01	10.0 (2891.47)	43.4 (787.52)	75.4 (422.42)	65.6 (971.57)	27.2 (1122.50)	32.0 (1852.19)	42.3 (-3.2).	1491.95 (-60.3%)
Mergenetic-TIES-02	13.3 (3499.37)	42.8 (777.64)	76.9 (414.79)	66.2 (1054.03)	24.3 (905.28)	31.3 (1765.11)	42.5 (-3.0)	1402.04 (-62.7%)
Mergenetic-TIES-03	10.0 (2102.50)	42.7 (755.97)	74.8 (380.73)	65.2 (910.50)	25.7 (825.56)	29.8 (1391.78)	41.4 (-4.1)	1227.84 (-67.3%)
Mergenetic-TIES-04	13.3 (2291.13)	42.9 (764.64)	74.8 (446.66)	65.2 (1021.58)	25.0 (1099.29)	33.0 (1778.88)	42.4 (-3.1)	1483.03 (-60.5%)
Mergenetic-TIES-05	10.0 (2721.03)	42.7 (781.91)	75.3 (410.29)	65.2 (902.64)	24.6 (819.16)	29.2 (1471.15)	41.2 (-4.3)	1224.36 (-67.4%)
Mergenetic-TIES-06	6.7 (2633.77)	43.2 (764.21)	75.9 (389.90)	65.2 (773.75)	26.8 (754.57)	30.2 (1351.81)	41.3 (-4.2)	1194.00. (-68.2%)
Mergenetic-TIES-07	10.0 (2836.53)	43.1 (765.97)	76.0 (419.59)	67.0 (1075.04)	24.6 (1065.32)	31.9 (1687.79)	42.1 (-3.4)	1341.04 (-64.3%)
Mergenetic-TIES-08	13.3 (3431.20)	42.7 (747.19)	76.2 (421.34)	67.4 (888.31)	23.2 (946.93)	31.3 (1749.26)	42.4 (-3.1)	1353.37 (-64.0%)
Mergenetic-TIES-09	13.3 (4290.83)	45.7 (863.73)	78.8 (616.45)	69.4 (1252.36)	29.8 (1405.45)	31.7 (2506.18)	44.8 (-0.7)	1822.50 (-51.5 %)
Mergenetic-TIES-10	10.0 (2680.37)	45.0 (881.3)	78.5 (532.89)	66.6 (1374.76)	28.7 (1364.69)	31.7 (2195.15)	43.4 (-2.1)	1504.53 (-59.9%)
Mergenetic-TIES-11	6.7 (3118.1)	45.4 (873.27)	78.5 (525.6)	68.2 (1241.96)	27.6 (1511.53)	31.7 (2264.88)	43.0 (-2.5)	1589.23 (-57.7%)
Mergenetic-TIES-12	16.7 (2718.97)	45.4 (883.43)	78.9 (544.14)	66.4 (1370.32)	30.1 (1231.92)	31.7 (2272.19)	44.9 (-0.6)	1669.83 (-52.7%)
Mergenetic-TA (single-obj: acc)	16.7 (6911.9)	43.6 (1385.35)	85.0 (588.31)	75.6 (1982.42)	27.9 (2097.53)	38.8 (4149.05)	47.9 (+2.4)	2852.43 (-24.0%)
Mergenetic-TA (single-obj: len)	6.7 (2466.63)	42.1 (719.15)	72.9 (392.17)	61.6 (817.31)	28.3 (841.39)	29.2 (1193.05)	40.1 (-5.4)	1071.6 (-71.5%)
Mergenetic-TIES (single-obj: acc)	10.0 (3253.43)	45.5 (857.39)	79.5 (522.96)	68.6 (1332.76)	29.8 (1260.41)	30.8 (2412.89)	44.0 (-1.5)	1606.6 (-57.2%)
Mergenetic-TIES (single-obj: len)	6.7 (2651.87)	42.4 (708.46)	73.4 (398.34)	62.2 (820.06)	27.6 (855.06)	28.7 (1237.55)	40.2 (-5.3)	1111.9 (-70.4%)

Figure 4. Evaluation Table for 1.5B Models

Model/method	AIME24	College Math	GSM8k	MATH500	Minerva Math	OlympiadBench	Avg. Accuracy (Δ)	Avg. Length. (Δ%)
Deepseek-R1-Distill-Qwen-7B	**40.0** (7975.00)	45.8 (3159.97)	90.0 (1833.92)	**86.6** (3536.35)	**42.6** (4695.65)	47.4 (6362.14)	**58.7**	4593.51
Qwen2.5-Math-7B	0.0 (1450.00)	26.2 (759.44)	83.7 (103.96)	59.4 (432.83)	9.6 (946.78)	17.0 (1228.25)	32.65	820.88
Average Merging	16.7 (2328.07)	49.6 (835.79)	91.4 (355.11)	84.0 (897.67)	38.6 (804.32)	47.7 (1584.44)	54.7 (-4.0)	1134.23 (-75.3%)
Task Arithmetic	30.0 (3467.23)	47.5 (1247.63)	92.0 (531.72)	85.2 (1467.19)	41.9 (1360.42)	46.7 (2577.03)	57.2 (-1.5)	1775.20 (-61.4%)
TIES Merging	26.7 (2644.23)	48.7 (881.54)	89.9 (347.79)	84.2 (976.55)	39.3 (844.61)	48.1 (1708.50)	56.2 (-2.5)	1233.87 (-73.1%)
ACM-TA	33.3 (3409.90)	40.1 (1494.50)	90.6 (652.40)	83.8 (1636.10)	38.6 (1514.10)	46.7 (2706.20)	55.5 (-3.2)	1902.20 (-58.6%)
ACM-TIES	33.3 (3076.40)	40.3 (1284.60)	**92.2** (538.30)	84.0 (1450.10)	38.6 (1153.10)	46.4 (2356.80)	55.8 (-2.9)	1643.88 (-64.2%)
Mergenetic-TA-01	20.0 (3273.2)	48.0 (826.1)	87.5 (356.8)	79.6 (1190.6)	37.1 (840.4)	41.5 (2055.2)	52.3 (-6.4)	1423.72 (-69.0%)
Mergenetic-TA-02	16.7 (3874.6)	48.2 (825.3)	87.4 (364.9)	82.2 (1044.7)	37.1 (733.0)	44.7 (2007.6)	52.7 (-6.0)	1475.01 (-67.9%)
Mergenetic-TA-03	23.3 (4308.2)	48.4 (918.6)	**92.2** (341.2)	83.8 (1290.6)	37.5 (900.0)	**49.5** (2388.1)	55.8 (-2.9)	1691.11 (-63.2%)
Mergenetic-TA-04	10.0 (2357.0)	47.5 (780.3)	85.4 (378.9)	76.8 (1004.8)	33.1 (753.1)	40.7 (1775.6)	48.9 (-9.8)	1174.95 (-74.4%)
Mergenetic-TA-05	26.7 (3623.4)	48.7 (916.7)	92.0 (351.7)	85.0 (1191.8)	38.6 (1045.3)	48.1 (2384.9)	56.5 (-2.0)	1585.62 (-65.5%)
Mergenetic-TA-06	20.0 (4309.2)	**49.1** (918.6)	91.9 (338.8)	84.4 (1076.5)	37.9 (912.5)	46.2 (2432.0)	54.9 (-3.8)	1664.57 (-63.8%)
Mergenetic-TA-07	20.0 (4149.8)	48.4 (932.3)	91.8 (364.8)	84.2 (1218.7)	40.8 (779.3)	47.3 (2628.3)	55.4 (-3.3)	1678.84 (-63.5%)
Mergenetic-TA-08	16.7 (3581.6)	47.6 (807.3)	86.1 (359.6)	79.6 (1134.1)	38.2 (842.3)	43.1 (1942.1)	51.9 (-6.8)	1444.51 (-68.6%)
Mergenetic-TIES-01	20.0 (2154.7)	42.9 (1274.8)	82.9 (599.8)	70.8 (934.8)	32.7 (1151.0)	36.4 (1609.8)	47.6 (-11.1)	1287.48 (-72.0%)
Mergenetic-TIES-02	33.3 (8514.4)	43.6 (3639.3)	82.6 (2752.2)	82.8 (3915.6)	35.3 (5619.5)	42.8 (6728.6)	53.4 (-5.3)	5194.93 (+13.1%)
Mergenetic-TIES-03	13.3 (2521.2)	42.3 (1270.7)	82.9 (656.3)	70.0 (888.0)	31.2 (968.4)	36.0 (1612.1)	45.9 (-12.8)	1319.45 (-71.3%)
Mergenetic-TIES-04	16.7 (9433.7)	37.3 (4860.4)	77.6 (3842.0)	66.4 (5651.1)	31.6 (6675.2)	27.9 (8352.5)	42.9 (-15.8)	6469.16 (+40.83%)
Mergenetic-TIES-05	23.3 (2445.0)	42.3 (1208.9)	82.6 (668.1)	70.0 (923.6)	34.2 (957.3)	36.6 (1554.4)	48.2 (-10.5)	1292.90 (-71.9%)

Figure 5. Evaluation Table for 7B Models

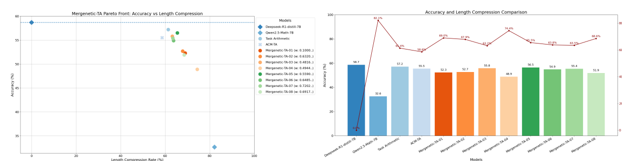

Figure 6. Evaluation plots for Mergenetic-TA (7B)