Unlocking Efficient Long-to-Short LLM Reasoning with Model Merging

Han Wu*, **Yuxuan Yao***, **Shuqi Liu***, **Zehua Liu***, **Xiaojin Fu, Xiongwei Han, Xing Li**
Hui-Ling Zhen, Tao Zhong, Mingxuan Yuan

Huawei Noah's Ark Lab

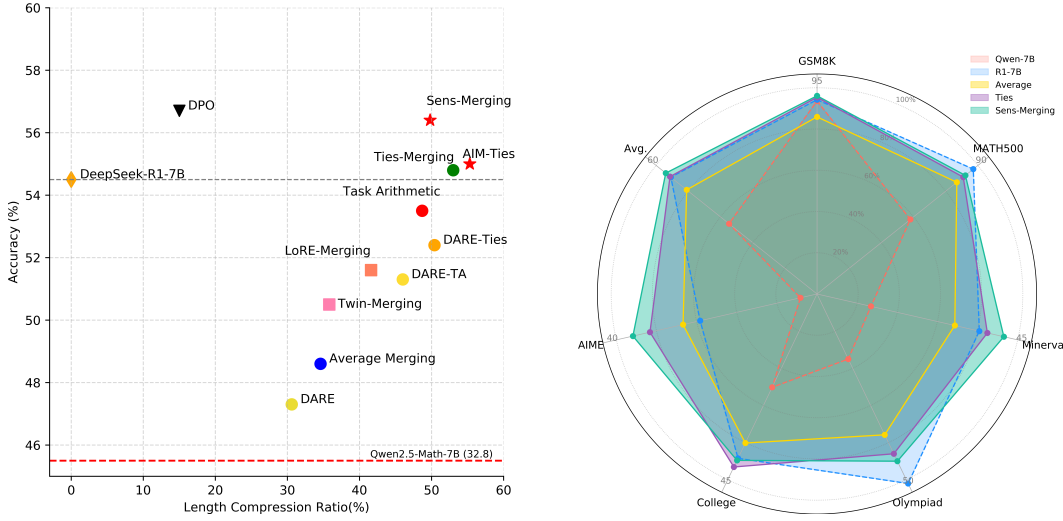{wu.han1,hanxiongwei,zhongtao5,Yuan.Mingxuan}@huawei.com

Figure 1: Performance of various model merging methods on DeepSeek-R1-7B and Qwen2.5-Math-7B models.

(TL;DR) Summary of our findings:

- Model merging is a highly efficient approach for long-to-short reasoning, as it directly operates on model parameters **without requiring additional training**.

- Task-vector based merging methods, especially like TA and Ties-Merging, can achieve long-to-short reasoning with around **50% length reduction** alongside **accuracy parity or even marginal gains** on 7B models.

- SVD-based merging methods exhibit limited effectiveness, delivering moderate performance and serving as viable alternatives only when task vectors inherently possess low-rank spectral characteristics.

- Activation-based merging is the future, as it demonstrates impressive performance in terms of both reasoning accuracy (+1.9) and response length compression ratios (-49.8%).

- Model merging methods applied to 1.5B-scale models remain effective on simple tasks. Smaller models struggle to learn long CoT reasoning ability through model merging.

- The merging of large-scale models (14B and 32B) poses significant challenges in simultaneously maintaining reasoning performance while substantially reducing response length.

Work in progress.
* indicates equal contribution.

ABSTRACT

The transition from System 1 to System 2 reasoning in large language models (LLMs) has marked significant advancements in handling complex tasks through deliberate, iterative thinking. However, this progress often comes at the cost of efficiency, as models tend to overthink, generating redundant reasoning steps without proportional improvements in output quality. Long-to-Short (L2S) reasoning has emerged as a promising solution to this challenge, aiming to balance reasoning depth with practical efficiency. While existing approaches, such as supervised fine-tuning (SFT), reinforcement learning (RL), and prompt engineering, have shown potential, they are either computationally expensive or unstable. Model merging, on the other hand, offers a cost-effective and robust alternative by integrating the quick-thinking capabilities of System 1 models with the methodical reasoning of System 2 models. In this work, we present a comprehensive empirical study on model merging for L2S reasoning, exploring diverse methodologies, including task-vector-based, SVD-based, and activation-informed merging. Our experiments reveal that model merging can reduce average response length by up to 55% while preserving or even improving baseline performance. We also identify a strong correlation between model scale and merging efficacy with extensive evaluations on 1.5B/7B/14B/32B models. Furthermore, we investigate the merged model's ability to self-critique and self-correct, as well as its adaptive response length based on task complexity. Our findings highlight model merging as a highly efficient and effective paradigm for L2S reasoning, offering a practical solution to the overthinking problem while maintaining the robustness of System 2 reasoning. This work can be found on Github `https://github.com/hahahawu/Long-to-Short-via-Model-Merging`.

1 INTRODUCTION

The development of large language models (LLMs) has transitioned from System 1 to System 2 reasoning (Yu et al., 2024b; Li et al., 2025b), marked by the advent of advanced reasoning models such as OpenAI's o1/o3 (OpenAI, 2024b; 2025), QwQ (Qwen, 2025), and DeepSeek-R1 (DeepSeek-AI et al., 2025). System 1 reasoning models, such as GPT-4o (OpenAI, 2024a), LLaMA-3 (Grattafiori & Team, 2024), and DeepSeek-V3 (DeepSeek-AI, 2025), are characterized by their capacity to generate straightforward, intuitive responses through rapid cognitive processing. However, this efficiency often limits their ability to handle intricate or multi-faceted tasks. In contrast, System 2 reasoning models prioritize deliberate, methodical analysis, engaging in iterative self-critique, error correction, and exhaustive evaluation prior to finalizing outputs. This systematic approach enhances their robustness, enabling superior performance on complex reasoning tasks.

Although significant advancements have been achieved in reasoning models, their practical efficiency remains limited by iterative thinking processes that often involve redundant reasoning steps and repetitive iterations. This inefficiency, commonly also referred to as the overthinking problem (Chen et al., 2025), occurs when models excessively extend their deliberation without yielding proportional improvements in output quality. To mitigate this challenge, Long-to-Short (L2S) reasoning has garnered significant attention in recent research (Ma et al., 2025; Xia et al., 2025). Existing studies have explored diverse methods, including SFT (Xia et al., 2025), reinforcement learning (Chen et al., 2025; Luo et al., 2025), token-budget-aware inference (Han et al., 2025) , prompt engineering (Luo et al., 2025), and model merging (Team et al., 2025). However, training-based methods are generally expensive, as they typically require the collection of aligned short and long answers for training. Prompting-based methods, on the other hand, have been shown to be unstable, as their performance is highly sensitive to changes in models and prompts. Research on model merging for long-to-short reasoning remains limited, with the only effort just focusing on the simplest average merging approach, which shows inferior performance (Team et al., 2025). Intuitively, if we view quick thinking and slow thinking as two distinct abilities developed through training on different tasks, model merging has been shown to effectively integrate these multi-task capabilities into a unified model (Ilharco et al., 2023; Yadav et al., 2023; Yu et al., 2024a). Therefore, we argue that the merging approach represents a promising solution to the long-to-short reasoning problem.

In this work, we conduct a systematic empirical analysis of diverse model merging methodologies for long-to-short reasoning tasks. For example, we directly merge the quick-thinking System 1 model (like Qwen2.5-Math-7B) with its System 2 counterpart (like DeepSeek-R1-Distill-Qwen-7B) using existing merging approaches. Remarkably, our findings reveal that model merging serves as a highly efficient and effective paradigm for long-to-short optimization: it can reduce average response length up to 53% while preserving baseline performance metrics, with certain configurations even achieving consistent performance improvements. Through a comprehensive analysis of various model merging methods on 7B models, we present several insightful observations: 1) task-vector-based merging methods demonstrate moderate effectiveness with minimal effort; 2) SVD-based merging methods generally yield unsatisfactory results; and 3) activation-informed merging shows promising performance. Furthermore, our analysis reveals a strong correlation between model scale and merging efficacy. Specifically, smaller models (e.g., Qwen2.5-Math-1.5B) struggle to obtain the long-CoT reasoning ability through parameter fusion, whereas larger-scale models (e.g., Qwen2.5-14B/32B) can largely preserve reasoning performance, though the reduction in response length remains relatively insignificant. More interestingly, we further investigate whether the merged model retains the ability for self-critique and self-correction, as well as how its response length correlates with the difficulty of the questions. The results reveal that the merged model essentially functions as a more intelligent slow-thinking model, capable of adjusting its output length based on the complexity of the problem and avoiding redundant reflections.

2 RELATED WORK

2.1 LONG-TO-SHORT REASONING

The evolution of large language models (LLMs) has transitioned from System 1 to System 2 reasoning (Li et al., 2025b). System 1 models excel at generating quick, intuitive responses but often struggle with complex tasks. In contrast, System 2 models emphasize deliberate and methodical analysis, incorporating self-critique and error correction to enhance their performance on intricate reasoning tasks. However, their iterative processes can lead to inefficiencies, commonly referred to as the "overthinking problem" (Chen et al., 2025; Yang et al., 2025), where excessive deliberation fails to significantly improve the quality of the output. To mitigate this challenge, Long-to-Short (L2S) (Team et al., 2025) reasoning has garnered significant attention in recent research.

An effective strategy for reducing inference costs without compromising accuracy involves compressing Chain of Thoughts (CoTs). TokenSkip (Xia et al., 2025) prioritizes tokens by semantic significance then conduct compression through selective omission. CoT-Valve (Ma et al., 2025) identifies a parameter space direction and adjusting step sizes systematically. Furthermore, to optimize computational resource allocation—prioritizing challenging tasks over simpler ones—O1-Pruner (Luo et al., 2025) integrates sampling and reinforcement learning (RL) fine-tuning, while DAST (Shen et al., 2025) introduces a difficulty metric based on token length budget to calibrate reward scores. Besides, there is a growing interest in leveraging target budgets to facilitate length control in conjunction with RL (Aggarwal & Welleck, 2025). Additionally, prompt-based approaches have been extensively explored in recent works(Team et al., 2025; Luo et al., 2025). Whether through short responses in few-shot settings or encouraging the model to provide "concise" replies in demonstrations (Munkhbat et al., 2025), these attempts have demonstrated instability, as their performance is highly sensitive to variations in models and prompts.

2.2 MODEL MERGING

Model merging (Yang et al., 2024), as a complementary approach to training-based methods, enables the integration of multiple task-specialized models into a unified model (Wortsman et al., 2022; Yadav et al., 2023). This approach enhances model performance on individual tasks by merging checkpoints without the need for additional training (Yadav et al., 2023; Ilharco et al., 2023), mitigates the issue of catastrophic forgetting (Alexandrov et al., 2024), and achieves the long-to-short reasoning (Team et al., 2025). In this work, we categorize existing model merging approaches into three groups: 1) vanilla task-vector-based merging methods (Wortsman et al., 2022; Ilharco et al., 2023; Yadav et al., 2023; Yu et al., 2024a), which represent fine-tuned features as task vectors and perform arithmetic operations on them to derive the merged model; 2) SVD-based merging methods (Lu et al., 2024; Liu et al., 2025d), characterized by their ability to identify and exploit the

low-rank features of task vectors; and 3) Activation-based methods (Nobari et al., 2025; Liu et al., 2025b), which utilize input activations to assign varying importance scores to the merging models.

Besides, MoE-based merging methods (Sukhbaatar et al., 2024; Liu et al., 2025c) also play an important role in this field. However, we exclude the discussion of these methods as they typically involve modifications to the model architecture and model size.

3   BACKGROUND

3.1   MODEL MERGING

Model merging seeks to integrate multiple fine-tuned (FT) models, derived from a pre-trained (PT) model θ_0, into a unified model that consolidates knowledge from diverse sources. Given K FT models to be merged, denoted as $\theta_1, \ldots, \theta_K$, the goal is to produce a single model θ_M that inherits the capabilities of the individual models.

Average Merging   Average merging (Wortsman et al., 2022) is a simple and effective method to enhance overall performance by performing an arithmetic average of the model weights. It reduces variance by smoothing random errors, especially when base models are diverse and exhibit low bias. However, its effectiveness depends on the quality and diversity of the base models; high bias across models limits its improvement potential.

Task Arithmetic (TA)   In most existing task-vector-based approaches, the base model θ_0 is essential for computing task vectors (Ilharco et al., 2023), which generally encapsulate the knowledge acquired during fine-tuning. A task vector is defined as the parameter shift between an FT model and its corresponding base model, expressed as $\delta_k = \theta_k - \theta_0$. The merged model θ_M is then obtained by aggregating the task vectors into the base model, as $\theta_M = \theta_0 + \sum_{k=1}^{K} \lambda_k \cdot \delta_k$, where λ_k represents the weight coefficient, which can either be manually set as a constant or determined through optimization.

Ties Merging   TIES Merging (Yadav et al., 2023) is an efficient method for integrating parameters from multiple FT models, addressing redundancy and conflicts. Its key steps include: (1) pruning parameters, retaining significant deviations from pre-trained weights; (2) resolving conflicts via majority voting or alignment; and (3) weighted aggregation of significant parameters to form the final model. This approach reduces noise and enhances generalization, particularly for integrating fine-tuned large language models (LLMs) across related tasks.

DARE   (Yu et al., 2024a) DARE Merging is a lightweight approach, whose core steps include: (1) randomly dropping redundant parameters (e.g., those with minimal gradient changes) to reduce noise; (2) adjusting the direction of retained parameters to resolve conflicts between models; and (3) performing weighted integration of key parameters to preserve essential knowledge. This method significantly reduces computational overhead while maintaining model performance, making it suitable for rapid fusion of multi-task fine-tuned models.

AIM   The core idea of AIM (Nobari et al., 2025) is to utilize activation space information to identify and protect critical weights in PT models, thereby minimizing alterations to these weights during the merging process. This approach is akin to weight regularization techniques in Continual Learning (CL), aiming to prevent catastrophic forgetting, which is the loss of performance on old tasks when learning new ones. The effectiveness of AIM across various merging methods and benchmarks underscores the importance of activation space information in model merging. As a supplementary method, AIM can be integrated with existing merging techniques to significantly enhance the performance and robustness of merged models.

LoRE-Merging   Existing merging methods rely on sparsely estimated task vectors but face two key limitations: dependence on base model parameters and task vector interference. LoRE-Merging (Liu et al., 2025d), a low-rank estimation framework, is proposed to address such issues. It constructs an approximate base model and low-rank task vectors via optimization, minimizing discrepancies between fine-tuned and merged models. Using coordinate descent and singular value

thresholding, LoRE-Merging reduces task vector interference, demonstrating the effectiveness of low-rank estimation in model merging.

Twin-Merging Performance gaps between merged and fine-tuned models stem from conflicts among models and diverse testing data. Twin-Merging (Lu et al., 2024) resolves this by categorizing expert knowledge into generalizable shared knowledge and task-specific knowledge. Through compression and difference extraction, this knowledge is modularized. A router then dynamically integrates shared and task-specific knowledge based on input, similar to the Mixture of Experts approach, allowing for flexible adjustments. In our study, we eliminate the router training and directly utilize its singular value decomposition (SVD) merging part.

Sens-Merging Sens-Merging (Liu et al., 2025b) focuses on the varying importance of parameters within and across tasks during model merging. It operates at two levels: (1) within individual tasks, where parameter sensitivity analysis identifies critical layers impacting performance, and (2) across tasks, where task sensitivity analysis prioritizes models that enhance others' performance. By combining these analyses, Sens-Merging derives merging coefficients for fine-grained parameter control, enabling effective layer-wise merging. It also serves as a plug-and-play enhancement to task vector-based merging, improving flexibility and performance.

3.2 LLM Reasoning

We divide the CoT-based LLM reasoning into two categories, including short-CoT (quick-thinking) reasoning and long-CoT (slow-thinking) reasoning. Given an input X, the quick-thinking reasoning models directly output the final answer, represented as $\{\theta, X\} \rightarrow A$. In contrast, slow-thinking models generate both the extensive thinking process and the final answer, denoted as $\{\theta, X\} \rightarrow [T, A]$. While the reasoning process in slow-thinking models often enhances the quality of the final answer, the lengthy context significantly impairs inference efficiency, particularly when the reasoning content contains substantial repetitions or unnecessary reflections.

4 Exploring Model Merging Methods for Long-to-Short Reasoning

In this section, we explore the effect of model merging on the most frequently used 7B reasoning models, i.e. Qwen2.5-Math-7B and DeepSeek-R1-Distill-Qwen-7B (DeepSeek-AI et al., 2025). Then, we reveal the effectiveness of model merging on achieving long-to-short reasoning.

4.1 Experiment Setup

We conduct evaluations on popular reasoning datasets: GSM8K, MATH500, Minerva Math, Olympiadbench, College Math and AIME24. We utilize the public evaluation toolkit[1] provided by Qwen for better reproducibility. The quick-thinking models are evaluated with few shots and the slow-thinking models are in a zero-shot setting, as recommended in DeepSeek-AI et al. (2025). For activation-based merging methods, we adopt the **s1K** (Muennighoff et al., 2025) dataset which contains the high-quality aligned quick-thinking answer and slow-thinking answer for each question. The maximum length for quick-thinking models and slow-thinking models are 8K and 10K, respectively. We also establish a training-based baseline using DPO, where DeepSeek-R1-7B is trained on the s1K dataset with the short answers as positive samples.

4.2 Task-Vector based Merging Works

We first explore the typical task-vector based merging methods: **Average Merging**, **Task Arithmetic**, **Ties-Merging** and **DARE**. We consider the average merging as a special case of task-arithmetic merging.

As shown in the second part of Table 1, task-vector based methods demonstrates significant efficacy in compressing output length while preserving reasoning accuracy. The baseline of average merg-

---

[1]https://github.com/QwenLM/Qwen2.5-Math

Method \ Bench	GSM8K	MATH500	Minerva Math	Olympiad Bench	College Math	AIME24	Avg.
Qwen2.5-Math-7B	88.9	52.2	12.1	17.5	22.6	3.3	32.8
	(130.2)	(526.2)	(956.7)	(1259.2)	(794.8)	(1528.0)	
DeepSeek-R1-7B	89.3	87.4	36.4	51.0	39.8	23.3	54.5
	(1062.0)	(2825.9)	(3055.9)	(5793.8)	(2461.6)	(8675.4)	
DPO	89.6	88.4	33.8	48.6	39.7	40.0	56.7
	(983.6)	(2587.1)	(2304.9)	(4932.3)	(2073.5)	(7029.5)	(-15.0%)
Task-vector based Merging Methods							
Average Merging	81.6	78.2	30.9	37.9	36.1	26.7	48.6
	(636.4)	(1416.7)	(2277.4)	(3202.2)	(2065.7)	(5964.8)	(-34.6%)
Task Arithmetic	90.5	83.4	41.9	45.0	40.3	20.0	53.5
	(617.5)	(1617.4)	(1416.2)	(2650.6)	(1443.9)	(3594.6)	(-48.7%)
Ties-Merging	90.6	81.8	38.2	43.0	41.9	33.3	54.8
	(552.2)	(1492.9)	(1349.2)	(2473.7)	(1287.8)	(3302.1)	(-53.0%)
DARE	84.0	75.4	29.4	35.7	36.2	23.3	47.3
	(815.6)	(2237.1)	(2317.8)	(3266.3)	(2072.1)	(3803.1)	(-30.6%)
DARE-TA	87.9	84.0	29.4	41.6	38.3	26.7	51.3
	(600.4)	(1703.9)	(1567.2)	(2774.8)	(1533.9)	(3454.2)	(-46.0%)
DARE-Ties	89.6	82.4	37.5	41.3	40.2	23.3	52.4
	(584.3)	(1589.4)	(1307.4)	(2655.5)	(1378.5)	(3579.9)	(-50.4%)
SVD-based Merging Methods							
LoRE-Merging	84.8	80.8	37.5	40.9	35.5	30	51.6
	(531.7)	(1819.1)	(1944.9)	(3072.3)	(1875.9)	(3691.6)	(-41.6%)
Twin-Merging	87.6	79.6	31.2	40.0	38.1	26.7	50.5
	(778.1)	(1997.3)	(2079.9)	(3098.3)	(1870.5)	(3768.9)	(-35.8%)
Activation-based Merging Methods							
AIM-Ties	90.8	83.0	40.8	46.4	42.3	26.7	55.0
	(540.6)	(1374.5)	(1229.8)	(2323.8)	(1249.6)	(3265.9)	(-55.3%)
Sens-Merging	91.3	83.0	41.9	45.0	40.3	36.7	56.4
	(600.7)	(1650.2)	(1345.9)	(2673.2)	(1446.4)	(3245.9)	(-49.8%)

Table 1: Evaluations of different model merging methods on Qwen-7B models. The number in () indicates the average response length on the dataset.

ing reduces response length by 34.6% relative to the reasoning model while improving accuracy by 15.8% over the quick-thinking baseline. Although this approach incurs marginal performance degradation compared to the pure reasoning model, advanced merging paradigms effectively mitigate this tradeoff. Notably, TA and Ties-Merging achieve comprehensive optimization, delivering 48-53% length reduction alongside accuracy parity (within 1.0%) or even marginal gains (+0.3% on the average score). This finding is particularly significant as it demonstrates that long-to-short reasoning can be accomplished with minimal computational effort. Furthermore, DARE, as a plug-and-play method, can be combined with other task-vector-based merging approaches. However, it consistently underperforms compared to the original merging methods. We hypothesize two potential reasons for this: 1) as noted by Yu et al. (2024a), the performance of DARE is constrained by parameter shifts being within 0.002, whereas the parameter shifts between Qwen-7B models significantly exceed this threshold; 2) the random dropping of certain task vectors, a key feature of DARE, might inadvertently remove critical parameters. Supporting this assumption, we empirically observe that the drop ratio in DARE should be less than 0.5, rather than the larger value of 0.9 recommended by Yu et al. (2024a).

Upon closely examining the performance scores across each dataset, we present the following observations: 1) Accuracy improvements over the reasoning model are more readily achieved on datasets where the quick-thinking model and the reasoning model exhibit similar performance levels, such as GSM8K and College Math; 2) The merged model demonstrates limited capacity to surpass the

reasoning model when substantial initial performance disparities exist between the base models, such as MATH500 and OlympiadBench; 3) The reduction in response length is more pronounced on complex datasets, such as AIME24 and OlympiadBench.

> **Takeaway 1**
>
> *Task-vector based merging methods, especially like TA and Ties-Merging, can achieve long-to-short reasoning with around 50% length reduction alongside accuracy parity or even marginal gains.*

4.3 SVD-based Merging Underperforms

Based on the observation that task vectors often exhibit a limited number of dominant singular values, recent SVD-based merging methods aim to address task vector interference through low-rank approximation. In our investigation, we examined the following SVD-based merging methods: **LoRE-Merging** and **Twin-Merging**.

As shown in the third section of Table 1, both LoRE-Merging and Twin-Merging outperform the average merging method in terms of both length compression and reasoning accuracy, highlighting their effectiveness in long-to-short reasoning tasks. However, their performance falls short compared to advanced task-vector-based methods, such as TA and Ties-Merging. We hypothesize that the effectiveness of these approaches is highly dependent on the distribution of the domain singular values of the task vectors. Notably, we observe that the singular value distributions of the candidate base models deviate from the ideal distribution. One more interesting finding is that, although the overall performance of SVD-based merging methods is moderate, they appear to consistently perform well on complex tasks, such as AIME24.

> **Takeaway 2**
>
> *SVD-based merging methods exhibit limited effectiveness, delivering moderate performance and serving as viable alternatives only when task vectors inherently possess low-rank spectral characteristics.*

4.4 Activation-based Merging is the Future

Inspired by prior research in model quantization (Lin et al., 2024) and pruning (Sun et al., 2024; Liu et al., 2025a), activation-based merging methods have also been extensively studied. As observed in task-based merging methods (Ilharco et al., 2023; Nobari et al., 2025), the choice of coefficients plays a crucial role in achieving effective merging. The most intuitive approach is to leverage activations to search for optimal coefficients. In this work, we evaluate two recent approaches: **AIM** and **Sens-Merging**. For our experiments, we utilize the s1K dataset as the calibration dataset, as it provides high-quality, well-aligned short and long responses for each question.

For Sens-Merging, we use quick-thinking answers as calibration data for the R1 model to enhance its ability to generate concise responses and reduce reasoning path length. Meanwhile, slow-thinking answers are selected as calibration data for the Qwen Math model to promote its deep reasoning capabilities. We randomly choose 100 samples from the s1K dataset for calibration, making the sensitivity score computation more efficient. Similarly, we also use the sampled data as the calibration data of AIM-Ties which operates on the merged model from Ties-Merging and Qwen base model.

As evidenced in Table 1, both activation-based methodologies surpass the DeepSeek-R1-7B baseline across comprehensive metrics, achieving superior average scores paired with significant response length reduction. Specifically, applying AIM to the merged model from Ties-Merging yields further enhancements - elevating performance by 0.2 points while raising the compression ratio to 55.3%. For Sens-Merging, it even achieves the comparable reasoning performance with DPO while significantly reducing the response length by approximately 50%.

However, there are certain limitations in existing activation-based methods. For instance, Sens-Merging requires gradient computation during the backward pass, which can impact its efficiency. In contrast, AIM utilizes activations solely during the forward pass, enhancing the efficiency of

Bench	Accuracy			Length & Reflections		
Method	GSM8K	MATH500	AIME24	GSM8K	MATH500	AIME24
Qwen2.5-Math-1.5B	75.9	36.2	0.0	118.1	411.0	864.9
				[0]	[0]	[0]
DeepSeek-R1-1.5B	76.8	69.6	20.0	2744.3	4510.0	8952.3
				[1315]	[499]	[30]
Average Merging	26.6	12.2	0.0	3428.7	3890.6	3977.3
				[774]	[306]	[16]
Task Arithmetic	74.5	62.6	10.0	549.7	1619.4	3395.3
				[24]	[51]	[9]
Ties-Merging	75.9	64.2	10.0	393.8	1102.4	2818.7
				[24]	[53]	[3]
LoRE-Merging	64.1	50.8	6.7	763.8	2248.4	3833.3
				[118]	[197]	[15]
Sens-Merging	81.3	69.8	16.7	340.3	852.8	2102.6
				[9]	[8]	[0]

Table 2: Evaluations of various model merging methods on Qwen-1.5B models. The number in [] indicates the number of reflective responses on the dataset.

activation usage. Nonetheless, the performance of activation-based methods is inherently reliant on the choice of calibration data. In our experiments, we also evaluated AIM and Sens-Merging with alternative calibration datasets but observed inferior performance. While existing activation-based methods have achieved notable success, further research into more robust and efficient activation-based merging techniques remains a critical direction for future work.

> **Takeaway 3**
>
> *Activation-based merging methods demonstrate superior performance in terms of both reasoning accuracy and response length compression rates; however, their effectiveness is certainly dependent on the choice of the calibration dataset.*

5 ANALYSIS ON MODELS WITH DIFFERENT SCALES

Apart from the evaluations on Qwen-7B models, we further investigate the effectiveness of model merging across different model scales, including the smaller Qwen-1.5B model and larger models such as Qwen-14B and Qwen-32B. To simplify the evaluation process, we only test on GSM8K, MATH500 and AIME24 datasets in this section.

5.1 SMALLER MODELS STRUGGLE TO LEARN FROM MODEL MERGING

The evaluation results are presented in Table 2. Firstly, we find the task-vector-based methods, such as TA and Ties-Merging, remain effective in reducing the response length while maintaining reasoning performance. The activation-based method, i.e. Sens-Merging, consistently achieves the best performance across the board. But unfortunately, the overall performance of model merging on 1.5B models lags behind that observed on 7B-scale models, particularly on more complex tasks. For instance, on the AIME24 dataset, consistent performance drops are observed across all merging methods. Besides, we unexpectedly find that the number of reflective responses[2] is even negatively correlated with the final performance. This suggests that the reflection actions learned by the merged models are "false reflections", which ultimately lead to incorrect final answers. Consequently, we argue that smaller 1.5B models face significant challenges in acquiring long CoT reasoning abilities

---

[2]The definition of reflective responses is provided in Section 6.

Bench	Accuracy			Length & Reflections			
Method	GSM8K	MATH500	AIME24	GSM8K	MATH500	AIME24	Avg.
Qwen2.5-14B	90.8	47.2	3.3	138.5	522.9	1493.8	-77.8%
				[2;1.0]	[0;0.0]	[0;0.0]	
DeepSeek-R1-14B	91.9	89.0	50.0	558.8	2314.1	7724.8	0
				[24;2.8]	[473;9.2]	[30;32.4]	
Average Merging	91.5	88.2	46.7	642.8	2925.1	7922.7	+14.7%
				[372;3.9]	[457;12.8]	[30;37.2]	
Task Arithmetic	94.1	85.4	33.3	565.9	2191.4	6873.6	-5.0%
				[861;3.0]	[472;6.9]	[29;17.3]	
Ties-Merging	93.4	84.6	26.7	503.1	2169.7	6953.7	-8.7%
				[991;2.4]	[455;6.9]	[29;23.6]	
AIM-Ties	93.0	86.4	50.0	534.4	2260.0	7129.0	-4.8%
				[568;2.8]	[462;8.1]	[30;30.5]	
Sens-Merging	94.2	89.4	53.3	654.4	2601.5	7435.1	+8.6%
				[714;3.1]	[483;8.8]	[29;29.9]	

Table 3: Evaluations of various model merging methods on Qwen-14B models. The number A in [A;B] indicates the number of reflective responses on the dataset and number B indicates the average frequency of reflection keywords appearing in each response.

from stronger reasoner models through model merging. This observation aligns with the findings reported in Li et al. (2025a).

> **Takeaway 4**
>
> *Model merging methods applied to 1.5B-scale models, such as TA, Ties-Merging and Sens-Merging, remain effective on simple tasks. Smaller models struggle to learn long CoT reasoning ability through model merging.*

5.2 LENGTH REDUCTION IS CHALLENGING ON LARGE-SCALE MODELS

We evaluate the performance of several advanced model merging methods on 14B and 32B models. Table 3 reports the results for Qwen 14B models. While reasoning performance is largely preserved after model merging, the length reduction ratios are less significant compared to those observed in smaller-scale models. Notably, methods such as average merging and Sens-Merging even lead to an increase in response length, although Sens-Merging consistently enhances reasoning accuracy across all benchmarks. Similar trends are observed for 32B models, as shown in Table 4. In the experiments with 14B and 32B models, we attempted to fine-tune the hyper-parameters of the merging methods to achieve a significant reduction in response length. For instance, a length reduction ratio of 58.6% was achieved by setting the coefficient of TA merging to 0.3. However, this configuration resulted in a substantial performance drop (GSM8K: +1.2; MATH500: -7.2; AIME24: -30.0). Conversely, a notable observation is that the performance on GSM8K remaines consistently comparable or even marginally higher under most settings. We attribute this to the slight performance variation of the base models on GSM8K, which differs by no more than 3.5 points. Since both Qwen2.5-14B and Qwen2.5-32B are general-purpose models, unlike domain-specific models such as Qwen2.5-Math-1.5B/7B that are extensively trained on mathematical tasks, there exists a significant performance disparity between these general-purpose models and the R1-distilled models on complex mathematical tasks like MATH500 and AIME, with gaps of approximately 40%. This substantial gap makes it challenging for the merged models to surpass the R1-distilled models, as they struggle to provide additional informative knowledge to the already superior models. This hypothesis is inline with the finding in model ensembling (Yao et al., 2024).

Bench Method	Accuracy			Length & Reflections			
	GSM8K	MATH500	AIME24	GSM8K	MATH500	AIME24	Avg.
Qwen2.5-32B	92.3	55.4	6.7	130.3	497.2	1127.9	-83.2%
				[3;0.0]	[0;0.0]	[0;0.0]	
DeepSeek-R1-32B	95.7	89.2	60.0	822.6	2621.8	7194.2	0
				[1287;3.3]	[497;10.2]	[30;34.3]	
Average Merging	93.6	89.2	66.7	662.2	2656.4	6966.8	-7.1%
				[570;3.8]	[489;10.5]	[30;27.1]	
Task Arithmetic	95.5	87.6	50.0	580.1	2168.7	6452.7	-19.0%
				[991;2.4]	[455;6.9]	[29;23.6]	
Ties-Merging	95.4	87.8	43.3	516.9	1898.7	5832.3	-27.9%
				[763;2.0]	[391;4.2]	[24;11.3]	

Table 4: Evaluations of various model merging methods on Qwen-32B models.

Bench Method	Accuracy			Length		
	GSM8K	MATH500	AIME24	GSM8K	MATH500	AIME24
Qwen2.5-32B	92.3	55.4	6.7	130.3	497.2	1127.9
DeepSeek-R1-32B	95.7	89.2	60.0	822.6	2621.8	7194.2
QwQ-32B	96.1	90.8	60.0	1419.8	3754.9	8369.9
Qwen + R1	93.6	89.2	66.7	662.2	2656.4	6966.8
Qwen + QwQ	96.0	90.4	56.7	1424.1	3754.3	8512.4
QwQ + R1	47.7	55.4	50.0	495.4	2288.3	7207.6
Qwen + R1 + QwQ	94.9	89.6	40.0	878.8	2558.5	7767.7

Table 5: Evaluations of average merging on Qwen-32B models.

Additionally, we conducted some preliminary experiments to merge QwQ-32B (Qwen, 2025), another variant of a System 2 reasoning model derived from Qwen2.5-32B. Using average merging, we observed, as shown in Table 5, that the merged model based on QwQ-32B tends to produce excessively lengthy responses, regardless of whether the other merging model is Qwen-32B or R1-32B. The worst performance occurs when directly merging two System 2 reasoning models, i.e., QwQ-32B and R1-32B, which highlights the significant parameter disparity between these two models. This disparity likely stems from their differing training strategies, as R1-32B is obtained through direct fine-tuning, whereas QwQ-32B is trained via reinforcement learning. However, when an intermediate base model is introduced into the merging process, the resulting merged model (Qwen+R1+QwQ) performs comparably, underscoring the critical role of a pre-trained model in facilitating effective model merging.

> **Takeaway 5**
>
> *The merging of large-scale models poses significant challenges in simultaneously maintaining reasoning performance while substantially reducing response length. The substantial performance gaps between the merging models likely contribute to this difficulty.*

6 FURTHER ANALYSIS

In this section, we further investigate some interesting questions, such as whether the merged model retains the ability for self-critic and self-correction, how the response length correlates with the difficulty of the questions, the special characteristics of model merging on long-to-short reasoning task, discussion about the relations between long-to-short and short-to-long, and some of our failure experience.

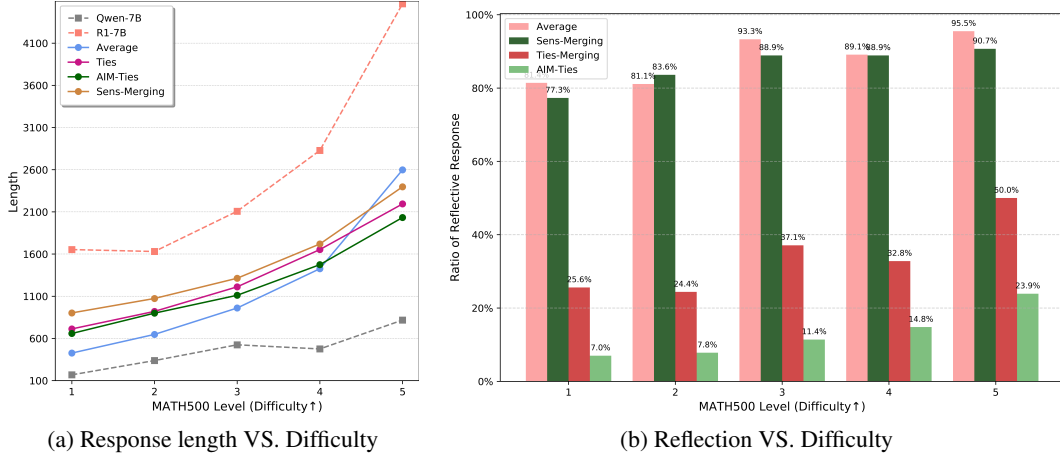| (a) Response length VS. Difficulty | (b) Reflection VS. Difficulty |

Figure 2: Changes in response length and the ratios of reflective responses corresponding to different difficulty levels on the Math500 dataset.

How the response length correlates to the difficulty of the question? From the perspective of datasets, as shown in Table 1, longer responses are generated for more challenging datasets. This trend is consistent across the System 1 quick-thinking model, the System 2 reasoning model, and the merged models. Specifically, when examining the MATH500 dataset with varying difficulty levels, the same conclusion can be drawn: response length positively correlates with the difficulty level of the questions, as shown in Figure 2(a).

Whether the merged model retains the ability for self-critique and self-correction? To address this question, we calculate the ratios of reflective responses. A response is defined as reflective if it contains any of the pre-defined keywords[3]. Our observations indicate that approximately 99.3% of the responses generated by the DeepSeek-R1-7B model are reflective, whereas the corresponding ratio for the Qwen2.5-Math-7B model is nearly 0. This result highlights both the accuracy and coverage of the selected keywords.

Method \ Bench	GSM8K (1319)	MATH500 (500)	Minerva Math (272)	Olympiad Bench (675)	College Math (2818)	AIME24 (30)	Avg.
DeepSeek-R1-7B	98.3	99.4	98.2	99.4	99.5	100	99.1
DPO	98.3	99.4	99.6	99.4	99.2	100	99.3
Average Merging	27.3	89.6	97.1	97.8	94.7	100	84.4
Task Arithmetic	56.7	83.6	73.9	81.2	81.7	76.7	75.6
Ties-Merging	19.2	36.2	16.9	42.4	18.2	50.0	30.5
LoRE-Merging	42.0	92.6	93.8	95.9	94.0	96.7	85.8
AIM-Ties	21.7	40.3	21.7	47.7	31.0	93.3	42.6
Sens-Merging	50.2	83.6	72.4	78.7	78.7	76.7	73.4

Table 6: Ratios (%) of responses containing reflective content across various datasets. Scores for Qwen2.5-Math-7B are not reported, as it produces almost no reflective responses.

As shown in Table 6, all models produced through various model merging methods exhibit the ability for self-critique and self-correction. Several observations can be made: 1) The reflection ratio does not correlate with reasoning accuracy. For instance, both Ties-Merging and AIM-Ties achieve excellent performance, yet their reflection ratios on simpler tasks, such as GSM8K (19.2%; 21.7%) and MATH500 (36.2%; 40.3%), remain relatively low. Conversely, Average Merging and

---

[3]The keywords include: wait, re-examine, recap, double-check, let me (just) check, and let me (just) verify.

LoRE-Merging achieve over 80% reflective responses while they exhibit inferior performance. 2) The reflection ratio is positively associated with response length, which aligns with an intuitive expectation. 3) As shown in Figure 2(b), with the increasing of difficulty level, more reflective responses are generated.

Based on these observations, we hypothesize how the merged models maintain reasoning accuracy while reducing response length. For merged models with a high proportion of reflective responses, such as LoRE-Merging and Sens-Merging, the reduction in response length is primarily achieved by eliminating redundant reasoning content and streamlining the reasoning process, while still performing reflections in most cases. This assumption is supported by the statistics on the MATH500 dataset, which indicate that the frequency of reflection keywords appearing in each response of Sens-Merging and LoRE-Merging is 2.5, compared to 11.1 per response of the DeepSeek-R1-7B model. In contrast, merged models like Ties-Merging and AIM-Ties inherently possess the ability to quickly respond to simpler questions and engage in more deliberate reasoning for complex ones, as evidenced by the distribution of reflective responses across various datasets.

Special Characteristics of Model Merging on Long-to-Short Reasoning   In traditional model merging, the merging objectives can be categorized into two main branches: merging models from different domains into a unified model to integrate multi-domain capabilities, and merging checkpoints from the training process to enhance the final performance on a single task. Unlike multi-domain model merging, which can typically decompose domain-specific knowledge from the task model, identifying feature vectors that represent quick-thinking or slow-thinking capacities poses a significant challenge. Moreover, the models being merged are often general-purpose models and are evaluated on general reasoning benchmarks rather than domain-specific datasets. In contrast to checkpoint merging, where the candidate models exhibit minimal parameter shifts, long-to-short merging involves models with substantial parameter differences, as well as variations in response style and downstream task performance, making the merging considerably more challenging.

From Long-to-Short to Short-to-Long   Unlike the training-based methods (Ma et al., 2025; Xia et al., 2025) which need to change the base model for training, model merging offers a more straightforward approach to achieving short-to-long adjustments on quick-thinking models by simply tuning the merging weights or selecting an appropriate base merging model, without the need for further training. For instance, in a simple attempt to achieve short-to-long reasoning, we assigned a small weight (e.g., 0.2) to the slow-thinking model while using the quick-thinking model as the base. This approach resulted in an approximate improvement of over 20 points in the overall score, accompanied by a 25%+ increase in response length. Model merging provides a more effective and efficient solution for enabling System 1 models to acquire System 2 reasoning abilities compared to model distillation (Yu et al., 2024b).

Failure Experience and Future Directions   Apart from the successful attempts, we also encountered numerous failure cases in this study. Here, we share some key insights from these critical failures and provide potential future directions in this field.

Sensitivity of merging hyper-parameters. Most merging methods are sensitive to hyper-parameters. For instance, task-vector-based merging methods involve a coefficient hyperparameter $\alpha \in [0, 1]$, which determines the contribution proportions of different merging models to the final model. To preserve overall reasoning accuracy, α is typically set to $\alpha > 0.5$, indicating that the System 2 model plays a more significant role. As expected, we observe a positive correlation between both reasoning accuracy and response length with increasing α. However, we regretfully find that even small variations in α (e.g., within a range of 0.1) result in a moderate performance gap (e.g., around 0.5-1.0 points on overall performance), highlighting a sensitivity issue. The sensitivity characteristics imposes additional experimental burdens to identify the optimal merging settings. Therefore, it is highly desirable to develop a hyperparameter-insensitive merging method or a framework capable of automatically determining the optimal parameters.

Sensitivity to calibration data in activation-based merging methods. Activation-based model merging methods demonstrate remarkable performance in maintaining reasoning accuracy while achieving response length reduction. However, we observe that the choice of calibration data significantly influences the merging performance. For example, when sampling calibration data from the s1K dataset, which contains aligned short and long CoT data for each question, different methods exhibit

varying preferences. AIM-Ties performs better when using the question paired with its short answer as activations, whereas Sense-Merging achieves superior results when incorporating both the short and long answers along with the question. Additionally, the sample size of calibration data also impacts merging effectiveness. In our experiments, we tested calibration with only 10 and 20 data samples, observing a performance gap of approximately 1.0 point in the average score. Further exploration of calibration data selection is left for future work.

Challenging of merging models having large performance dispensary. The moderate merging performance observed on 14B and 32B models indicates that not all quick- and slow-thinking model pairs can be effectively merged using existing model merging methods, particularly when evaluated on tasks where the candidate models exhibit significant performance disparities. In this context, exploring strategies to enable the merged model to compare with even surpass the upper bound of the superior model with shorter response lengths, especially when there is a substantial performance gap between the candidate models, presents an intriguing research direction.

7 CONCLUSION

Our study demonstrates model merging effectively addresses Long-to-Short reasoning in LLMs by integrating System 1's efficiency with System 2's rigor, reducing response lengths by up to 55% without compromising performance. Task-vector merging balanced simplicity and effectiveness, while activation-informed methods showed promising improvement potential. These findings position model merging as a cost-efficient solution to mitigate overthinking while preserving reasoning quality. Future work should explore theoretical foundations and refine merging strategies to broaden real-world applicability.

REFERENCES

Pranjal Aggarwal and Sean Welleck. L1: Controlling how long a reasoning model thinks with reinforcement learning, 2025. URL `https://arxiv.org/abs/2503.04697`.

Anton Alexandrov, Veselin Raychev, Mark Mueller, Ce Zhang, Martin T. Vechev, and Kristina Toutanova. Mitigating catastrophic forgetting in language transfer via model merging. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, pp. 17167–17186. Association for Computational Linguistics, 2024. URL `https://aclanthology.org/2024.findings-emnlp.1000`.

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 o1-like llms, 2025. URL `https://arxiv.org/abs/2412.21187`.

DeepSeek-AI. Deepseek-v3 technical report, 2025. URL `https://arxiv.org/abs/2412.19437`.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, and S. S. Li. Deepseek-r1: Incentivizing

reasoning capability in llms via reinforcement learning. *CoRR*, abs/2501.12948, 2025. doi: 10.48550/ARXIV.2501.12948. URL `https://doi.org/10.48550/arXiv.2501.12948`.

Aaron Grattafiori and LLaMA Team. The llama 3 herd of models, 2024. URL `https://arxiv.org/abs/2407.21783`.

Tingxu Han, Zhenting Wang, Chunrong Fang, Shiyu Zhao, Shiqing Ma, and Zhenyu Chen. Token-budget-aware llm reasoning, 2025. URL `https://arxiv.org/abs/2412.18547`.

Gabriel Ilharco, Marco Túlio Ribeiro, Mitchell Wortsman, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL `https://openreview.net/forum?id=6t0Kwf8-jrj`.

Yuetai Li, Xiang Yue, Zhangchen Xu, Fengqing Jiang, Luyao Niu, Bill Yuchen Lin, Bhaskar Ramasubramanian, and Radha Poovendran. Small models struggle to learn from strong reasoners, 2025a. URL `https://arxiv.org/abs/2502.12143`.

Zhong-Zhi Li, Duzhen Zhang, Ming-Liang Zhang, Jiaxin Zhang, Zengyan Liu, Yuxuan Yao, Haotian Xu, Junhao Zheng, Pei-Jie Wang, Xiuyi Chen, Yingying Zhang, Fei Yin, Jiahua Dong, Zhijiang Guo, Le Song, and Cheng-Lin Liu. From system 1 to system 2: A survey of reasoning large language models, 2025b. URL `https://arxiv.org/abs/2502.17419`.

Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Wei-Ming Chen, Wei-Chen Wang, Guangxuan Xiao, Xingyu Dang, Chuang Gan, and Song Han. AWQ: activation-aware weight quantization for on-device LLM compression and acceleration. In Phillip B. Gibbons, Gennady Pekhimenko, and Christopher De Sa (eds.), *Proceedings of the Seventh Annual Conference on Machine Learning and Systems, MLSys 2024, Santa Clara, CA, USA, May 13-16, 2024*. mlsys.org, 2024. URL `https://proceedings.mlsys.org/paper_files/paper/2024/hash/42a452cbafa9dd64e9ba4aa95cc1ef21-Abstract-Conference.html`.

Shuqi Liu, Bowei He, Han Wu, and Linqi Song. Optishear: Towards efficient and adaptive pruning of large language models via evolutionary optimization, 2025a. URL `https://arxiv.org/abs/2502.10735`.

Shuqi Liu, Han Wu, Bowei He, Xiongwei Han, Mingxuan Yuan, and Linqi Song. Sens-merging: Sensitivity-guided parameter balancing for merging large language models, 2025b. URL `https://arxiv.org/abs/2502.12420`.

Shuqi Liu, Han Wu, Bowei He, Zehua Liu, Xiongwei Han, Mingxuan Yuan, and Linqi Song. 1bit-merging: Dynamic quantized merging for large language models, 2025c. URL `https://arxiv.org/abs/2502.10743`.

Zehua Liu, Han Wu, Yuxuan Yao, Ruifeng She, Xiongwei Han, Tao Zhong, and Mingxuan Yuan. Lore-merging: Exploring low-rank estimation for large language model merging, 2025d. URL `https://arxiv.org/abs/2502.10749`.

Zhenyi Lu, Chenghao Fan, Wei Wei, Xiaoye Qu, Dangyang Chen, and Yu Cheng. Twin-merging: Dynamic integration of modular expertise in model merging. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL `http://papers.nips.cc/paper_files/paper/2024/hash/8fcd17eb91bae20d9826786d7d6be799-Abstract-Conference.html`.

Haotian Luo, Li Shen, Haiying He, Yibo Wang, Shiwei Liu, Wei Li, Naiqiang Tan, Xiaochun Cao, and Dacheng Tao. O1-pruner: Length-harmonizing fine-tuning for o1-like reasoning pruning, 2025. URL `https://arxiv.org/abs/2501.12570`.

Xinyin Ma, Guangnian Wan, Runpeng Yu, Gongfan Fang, and Xinchao Wang. Cot-valve: Length-compressible chain-of-thought tuning, 2025. URL `https://arxiv.org/abs/2502.09601`.

Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel J. Candès, and Tatsunori Hashimoto. s1: Simple test-time scaling. *CoRR*, abs/2501.19393, 2025. doi: 10.48550/ARXIV.2501.19393. URL `https://doi.org/10.48550/arXiv.2501.19393`.

Tergel Munkhbat, Namgyu Ho, Seo Hyun Kim, Yongjin Yang, Yujin Kim, and Se-Young Yun. Self-training elicits concise reasoning in large language models, 2025. URL `https://arxiv.org/abs/2502.20122`.

Amin Heyrani Nobari, Kaveh Alimohammadi, Ali ArjomandBigdeli, Akash Srivastava, Faez Ahmed, and Navid Azizan. Activation-informed merging of large language models, 2025. URL `https://arxiv.org/abs/2502.02421`.

OpenAI. Gpt-4o system card. `https://openai.com/index/gpt-4o-system-card/`, 2024a. Accessed: August 8, 2024.

OpenAI. Introducing openai o1. `https://openai.com/o1/`, 2024b. Accessed: December 5, 2024.

OpenAI. Openai o3-mini: Pushing the frontier of cost-effective reasoning. `https://openai.com/index/openai-o3-mini/`, 2025. Accessed: January 31, 2025.

Qwen. Qwq-32b: Embracing the power of reinforcement learning. `https://qwenlm.github.io/blog/qwq-32b/`, 2025. Accessed:March 6, 2025.

Yi Shen, Jian Zhang, Jieyun Huang, Shuming Shi, Wenjing Zhang, Jiangze Yan, Ning Wang, Kai Wang, and Shiguo Lian. Dast: Difficulty-adaptive slow-thinking for large reasoning models, 2025. URL `https://arxiv.org/abs/2503.04472`.

Sainbayar Sukhbaatar, Olga Golovneva, Vasu Sharma, Hu Xu, Xi Victoria Lin, Baptiste Rozière, Jacob Kahn, Daniel Li, Wen tau Yih, Jason Weston, and Xian Li. Branch-train-mix: Mixing expert llms into a mixture-of-experts llm, 2024. URL `https://arxiv.org/abs/2403.07816`.

Mingjie Sun, Zhuang Liu, Anna Bair, and J. Zico Kolter. A simple and effective pruning approach for large language models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL `https://openreview.net/forum?id=PxoFut3dWW`.

Kimi Team, Angang Du, Bofei Gao, Bowei Xing, Changjiu Jiang, Cheng Chen, Cheng Li, Chenjun Xiao, Chenzhuang Du, Chonghua Liao, Chuning Tang, Congcong Wang, Dehao Zhang, Enming Yuan, Enzhe Lu, Fengxiang Tang, Flood Sung, Guangda Wei, Guokun Lai, Haiqing Guo, Han Zhu, Hao Ding, Hao Hu, Hao Yang, Hao Zhang, Haotian Yao, Haotian Zhao, Haoyu Lu, Haoze Li, Haozhen Yu, Hongcheng Gao, Huabin Zheng, Huan Yuan, Jia Chen, Jianhang Guo, Jianlin Su, Jianzhou Wang, Jie Zhao, Jin Zhang, Jingyuan Liu, Junjie Yan, Junyan Wu, Lidong Shi, Ling Ye, Longhui Yu, Mengnan Dong, Neo Zhang, Ningchen Ma, Qiwei Pan, Qucheng Gong, Shaowei Liu, Shengling Ma, Shupeng Wei, Sihan Cao, Siying Huang, Tao Jiang, Weihao Gao, Weimin Xiong, Weiran He, Weixiao Huang, Wenhao Wu, Wenyang He, Xianghui Wei, Xianqing Jia, Xingzhe Wu, Xinran Xu, Xinxing Zu, Xinyu Zhou, Xuehai Pan, Y. Charles, Yang Li, Yangyang Hu, Yangyang Liu, Yanru Chen, Yejie Wang, Yibo Liu, Yidao Qin, Yifeng Liu, Ying Yang, Yiping Bao, Yulun Du, Yuxin Wu, Yuzhi Wang, Zaida Zhou, Zhaoji Wang, Zhaowei Li, Zhen Zhu, Zheng Zhang, Zhexu Wang, Zhilin Yang, Zhiqi Huang, Zihao Huang, Ziyao Xu, and Zonghan Yang. Kimi k1.5: Scaling reinforcement learning with llms, 2025. URL `https://arxiv.org/abs/2501.12599`.

Mitchell Wortsman, Gabriel Ilharco, Samir Yitzhak 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 Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato (eds.), *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pp. 23965–23998. PMLR, 2022. URL `https://proceedings.mlr.press/v162/wortsman22a.html`.

Heming Xia, Yongqi Li, Chak Tou Leong, Wenjie Wang, and Wenjie Li. Tokenskip: Controllable chain-of-thought compression in llms, 2025. URL `https://arxiv.org/abs/2502.12067`.

Prateek Yadav, Derek Tam, Leshem Choshen, Colin A. Raffel, and Mohit Bansal. Ties-merging: Resolving interference when merging models. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL `http://papers.nips.cc/paper_files/paper/2023/hash/1644c9af28ab7916874f6fd6228a9bcf-Abstract-Conference.html`.

Enneng Yang, Li Shen, Guibing Guo, Xingwei Wang, Xiaochun Cao, Jie Zhang, and Dacheng Tao. Model merging in llms, mllms, and beyond: Methods, theories, applications and opportunities. *CoRR*, abs/2408.07666, 2024. doi: 10.48550/ARXIV.2408.07666. URL `https://doi.org/10.48550/arXiv.2408.07666`.

Wenkai Yang, Shuming Ma, Yankai Lin, and Furu Wei. Towards thinking-optimal scaling of test-time compute for llm reasoning, 2025. URL `https://arxiv.org/abs/2502.18080`.

Yuxuan Yao, Han Wu, Mingyang Liu, Sichun Luo, Xiongwei Han, Jie Liu, Zhijiang Guo, and Linqi Song. Determine-then-ensemble: Necessity of top-k union for large language model ensembling. *CoRR*, abs/2410.03777, 2024. doi: 10.48550/ARXIV.2410.03777. URL `https://doi.org/10.48550/arXiv.2410.03777`.

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 *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024a. URL `https://openreview.net/forum?id=fq0NaiU8Ex`.

Ping Yu, Jing Xu, Jason Weston, and Ilia Kulikov. Distilling system 2 into system 1, 2024b. URL `https://arxiv.org/abs/2407.06023`.

A APPENDIX

A.1 EXPERIMENTS CONFIGURATION

DPO Training We use the full set of s1K dataset for the DPO training. Due to the limited computational resource, we conducted the LoRA training with lora rank of 16. The maximum length is set to 4096 and the learning rate is 5e-5. We train the dataset for 3 epoch and report the best performance of the saved checkpoints.

Configurations We report our merging configurations on various merging methods as Table 7. All merged models are saved and evaluated in BF16, except for those produced by Sens-Merging, as the sensitivity factors are calculated on the CPUs, which does not support BF16 computations. To preserve the accuracy of the calculated sensitivity scores, the merged models generated through Sens-Merging are saved in FP32. For evaluations, we use the standard Qwen evaluation toolkit. For System 1 models, we use the 'cot' prompt with few-shots demonstrations and a maximum length of 8192. For System 2 models, we use 'qwen25-math-cot' prompt in a zero-shot setting with a maximum length of 10240. We also conducted experiments on LLaMA3.1-8B models; however, we were unable to reproduce the reasoning accuracy reported in their work using the Qwen evaluation toolkit. As a result, we do not report the results for the LLaMA-series models.

Method	Hyper-parameters			
	1.5B	7B	14B	32B
Task Arithmetic	$\alpha = 0.7$	$\alpha = 0.7$	$\alpha = 0.7$	$\alpha = 0.7$
Ties-Merging	$k = 0.8, \alpha = 1.0$	$k = 0.8, \alpha = 1.0$	$k = 0.2, \alpha = 0.5$	$k = 0.25, \alpha = 0.55$
DARE	$p = 0.3$	$p = 0.3$	$p = 0.4$	-
AIM-Ties	$\omega = 0.4$	$\omega = 0.4$	$\omega = 0.4$	-
Sens-Merging	$\alpha = 0.4, T = 3.0$	$\alpha = 0.7, T = 2.0$	$\alpha = 0.8, T = 6.0$	-

Table 7: The configurations of various merging methods. α means the coefficient in TA merging. p means the drop rate in DARE merging. k denotes the trim ratio in Ties-Merging. ω means the balance factor in AIM. T is the temperature in Sens-Merging.