Outlier Matters: Efficient Long-to-Short Reasoning via Outlier-Guided Model Merging

Qiyuan Zhu [1*], Dezhi Li[1*], Lujun Li[1*], Xiaoyu Qin[2], Wei Li[3], Hao Gu[1], Hua Xu[1], Sirui Han[1†], Yike Guo [1†]

[1]The Hong Kong University of Science and Technology
[2]Tsinghua University
[3]University of Birmingham
{qzhuat,lliee,hguam}@connect.ust.hk, {huaxu,siruihan,yikeguo}@ust.hk, {ldz978560355,xiao.y.qin}@gmail.com, wxl885@student.ham.ac.uk

Abstract

Large Reasoning Language Models (LRMs) have recently shown remarkable performance in complex reasoning tasks, but their extensive reasoning chains incur substantial computational overhead. To address this challenge, we propose Outlier-aware Reasoning Conciseness Adaptive Merge (ORCA), a novel plug-and-play model merging framework that leverages outlier activation patterns to fuse base models with reasoning models. Our ORCA introduces three key innovations: (1) adaptive alignment that reduces conflicts between disparate activation patterns during merging, (2) outlier-guided allocation that assigns merging coefficients proportional to each layer's reasoning importance as indicated by outlier concentrations, and (3) dynamic probe-based adjustment that adapts merging coefficients during inference based on input-specific activation characteristics. These strategies allow seamless integration into existing merging pipelines while creating unified models that maintain reasoning accuracy with significantly reduced response verbosity. Comprehensive evaluation across six benchmarks using Qwen and LLaMA models shows ORCA reduces average response length by 55% while improving accuracy by 2.4∼5.7% over existing methods.

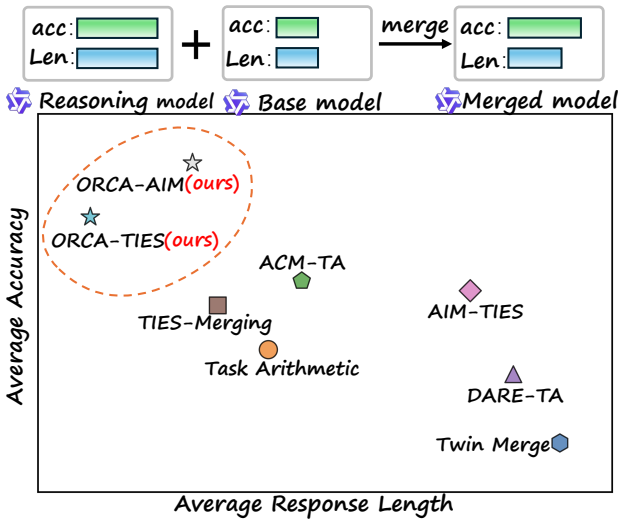


Figure 1: Average accuracy/length results of merged Qwen2.5-7B on 6 reasoning tasks.

1 Introduction

Large Reasoning Language Models (LRMs), such as DeepSeek-R1 (DeepSeek-AI et al. 2025) and OpenAI O3 (OpenAI 2025), have demonstrated remarkable performance in complex reasoning tasks, including advanced mathematics, logical deduction, and multi-step problem-solving. A key driver of their success is the adoption of Chain-of-Thought (CoT) techniques, which break down problems into intermediate reasoning steps (Wei et al. 2022). However, this approach incurs substantial computational overhead: even simple problems (e.g., "2+2=?") can trigger excessively long reasoning chains, generating thousands of tokens. Since inference latency and resource consumption scale linearly with output length (Chen et al. 2025), the growing size of modern LRMs—often reaching billions of parameters—makes their computational and energy demands unsustainable. For real-world deployment, efficiency will be prioritized alongside accuracy to ensure scalability on edge devices or high-throughput cloud systems.

To address these efficiency challenges, researchers have developed three primary categories of solutions: prompt-driven strategies, model-based approaches, and model merging techniques, each exhibiting distinct characteristics and limitations. Prompt-driven methods like CCoT (Renze and Guven 2024) or Token Complexity (Lee, Che, and Peng 2025) employ instructions to reduce output length but struggle with consistency, often degrading accuracy on complex problems (Han et al. 2025). Model-based approaches, such as length-penalized reinforcement learning (Aggarwal and Welleck 2025) or supervised fine-tuning on shortened CoTs (Munkhbat et al. 2025), require extensive retraining and may overfit to specific reasoning patterns. Model merging techniques (Ilharco et al. 2023; Wu et al. 2025) offer a

[*]These authors contributed equally.
[†]Corresponding author

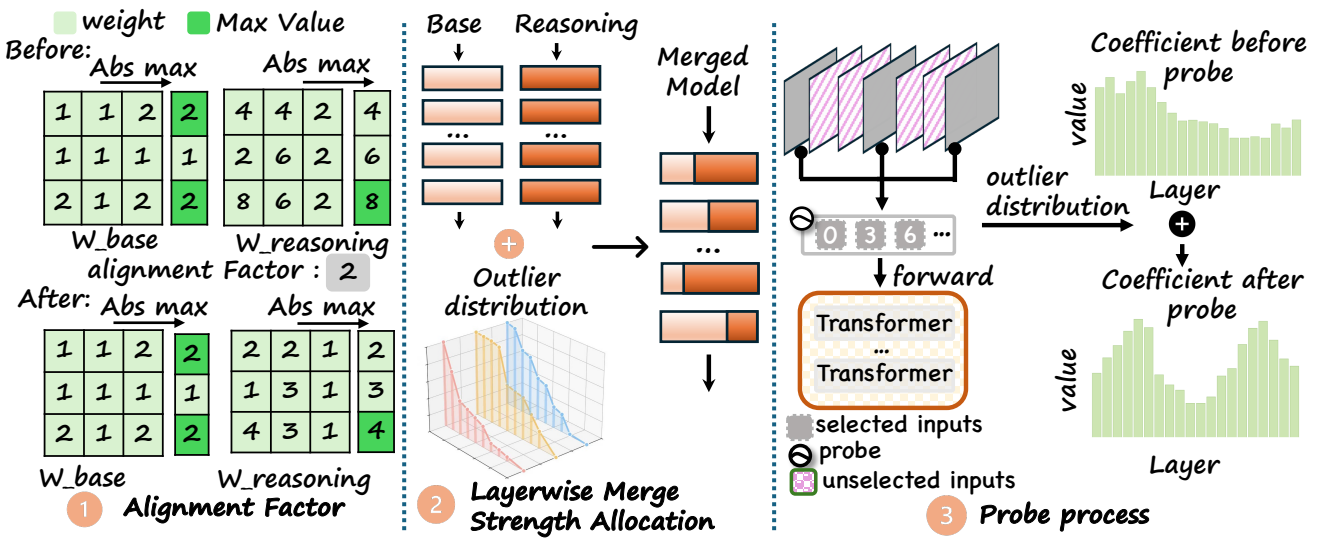


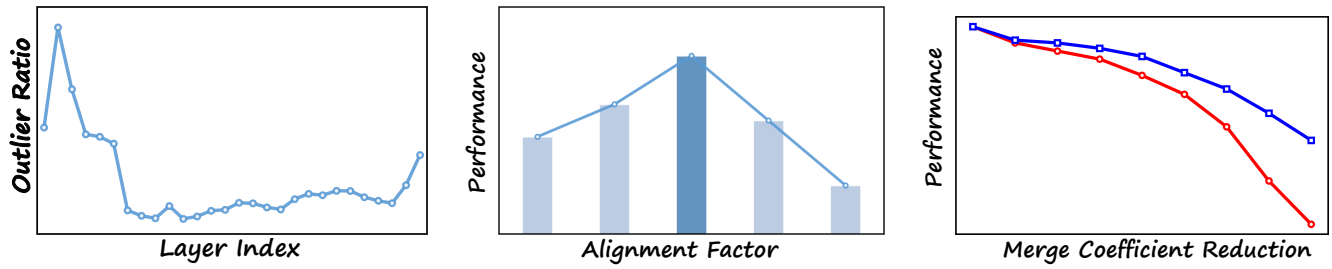
Figure 2: Illustration of the ORCA framework: (1) Alignment Factor adaptively suppresses outlier activations in the reasoning model to reduce merge conflicts. (2) Layer-wise Allocation assigns merge coefficients based on each layer's outlier ratio, emphasizing reasoning-critical layers. (3) Probe Process dynamically adjusts merge coefficients during inference using lightweight probes, enabling input-adaptive reasoning.

promising training-free alternative by interpolating parameters between a base model and its fine-tuned reasoning model, achieving near 50% reductions in inference time and response length. However, these efficiency gains come at substantial accuracy costs, with traditional merging methods suffering around 10% performance drops. This motivates our core research question: ***(RQ) How can we preserve the efficiency benefits of model merging while minimizing performance degradation?***

To answer this question, we began by analyzing the fundamental differences between base and reasoning models, leading to three critical findings that reveal the underlying mechanisms of merging failures: **(1) Outlier Pattern Conflicts:** reasoning models exhibit significantly larger and more extreme outlier activations (>5× mean) compared to base models for identical inputs, creating destructive interference during parameter interpolation; **(2) Layer-wise Sensitivity Variations:** high-outlier layers demonstrate disproportionate importance for reasoning performance, with 40% accuracy degradation when merge coefficients are reduced compared to minimal impact in low-outlier layers; **(3) Input-dependent Outlier Dynamics:** outlier distributions vary greatly across different inputs, indicating that static merging strategies fail to adapt to the dynamic reasoning demands of individual problems. These observations reveal that traditional merging approaches treat all parameters uniformly, ignoring the heterogeneous activation patterns that are crucial for preserving specialized reasoning capabilities while achieving conciseness.

Based on these insights, we propose the Outlier-aware Reasoning Conciseness Adaptive Merge (ORCA) framework, which uses activation information to enhance existing merging pipelines through three novel components (see Figure 2). **First, our Adaptive Alignment Factor** addresses outlier pattern conflicts by dynamically scaling reasoning model parameters before merging. Since reasoning models generate significantly larger outlier activations than base models for identical inputs, we automatically adjust scaling strength based on magnitude relationships between corresponding layers, reducing destructive interference while preserving essential reasoning information encoded in outliers. **Second, our Outlier-Guided Layer-wise Allocation** exploits layer-wise sensitivity variations by differentially weighting layers based on their outlier concentrations. Recognizing that high-outlier layers are disproportionately critical for reasoning performance, we prioritize outlier-rich layers to preserve reasoning capabilities while allowing less critical layers to adopt the base model's concise characteristics, creating a balanced integration strategy. **Third, our Dynamic Merge Coefficient Adjustment** tackles input-dependent outlier dynamics by adapting merge coefficients during inference. Since outlier distributions vary greatly across different problems, lightweight probes capture input-specific activation patterns and dynamically adjust parameters accordingly, ensuring optimal performance for each individual reasoning task. We evaluate ORCA across six benchmarks using the Qwen and LLaMA models. On Qwen2.5-7B models, ORCA reduces average response length by 54∼58% , achieving state-of-the-art results. On LLaMA-3.1-8B models, ORCA demonstrates consistent improvements with up to 2% accuracy gains and 5∼10% shorter responses, confirming its effectiveness across different architectures.

Our contributions can be summarized as follows: (1) We identify the role of outlier activations in reasoning model merging, demonstrating their U-shaped layer-wise distribu-



(a) Outlier Distribution (b) Outlier Alignment Effect (c) High-Outlier vs. Low-Outlier Layers

Figure 3: Analysis of outlier patterns and their impact on model merging in Qwen2.5-7B. (a) Layer-wise outlier percentage distribution showing a U-shaped pattern with higher outlier ratios at beginning and end layers. (b) Impact of alignment factor(1.0-1.08) on model merging performance, showing that strategically aligning outlier patterns between the base model and reasoning model can improve merging effectiveness within an optimal range. (c) Performance comparison between high-outlier layers (red) and low-outlier layers (blue) when reducing merge coefficients, demonstrating the critical importance of high-outlier layers for maintaining reasoning performance.

tion and dynamic input dependence. (2) We propose ORCA, a plug-and-play framework that can enhance existing merging methods by explicitly optimizing for outlier preservation via adaptive alignment, layer-wise allocation, and dynamic probing, achieving superior length-accuracy trade-offs. (3) We conduct extensive experiments across 7 model variants and 6 benchmarks, showing ORCA outperforms state-of-the-art merging techniques by clear margins.

2 Related Works

Efficient Reasoning: Prompt-driven strategies achieve concise reasoning by adjusting prompts or leveraging routing mechanisms. For example, CCoT (Renze and Guven 2024) prompts models with direct instructions to reduce token usage, while TALE-EP (Han et al. 2025) and Token Complexity (Lee, Che, and Peng 2025) explore the impact of explicitly specifying token budgets in input prompts. RouteLLM (Ong et al. 2025) uses learned routers to allocate queries across LLMs for cost-efficient inference, and SoT (Aytes, Baek, and Hwang 2025) utilizes routing mechanisms inspired by cognitive science to choose optimal reasoning approaches. Model-based strategies include methods that introduce length penalties into reinforcement learning frameworks (Luo et al. 2025; Yeo et al. 2025; Aggarwal and Welleck 2025), designing reward functions that penalize redundant reasoning steps. Alternative approaches like Self-Training (Munkhbat et al. 2025) and C3oT (Kang et al. 2025) construct variable-length CoT datasets containing shorter reasoning paths and employ supervised fine-tuning to encourage concise reasoning behaviors. **Unlike these approaches that either require prompt engineering, specialized training, or architectural modifications, our ORCA framework achieves efficient reasoning through training-free model merging without many additional computational overhead.**

Model Merging (Zhu et al. 2025) has emerged as a compelling alternative to traditional Multi-task Learning, operating directly on pre-trained model parameters to offer training-free capability combination. Early work like model soup (Wortsman et al. 2022) demonstrated potential by averaging fine-tuned model parameters, while Fisher Merging (Matena and Raffel 2022) uses Fisher information matrices for more effective fusion. Following the pioneering concept of task vectors (Ilharco et al. 2023), TIES (Yadav et al. 2023b) preserves significant values while resolving sign conflicts, and DARE (Yu et al. 2024) with DAREx (Deng et al. 2025) randomly zeros certain parameters and rescales remaining ones to eliminate redundancy. Recent methods have recognized weight-only limitations and begun integrating activation information. **Our ORCA distinctly differs from weight-only approaches by addressing activation interference through outlier-aware analysis and ensuring dynamic adaptation to input-specific reasoning demands.** AdaMerging (Yang et al. 2024b) and AIM (Nobari et al. 2025) utilize activation information from calibration data to identify important neurons and adjust merging coefficients accordingly. **In sharp contrast to these methods, our framework explicitly optimizes for outlier preservation through adaptive alignment and layer-wise allocation, achieving superior length-accuracy trade-offs.**

3 Methodology

3.1 Problem Formulation

Given a common base model with parameters θ_0 and its N fine-tuned variants $\{\theta_1, \theta_2, \ldots, \theta_N\}$, we aim to obtain a unified model θ_{merged} capable of performing all tasks effectively. We utilize task vectors (Ilharco et al. 2023) defined as $\delta_i = \theta_i - \theta_0$ for $i \in \{1, \ldots, N\}$ to encapsulate the capabilities of fine-tuned models and employ merge coefficients λ_i to control the integration strength of each model. We formulate the final merged model as:

$$\theta_{\text{merged}} = \theta_0 + \sum_{i=1}^{N} \lambda_i \cdot \delta_i. \quad (1)$$

For long-to-short reasoning tasks, we interpret the base

model θ_0 as providing concise responses, while the fine-tuned model represents a reasoning model that performs deliberate analytical thinking. Our merging goal is to combine the advantages of both models, enabling the merged model to output accurate answers while minimizing reasoning chain length.

3.2 Observations & Motivations: Outlier Matters

Previous works on LLM compression (Lin et al. 2024; Xiao et al. 2023; Yin et al. 2024; Le et al. 2025) have demonstrated that outliers—activation values greater than n times the mean (with $n = 5$ in our work)—significantly impact model performance. Meanwhile, reasoning models exhibit unique activation patterns compared to non-reasoning models for identical inputs (Zhao et al. 2025), suggesting that activation manifestations play a critical role in complex reasoning tasks. Inspired by these backgrounds, we analyze the outlier distribution of the DeepSeek-R1-Distill-Qwen-7B model and observe a distinctive pattern: as illustrated in Figure 3 (a), the outlier distribution is highly non-uniform, with outliers being rare overall and their density across layers forming a distinct 'U' shape.

Empirical Experiment I: Reducing Outlier Pattern Inconsistencies During Merge. When merging models with distinct functional focuses, such as a general base model and a reasoning-specialized model, a key challenge arises from the inconsistency in behavioral patterns between the two models. Specifically, for identical input sequences, the reasoning model tends to generate significantly larger and more extreme outlier activations in certain channels or layers due to its specialized reasoning mechanisms, while the base model produces more moderate and constrained activations for the same inputs. This discrepancy could lead to *outlier pattern conflicts* during model merging.

Setup. We investigate whether aligning the output activation patterns of corresponding layers between two models prior to merging can help mitigate potential outlier conflicts while preserving the valuable information encoded in these outliers (Turner et al. 2024). We apply an alignment factor $s > 1$ to adjust the weight matrices of selected linear layers in the reasoning model as $W' = W/s$. For the same input x, the layer's output becomes $W'x = Wx/s$, making this layer's behavior more moderate and aligned with the base model's activation patterns. We then evaluate the merging performance across different values of s.

Result. As shown in Figure 3 (b), merging the base model with the aligned reasoning model can improve performance within a suitable range of s. This suggests that ***strategically aligning patterns in the reasoning model can enhance merging effectiveness***, thereby providing empirical support for developing outlier-aware merging strategies.

Empirical Experiment II: Effect of Merge Coefficient in High-Outlier vs. Low-Outlier Layers. To further elucidate the role of outliers in model merging, we investigate whether the sensitivity of model performance to the merge coefficient varies across layers with different outlier characteristics. Specifically, we ask: *Do layers with a higher proportion of outliers contribute more critically to the merged model's reasoning performance?*

Setup. We isolate two groups of layers—those with a high outlier ratio and those with a low outlier ratio. For each group, we independently reduce the merge coefficient while keeping all other layers fixed. This design allows us to directly assess the impact of outlier prevalence on merge effectiveness by comparing the performance degradation in each case.

Result. As shown in Figure 3 (c), performance drops much more rapidly when the merge coefficient is reduced in high-outlier layers compared to low-outlier layers. This demonstrates that not all layers contribute equally to preserving complex reasoning ability in the merged model. These results indicate that ***layers with more outliers are more critical for reasoning performance, and that layer-wise outlier statistics provide a reliable basis for guiding merge coefficient allocation***.

3.3 ORCA Framework Components

Based on our observations, we present the ORCA framework with three key components: adaptive alignment factors, outlier-guided layer-wise merge coefficient allocation, and dynamic merge coefficient adjustment via probe signals.

Adaptive Alignment Factor. Empirical Experiment I reveals that behavioral inconsistencies between models indeed pose significant challenges during merging, but using a constant alignment factor is obviously suboptimal as it ignores the unique weight characteristics of different layers and the relationships between corresponding layers in the base and reasoning models. Here, we propose adaptive alignment factors that consider both the varying weight distributions across model components and the magnitude relationships between corresponding layers across the two model. We define the layer-wise adaptive alignment factor s_j as:

$$s_j = \sqrt{\frac{\max(|W_{i,j}|)}{\max(|W_{0,j}|)}}, \tag{2}$$

where j denotes the layer index, and $W_{0,j}$ and $W_{i,j}$ represent the weight matrices of layer j in the base model and reasoning model, respectively. Our adaptive factor automatically adjusts the scaling strength based on the magnitude relationship between corresponding layers in the two models, ensuring better parameter alignment during the merging process. We apply this factor to self-attention and feed-forward layers in the reasoning model, as these components dominate both parameter count and computation in LRMs.

Outlier-Guided Layer-wise Allocation. Our analysis in Empirical Experiment II demonstrates that different layers exhibit varying impacts on merge performance, with outlier-rich layers playing a more crucial role in maintaining the reasoning model's capabilities. We integrate the outlier proportion information into layer-wise adaptive allocation during merging. Given an L-layer large language model with a target merge coefficient λ, we calculate the target layer-wise merge coefficients $[\lambda_1, \lambda_2, \ldots, \lambda_L]$. We represent the outlier proportion for each layer as $\mathbf{D} = [D_1, D_2, \ldots, D_L]$. Based on our observation that layers with higher outlier concentrations should be more prominently integrated into the merged model, we establish the relationship $\lambda_i \propto D_i$. We

implement this principle through the following algorithmic approach to ensure $\lambda_i \in [\lambda - \alpha, \lambda + \alpha]$:

$$\tilde{D}i = \frac{D_i - D\text{min}}{D_{\max} - D_{\min}} \cdot 2\alpha, \tag{3}$$

$$\lambda_i = \tilde{D}_i - \bar{\tilde{D}} + \lambda, \tag{4}$$

Under our design paradigm, we prioritize layers critical to the reasoning model to enhance the model's reasoning capabilities, while relatively less important layers allow the merged model to learn the base model's concise and precise characteristics. Our approach creates a balanced integration strategy that leverages the complementary strengths of both models while maintaining the essential reasoning patterns we identify through outlier analysis.

Dynamic Merge Coefficient Adjustment. We observe that the distribution of outliers in the reasoning model varies significantly across different inputs. Since decisions based solely on calibration datasets may fail to account for these dynamic outlier patterns during inference, we propose a probe-driven dynamic adjustment strategy for merge coefficients. During inference, when input hidden states $\mathbf{x}_l$ reach layer f_l, we generate a probe p_l via a magnitude-based probe sampling strategy, selecting the most significant tokens based on absolute magnitude. We propagate this probe through the next n layers to collect activation distributions $A^{(k)}_{probe} = f_{l:k}(p_l)$ and compute dynamic outlier proportions D'_k:

$$D'_k = \frac{\sum_{a \in A^{(k)}_{probe}} I(|a| > 5 \cdot \text{mean}(A^{(k)}_{probe}))}{|A^{(k)}_{probe}|}, \tag{5}$$

Then, we apply D'_k to Equation 4 across these n layers to perform data-driven dynamic allocation. This re-allocation is performed while keeping the local total merge coefficients fixed. The corresponding merged parameters are subsequently updated, achieving a more input-aware and locally balanced parameter configuration. During implementation, we apply this probing technique to only a small subset of layers for efficiency.

3.4 Integration with Existing Merge Frameworks

Our ORCA framework can seamlessly integrate into into existing merge frameworks such as AIM (Nobari et al. 2025) and TIES (Yadav et al. 2023a) through a simple preprocessing and runtime adjustment pipeline. We demonstrate integration with the general merging formulation:

$$\theta_{\text{merged}} = \theta_0 + \sum_{i=1}^{N} \mathbf{\Lambda}_i \odot \delta_i. \tag{6}$$

where $\mathbf{\Lambda}_i$ represents our ORCA-enhanced layer-wise merge coefficients that replace scalar coefficients λ_i in Equation 1.

4 Experiments

4.1 Experiment Setup

To validate our approach across different architectures, we perform experiments using two distinct model families for base and reasoning models. Specifically, we select Qwen2.5-Math (7B), Qwen2.5 (32B) (Yang et al. 2024a) and LLaMA-3.1(8B) (Dubey et al. 2024) as base models, responsible for fast mathematical problem-solving, and DeepSeek-R1-Distill-Qwen (7B,32B), DeepSeek-R1-Distill-LLaMA-8B (DeepSeek-AI et al. 2025), and QwQ-32B (Yang et al. 2024a) as reasoning models, designed for deliberate analytical tasks. We compare our proposed approach with multiple advanced model merging methods, including Model Soup (Wortsman et al. 2022), Task Arithmetic (Ilharco et al. 2023), TIES (Yadav et al. 2023a), DARE (Yu et al. 2024), Twin Merge (Lu et al. 2024), AIM (Nobari et al. 2025), and ACM (Yao et al. 2025). These state-of-the-art baselines provide robust references for evaluating our method.

Datasets. We evaluate our proposed method using several widely recognized mathematical reasoning benchmarks, including GSM8K (Cobbe et al. 2021), MATH500 (Lightman et al. 2024), Minerva Math (Lewkowycz et al. 2022), OlympiadBench (He et al. 2024), CollegeMath (Tang et al. 2024), and AIME24. These datasets cover various mathematical domains and difficulty levels, enabling a comprehensive assessment of reasoning abilities.

Implementation. For hyperparameters, the variation bound (α) controls the allocation of merge coefficients across layers to ensure layer-wise coefficients remain within a reasonable range, and is set to $\alpha = 0.1$ for stable optimization. Additionally, we apply the probing technique to only three layers during inference, and select only 25% of tokens for probing based on their activation magnitudes to further reduce computational overhead. Finally, we use a calibration set of 128 randomly selected samples from the **s1K** dataset (Muennighoff et al. 2025) to compute necessary outlier statistics. All experiments are conducted under zero-shot evaluation conditions with consistent protocols.

4.2 Long-to-Short Merging Results

Our main experimental results demonstrate the effectiveness of our outlier-aware model merging approach, which we denote as ORCA, across different model scales and mathematical reasoning benchmarks. We present comprehensive comparisons with existing state-of-the-art merging methods on both Qwen and LLaMA models.

Qwen-series Model Results. Table 1 presents the results of our proposed ORCA method and other state-of-the-art merging approaches in Qwen2.5-7B models. Our method consistently achieves superior accuracy while significantly reducing response lengths. ORCA-AIM achieves the highest average accuracy of 57.4%, outperforming the best baseline ACM-TA by 2.4%. Of particular significance, on the challenging AIME24 dataset, ORCA-AIM demonstrates the highest accuracy of 33.3%, surpassing all baseline methods. Compared with other plug-and-play merging methods, such as DARE-based approaches (DARE-TA), ORCA variants exhibit substantial improvements in reasoning accuracy while achieving considerable reductions in response length. Moreover, our ORCA approach demonstrates substantial mitigation of reasoning verbosity across all variants, with our methods achieving average response lengths ranging from 1599 to 1734 tokens compared to DeepSeek-

Model	Accuracy (%) ↑							Response Length ↓						
	GSM	M500	Min.	Oly.	Coll.	AIME	**Avg.**	GSM	M500	Min.	Oly.	Coll.	AIME	**Avg.**
Qwen2.5-Math-7B	57.4	51.0	11.4	16.7	22.1	16.7	29.2	1012	1293	1363	1930	1439	1385	1404
R1-Qwen-7B	89.5	87.8	36.0	48.3	45.0	43.3	58.3	1050	2901	3169	5781	2456	7547	3817
Average Merging	86.7	80.6	31.6	41.8	41.6	33.3	52.6	646	1941	1911	3014	1850	3641	2167
Task Arithmetic	89.5	84.6	32.7	45.9	45.7	23.3	53.6	557	1485	1340	2398	1353	3481	1769
TIES-Merging	90.2	83.4	33.8	45.5	44.2	30.0	54.5	530	**1189**	**1130**	2546	1935	2994	1721
DARE-TA	91.7	82.8	34.6	47.1	48.8	13.3	53.1	489	1606	1466	2681	2542	3234	2003
Twin Merge	86.1	80.2	31.6	44.0	41.7	26.7	51.7	595	1800	1675	2815	2079	3321	2048
AIM-TIES	91.5	84.6	36.0	48.4	48.7	20.0	54.9	598	1665	1433	2428	1743	3904	1962
ACM-TA	90.9	84.7	37.0	46.4	44.2	26.7	55.0	579	1500	1300	2342	1614	3469	1801
ORCA-TA	88.9	84.8	36.0	48.3	46.0	30.0	55.7	**458**	1356	1299	2299	1254	3225	1649
ORCA-DARE	91.3	**86.0**	34.9	47.0	47.3	23.3	55.0	599	1482	1265	2363	1331	3363	1734
ORCA-TIES	91.1	83.0	36.4	**49.0**	48.4	30.0	56.3	554	1373	1492	2006	1242	**2926**	**1599**
ORCA-AIM	**91.8**	85.2	**37.1**	48.3	**48.8**	**33.3**	**57.4**	540	1467	1331	**1951**	**1167**	3723	1696

Table 1: Results of merging methods on Qwen2.5-7B. Abbreviations: Min. (Minerva Math), Oly. (Olympiad Bench), Coll. (College Math). "Avg." columns represent the mean across all benchmarks. Response Length denotes the mean number of tokens.

R1-Distill-Qwen-7B's 3817 tokens–representing a reduction exceeding 50%–while consistently maintaining superior accuracy performance, thereby validating the effectiveness of our outlier-aware merging strategy. Additionally, we conduct three-model merging experiments using Qwen2.5-32B (mathematical base), DeepSeek-R1-Distill-Qwen-32B (reasoning), and QwQ-32B (additional reasoning) to evaluate our method's effectiveness in multi-model merging scenarios. Table 2 shows results across three mathematical reasoning benchmarks. Our ORCA variants consistently outperform baseline methods, achieving both superior accuracy and improved response efficiency. ORCA-AIM attains the highest accuracy across all benchmarks (95.7% GSM8K, 46.3% Minerva, 53.3% Olympia) while generating more concise responses than baseline methods, demonstrating that our outlier-aware approach effectively balances performance and efficiency.

Model	Accuracy (%) ↑			Response Length ↓		
	GSM	Min.	Oly.	GSM	Min.	Oly.
Qwen2.5-32B	81.9	22.8	28.9	425	1314	1294
R1-Qwen-32B	95.7	43.8	53.8	823	2582	5274
QwQ-32B	96.4	51.7	54.6	1420	4946	6713
TA	93.3	44.5	50.7	1013	2965	4583
AIM	94.8	45.2	52.6	976	2183	3981
ORCA-TA	93.6	46.0	52.0	1005	2577	4271
ORCA-AIM	95.7	46.3	53.3	900	1979	3828

Table 2: Performance on Qwen2.5-32B models.

Model	Accuracy (%) ↑			Response Length ↓		
	GSM	Min.	Oly.	GSM	Min.	Oly.
Llama-3.1-8B	83.1	22.1	16.1	216	573	911
R1-Llama-8B	73.1	13.2	19.7	583	1554	1844
TA	73.5	15.1	21.3	284	784	1438
DARE	70.7	14.0	21.8	425	870	1821
TIES	73.7	15.1	21.2	262	820	1497
AIM	74.2	14.3	21.3	337	850	1597
ORCA-TA	74.6	15.8	22.2	**242**	**721**	1404
ORCA-DARE	72.9	**17.6**	22.5	383	800	1644
ORCA-TIES	**75.0**	14.3	**23.3**	247	804	1393
ORCA-AIM	74.6	16.5	22.5	320	816	**1383**

Table 3: Performance on Llama models

LLaMA-series Model Results. Table 3 demonstrates ORCA's effectiveness on LLaMA architectures, achieving consistent performance gains across all mathematical reasoning benchmarks. ORCA variants outperform their baseline counterparts while maintaining shorter response lengths, with ORCA-TIES and ORCA-AIM leading in different tasks. These results confirm that our outlier-aware approach successfully transfers to different model architectures, effectively balancing reasoning accuracy with response efficiency on LLaMA-based models.

The consistent improvements across different model architectures and multiple benchmarks validate the effectiveness of our outlier-aware merging approach. Our method successfully addresses the core challenge of long-to-short

reasoning by leveraging outlier distribution patterns to guide the merging process, resulting in models that maintain reasoning accuracy while generating significantly more concise responses.

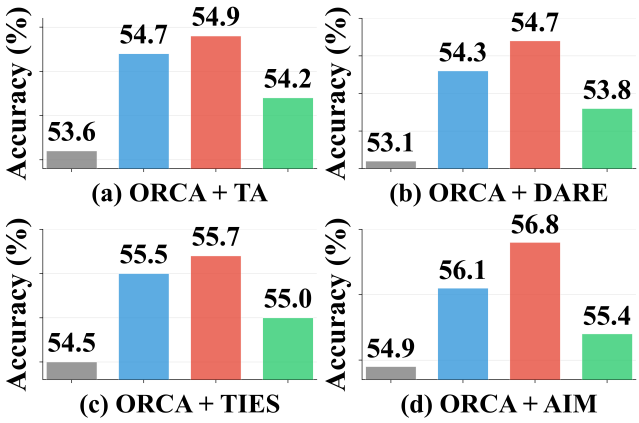


Figure 4: Component ablation in Qwen2.5-7B showing the contribution of each component in our ORCA framework. The four bars in each sub-figure represent Vanilla, Alignment Factor, Adaptive Coeff, and Probe Signals, respectively.

Allocation Strategy	Average Accuracy (%)
Task Arithmetic	53.6
Increasing Allocation	45.6
Decreasing Allocation	47.2
Random Allocation	52.7
ORCA Allocation	**54.9**

Table 4: Ablation study in Qwen2.5-7B on layer-wise merge coefficient allocation strategies.

Probe Selection Method	Average Accuracy (%)
Task Arithmetic	53.6
Minimum Value Selection	53.1
Random Selection	53.9
ORCA Probe	**54.2**

Table 5: Ablation study in Qwen2.5-7B on probe selection strategies for dynamic merge coefficient adjustment.

4.3 Ablation Study

Component Ablation Analysis. As shown in Figure 4, we add each of the Adaptive Alignment Factor, Outlier-Guided Coefficient, and Probe Signals to four merging baselines (TA, DARE, TIES, AIM) on the Qwen 2.5-7B model, inserting only one component at a time. Averaged across baselines, the Adaptive Alignment Factor increases accuracy by roughly 1.2%, the Outlier-Guided Coefficient by about 1.5 %, and Probe Signals by approximately 0.6%. The Outlier-Guided Coefficient provides the largest individual gain, and combining all three components yields the best overall performance.

Stage / Operation	Time
Offline Operations	
AIM	210.01 seconds
Outlier-based Coefficient Analysis	30.49 seconds
Adaptive Alignment Factor Computation	3.22 seconds
Online Inference (per input)	
AIM	399 ms
ORCA w/o probe	368 ms
ORCA	378 ms (+2.7%)

Table 6: Computational overhead on merged Qwen2.5-7B.

Layer-wise Allocation Strategy Analysis. We investigated how different layer-wise merge strategies impact mathematical reasoning. Our proposed outlier-guided ORCA allocation significantly outperforms baseline methods, including uniform, increasing, decreasing, and random allocations. This demonstrates that reasoning capabilities are unevenly distributed across layers. Effective model merging must therefore prioritize critical layers identified by outliers, rather than applying arbitrary or uniform coefficients.

Probe Sampling Strategy Analysis. To validate the effectiveness of our magnitude-based probe sampling strategy, we conduct an ablation study comparing different probe selection methods on the TA baseline. Our approach, ORCA Probe, selects samples based on high activation magnitudes to identify critical outlier patterns affecting the merge. Results show this method achieves the best performance, significantly outperforming random selection and a strategy targeting low-magnitude activations. This validates that high-magnitude activations are reliable indicators for adapting merge coefficients dynamically and effectively.

Computational Efficiency. A key advantage of ORCA is its efficiency. The pre-merge operations are one-time, offline costs, and the dynamic probe is designed to be lightweight. As shown in Table 6, the offline operations complete in approximately 34 seconds, a marginal time compared to the baseline AIM merging time of 210 seconds. During inference, the dynamic probe adds only 2.7% latency overhead, demonstrating that ORCA achieves substantial accuracy improvements without compromising deployment efficiency.

5 Conclusion

In this paper, we demonstrate that outlier activation patterns are critical for effective model merging and introduce ORCA as a comprehensive framework addressing the urgent need for efficient reasoning model deployment. Our approach delivers production-ready solutions that seamlessly bridge the gap between powerful reasoning models and practical deployment constraints while making sophisticated reasoning accessible in resource-limited environments (Li et al. 2023, 2024b,c, 2025c, 2024a, 2025d,a,b,d; Gu et al. 2025).

Acknowledgments

This work is funded in part by the HKUST Start-up Fund (R9911), Theme-based Research Scheme grant (T45-205/21-N), the InnoHK funding for Hong Kong Generative AI Research and Development Center, Hong Kong SAR, and the research funding under HKUST-DXM AI for Finance Joint Laboratory (DXM25EG01).

References

Aggarwal, P.; and Welleck, S. 2025. L1: Controlling How Long A Reasoning Model Thinks With Reinforcement Learning. arXiv:2503.04697.

Aytes, S. A.; Baek, J.; and Hwang, S. J. 2025. Sketch-of-Thought: Efficient LLM Reasoning with Adaptive Cognitive-Inspired Sketching. arXiv:2503.05179.

Chen, X.; Xu, J.; Liang, T.; He, Z.; Pang, J.; Yu, D.; Song, L.; Liu, Q.; Zhou, M.; Zhang, Z.; Wang, R.; Tu, Z.; Mi, H.; and Yu, D. 2025. Do NOT Think That Much for 2+3=? On the Overthinking of o1-Like LLMs. arXiv:2412.21187.

Cobbe, K.; Kosaraju, V.; Bavarian, M.; Chen, M.; Jun, H.; Kaiser, L.; Plappert, M.; Tworek, J.; Hilton, J.; Nakano, R.; Hesse, C.; and Schulman, J. 2021. Training Verifiers to Solve Math Word Problems. .

DeepSeek-AI; Guo, D.; Yang, D.; Zhang, H.; Song, J.; Zhang, R.; Xu, R.; Zhu, Q.; Ma, S.; Wang, P.; et al. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. arXiv:2501.12948.

Deng, W.; Zhao, Y.; Vakilian, V.; Chen, M.; Li, X.; and Thrampoulidis, C. 2025. DARE the Extreme: Revisiting Delta-Parameter Pruning For Fine-Tuned Models. arXiv:2410.09344.

Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Yang, A.; Fan, A.; et al. 2024. The Llama 3 Herd of Models. .

Gu, H.; Li, W.; Li, L.; Qiyuan, Z.; Lee, M.; Sun, S.; Xue, W.; and Guo, Y. 2025. Delta Decompression for MoE-based LLMs Compression. *arXiv preprint arXiv:2502.17298*.

Han, T.; Wang, Z.; Fang, C.; Zhao, S.; Ma, S.; and Chen, Z. 2025. Token-Budget-Aware LLM Reasoning. arXiv:2412.18547.

He, C.; Luo, R.; Bai, Y.; Hu, S.; Thai, Z. L.; Shen, J.; Hu, J.; Han, X.; Huang, Y.; Zhang, Y.; Liu, J.; Qi, L.; Liu, Z.; and Sun, M. 2024. OlympiadBench: A Challenging Benchmark for Promoting AGI with Olympiad-Level Bilingual Multimodal Scientific Problems. In *ACL*, 3828–3850. ACL.

Ilharco, G.; Ribeiro, M. T.; Wortsman, M.; Schmidt, L.; Hajishirzi, H.; and Farhadi, A. 2023. Editing models with task arithmetic. In *The Eleventh ICLR, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Kang, Y.; Sun, X.; Chen, L.; and Zou, W. 2025. C3oT: Generating Shorter Chain-of-Thought Without Compromising Effectiveness. In Walsh, T.; Shah, J.; and Kolter, Z., eds., *AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence, February 25 - March 4, 2025, Philadelphia, PA, USA*, 24312–24320. AAAI Press.

Le, Q.; Diao, E.; Wang, Z.; Wang, X.; Ding, J.; Yang, L.; and Anwar, A. 2025. Probe Pruning: Accelerating LLMs through Dynamic Pruning via Model-Probing. arXiv:2502.15618.

Lee, A.; Che, E.; and Peng, T. 2025. How Well do LLMs Compress Their Own Chain-of-Thought? A Token Complexity Approach. arXiv:2503.01141.

Lewkowycz, A.; Andreassen, A.; Dohan, D.; Dyer, E.; Michalewski, H.; Ramasesh, V.; Slone, A.; Anil, C.; Schlag, I.; Gutman-Solo, T.; Wu, Y.; Neyshabur, B.; Gur-Ari, G.; and Misra, V. 2022. Solving Quantitative Reasoning Problems with Language Models. In *NeurIPS*, volume 35, 3843–3857. Curran Associates, Inc.

Li, L.; Bao, Y.; Dong, P.; Yang, C.; Li, A.; Luo, W.; Liu, Q.; Xue, W.; and Guo, Y. 2024a. DetKDS: Knowledge Distillation Search for Object Detectors. In *ICML*.

Li, L.; Dong, P.; Li, A.; Wei, Z.; and Yang, Y. 2024b. Kd-zero: Evolving knowledge distiller for any teacher-student pairs. *NeuIPS*.

Li, L.; Dong, P.; Wei, Z.; and Yang, Y. 2023. Automated knowledge distillation via monte carlo tree search. In *ICCV*.

Li, L.; Li, D.; Lin, C.; Li, W.; Xue, W.; Han, S.; and Guo, Y. 2025a. AIRA: Activation-Informed Low-Rank Adaptation for Large Models. In *ICCV*.

Li, L.; Lin, C.; Li, D.; Huang, Y.-L.; Li, W.; Wu, T.; Zou, J.; Xue, W.; Han, S.; and Guo, Y. 2025b. Efficient Fine-Tuning of Large Models via Nested Low-Rank Adaptation. In *ICCV*.

Li, L.; Peijie; Tang, Z.; Liu, X.; Wang, Q.; Luo, W.; Xue, W.; Liu, Q.; Chu, X.; and Guo, Y. 2024c. Discovering Sparsity Allocation for Layer-wise Pruning of Large Language Models. In *NeuIPS*.

Li, L.; Qiyuan, Z.; Wang, J.; Li, W.; Gu, H.; Han, S.; and Guo, Y. 2025c. Sub-MoE: Efficient Mixture-of-Expert LLMs Compression via Subspace Expert Merging. *arXiv preprint arXiv:2506.23266*.

Li, W.; Li, L.; Huang, Y.-L.; Lee, M. G.; Sun, S.; Xue, W.; and Guo, Y. 2025d. Structured Mixture-of-Experts LLMs Compression via Singular Value Decomposition. In *ICML*.

Lightman, H.; Kosaraju, V.; Burda, Y.; Edwards, H.; Baker, B.; Lee, T.; Leike, J.; Schulman, J.; Sutskever, I.; and Cobbe, K. 2024. Let's Verify Step by Step. In *ICLR*. OpenReview.net.

Lin, J.; Tang, J.; Tang, H.; Yang, S.; Chen, W.; Wang, W.; Xiao, G.; Dang, X.; Gan, C.; and Han, S. 2024. AWQ: Activation-aware Weight Quantization for On-Device LLM Compression and Acceleration. In *Proceedings of the Seventh Annual Conference on Machine Learning and Systems, MLSys 2024, Santa Clara, CA, USA, May 13-16, 2024*. mlsys.org.

Lu, Z.; Fan, C.; Wei, W.; Qu, X.; Chen, D.; and Cheng, Y. 2024. Twin-Merging: Dynamic Integration of Modular Expertise in Model Merging. arXiv:2406.15479.

Luo, H.; Shen, L.; He, H.; Wang, Y.; Liu, S.; Li, W.; Tan, N.; Cao, X.; and Tao, D. 2025. O1-Pruner: Length-Harmonizing Fine-Tuning for O1-Like Reasoning Pruning. arXiv:2501.12570.

Matena, M.; and Raffel, C. 2022. Merging Models with Fisher-Weighted Averaging. arXiv:2111.09832.

Muennighoff, N.; Yang, Z.; Shi, W.; Li, X. L.; Fei-Fei, L.; Hajishirzi, H.; Zettlemoyer, L.; Liang, P.; Candès, E.; and Hashimoto, T. B. 2025. s1: Simple test-time scaling. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 20286–20332.

Munkhbat, T.; Ho, N.; Kim, S. H.; Yang, Y.; Kim, Y.; and Yun, S.-Y. 2025. Self-Training Elicits Concise Reasoning in Large Language Models. arXiv:2502.20122.

Nobari, A. H.; Alimohammadi, K.; ArjomandBigdeli, A.; Srivastava, A.; Ahmed, F.; and Azizan, N. 2025. Activation-Informed Merging of Large Language Models. arXiv:2502.02421.

Ong, I.; Almahairi, A.; Wu, V.; Chiang, W.-L.; Wu, T.; Gonzalez, J. E.; Kadous, M. W.; and Stoica, I. 2025. RouteLLM: Learning to Route LLMs with Preference Data. arXiv:2406.18665.

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

Renze, M.; and Guven, E. 2024. The Benefits of a Concise Chain of Thought on Problem-Solving in Large Language Models. In *2nd International Conference on Foundation and Large Language Models, FLLM 2024, Dubai, United Arab Emirates, November 26-29, 2024*, 476–483. IEEE.

Tang, Z.; Zhang, X.; Wang, B.; and Wei, F. 2024. MathScale: Scaling Instruction Tuning for Mathematical Reasoning. In *ICML*. OpenReview.net.

Turner, A. M.; Thiergart, L.; Leech, G.; Udell, D.; Vazquez, J. J.; Mini, U.; and MacDiarmid, M. 2024. Steering Language Models With Activation Engineering. arXiv:2308.10248.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Ichter, B.; Xia, F.; Chi, E. H.; Le, Q. V.; and Zhou, D. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In *NeurIPS*, 24824–24837. Curran Associates, Inc.

Wortsman, M.; Ilharco, G.; Gadre, S. Y.; Roelofs, R.; Lopes, R. G.; Morcos, A. S.; Namkoong, H.; Farhadi, A.; Carmon, Y.; Kornblith, S.; and Schmidt, L. 2022. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *ICML*, volume 162, 23965–23998. PMLR.

Wu, H.; Yao, Y.; Liu, S.; Liu, Z.; Fu, X.; Han, X.; Li, X.; Zhen, H.-L.; Zhong, T.; and Yuan, M. 2025. Unlocking Efficient Long-to-Short LLM Reasoning with Model Merging. arXiv:2503.20641.

Xiao, G.; Lin, J.; Seznec, M.; Wu, H.; Demouth, J.; and Han, S. 2023. SmoothQuant: Accurate and Efficient Post-Training Quantization for Large Language Models. In *ICML, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202, 38087–38099. PMLR.

Yadav, P.; Tam, D.; Choshen, L.; Raffel, C.; and Bansal, M. 2023a. TIES-Merging: Resolving Interference When Merging Models. arXiv:2306.01708.

Yadav, P.; Tam, D.; Choshen, L.; Raffel, C. A.; and Bansal, M. 2023b. TIES-Merging: Resolving Interference When Merging Models. In Oh, A.; Naumann, T.; Globerson, A.; Saenko, K.; Hardt, M.; and Levine, S., eds., *NeurIPS*.

Yang, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Li, C.; Liu, D.; Huang, F.; Wei, H.; et al. 2024a. Qwen2.5 Technical Report. .

Yang, E.; Wang, Z.; Shen, L.; Liu, S.; Guo, G.; Wang, X.; and Tao, D. 2024b. AdaMerging: Adaptive Model Merging for Multi-Task Learning. In *The Twelfth ICLR, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Yao, Y.; Liu, S.; Liu, Z.; Li, Q.; Liu, M.; Han, X.; Guo, Z.; Wu, H.; and Song, L. 2025. Activation-Guided Consensus Merging for Large Language Models. arXiv:2505.14009.

Yeo, E.; Tong, Y.; Niu, M.; Neubig, G.; and Yue, X. 2025. Demystifying Long Chain-of-Thought Reasoning in LLMs. arXiv:2502.03373.

Yin, L.; Wu, Y.; Zhang, Z.; Hsieh, C.; Wang, Y.; Jia, Y.; Li, G.; Jaiswal, A. K.; Pechenizkiy, M.; Liang, Y.; Bendersky, M.; Wang, Z.; and Liu, S. 2024. Outlier Weighed Layerwise Sparsity (OWL): A Missing Secret Sauce for Pruning LLMs to High Sparsity. In *Forty-first ICML, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

Yu, L.; Yu, B.; Yu, H.; Huang, F.; and Li, Y. 2024. Language Models are Super Mario: Absorbing Abilities from Homologous Models as a Free Lunch. In *Forty-first ICML, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

Zhao, Z.; Liu, Q.; Zhou, K.; Liu, Z.; Shao, Y.; Hu, Z.; and Huang, B. 2025. Activation Control for Efficiently Eliciting Long Chain-of-thought Ability of Language Models. arXiv:2505.17697.

Zhu, Q.; Li, L.; Li, D.; Liu, J.; Cheng, P.; Xu, Y.; Han, S.; and Guo, Y. 2025. Outlier-Aware Model Merging for Efficient Multitask Inference. In *Proceedings of the 33rd ACM International Conference on Multimedia*.