DALSA: Difficulty-Aware Long-to-Short LLM Reasoning with Model Merging

Dongwook Min[1] **Jeonhwi Kang**[1] **Donghyeon Jeon**[2]
Inho Kang[2] **Seung-Hoon Na**[1*]
[1]Ulsan National Institute of Science and Technology
[2]Naver Corporation
mindw96@unist.ac.kr, kangjeonhwi@unist.ac.kr
donghyeon.jeon@navercorp.com, once.ihkang@navercorp.com
nash@unist.ac.kr

Abstract

Recent advances in large language models (LLMs) have given rise to *large reasoning models* that achieve strong performance through explicit multi-step reasoning. However, models often generate long reasoning chains even for simple inputs. We present *Difficulty-Aware Long-to-Short LLM Reasoning with Model Merging* (**DALSA**), an inference-time framework that adaptively controls reasoning depth based on input difficulty. DALSA integrates (i) representation-based difficulty estimation, which predicts a soft failure-rate score from the prompt hidden state, and (ii) difficulty-aware model merging, which uses a quantile-remapped score to modulate the interpolation coefficient between short- and long-reasoning models. DALSA trains a lightweight probe while keeping both LLMs frozen. On mathematical benchmarks with the evaluated Qwen pairs, the accuracy–length trade-off depends on scale: at 7B, DALSA preserves static ACM-TA average accuracy (66.60% vs. 66.61%) while reducing average response length by 22.35%; at 1.5B, it improves average accuracy from 56.75% to 58.24% with a 2.27% length increase. Additional merging baselines, routing-threshold and merge-range sweeps, and component ablations show sensitivity to the operating configuration. Domain-matched commonsense and Llama-family diagnostics do not establish broader efficiency gains, limiting the primary empirical claim to the evaluated Qwen mathematical setting.

1 Introduction

Recent advances in large language models (LLMs), particularly through techniques such as Chain-of-Thought (CoT) prompting (Wei et al., 2022), have substantially enhanced their reasoning capabilities (Yao et al., 2023). In domains such as complex mathematical problem solving and logical inference, these models now perform explicit multi-step reasoning, yielding marked improvements in accuracy over prior approaches. These developments have given rise to so-called large reasoning models, which enable LLMs to solve classes of problems that were previously beyond the reach of conventional architectures, significantly expanding the scope of AI problem-solving.

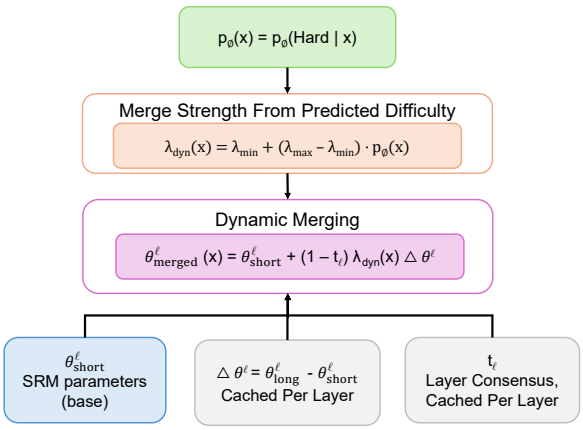


Figure 1: **DALSA Highlight.** DALSA converts the estimated difficulty score into an instance-specific merge strength. The final merged parameters combine the short-reasoning base, cached long-reasoning delta, and cached layer-consensus factor. In the schematic, the difficulty label $p_\phi(\text{Hard} \mid x)$ denotes the remapped soft-failure score $\tilde{p}_\phi(x)$ used here, rather than a calibrated hard-label probability.

However, these gains in reasoning performance come at a substantial cost: inefficiency. Existing reasoning models typically generate long inference chains regardless of problem difficulty. As a result, even trivial instances incur excessive reasoning,

*Corresponding author.

Proceedings of the 15th International Joint Conference on Natural Language Processing and the 5th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 6023–6040
November 6-10, 2026 ©2026 Association for Computational Linguistics

leading to unnecessary computational overhead and increased latency. Addressing this inefficiency—while preserving the benefits of extended reasoning for genuinely hard problems—has therefore become a central challenge for deploying large reasoning models in practice.

To mitigate this inefficiency, *Long-to-Short* (L2S) reasoning has become an active area of research (Luo et al., 2025a; Wu et al., 2025a). As summarized in recent surveys (Yue et al., 2025; Feng et al., 2025; Qu et al., 2025; Sui et al., 2025), the majority of existing approaches adopt *single-model* solutions that regulate reasoning length through early stopping, reasoning compression, prompt engineering, or reinforcement learning–based adaptive control. While effective to some extent, these methods operate within a single-model paradigm and often incur additional training or inference overhead. More recent work has explored *multi-model* strategies that combine complementary reasoning capabilities. Such approaches are commonly realized through routing mechanisms, speculative decoding, or parameter-space model merging (Wu et al., 2025a; Hu et al., 2025; Jin et al., 2025). We argue that *model merging* is particularly well suited for the L2S setting: fast and slow reasoning can be naturally interpreted as distinct but complementary abilities learned under different objectives, and merging provides a direct means of integrating these abilities without retraining. Despite this promise, most existing merging-based L2S methods rely on a *static* merging strategy, applying a fixed coefficient (λ) uniformly to all inputs. In contrast, activation-guided consensus merging has shown that inference behavior can be modulated by exploiting layer-wise interactions through task arithmetic, suggesting a pathway toward more adaptive reasoning control (Yao et al., 2025).

Inspired by activation-guided consensus merging (Yao et al., 2025) and recent advances in difficulty estimation for Long-to-Short reasoning, we introduce *Difficulty-Aware Long-to-Short LLM Reasoning with Model Merging* (**DALSA**)[1], an inference-time framework that dynamically controls reasoning depth through adaptive model merging. DALSA operates in two stages. First, it performs **representation-based difficulty estimation** by analyzing the LLM's initial hidden state to predict a soft failure-rate score for an input prompt. Second, it conducts **difficulty-aware model merging**, where the predicted difficulty directly parameterizes the interpolation coefficient (λ) in task arithmetic merging. This design selectively emphasizes short-reasoning behavior for easy instances while progressively incorporating long-reasoning capabilities as problem difficulty increases. This enables instance-level control over inference length without updating either LLM; only the lightweight difficulty probe is trained. Figure 1 summarizes this central mechanism: the predicted difficulty score is quantile-remapped and converted into an instance-specific merge strength $\lambda_{\text{dyn}}(x)$, which scales the cached long-reasoning delta during dynamic merging.

On the evaluated Qwen mathematical benchmark suite, DALSA at 7B matches the static ACM-TA baseline in average accuracy (66.60% vs. 66.61%) while reducing average response length from 2,589.90 to 2,011.09 tokens, a 22.35% reduction. On GSM8K and MATH500, DALSA improves both accuracy and response length over static ACM-TA, suggesting that instance-adaptive model merging can reduce unnecessary computation while preserving the benefits of long-reasoning behavior when it is useful. At 1.5B, DALSA instead improves average accuracy from 56.75% to 58.24% with a 2.27% increase in response length. We also evaluate a same-probe Simple Routing baseline and extend the comparison with routing-threshold sweeps, merge-range sweeps, additional merging operators, and separate component ablations. These results show configuration-dependent accuracy–length trade-offs. Because routing and merging use different prompt protocols, their comparison does not isolate interpolation alone. Non-mathematical and Llama-family diagnostics do not establish broader efficiency gains; our primary empirical claim is therefore limited to mathematical reasoning on the evaluated Qwen pairs.

2 Related Works

2.1 Difficulty Estimation

Accurate difficulty estimation is crucial for efficient resource allocation. Sampling-based approaches use repeated generations for consistency estimation or compute allocation (Wang et al., 2023; Snell et al., 2025), but incur costs from repeated generation. Evidence-based uncertainty estimation can instead use token logits without multiple sampling (Ma et al., 2025). Auxiliary approaches use dedicated question-difficulty predictors (Venktesh

[1]Code: https://github.com/mindw96/DALSA.

et al., 2022), while performance-based approaches use historical outcomes to estimate difficulty for reinforcement-learning data selection (Kong et al., 2025).

2.2 Long-to-Short Reasoning Frameworks

To reconcile the depth of System 2 reasoning (Ziabari et al., 2025; Wei et al., 2022; Madaan et al., 2023) with the efficiency of System 1, Long-to-Short frameworks compress reasoning processes (Luo et al., 2025a; Zhang et al., 2025). Model merging offers a training-free alternative (Yang et al., 2025). Recent adaptive reasoning approaches dynamically modulate thinking depth based on instance difficulty. These include adaptive reasoning-mode selection (LHRMs) (Jiang et al., 2025), bi-level preference optimization (Ada-R1) (Luo et al., 2025b), and difficulty- or certainty-aware CoT adaptation techniques such as DAST (Shen et al., 2025), CAR (Lu et al., 2025), AdaCtrl (Huang et al., 2025b), Thinkless (Fang et al., 2025), AdaCoT (Lou et al., 2025), and ARM (Wu et al., 2025b). DALSA differs from hard model switching by using the difficulty signal to modulate a parameter-space merge rather than selecting a single model. Furthermore, reinforcement learning-based methods have been introduced for adaptive reasoning and budget control, including ACPO (Cheng et al., 2025b), LASER (Liu et al., 2025), HAPO (Huang et al., 2025a), and SelfBudgeter (Li et al., 2025).

2.3 Activation-Aware Model Merging

Model merging combines base and fine-tuned parameters (Ilharco et al., 2023; Yang et al., 2025). Advanced techniques address interference and redundancy through task arithmetic (Ilharco et al., 2023), task-vector geometry (Cheng et al., 2025a), sign resolution (TIES) (Yadav et al., 2023), random dropping (DARE) (Yu et al., 2024), and magnitude-based sampling (DELLA) (Deep et al., 2024). Activation-guided merging further refines coefficients using activation statistics or geometry (Yao et al., 2025; Nobari et al., 2025). However, these methods typically apply a static merge configuration across inputs. We extend this by proposing dynamic, instance-level merging driven by real-time difficulty estimates, aligning with emerging test-time compute scaling strategies (Snell et al., 2025; Bi et al., 2025; Zhu et al., 2025a; Light et al., 2025).

3 Methodology

Figure 2 summarizes DALSA: offline calibration caches the task-vector delta and layer-consensus factors; before generation, a prompt-state probe predicts difficulty, which controls the cached delta's contribution to the merged model.

3.1 Difficulty Estimation

DALSA estimates input difficulty before generation, using only the final hidden state of the prompt. This follows the central observation of prior representation-based difficulty estimation work (Zhu et al., 2025b): the initial state already contains useful information about whether an input is likely to require extended reasoning. Unlike Temporal Difference (TD) value estimation, however, our estimator does not train over full generation trajectories. This avoids storing token-level hidden states and performing token-level bootstrapping over long reasoning trajectories.

Let $\mathcal{M}_{\text{short}}$ denote the short-reasoning model. Given an input x, we run a forward pass over the prompt and extract the final prompt-token hidden representation $\mathbf{h}_0(x) \in \mathbb{R}^{d_{\text{model}}}$. We then train a lightweight probe f_ϕ to map $\mathbf{h}_0(x)$ to a continuous difficulty score. In our final DALSA controller, this probe is a linear model followed by a sigmoid:

$$p_\phi(x) = \sigma(\mathbf{w}^\top \mathbf{h}_0(x) + b), \tag{1}$$

where $p_\phi(x) \in [0, 1]$ is interpreted as a soft failure-rate estimate rather than a hard routing decision.

Soft-Failure Target For each calibration input x, we run the short model K times ($K = 3$ in the main experiments) and record the number of correct rollouts $c(x)$. Rather than collapsing this signal into only an easy/hard label, we define a soft-failure target:

$$y_{\text{sf}}(x) = 1 - \frac{c(x)}{K}. \tag{2}$$

Thus, examples that are solved consistently receive scores near zero, examples that fail consistently receive scores near one, and partially solved examples retain intermediate information. The final probe is trained with a sigmoid-MSE objective:

$$\mathcal{L}_{\text{sf}} = \mathbb{E}_{x \sim \mathcal{D}_{\text{probe}}} \left(p_\phi(x) - y_{\text{sf}}(x)\right)^2. \tag{3}$$

This differs from our earlier hard-label binary classifier. In preliminary ablations, hard-BCE training produced a usable hard/easy ranking but inflated the absolute hardness scale, which in turn

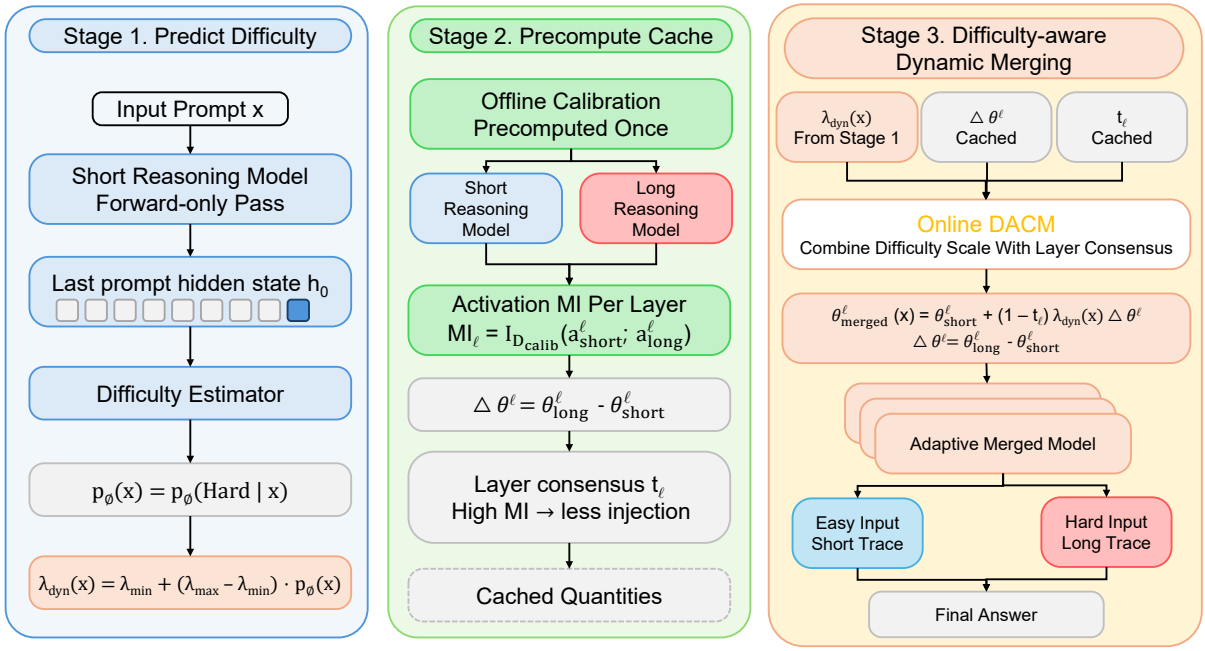


Figure 2: **Overview of DALSA.** DALSA predicts input difficulty from the final prompt hidden state and combines the predicted merge strength with layer-wise activation-consensus factors and task vectors precomputed offline. In the schematic, the difficulty label $p_\phi(\text{Hard} \mid x)$ denotes the remapped soft-failure score $\tilde{p}_\phi(x)$ used here, rather than a calibrated hard-label probability.

pushed many examples to large λ values and increased generated length. Soft-failure regression provides a better calibrated continuous controller for DALSA. Appendix tables report historical held-out and downstream comparisons of hard-label linear and residual-MLP probes; these are distinct from the final soft-failure controller.

Quantile Remapping The raw soft-failure score is not used directly as the merge coefficient because its scale can be compressed when the probe is trained on relatively easy calibration data. We therefore remap $p_\phi(x)$ using score quantiles computed on the GSM8K training-split reference set:

$$\tilde{p}_\phi(x) = \text{clip}\left(\frac{p_\phi(x) - Q_{0.05}}{Q_{0.95} - Q_{0.05}}, 0, 1\right), \quad (4)$$

where $Q_{0.05}$ and $Q_{0.95}$ are the 5th and 95th percentiles of the reference scores. This remapped score is the continuous difficulty signal used by the dynamic merging rule below.

3.2 Difficulty-Aware Activation Consensus Merging

Difficulty-Aware Activation Consensus Merging (DACM) combines predicted difficulty with cached layer-consensus factors (Stages 2 and 3 in Figure 2). Following Activation-Guided Consensus Merging (ACM), we treat the Short Reasoning Model (SRM) as the base model and represent the Long Reasoning Model (LRM) as a task vector relative to this base:

$$\Delta\theta^l = \theta^l_{long} - \theta^l_{short}. \quad (5)$$

All main static-merging and DALSA experiments use this short-base formulation; the appendix reports a separate long-base diagnostic.

3.2.1 Offline Calibration via Activation Consensus

The indiscriminate merging of two models may result in performance degradation due to interference within the parameter space. To address this issue, we emphasize layers exhibiting significant informational discrepancies between the short model (θ_{short}) and the long model (θ_{long}), utilizing Activation-Guided Consensus Merging (ACM). We collect the activation distributions for each layer l of both models using a small calibration dataset, denoted as $\mathcal{D}_{calib}$ (e.g., `simplescaling/s1K` (Muennighoff et al., 2025)). Using the short model's tokenizer for both models, we quantify their similarity by calculating the Mutual Information (MI) between the output features of each layer. For each calibrated module, we cluster its prompt-activation vectors across calibration examples separately for

the short and long models, then compute mutual information between the two cluster assignments. We average these module-level values within each layer to obtain MI_l, the calibration-set statistic summarized in Figure 2. A high mutual information value indicates that the two models are processing similar information at this layer (i.e., exhibiting high consensus), thereby diminishing the necessity for merging. Conversely, a low value indicates distinct roles for the two models, thereby necessitating the incorporation of the long model's inference capabilities. To quantify this, we calculate a layer-wise consensus factor t_l as follows:

$$t_l = \sigma\left(\beta\left(\mathrm{MI}_l - \max_j \mathrm{MI}_j\right)\right), \tag{6}$$

where $\beta > 0$ controls the sigmoid scale and the maximum is taken across calibrated layers. This implementation gives $0 < t_l \leq 0.5$ for calibrated layers; a larger value indicates greater relative consensus. During the merging process, we multiply by the term $(1 - t_l)$, thereby assigning greater weight to layers exhibiting low consensus (i.e., significant discrepancies).

3.2.2 Dynamic Merging

With the offline consensus factors fixed, we map the remapped difficulty score to an input-specific merge strength:

$$\lambda_{\mathrm{dyn}}(x) = \lambda_{\min} + (\lambda_{\max} - \lambda_{\min})\tilde{p}_\phi(x), \tag{7}$$

where the main DALSA setting uses $\lambda_{\min} = 0.1$ and $\lambda_{\max} = 0.7$. The upper bound matches the fixed ACM-TA coefficient, while the lower bound is an engineering default retaining a nonzero long-model delta. These bounds are reference settings, not held-out-selected optima; Appendix B.2 reports their sensitivity. Thus, harder inputs receive a larger contribution from the long-reasoning model, while easier inputs remain closer to the short-reasoning model. For efficient vLLM inference, we instantiate this continuous rule with a small set of reusable merge buckets, assigning each example to the nearest value in $\{0.1, 0.2, \ldots, 0.7\}$.

For the Simple Routing baseline, we reuse the same remapped score $\tilde{p}_\phi(x)$ but replace parameter-space merging with hard model selection: examples below a threshold of 0.5 are sent to the short model, and the remaining examples are sent to the long model. This baseline shares the difficulty signal with DALSA, but its native endpoint prompts differ from the shared merged-model prompt; the comparison therefore evaluates the complete inference protocols.

Ultimately, the parameters θ^l_{merged} for layer l are updated according to the following equation:

$$\mathcal{S}_l(x) = \underbrace{(1 - t_l) \cdot \lambda_{\mathrm{dyn}}(x)}_{\text{Effective Scale}} \tag{8}$$

$$\theta^l_{merged}(x) = \theta^l_{short} + \mathcal{S}_l(x) \cdot \Delta\theta^l \tag{9}$$

At $\lambda = 0$, the formula recovers the short base. At $\lambda = 1$, the layer factor $(1 - t_l)$ remains, so the formula does not generally recover the native long model. The factor $(1 - t_l)$ up-weights lower-consensus layers rather than selecting a subset of layers; $\lambda_{\mathrm{dyn}}(x)$ controls the overall contribution according to predicted difficulty.

4 Experiments

4.1 Experimental Setup

To validate the proposed DALSA framework, we employ six mathematical test sets grouped into four benchmark families characterized by varying levels of difficulty:

- GSM8K (Cobbe et al., 2021): This dataset consists of mathematics problems at the elementary school level that require relatively straightforward reasoning.
- MATH500 (Hendrycks et al., 2021): Comprising 500 problems selected from the larger MATH dataset, MATH500 addresses more complex and diverse mathematical concepts than GSM8K.
- AMC23: We include `math-ai/amc23` as an intermediate contest-style benchmark between MATH500 and AIME-level problems.
- American Invitational Mathematics Examination 2024–2026 (AIME24–26): We use the 30-problem `math-ai/aime24`, `math-ai/aime25`, and `math-ai/aime26` test sets to evaluate highly challenging Olympiad-level reasoning.

We evaluate two matched Qwen/DeepSeek model pairs: Qwen-2.5-Math-1.5B-Instruct and Qwen-2.5-Math-7B-Instruct (Yang et al., 2024) with DeepSeek-R1-Distill-Qwen-1.5B and DeepSeek-R1-Distill-Qwen-7B (Guo et al., 2025). The following systems serve as comparative baselines:

- Short Reasoning Model: Qwen-2.5-Math-1.5B-Instruct or Qwen-2.5-Math-7B-Instruct, depending on the scale. This model is distinguished by its capacity to produce concise responses and demonstrate essential mathematical abilities.

- Long Reasoning Model: DeepSeek-R1-Distill-Qwen-1.5B or DeepSeek-R1-Distill-Qwen-7B, depending on the scale. This model is specialized in addressing complex problems through prolonged Chain-of-Thought (CoT) reasoning.

- Simple Routing: This non-merging baseline uses the same linear soft-failure regression score and GSM8K-H0 quantile remapping as DALSA, then sends examples with remapped difficulty below 0.5 to the short model and the remaining examples to the long model.

- Static Merging (ACM-TA Baseline Model): This model uses the Short Reasoning Model as the base and adds the task vector $\theta_{long} - \theta_{short}$ with a fixed merging ratio of $\lambda = 0.7$, regardless of the dataset or problem complexity, following ACM.

Evaluation Metrics: Model performance was evaluated using accuracy, whereas efficiency was measured by the average number of tokens in the generated responses (Avg. Response Length). These token counts do not measure end-to-end latency or include offline calibration and merge construction. Native endpoints and routing use native prompts; merged models use a shared prompt. Main and additional diagnostic runs are reported separately, with their settings detailed in Appendix A and Appendix B.

DALSA Controller The main DALSA setting uses the linear soft-failure probe described in Section 3 and applies quantile remapping to convert the raw probe score into a merge-strength range of $\lambda \in [0.1, 0.7]$. To make inference practical with vLLM (Kwon et al., 2023), examples are assigned to the nearest reusable merge bucket in $\{0.1, 0.2, \ldots, 0.7\}$ and generated in bucketed batches. This implementation preserves the continuous difficulty-aware formulation while avoiding per-example checkpoint materialization.

5 Results

5.1 Main Results

Tables 1 and 2 display accuracy and average response length across the main mathematical benchmark suite. We aggregate AIME24–26 into a single benchmark-family column and add AMC23 as an intermediate contest-style benchmark between MATH500 and AIME. The DALSA rows at both scales use the current linear soft-failure probe with quantile-remapped bucketed merge strengths, including the finalized 1.5B AMC23 result. The Simple Routing rows were rerun with the same remapped difficulty score as DALSA, but the native routing prompts and shared merging prompt remain different. The main merging results use a short-base formulation, which keeps DALSA aligned with the long-to-short efficiency objective but also inherits the short model's 4096-token context budget. On the completed 7B suite, DALSA improves over the short model on average while substantially reducing response length relative to the native long-reasoning model.

5.1.1 Scale-Dependent Accuracy–Length Trade-offs

In the 7B setting, DALSA attained an average accuracy of 66.60% across the displayed benchmark families, surpassing the Short Model average of 59.45% while using substantially fewer tokens than the native long-reasoning model. This matches the static ACM-TA baseline within 0.01 percentage points while reducing average response length from 2,589.90 to 2,011.09 tokens. In the 1.5B setting, DALSA achieves the highest method-block average accuracy (58.24%), an increase of 1.49 percentage points over ACM-TA, with a 2.27% increase in average length (2,762.98 vs. 2,701.65 tokens). Because Simple Routing uses the same difficulty score but remains below DALSA in average accuracy at both scales (53.10% vs. 58.24% for 1.5B, and 57.22% vs. 66.60% for 7B), the systems occupy different accuracy–length operating points. The different prompt protocols prevent attributing this difference solely to merging.

- Notably, on GSM8K, DALSA achieved 95.53%, improving over the Static (92.95%) and Original Short (88.9%) models.

- Furthermore, on MATH500, DALSA achieved a score of 85.6%, representing an

Method	GSM8K		MATH500		AMC23		AIME24–26		Avg.	
	Acc	Length	Acc	Length	Acc	Length	Acc	Length	Acc	Length
Qwen-2.5-Math-1.5B	83.02	(316.5)	62.40	(902.5)	47.50	(826.4)	10.00	(1,125.3)	50.73	(792.68)
DeepSeek-R1-Distill-Qwen-1.5B	68.69	(440.3)	78.40	(4,371.5)	67.50	(8,066.1)	25.56	(15,115.1)	60.04	(6,998.25)
Simple Routing	81.27	**(355.4)**	70.00	**(1,713.3)**	50.00	(2,785.5)	11.11	(3,730.1)	53.10	**(2,146.04)**
TA	71.49	(1,884.1)	65.80	(2,298.8)	37.50	(2,829.8)	7.78	(3,457.8)	45.64	(2,617.63)
ACM-TA (Baseline)	82.87	(2,460.3)	77.20	(2,835.1)	52.50	**(2,535.7)**	**14.44**	**(2,975.5)**	56.75	(2,701.65)
DALSA (Ours)	**84.84**	(2,457.5)	**78.40**	(2,653.2)	**57.50**	(2,766.2)	12.22	(3,175.0)	**58.24**	(2,762.98)

Table 1: Main mathematical results at 1.5B: accuracy (%) and average generated tokens (parentheses). AIME24–26 averages the three yearly test sets; Avg. averages the four displayed benchmark families. TA is full-model task arithmetic; ACM-TA uses fixed $\lambda = 0.7$. Routing uses the same difficulty score as DALSA with threshold 0.5. Bold marks best accuracy and shortest length within the method block.

Method	GSM8K		MATH500		AMC23		AIME24–26		Avg.	
	Acc	Length	Acc	Length	Acc	Length	Acc	Length	Acc	Length
Qwen-2.5-Math-7B	88.9	(130.2)	77.80	(636.9)	60.00	(1,002.8)	11.11	(1,375.1)	59.45	(786.26)
DeepSeek-R1-Distill-Qwen-7B	85.06	(641.1)	92.40	(3,366.5)	95.00	(4,888.8)	47.78	(12,453.8)	80.06	(5,342.54)
Simple Routing	91.96	**(410.2)**	74.40	(1,583.2)	52.50	(2,308.2)	10.00	(2,918.7)	57.22	(1,805.06)
TA	92.65	(598.9)	84.00	**(1,246.9)**	70.00	**(1,658.6)**	**17.78**	**(2,545.5)**	66.11	**(1,512.48)**
ACM-TA (Baseline)	92.95	(2,072.3)	81.80	(2,588.7)	**75.00**	(2,789.2)	16.67	(2,909.4)	**66.61**	(2,589.90)
DALSA (Ours)	**95.53**	(974.8)	**85.60**	(1,805.0)	67.50	(2,581.3)	**17.78**	(2,683.4)	66.60	(2,011.09)

Table 2: Main mathematical results at 7B, using the metrics, averaging, and highlighting conventions of Table 1.

improvement over the Static method, which scored 81.8%.

5.1.2 Efficiency (Response Length Reduction)

The efficiency benefits of DALSA are most visible relative to the native long-reasoning model and the static ACM-TA baseline.

- In the GSM8K category, DALSA generated 974.8 tokens on average in the latest 7B run while improving accuracy over both the Static and Original Short baselines. Compared with the static ACM-TA baseline, DALSA also reduced GSM8K response length from 2,072.3 tokens.

- Comparison of Response Lengths: Across the completed displayed 7B columns, DALSA used 2,011.09 tokens on average, far below the 5,342.54-token average of the native long-reasoning model and below the 2,589.90-token average of the static ACM-TA baseline. This disparity suggests that adaptive merging can retain useful long-reasoning behavior without imposing full long-form reasoning on every problem.

5.1.3 Adaptive Reasoning Capability

A comparison of token lengths across various datasets underscores DALSA's capacity for dynamic adaptation. For example, in the 7B setting, the token count for GSM8K was 974.8, while the token count for the more challenging AIME24–26 aggregate increased to 2,683.4 tokens. This is consistent with difficulty-dependent allocation, although aggregate lengths alone do not isolate the effect of the controller. At the same time, the AIME24–26 result should be read under the short-base context constraint: many contest-style examples are assigned to high merge buckets but still reach generation limits within the 4096-token context window. Appendix Table 6 reports a long-base diagnostic with higher AIME accuracy and longer responses. Because it changes both the merge direction and context budget, it does not isolate the effect of context length. The default setting also lowers AMC23 accuracy relative to ACM-TA at 7B (67.50% vs. 75.00%) and AIME24–26 accuracy at 1.5B (12.22% vs. 14.44%), illustrating that the gains are not uniform.

5.2 Analysis of Difficulty Estimator

The difficulty estimator changed substantially from the initial hard-label design. A direct hard-BCE classifier is useful for separating easy and hard examples, but its raw scores tend to be too large for a length-sensitive DALSA controller. In downstream generation, this pushes many examples toward stronger long-model interpolation and can erase the intended efficiency gains. We therefore use the probe as a continuous controller rather than a binary router.

Our final DALSA setting adopts a linear soft-

failure regression probe. The soft-failure target preserves information from multiple rollouts, so partially solved examples are not forced into a brittle easy/hard label. Because the raw regression score is compressed in scale, we apply quantile remapping before converting it to λ. This combination is associated with the GSM8K and MATH500 gains in Tables 1 and 2: the controller can assign more long-reasoning capacity where it helps, while avoiding the overly long generations produced by a uniformly large static merge coefficient. Appendix tables provide the historical hard-label probe comparisons and downstream architecture ablations. Separate ablations of the dynamic coefficient and consensus factor are reported in Appendix B.3; their benefits are not uniformly additive.

5.3 Additional Comparisons and Sensitivity

Appendix B adds TIES-style trimming, DARE, and an approximate FIM-weighted TA proxy, full routing-threshold and merge-range sweeps, and component ablations. Static settings are retrospective best-of-grid diagnostics; regenerated controller runs are kept separate from Tables 1–2. FIM-TA at 7B achieves 66.11% accuracy with 1,104.6 tokens, illustrating a competitive alternative operating point rather than uniform DALSA superiority. The range sweep is non-monotonic: in the regenerated 7B suite, $[0.1, 0.3]$ gives 65.28%/1,265.3 tokens and improves both axes over all nine interior routing thresholds, but the always-short endpoint remains shorter. Removing consensus hurts the 1.5B rerun but improves the 7B rerun; neither component has a uniformly additive benefit. Four-generation checks reduce the apparent range-sensitivity gap. These results motivate held-out selection of the operating range and repeated evaluation. The Llama-family diagnostic (Appendix B.5) shows partial accuracy transfer but no efficiency gain, with prompt differences limiting causal interpretation.

5.4 Non-Mathematical Diagnostics

Table 3 evaluates CommonsenseQA (Talmor et al., 2019) and ARC-Challenge (Clark et al., 2018) with the 1.5B pair. The earlier setting already used commonsense training data and activation calibration, but used a residual-MLP hard-BCE controller, raw scores, and $\lambda \in [0.1, 0.5]$. We replace its DALSA and routing rows with reruns using domain-trained linear soft-failure probes, per-domain quantile remapping, and the main $[0.1, 0.7]$ range; Appendix B.4 preserves the earlier results.

The aligned reruns have nearly identical average accuracy (DALSA 52.15%, routing 52.13%), while DALSA uses 2.86 times as many tokens. Moreover, 85.3% of CSQA and 90.0% of ARC-C outputs reach the generation limit. Thus, domain-matched calibration does not establish an efficiency advantage over routing. Since the probes are retrained on commonsense data, this is not zero-shot transfer. Bucket-wise accuracies are non-monotonic and confounded by different merged checkpoints; fixed-model validation and termination control remain necessary.

5.5 Discussion

The results support scale-dependent accuracy–length trade-offs on the evaluated Qwen mathematical benchmarks. DALSA improves over static ACM-TA on both axes for GSM8K and MATH500 at 7B, while contest-style results remain mixed. The long-base diagnostic in Appendix Table 6 obtains higher AIME accuracy at the cost of longer responses and weaker performance on easier benchmarks. Controlled evaluations that vary the context budget while holding the merge formulation fixed are needed to separate these effects.

6 Conclusion

This study introduces DALSA (Difficulty-Aware Long2Short), an inference-time framework combining a lightweight hidden-state difficulty probe trained on rollout-derived soft-failure targets with dynamic model merging guided by activation consensus. On the evaluated Qwen mathematical benchmarks, DALSA reaches nearly identical average accuracy to static ACM-TA at 7B while reducing response length by 22.35%. On GSM8K and MATH500, it improves both accuracy and response length. At 1.5B, it improves average accuracy by 1.49 percentage points with a 2.27% length increase. Routing-threshold and merge-range sweeps reveal configuration-dependent operating points, and consensus ablations show different effects at the two scales. For AIME-style problems, the current short-base implementation frequently reaches generation limits. The long-base diagnostic achieves higher accuracy with longer responses but changes both merge direction and context budget, so it does not isolate the cause of the gap. Domain-matched commonsense and Llama-family diagnostics do not establish broader effi-

method	CSQA		ARC-C		Avg.	
	Acc	Length	Acc	Length	Acc	Length
Qwen-2.5-Math-1.5B	41.69	(332.1)	54.10	(361.3)	47.89	(346.7)
DeepSeek-R1-Distill-Qwen-1.5B	48.57	(1,386.3)	64.08	(1,608.9)	56.32	(1,497.6)
TA	39.80	(1,614.4)	46.25	(1,551.3)	43.02	(1,582.9)
ACM-TA (original)	46.60	(1,424.7)	59.13	(1,397.0)	52.87	(1,410.8)
Simple Routing (aligned rerun)	44.88	(598.1)	59.39	(686.4)	52.13	(642.2)
DALSA (aligned rerun)	46.36	(1,794.4)	57.94	(1,879.2)	52.15	(1,836.8)

Table 3: Non-mathematical diagnostics at 1.5B: accuracy (%) and average generated tokens. The upper block retains the original native and static results; the lower block replaces the earlier DALSA/routing rows with domain-matched linear soft-failure reruns. CSQA uses CommonsenseQA validation; ARC-C uses ARC-Challenge test. Routing/native models use native prompts; merged models use the shared prompt. Both rerun methods use a 2,048-generation-token limit and 4,096-token context. Cross-block differences are not controlled same-run comparisons.

ciency gains. The strongest evidence thus supports adaptive merging for the evaluated Qwen mathematical setting, with model-family calibration, generation termination, and held-out range selection remaining important directions.

Limitations

The current difficulty probe is calibrated from rollout-derived soft-failure targets and remains sensitive to the calibration distribution. Our strongest empirical evidence concerns mathematical reasoning on the evaluated Qwen pairs; benefits depend on model scale, merge operator, and coefficient range. Domain-matched commonsense calibration does not yield a favorable accuracy–length tradeoff over routing, and the Llama-family diagnostic does not show an efficiency gain. Generation termination and family-specific calibration require further study. Contest-style problems can be mapped to high merge buckets while remaining constrained by generation and context limits. The long-base diagnostic changes both the merge direction and context budget, so it does not isolate the cause of the AIME gap. Future work should evaluate matched prompts and budgets across additional model families, select merge ranges on held-out data, validate difficulty rankings with a fixed model, and report repeated-run uncertainty and end-to-end latency, memory, and offline calibration costs. The current token-count results alone do not establish deployment speedups.

Acknowledgments

This work was supported by the Institute of Information & communications Technology Planning & Evaluation (IITP) grants funded by the Korea government (MSIT): No. RS-2020-II201336, Artificial Intelligence graduate school support (UNIST); and No. RS-2026-25523906, Hyper-scale Industrial AI Research Support (R&D) Program, Development of an industry-specified multimodal hyper-scale foundation model.

References

Zhenni Bi, Kai Han, Chuanjian Liu, Yehui Tang, and Yunhe Wang. 2025. Forest-of-thought: Scaling test-time compute for enhancing LLM reasoning. In *Forty-second International Conference on Machine Learning*.

Runxi Cheng, Feng Xiong, Yongxian Wei, Wanyun Zhu, and Chun Yuan. 2025a. Whoever started the interference should end it: Guiding data-free model merging via task vectors. In *Forty-second International Conference on Machine Learning*.

Xiaoxue Cheng, Junyi Li, Zhenduo Zhang, Xinyu Tang, Wayne Xin Zhao, Xinyu Kong, and Zhiqiang Zhang. 2025b. Incentivizing dual process thinking for efficient large language model reasoning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? Try ARC, the AI2 reasoning challenge. *Preprint*, arXiv:1803.05457.

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

Pala Tej Deep, Rishabh Bhardwaj, and Soujanya Poria. 2024. Della-merging: Reducing interference in

model merging through magnitude-based sampling. *Preprint*, arXiv:2406.11617.

Gongfan Fang, Xinyin Ma, and Xinchao Wang. 2025. Thinkless: LLM learns when to think. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Sicheng Feng, Gongfan Fang, Xinyin Ma, and Xinchao Wang. 2025. Efficient reasoning models: A survey. *Preprint*, arXiv:2504.10903.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, and 175 others. 2025. Deepseek-r1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645:633–638.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the MATH dataset. *Preprint*, arXiv:2103.03874.

Zhiyuan Hu, Yibo Wang, Hanze Dong, Yuhui Xu, Amrita Saha, Caiming Xiong, Bryan Hooi, and Junnan Li. 2025. Beyond 'aha!': Toward systematic meta-abilities alignment in large reasoning models. *Preprint*, arXiv:2505.10554.

Chengyu Huang, Zhengxin Zhang, and Claire Cardie. 2025a. Hapo: Training language models to reason concisely via history-aware policy optimization. *Preprint*, arXiv:2505.11225.

Shijue Huang, Hongru Wang, Wanjun Zhong, Zhaochen Su, Jiazhan Feng, Bowen Cao, and Yi R. Fung. 2025b. Adactrl: Towards adaptive and controllable reasoning via difficulty-aware budgeting. *Preprint*, arXiv:2505.18822.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. 2023. Editing models with task arithmetic. In *The Eleventh International Conference on Learning Representations*.

Lingjie Jiang, Xun Wu, Shaohan Huang, Qingxiu Dong, Zewen Chi, Li Dong, Xingxing Zhang, Tengchao Lv, Lei Cui, and Furu Wei. 2025. Think only when you need with large hybrid-reasoning models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Zhensheng Jin, Xinze Li, Yifan Ji, Chunyi Peng, Zhenghao Liu, Qi Shi, Yukun Yan, Shuo Wang, Furong Peng, and Ge Yu. 2025. ReCUT: Balancing reasoning length and accuracy in LLMs via stepwise trails and preference optimization. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 14269–14282, Suzhou, China. Association for Computational Linguistics.

Deyang Kong, Qi Guo, Xiangyu Xi, Wei Wang, Jingang Wang, Xunliang Cai, Shikun Zhang, and Wei Ye. 2025. Rethinking the sampling criteria in reinforcement learning for llm reasoning: A competence-difficulty alignment perspective. *Preprint*, arXiv:2505.17652.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with PagedAttention. In *Proceedings of the 29th Symposium on Operating Systems Principles*.

Zheng Li, Qingxiu Dong, Jingyuan Ma, Di Zhang, Kai Jia, and Zhifang Sui. 2025. Selfbudgeter: Adaptive token allocation for efficient llm reasoning. *Preprint*, arXiv:2505.11274.

Jonathan Light, Wei Cheng, Benjamin Riviere, Yue Wu, Masafumi Oyamada, Mengdi Wang, Yisong Yue, Santiago Paternain, and Haifeng Chen. 2025. DISC: Dynamic decomposition improves LLM inference scaling. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Wei Liu, Ruochen Zhou, Yiyun Deng, Yuzhen Huang, Junteng Liu, Yuntian Deng, Yizhe Zhang, and Junxian He. 2025. Learn to reason efficiently with adaptive length-based reward shaping. *Preprint*, arXiv:2505.15612.

Llama Team. 2024. The Llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Chenwei Lou, Zewei Sun, Xinnian Liang, Meng Qu, Wei Shen, Wenqi Wang, Yuntao Li, Qingping Yang, and Shuangzhi Wu. 2025. Adacot: Pareto-optimal adaptive chain-of-thought triggering via reinforcement learning. *Preprint*, arXiv:2505.11896.

Jinghui Lu, Haiyang Yu, Siliang Xu, Shiwei Ran, Guozhi Tang, Siqi Wang, Bin Shan, Teng Fu, Hao Feng, Jingqun Tang, Han Wang, and Can Huang. 2025. Prolonged reasoning is not all you need: Certainty-based adaptive routing for efficient llm/mllm reasoning. *Preprint*, arXiv:2505.15154.

Feng Luo, Yu-Neng Chuang, Guanchu Wang, Hoang Anh Duy Le, Shaochen Zhong, Hongyi Liu, Jiayi Yuan, Yang Sui, Vladimir Braverman, Vipin Chaudhary, and Xia Hu. 2025a. Autol2s: Auto long-short reasoning for efficient large language models. *Preprint*, arXiv:2505.22662.

Haotian Luo, Haiying He, Yibo Wang, Jinluan Yang, Rui Liu, Naiqiang Tan, Xiaochun Cao, Dacheng Tao, and Li Shen. 2025b. Ada-r1: Hybrid-cot via bi-level adaptive reasoning optimization. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Huan Ma, Jingdong Chen, Joey Tianyi Zhou, Guangyu Wang, and Changqing Zhang. 2025. Estimating llm uncertainty with evidence. *Preprint*, arXiv:2502.00290.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Michael S Matena and Colin A Raffel. 2022. Merging models with Fisher-weighted averaging. In *Advances in Neural Information Processing Systems*, volume 35.

Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. 2025. s1: Simple test-time scaling. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 20275–20321, Suzhou, China. Association for Computational Linguistics.

Amin Heyrani Nobari, Kaveh Alim, Ali Arjomand-Bigdeli, Akash Srivastava, Faez Ahmed, and Navid Azizan. 2025. Activation-informed merging of large language models. *Preprint*, arXiv:2502.02421.

Xiaoye Qu, Yafu Li, Zhao-Chen Su, Weigao Sun, Jianhao Yan, Dongrui Liu, Ganqu Cui, Daizong Liu, Shuxian Liang, Junxian He, Peng Li, Wei Wei, Jing Shao, Chaochao Lu, Yue Zhang, Xian-Sheng Hua, Bowen Zhou, and Yu Cheng. 2025. A survey of efficient reasoning for large reasoning models: Language, multimodality, and beyond. *Preprint*, arXiv:2503.21614.

Yi Shen, Jian Zhang, Jieyun Huang, Shuming Shi, Wenjing Zhang, Jiangze Yan, Ning Wang, Kai Wang, Zhaoxiang Liu, and Shiguo Lian. 2025. DAST: Difficulty-adaptive slow-thinking for large reasoning models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 2322–2331, Suzhou (China). Association for Computational Linguistics.

Charlie Victor Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. 2025. Scaling LLM test-time compute optimally can be more effective than scaling parameters for reasoning. In *The Thirteenth International Conference on Learning Representations*.

Yang Sui, Yu-Neng Chuang, Guanchu Wang, Jiamu Zhang, Tianyi Zhang, Jiayi Yuan, Hongyi Liu, Andrew Wen, Shaochen Zhong, Na Zou, Hanjie Chen, and Xia Hu. 2025. Stop overthinking: A survey on efficient reasoning for large language models. *Preprint*, arXiv:2503.16419.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

V. Venktesh, Md. Shad Akhtar, Mukesh Mohania, and Vikram Goyal. 2022. Auxiliary task guided interactive attention model for question difficulty prediction. In *Artificial Intelligence in Education: 23rd International Conference, AIED 2022, Durham, UK, July 27–31, 2022, Proceedings, Part I*, pages 477–489, Berlin, Heidelberg. Springer-Verlag.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*.

Han Wu, Yuxuan Yao, Shuqi Liu, Zehua Liu, Xiaojin Fu, Xiongwei Han, Xing Li, Hui-Ling Zhen, Tao Zhong, and Mingxuan Yuan. 2025a. Unlocking efficient long-to-short llm reasoning with model merging. *Preprint*, arXiv:2503.20641.

Siye Wu, Jian Xie, Yikai Zhang, Aili Chen, Kai Zhang, Yu Su, and Yanghua Xiao. 2025b. ARM: Adaptive reasoning model. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Prateek Yadav, Derek Tam, Leshem Choshen, Colin Raffel, and Mohit Bansal. 2023. TIES-merging: Resolving interference when merging models. In *Thirty-seventh Conference on Neural Information Processing Systems*.

An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu, Jingren Zhou, Junyang Lin, Keming Lu, Mingfeng Xue, Runji Lin, Tianyu Liu, Xingzhang Ren, and Zhenru Zhang. 2024. Qwen2.5-math technical report: Toward mathematical expert model via self-improvement. *Preprint*, arXiv:2409.12122.

Enneng Yang, Li Shen, Guibing Guo, Xingwei Wang, Xiaochun Cao, Jie Zhang, and Dacheng Tao. 2025. Model merging in llms, mllms, and beyond: Methods, theories, applications and opportunities. *Preprint*, arXiv:2408.07666. Version 5.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik R Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Yuxuan Yao, Shuqi LIU, Zehua Liu, Qintong Li, Mingyang LIU, Xiongwei Han, Zhijiang Guo, Han

Wu, and Linqi Song. 2025. Activation-guided consensus merging for large language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Le Yu, Bowen Yu, Haiyang Yu, Fei Huang, and Yongbin Li. 2024. Language models are super mario: Absorbing abilities from homologous models as a free lunch. In *Forty-first International Conference on Machine Learning*.

Linan Yue, Yichao Du, Yizhi Wang, Weibo Gao, Fangzhou Yao, Li Wang, Ye Liu, Ziyu Xu, Qi Liu, Shimin Di, and Min-Ling Zhang. 2025. Don't overthink it: A survey of efficient r1-style large reasoning models. *Preprint*, arXiv:2508.02120.

Yuxiang Zhang, Zhengxu Yu, Weihang Pan, Zhongming Jin, Qiang Fu, Deng Cai, Binbin Lin, and Jieping Ye. 2025. Tokensqueeze: Performance-preserving compression for reasoning LLMs. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

King Zhu, Hanhao Li, Siwei Wu, Tianshun Xing, Dehua Ma, Xiangru Tang, Minghao Liu, Jian Yang, Jiaheng Liu, Yuchen Eleanor Jiang, Changwang Zhang, Chenghua Lin, Jun Wang, Ge Zhang, and Wangchunshu Zhou. 2025a. Scaling test-time compute for LLM agents. *Preprint*, arXiv:2506.12928.

Yubo Zhu, Dongrui Liu, Zecheng Lin, Wei Tong, Sheng Zhong, and Jing Shao. 2025b. The LLM already knows: Estimating LLM-perceived question difficulty via hidden representations. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 1160–1176, Suzhou, China. Association for Computational Linguistics.

Alireza S. Ziabari, Nona Ghazizadeh, Zhivar Sourati, Farzan Karimi-Malekabadi, Payam Piray, and Morteza Dehghani. 2025. Reasoning on a spectrum: Aligning LLMs to system 1 and system 2 thinking. *Preprint*, arXiv:2502.12470.

A Additional Experimental Details

A.1 Original Mathematical Evaluation Settings

Table 4 records the decoding settings underlying the original mathematical results. Each evaluation uses one generation per example and seed 42. Native baselines and routing use each selected model's chat template; native DeepSeek generation includes a forced `<think>` prefix. Merging methods use a shared plain prompt requesting step-by-step reasoning and a final boxed answer. Rollout labeling, hidden-state extraction, and difficulty scoring use the shared short-model prompt. Response length counts generated tokens and excludes probe scoring, calibration, checkpoint materialization, and loading costs.

The main soft-failure probes use $K = 3$ short-model rollouts per GSM8K training example and train for 30 epochs with batch size 64, learning rate 10^{-4}, and AdamW. The 1.5B probe uses all 7473 training examples; the 7B probe uses a stratified split of 5232 training and 2241 validation examples. At both scales, inference-time quantile remapping uses all 7473 GSM8K training-split hidden states as its reference, including the 7B probe-validation partition but excluding benchmark test examples. The recorded 1.5B activation cache uses 20 `simplescaling/s1K` examples, eight clusters, and $\beta = 1$. Calibration features pool the maximum absolute activation over prompt tokens, and mutual information is computed between the two models' cluster assignments across calibration examples. The original 7B runs reuse a precomputed activation cache; its hyperparameters are not independently preserved in the available run summaries.

A.2 Bucketed Inference and Context-Limit Diagnostics

The main DALSA results use a short-base merge, where the merged checkpoint starts from the short-reasoning model and injects a controlled amount of the long-reasoning task vector. This setting preserves the intended long-to-short efficiency objective, but it also inherits the short model's context configuration. For the Qwen-based merged checkpoints used in our main experiments, the context limit is 4096 tokens; the maximum generated length is 3840 tokens at 1.5B and 3072 tokens at 7B.

For efficient vLLM inference, we quantize the continuous DALSA score into a small set of reusable merge buckets. In the 7B soft-quantile setting, the raw difficulty score $p_\phi(x)$ is first normalized using the 5th and 95th percentiles of the GSM8K train H0 reference scores and then clipped to $[0, 1]$. We then map the normalized score to the λ range $[0.1, 0.7]$ and assign it to the nearest bucket in $\{0.1, 0.2, \ldots, 0.7\}$. This is a practical approximation of the continuous formulation that enables batched generation; it is not intended as a separate modeling assumption.

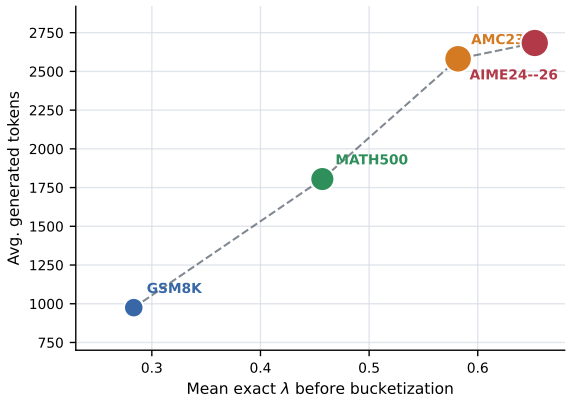


Figure 3: Generated-token profile for the 7B soft-quantile DALSA setting. Each point plots benchmark-family average generated tokens against the mean exact λ before bucketization; marker area indicates the length-stop rate.

Table 5 shows that harder benchmark families are increasingly assigned to upper merge buckets. However, generation remains budget-limited: on AIME24–26, 74 of 90 examples are assigned to the highest bucket and 70 of 90 terminate by a length limit. The recorded length-stop flag does not by itself distinguish the generated-token cap from the total context limit.

Figure 3 visualizes the same pattern. The generated-token cost rises with the mean selected merge strength across benchmark families. This association is consistent with the intended difficulty ordering, while the generation and context limits constrain the available reasoning budget.

A.3 Long-Base Diagnostic

We additionally report an earlier diagnostic comparing short- and long-base merging. Both variants use the earlier hard-label probe and retain the layer-consensus factor. The short-base variant uses $\theta^l_{\text{short}} + (1 - t_l)\lambda(x)(\theta^l_{\text{long}} - \theta^l_{\text{short}})$, with a 4096-token context and a 3072-token generation cap. The long-base variant uses $\theta^l_{\text{long}} + (1 -$

Scale	System	Max. generated	Context	Temp.	Top-p	Prompt
Both	Native Qwen	3072	4096	0.0	1.0	Native
1.5B	Native DeepSeek	32768	32768	0.6	0.95	Native
7B	Native DeepSeek	32768	65536*	0.6	0.95	Native
1.5B	Simple Routing	3840	4096	0.6	0.95	Native
7B	Simple Routing	3072	4096	0.6	0.95	Native
1.5B	TA / ACM-TA / DALSA	3840	4096	0.6	0.95	Shared
7B	TA / ACM-TA / DALSA	3072	4096	0.6	0.95	Shared

Table 4: Original mathematical decoding settings. *The 65,536-token context is explicitly recorded for the 7B native DeepSeek AMC23 and AIME25–26 runs; the earlier native summaries do not preserve an explicit context setting. Native endpoints therefore provide reference operating points under different prompts and budgets, rather than a fully controlled runtime comparison.

Dataset	$\lambda = .1$	.2	.3	.4	.5	.6	.7	Length-stop	Avg. tokens	Acc.
GSM8K	410	374	188	91	66	44	146	23.12%	974.8	95.53
MATH500	48	86	67	44	33	32	190	49.20%	1,805.0	85.60
AMC23	0	2	5	2	3	4	24	72.50%	2,581.3	67.50
AIME24–26	1	2	2	4	1	6	74	77.78%	2,683.4	17.78

Table 5: Bucket distribution for the 7B soft-quantile DALSA setting. Bucket counts are recomputed from per-sample outputs. Length-stop denotes the fraction of examples whose generation terminates by the length limit.

$t_l)\lambda(x)(\theta^l_{\text{short}} - \theta^l_{\text{long}})$, with a 32768-token context and a 16384-token generation cap. For the long-base variant, the coefficient decreases with predicted difficulty, keeping harder inputs closer to the long model. Both variants use buckets from 0 to 1 in steps of 0.1. This diagnostic changes the merge base and generation budgets together, so it does not isolate the causal effect of context length; it also differs from the main soft-failure controller.

In Table 6, moving to a long-base checkpoint raises AIME24–26 accuracy from 16.67% to 31.11% on average, but the five-dataset average response length increases from 1,696.6 to 8,778.3 tokens and performance on GSM8K and MATH500 drops sharply. These results motivate a controlled context-length study while illustrating the cost of this alternative operating point.

A.4 Difficulty Probe Architecture Ablations

Table 7 reports held-out validation of earlier hard-label probes in the 1.5B setting. These probes use binary cross-entropy and are distinct from the soft-failure controller used in the main results. For each dataset, we split the original train split into 70% probe-training and 30% validation partitions with stratification by the rollout-derived hard/easy label. GSM8K uses `openai/gsm8k`; the MATH500 row uses the 12,000-example train split of `ricdomolm/MATH-500`, not the 500-example evaluation set.

Table 8 reports a downstream ablation of these earlier hard-label probe architectures in the 1.5B DALSA setting. We separate this downstream generation result from the held-out validation metrics above and from the main soft-failure controller.

B Additional Evaluations and Sensitivity Analyses

The following analyses were added during the review period. We distinguish the submitted main results from the regenerated controller suite and from retrospective best-of-grid diagnostics. Unless stated otherwise, accuracy and generated length are averaged equally over GSM8K, MATH500, AMC23, and the AIME24–26 family; the latter is first averaged over its three years. The primary additional evaluations use one generation per prompt. They are point estimates, not significance tests. Regenerated results are compared within their own suite and do not replace the submitted mathematical results in Tables 1 and 2.

B.1 Additional Merging Operators

We evaluated single-vector TIES-style magnitude trimming (Yadav et al., 2023) with density 0.2, DARE (Yu et al., 2024) with drop rate $p = 0.5$, $1/(1-p)$ rescaling and mask seed 42, and an approximate Fisher-information-weighted task-arithmetic proxy (FIM-TA), inspired by Fisher-weighted model merging (Matena and Raffel, 2022). With one task vector, TIES sign election does not change the result. The FIM-TA proxy

Variant	GSM8K		MATH500		AIME24		AIME25		AIME26		Avg.	
	Acc	Length	Acc	Length	Acc	Length	Acc	Length	Acc	Length	Acc	Length
Short-base DALSA, $\lambda \in [0, 1]$	95.30	(826.6)	84.40	(1,243.5)	16.67	(2,163.8)	20.00	(2,172.4)	13.33	(2,076.8)	45.94	(1,696.6)
Long-base diagnostic, $\lambda \in [0, 1]$	81.27	(5,388.6)	67.00	(7,724.0)	40.00	(11,091.4)	30.00	(8,257.0)	23.33	(11,430.6)	48.32	(8,778.3)

Table 6: Earlier short-base versus long-base diagnostic in the 7B setting, using the hard-label probe. The long-base variant has higher AIME accuracy, lower GSM8K/MATH500 accuracy, and longer responses. Merge base and generation budgets differ jointly. Avg. is over the five displayed datasets, unlike the four-family average in the main tables.

Dataset	Probe	N	Hard%	Pred. Hard%	Acc	Macro-F1	ROC-AUC
GSM8K	Residual MLP	2242	11.06	17.44	77.03	53.01	57.92
GSM8K	Linear Probe	2242	11.06	36.84	64.85	51.76	63.59
MATH500	Residual MLP	3600	34.78	51.64	69.36	68.79	80.76
MATH500	Linear Probe	3600	34.78	45.69	75.92	74.96	84.95

Table 7: Held-out validation of earlier hard-label probe architectures at 1.5B; these metrics do not validate the main soft-failure controller.

estimates layer-wise weights from eight random sequences of length 128; it is not an official FIM implementation. All three methods use the short model as the base. These additional diagnostics use the same benchmark splits and shared merged-model prompt, with temperature 0.6, top-p 0.95, a 3,840-generation-token limit, and a 4,096-token context window. In particular, their generation cap differs from the original 7B merging cap of 3,072 tokens in Table 4.

Table 9 reports the most accurate tested setting from $\lambda \in \{0.3, 0.5, 0.7\}$ for each additional operator. These settings were selected retrospectively on the reported benchmarks, not on held-out validation data. None exceeds DALSA's observed average accuracy, but the gaps to FIM-TA are only 0.43 and 0.49 percentage points at 1.5B and 7B, respectively, and FIM-TA generates substantially shorter responses. These differences do not establish statistically significant superiority. At $\lambda = 0.7$, DARE accuracy drops to 9.04% at 1.5B and 0.09% at 7B; the cause of this sensitivity remains undetermined.

We also combine the dynamic coefficient with the FIM-weighted proxy delta, using $\lambda_{\text{dyn}}(x)\alpha_l\Delta\theta^l$ in place of $\lambda_{\text{dyn}}(x)(1 - t_l)\Delta\theta^l$. This variant removes the ACM factor and uses the same approximate FIM construction. It reuses the controller, reference hidden states, and evaluation settings of the regenerated ablation suite. Table 10 shows compatibility with this additional operator, but not operator-independent gains. At 7B, the $[0.1, 0.7]$ FIM variant improves observed accuracy over the same-range regenerated ACM-based DALSA (64.34 versus 63.51) and reduces length (1,087.3 versus 2,291.0). With $[0.1, 0.9]$, it instead trades lower accuracy for shorter length (66.20/1,138.5 versus 67.87/1,975.5). The 1.5B variants have lower accuracy than their corresponding ACM-based variants. The static FIM-TA results in Table 9 come from a separate run and do not provide a controlled static-versus-dynamic FIM comparison.

B.2 Merge-Range and Routing-Threshold Sensitivity

Table 11 presents the full additional range and threshold sweeps, visualized in Figure 4. The controller and generations were regenerated once per scale, so these rows should be interpreted separately from the submitted main results. The router reuses one set of native short/long generations and changes only their assignment across thresholds. The capped endpoints use temperature 0.6, top-p 0.95, a 3,840-generation-token limit, and a 4,096-token context window. Routing uses native endpoint prompts, whereas DALSA uses the shared merged-model prompt. The comparison therefore measures the established end-to-end protocols and does not isolate continuous interpolation as the only causal difference.

At 1.5B, the capped always-short endpoint is more accurate and shorter than every interior routing threshold. DALSA's highest-accuracy points improve accuracy at a substantial length cost, rather than dominating that endpoint. At 7B, DALSA $[0.1, 0.3]$ exceeds the best interior threshold by 3.22 percentage points while reducing generated length by 34.0%, and is better on both axes than all nine

Probe	GSM8K		MATH500		AIME24		Avg.	
	Acc	Length	Acc	Length	Acc	Length	Acc	Length
Residual MLP	83.02	(2002.9)	72.20	(2721.2)	13.33	(3669.7)	56.18	(2798.0)
Linear Probe	81.43	(3031.2)	71.40	(3078.6)	13.33	(3671.4)	55.39	(3260.4)

Table 8: Downstream ablation of earlier hard-label probe architectures in the 1.5B DALSA setting.

Scale	Method	λ	Avg. Acc.	Avg. Length
1.5B	TIES (single-vector)	0.3	57.11	2,302.4
1.5B	DARE	0.3	54.83	3,511.1
1.5B	FIM-TA proxy	0.5	57.81	2,422.6
7B	TIES (single-vector)	0.5	63.74	1,514.4
7B	DARE	0.5	65.77	3,239.6
7B	FIM-TA proxy	0.7	66.11	1,104.6
1.5B	ACM-TA (submitted reference)	0.7	56.75	2,701.7
1.5B	DALSA (submitted reference)	$[0.1, 0.7]$	58.24	2,763.0
7B	ACM-TA (submitted reference)	0.7	66.61	2,589.9
7B	DALSA (submitted reference)	$[0.1, 0.7]$	66.60	2,011.1

Table 9: Retrospective best-of-grid static merging diagnostics, with submitted references shown separately. Accuracy is in percent and length is average generated tokens. The additional grid uses one generation per prompt; the selected settings are not validation-selected.

Scale	Range	Avg. Acc.	Avg. Length
1.5B	$[0.1, 0.7]$	56.16	2,227.1
1.5B	$[0.1, 0.9]$	53.66	2,545.3
7B	$[0.1, 0.7]$	64.34	1,087.3
7B	$[0.1, 0.9]$	66.20	1,138.5

Table 10: DALSA-on-FIM proxy in the regenerated controller suite; one generation per prompt.

interior thresholds. The always-short endpoint nevertheless remains shorter. These results do not establish universal Pareto dominance.

Range sensitivity is non-monotonic. At 1.5B, $[0.1, 0.3]$ has the highest observed accuracy, while $[0.1, 0.9]$ is shorter but 3.36 points less accurate. At 7B, the one-generation accuracy difference between $[0.1, 0.9]$ and $[0.1, 0.7]$ is 4.36 points; with four generations per prompt, it decreases to 0.70 points (Table 14). Further repeated evaluation is needed to assess robustness. We retain $[0.1, 0.7]$ for continuity with the submitted, reference-aligned configuration; the sweep does not establish it as optimal. Future range selection should use held-out validation data.

B.3 Difficulty and Consensus Contributions

We report two separate contrasts, rather than a same-run complete 2×2 factorial experiment. First, the submitted ACM-TA/DALSA comparison keeps the consensus form fixed and replaces the fixed coefficient $\lambda = 0.7$ with $\lambda_{\text{dyn}}(x)$. At 1.5B, accuracy increases from 56.75% to 58.24% and length increases by 2.27%; at 7B, accuracy differs by 0.01 points and length decreases by 22.35% (Tables 1 and 2).

Second, Table 12 holds the regenerated dynamic controller fixed and removes the layer-consensus factor. Each scale uses the same controller, score remapping, $[0.1, 0.7]$ range, prompts, and generation limits within this comparison. Removing consensus reduces average accuracy by 9.81 points at 1.5B and increases length by 281.0 tokens. At 7B, the direction reverses; the no-consensus variant is more accurate and shorter. The four-generation 7B evaluation retains this pattern, with a 2.12-point accuracy difference (Table 14). The observations suggest an interaction between the components and evaluated settings. They do not establish whether model capacity, parameter interference, calibration, or another factor causes the difference.

B.4 Commonsense Controller and Termination Details

The submitted commonsense evaluation used CSQA/ARC training rollouts and a commonsense activation-calibration pool. Its DALSA controller was an earlier residual-MLP hard-BCE probe with raw, unremapped scores and range $[0.1, 0.5]$; it was not the linear soft-failure controller used in the main mathematical results. Table 15 preserves the submitted comparison for provenance. Table 3

Operating point	1.5B Avg. Acc.	1.5B Avg. Length	7B Avg. Acc.	7B Avg. Length
Capped always-long endpoint	50.91	2,264.8	60.52	2,384.9
Routing, $\tau = 0.1$	52.07	2,243.3	61.38	2,326.3
Routing, $\tau = 0.2$	52.43	2,227.6	61.36	2,195.8
Routing, $\tau = 0.3$	52.90	2,215.0	61.65	2,085.7
Routing, $\tau = 0.4$	52.75	2,177.1	61.16	2,017.3
Routing, $\tau = 0.5$	53.19	2,140.1	62.06	1,917.2
Routing, $\tau = 0.6$	53.38	2,125.2	62.01	1,860.1
Routing, $\tau = 0.7$	53.60	2,115.3	61.30	1,797.4
Routing, $\tau = 0.8$	53.67	2,109.2	60.70	1,733.5
Routing, $\tau = 0.9$	53.40	2,099.7	60.17	1,680.1
Capped always-short endpoint	56.30	657.7	61.64	865.4
DALSA $[0.1, 0.3]$	58.07	2,740.9	65.28	1,265.3
DALSA $[0.1, 0.5]$	55.79	3,014.7	63.83	2,057.5
DALSA $[0.1, 0.7]$ (reference range)	57.65	2,623.8	63.51	2,291.0
DALSA $[0.1, 0.9]$	54.71	2,446.8	67.87	1,975.5

Table 11: Complete routing-threshold and merge-range sweeps in the regenerated suite. Each point uses one generation per prompt; endpoint prompts differ from merged-model prompts. The four-family averages are not comparable to the five-column averages in the separate $[0, 1]$ long-base diagnostic (Table 6).

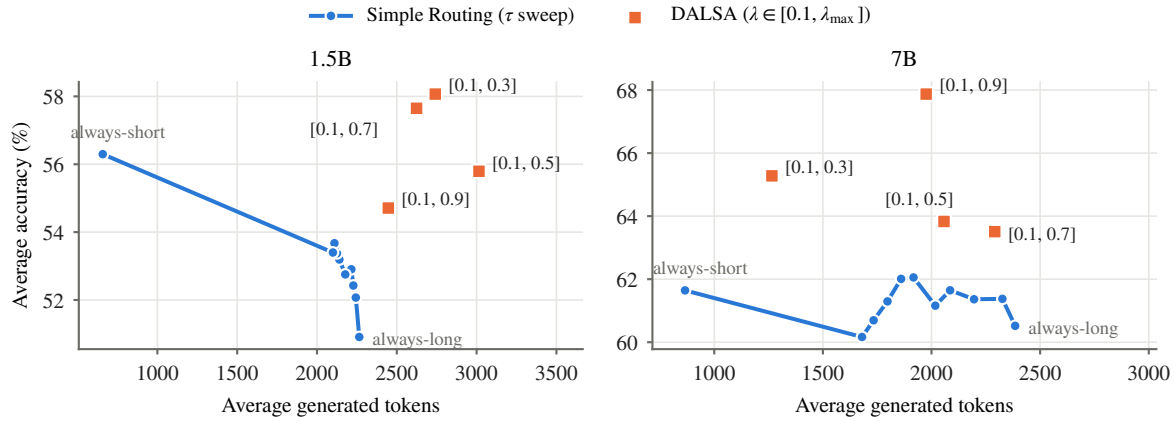


Figure 4: Accuracy–length trade-offs for the complete routing-threshold and DALSA merge-range sweeps in Table 11. Each point averages the four benchmark families in the regenerated suite, with one generation per prompt. Routing uses native endpoint prompts; DALSA uses the shared merged-model prompt. The capped endpoints use the same generation limits as the sweep and differ from the native rows in Tables 1 and 2.

presents the aligned rerun with a linear soft-failure probe retrained on the corresponding commonsense training data, per-domain quantile remapping, and range $[0.1, 0.7]$. Both commonsense evaluations use a 2,048-generation-token limit and a 4,096-token context limit.

The aligned controller reduces DALSA's average length by 4.79%, but DALSA and routing have nearly identical accuracy (52.15% versus 52.13%) while DALSA generates approximately 2.86 times as many tokens. In the new DALSA outputs, 1,042/1,221 CSQA responses (85.3%) and 1,055/1,172 ARC-Challenge responses (90.0%) hit the generation limit. The submitted ACM-TA run also reaches that limit, at lower rates of 64.1% and 61.4%, respectively. Termination is therefore an important efficiency limitation; these observations do not establish the cause of the accuracy differences.

All seven DALSA buckets are populated. The lowest-score bucket has the highest observed accuracy (57.9% on CSQA and 79.1% on ARC-Challenge), but the remaining bucket accuracies are not monotonic. Each bucket uses a different merged checkpoint, so these results jointly reflect input difficulty and checkpoint-specific generation quality. The probe is also retrained on commonsense data. Accordingly, this analysis establishes neither zero-shot cross-domain transfer nor isolated validity of the difficulty ordering. Fixed-model

Scale	Variant	GSM8K	MATH500	AMC23	AIME24–26	Avg.
1.5B	With consensus	84.69 (2,429.5)	74.80 (2,535.8)	60.00 (2,582.2)	11.11 (2,947.9)	57.65 (2,623.8)
1.5B	Without consensus	84.23 (2,738.0)	68.80 (2,591.1)	35.00 (2,847.9)	3.33 (3,442.2)	47.84 (2,904.8)
7B	With consensus	95.30 (1,156.0)	85.40 (1,982.2)	60.00 (2,699.5)	13.33 (3,326.4)	63.51 (2,291.0)
7B	Without consensus	95.38 (1,199.3)	86.20 (1,720.2)	67.50 (2,450.0)	22.22 (2,939.1)	67.82 (2,077.2)

Table 12: Consensus ablation in the regenerated suite at $\lambda \in [0.1, 0.7]$, with one generation per prompt. Entries give accuracy (%) and average generated tokens in parentheses. These are matched within each scale, but are separate from the submitted fixed-versus-dynamic comparison.

Method	GSM8K	MATH500	AMC23	AIME24–26	Avg.
Llama-3.1-8B-Instruct (short)	83.62 (230.7)	47.60 (607.2)	22.50 (1,005.4)	5.56 (928.7)	39.82 (693.0)
DeepSeek-R1-Distill-Llama-8B (long)	77.56 (539.7)	73.40 (2,056.8)	45.00 (2,751.4)	15.56 (3,631.5)	52.88 (2,244.8)
DALSA $[0.1, 0.7]$	82.41 (3,692.2)	60.60 (3,535.0)	35.00 (3,465.8)	6.67 (3,572.4)	46.17 (3,566.4)
DALSA $[0.1, 0.9]$	83.17 (3,570.5)	64.20 (3,277.3)	45.00 (3,514.2)	8.89 (3,364.8)	50.31 (3,431.7)
DALSA w/o consensus $[0.1, 0.7]$	83.47 (3,248.6)	68.00 (2,867.0)	45.00 (2,895.9)	11.11 (3,227.8)	51.90 (3,059.8)

Table 13: Llama-family diagnostic with capped native endpoints. Entries give accuracy (%) and average generated tokens in parentheses. Native and merged rows use different prompt protocols; the no-consensus variant is a separate ablation.

Consensus	Range	Acc.	Length
With	$[0.1, 0.7]$	66.07	2,277.0
With	$[0.1, 0.9]$	66.77	1,987.4
Without	$[0.1, 0.7]$	68.19	2,087.8

Table 14: Additional 7B evaluation averaging four generations per prompt. These averages are reported separately from the one-generation sweep; they are not confidence intervals.

Submitted method	Avg. Acc.	Avg. Length
DALSA	52.47	1,929.3
Simple Routing	53.52	833.1
ACM-TA	52.87	1,410.8

Table 15: Submitted CSQA/ARC-Challenge averages, retained to distinguish the earlier controller from the aligned rerun in Table 3.

score validation and termination control remain future work.

B.5 Llama-Family Diagnostic

Parameter-space interpolation requires tensor-compatible endpoints. To examine an additional family, we applied the controller pipeline to Meta-Llama-3.1-8B-Instruct (Llama Team, 2024) (the NousResearch mirror) and DeepSeek-R1-Distill-Llama-8B (Guo et al., 2025). All rows in Table 13 use a 4,096-token context limit. Native endpoints use their native prompts, whereas merged checkpoints use the shared merging prompt. These regenerated results are therefore a model-family diagnostic, not a controlled isolation of the merging effect.

Full DALSA improves observed accuracy over the short endpoint by 6.35 points with $[0.1, 0.7]$ and 10.49 points with $[0.1, 0.9]$. The latter remains below the capped long endpoint (50.31% versus 52.88%) and produces longer responses (3,431.7 versus 2,244.8 tokens). Removing consensus reaches 51.90%, but removes a core component and does not demonstrate an efficiency gain. The main accuracy–length evidence thus remains limited to the evaluated Qwen pairs. The Llama results motivate family-specific activation-consensus calibration and broader evaluation of aligned model pairs.