

RCP-Merging: Merging Long Chain-of-Thought Models with Domain-Specific Models by Considering Reasoning Capability as Prior

Junyao Yang¹, Jianwei Wang¹, Huiping Zhuang¹, Cen Chen¹, Ziqian Zeng^{1*},

¹South China University of Technology, China

Abstract

Large Language Models (LLMs) with long chain-of-thought (CoT) capability, termed Reasoning Models, demonstrate superior intricate problem-solving abilities through multi-step long CoT reasoning. To create a dual-capability model with long CoT capability and domain-specific knowledge without substantial computational and data costs, model merging emerges as a highly resource-efficient method. However, significant challenges lie in merging domain-specific LLMs with long CoT ones since nowadays merging methods suffer from reasoning capability degradation, even gibberish output and output collapse. To overcome this, we introduce **RCP-Merging**: Merging Long Chain-of-Thought Models with Domain-Specific Models by Considering Reasoning Capability as Prior, a novel merging framework designed to integrate domain-specific LLMs with long CoT capability, meanwhile maintaining model performance in the original domain. Treating reasoning model weights as foundational prior, our method utilizes a reasoning capability indicator to preserve core long CoT capability model weights while selectively merging essential domain-specific weights. We conducted extensive experiments on Qwen2.5-7B, Llama3.1-8B, and Qwen2.5-1.5B models in BioMedicine and Finance domains. Our results show that RCP-Merging successfully merges a reasoning model with domain-specific ones, improving domain task performance by 9.5% and 9.2% over state-of-the-art methods, without significantly harming the original long CoT reasoning capability.

Introduction

Large Language Models (LLMs) with long chain-of-thought (CoT) capability, termed Reasoning Models, have demonstrated exceptional performance on complex reasoning tasks (Jaech et al. 2024; OpenAI 2025; Guo et al. 2025; xAI 2025). Mostly trained on verifiable tasks like code generation and mathematical reasoning, the results in Table 1 show that the reasoning model demonstrates relatively weak performance compared with models that specifically fine-tune on a certain domain. However, long CoT’s multi-step reasoning deduction is critical for complex problem-solving in specific domains like BioMedicine and Finance, extending beyond simple information retrieval (Cui et al. 2025; Tang et al. 2025). Moreover, the scarcity of models specifically trained for these fields remains a key challenge. This

*Corresponding author.

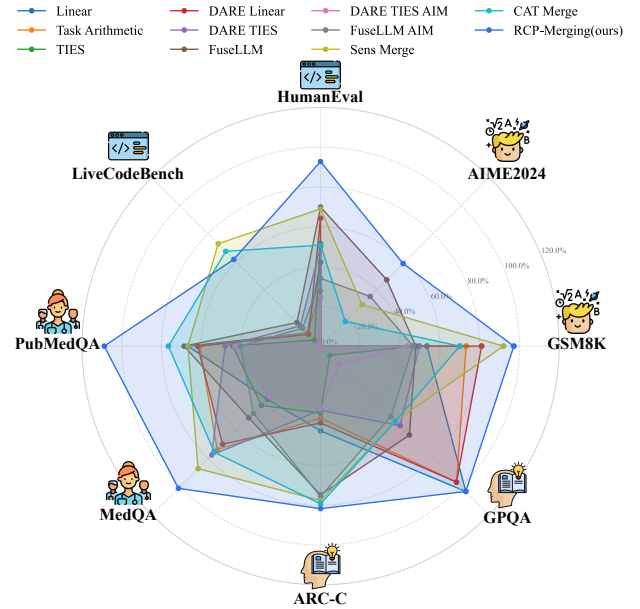


Figure 1: Performance comparison of RCP-Merging and other methods in merging Qwen2.5-7B, Meditron3-Qwen2.5-7B, and DeepSeek-R1-Distill-Qwen-7B on eight datasets in Math, Code, BioMedicine, and Knowledge areas.

difficulty stems from current long CoT realization relying on additional training, which introduces challenges like catastrophic forgetting, inefficient resource allocation, not to mention the inherent difficulty in obtaining high-quality domain reasoning data (Dong et al. 2025; Zhang et al. 2025; Zeng et al. 2025).

Fortunately, model merging (Li et al. 2023; Ilharco et al. 2023; Yang et al. 2024a) has recently emerged as a resource-efficient technique to create a single model with dual capabilities without requiring extra training data. However, a significant gap exists that current model merging focuses on combining models for certain domains, such as merging a model specialized in General Knowledge with one for Chinese. As results show in the LiveCodeBench (Jain et al. 2024) and AIME (Veeraboina 2023) datasets in Figure 1,

trying to merge a reasoning model with a domain-specific one often leads to a collapse of the output and a sharp performance decline. Therefore, it is highly valuable to find a method that can successfully integrate a domain-specific model with a reasoning model and subsequently boost the merged model’s performance on its original domain-specific tasks.

To tackle this problem, existing merging methods often struggle to preserve long CoT capabilities when integrating reasoning models with domain-specific ones. For instance, some methods (Ilharco et al. 2023; Wan et al. 2024) operate under the assumption that larger weights are more important. By trimming the smaller weights (Yadav et al. 2023) or rescaling the larger weights (Yu et al. 2024), these methods create significant risks as the large-magnitude weights from a domain-specific model can easily overwrite the smaller, yet more critical weights for long CoT capability. Other works (Liu et al. 2025; Nobari et al. 2025) utilize the product of weight magnitude and its gradient on a certain domain to identify how crucial the model weight is. Some do this by identifying key neurons to preserve crucial knowledge (Ma et al. 2025) while others resolve knowledge conflicts before merging (Sun et al. 2025a,b). However, domain-specific gradient is not a suitable proxy for long CoT, as they often track performance adjustments on certain domains instead of the multi-step reasoning deduction that is crucial for long CoT capability. These superficial gradients make it challenging to identify and preserve the specific weights that are essential to long CoT capability (Thapa et al. 2025; Hao et al. 2025; Zeng et al. 2025). Consequently, merged models through these methods inadvertently compromise long CoT capability. Moreover, as shown in Figure 3, these models lead to the generation of non-sensical gibberish outputs, highlighting the primary challenge of improving performance in a specific domain without sacrificing long CoT capability.

Motivated by this objective, we propose our core method: **Merging Long Chain-of-Thought Models with Domain-Specific Models by Considering Reasoning Capability as Prior (RCP-Merging)**. RCP-Merging is a framework designed to equip a domain-specific model with long CoT capability by merging with a reasoning model and further boosting the merged model’s performance on its original domain-specific tasks. The cornerstone of our method is the **Reasoning Preservation Indicator**. Instead of relying on conventional methods focusing on the gradient of loss on a certain domain and the magnitude of model weight, our method treats the model’s long CoT capabilities as a guiding principle for the merge. It conceptually views reasoning model’s parameters as a stable prior, constraining updates that would significantly deviate from this established reasoning foundation using the Fisher Information Matrix (Fisher 1925) gained from each calibration data. This ensures that as the model acquires new domain-specific knowledge, it is given an indicator for each model weight to ensure the merged weight does not greatly harm the long CoT capability, consistently preventing catastrophic forgetting, gibberish output, and long CoT capability degradation that emerged from previous methods. Our framework complements this with **Domain Knowledge Sensitivity** to identify and re-

tain important domain-specific weights. Finally, **Reasoning-preserved Merging** step synthesizes these factors, utilizing both the reasoning preservation matrix and domain sensitivity as a comprehensive guide to select the most critical parameters for the final model, achieving a robust balance between domain-specific knowledge and long CoT capability.

We demonstrate that RCP-Merging, requiring only a small number of open-source calibration samples, can effectively integrate long CoT reasoning capabilities into a domain-specific model. Through extensive experiments across various tasks and model architectures like Qwen2.5 (Qwen 2024) and Llama3.1 (Grattafiori et al. 2024), our method consistently produces merged models that not only preserve domain-specific expertise but also exhibit surprisingly long CoT capabilities when addressing domain-specific questions, ultimately elevating their performance in certain domains. Notably, the average performance of the merged model on eight datasets improves by 9.5% and 9.2% compared with the state-of-the-art method on BioMedicine and Finance domains, respectively. Moreover, though model merging aims to find a comprehensive model that compromises the performance of original models, our method improved performance by 4.5% and 0.7% on PubMedQA and MedQA datasets (Jin et al. 2019, 2020), respectively, and improved performance by 0.5% on ConvFinQA dataset (Cheng, Huang, and Wei 2024) compared to the original domain-specific models. To sum up, our contributions include:

- We propose a novel model merging framework, RCP-Merging, which effectively integrates a domain-specific model with a long CoT reasoning model by treating reasoning ability as a prior.
- We conduct extensive experiments across multiple benchmarks, demonstrating that RCP-Merging surpasses existing methods by preserving both specialized knowledge and long-CoT reasoning capabilities.
- Results surprisingly demonstrate that models merged via RCP-Merging exhibit emergent long CoT reasoning capabilities within model outputs when handling domain-specific problems.

Related Work

Model merging (Goddard et al. 2024a; Yang et al. 2024a; Ruan et al. 2025; Li et al. 2023; Lu et al. 2024) aims to combine multiple specialized models into a single, powerful model without costly retraining (Ilharco et al. 2023; Yadav et al. 2023; Yang et al. 2024b; Alexandrov et al. 2024). Existing approaches can be broadly categorized based on the information they use to determine how parameters are combined: magnitude-based methods that operate directly on parameter values, and activation-based methods that leverage model outputs or gradients on calibration data.

Magnitude-Based Methods

Magnitude-based methods merge models by performing arithmetic operations directly on their weight parameters or

task vectors, often using parameter magnitude as a proxy for importance.

A foundational approach is simple Linear or weight averaging, which calculates the element-wise mean of the parameters of all models to be merged (Izmailov et al. 2018; Wortsman et al. 2022). Task Arithmetic (Ilharco et al. 2023) refines this by first computing task vectors, defined as the difference between fine-tuned and pre-trained weights ($\delta_{ft} = \theta_{ft} - \theta_{pre}$). These vectors, representing task-specific knowledge, are then combined through arithmetic operations like addition or negation before being applied to the base model.

To mitigate interference between task vectors, several methods have been proposed. TIES-Merging (Yadav et al. 2023) introduces a three-step process: it trims each task vector by retaining only a top-k of high-magnitude parameters and resetting the rest to zero, then elects a single, dominant sign for each parameter across all task vectors. DARE (Yu et al. 2024) and PCB-Merging (Du et al. 2024) adjust model weights to reduce task conflicts by randomly dropping a ratio of weights and rescaling the remaining ones. FuseLLM (Wan et al. 2024) operates by leveraging the generative probability distributions of diverse source LLMs to externalize their knowledge, which is then transferred to a single target model through a lightweight continual training phase.

A primary drawback of magnitude-based methods is their assumption that parameter magnitude equates to importance. This can lead to the retention of high-magnitude parameters that are harmful to other models, causing significant knowledge conflicts and degrading the performance of the merged model.

Activation-Based Methods

To address the limitations of magnitude-based approaches, activation-based methods leverage data-driven signals, such as model activations or gradients on a small calibration set, to obtain a more nuanced understanding of parameter importance (Springenberg et al. 2015; Shrikumar, Greenside, and Kundaje 2017; Sundararajan, Taly, and Yan 2017; Michel, Levy, and Neubig 2019; Maini et al. 2023; Wang et al. 2023; Liu et al. 2024).

Sens-Merging (Liu et al. 2025) operates at two levels to perform task-specific analysis to identify the sensitivity of each layer and evaluate cross-task transferability between different models on a calibration dataset. CAT-Merging (Sun et al. 2025a) directly tackles knowledge conflict (Sun et al. 2025b) by identifying and trimming conflict-prone components from task vectors. Using a few unlabeled examples, it computes layer-specific projection operators for linear weights and masks for normalization parameters to resolve interference before merging.

Moreover, Fisher Merging (Matena and Raffel 2022) and RegMean (Jin et al. 2023) using Fisher Information Matrix to determine parameter importance or utilizing local regression for model merging; however, these approaches are characterized by high computational complexity. Other methods, such as Activation-Informed Merging (AIM) (Nobari et al. 2025) and LED-Merging (Ma et al. 2025) utilize activations

to guide the merging process, offering ways to find neurons that are crucial to certain domains.

While these activation-based methods can more effectively mitigate the knowledge conflicts seen in magnitude-based approaches, they have their own limitations since the gradient-based evaluation is hard to capture the complex, sequential reasoning patterns within the model’s weight.

Preliminary

Task Vector. We adopt the concept of task vectors from the field of model merging. A task vector, δ , represents the knowledge acquired by a model during fine-tuning for a specific task. It is computed as the difference between the weights of the fine-tuned model and base model, θ_t , where t represents the domain-specific task. The weights of the original pre-trained base model is represented by θ_{pre} :

$$\delta_t = \theta_t - \theta_{pre}, \text{ for } t \in \{1, \dots, T\}. \quad (1)$$

In our framework, we define a task vector for each domain-specific model, $\delta_t = \theta_t - \theta_{pre}$, and a task vector for the reasoning model, $\delta_r = \theta_r - \theta_{pre}$, where θ_t and θ_r are the weights of the domain-specialized model and the long-chain reasoning model, respectively. Task vector-based merging combines these task vectors into a single, static model:

$$\theta_{merged} = \theta_{pre} + \sum_{t=1}^T \lambda \cdot \delta_t, \quad (2)$$

where the coefficient λ represents the importance of each merged task vector.

Fisher Information Matrix. The Fisher Information Matrix (FIM) is a fundamental concept in information geometry that quantifies the amount of information an observable random variable, x , carries about an unknown parameter, θ , of a statistical model. For a model with parameters θ , the FIM element $F(\theta)_{ij}$ is defined as the expected value of the outer product of the gradients of the log-likelihood function, the (i, j) -th element of the matrix can be denoted as:

$$F(\theta)_{ij} = E_{x \sim p(x|\theta)} \left[\left(\frac{\partial}{\partial \theta_i} \log p(x|\theta) \right) \left(\frac{\partial}{\partial \theta_j} \log p(x|\theta) \right) \right]. \quad (3)$$

This can also be expressed as the negative expected value of the Hessian of the log-likelihood:

$$F(\theta)_{ij} = -E_{x \sim p(x|\theta)} \left[\frac{\partial^2}{\partial \theta_i \partial \theta_j} \log p(x|\theta) \right]. \quad (4)$$

In the context of autoregressive tasks where the loss function $\mathcal{L}(\theta, x)$ is the negative log-likelihood, i.e., $\mathcal{L}(\theta, x) = -\log p(x|\theta)$, the diagonal elements F_i of the FIM can be approximated by the expected squared gradient of the loss function. For a single i -th diagonal parameter θ_i and a dataset D_r , this approximation is:

$$F(\theta)_i \approx E_{d \sim D_r} \left[\left(\frac{\partial \mathcal{L}(\theta, d)}{\partial \theta_i} \right)^2 \right] = E_{d \sim D_r} [(g_{i,d})^2], \quad (5)$$

where $g_{i,d}^r$ is the gradient of the loss with respect to the parameter θ_i^r for a given data sample d . This approximation is pivotal for calculating our reasoning capability indicator.

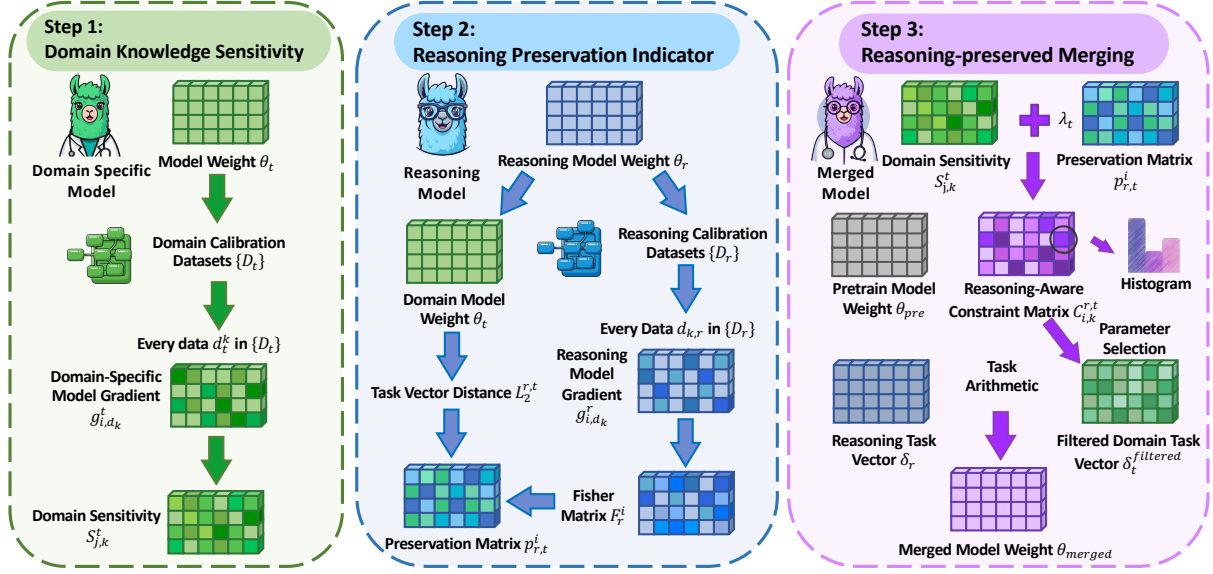


Figure 2: RCP-Merging consists of three stages. (1) **Domain Knowledge Sensitivity**. This step quantifies each weight’s importance for a specific domain by measuring the change in model loss when that weight is removed. (2) **Reasoning Preservation Indicator**. To protect the model’s core reasoning capabilities, this stage applies a preservation term to weights that are crucial for reasoning. (3) **Reasoning-preserved Merging**. The final stage balances domain sensitivity and the reasoning preserving matrix, merging only the weights that enhance domain knowledge without harming reasoning capabilities.

Methodology

Our methodology is designed to merge models by integrating domain-specific knowledge while preserving long CoT capability. This is achieved by first identifying parameters crucial for domain-specific tasks and then applying a preservation term derived from the Bayesian rule to mitigate the degradation of reasoning abilities. The final model is constructed by selectively merging domain-specific task vectors based on a reasoning-aware constraint matrix, as shown in Figure 2.

Domain Knowledge Sensitivity

To quantify the importance of each parameter on the domain-specific model for a task t , by setting the corresponding model as the domain-specific task model θ_t , we introduce the concept of Domain Knowledge Sensitivity, $S_{i,k}^t$. This metric measures the impact on the model’s performance when a particular weight is nullified.

Given a domain-specific model with parameters $\theta_t = [\theta_1, \theta_2, \dots, \theta_N]$ and a calibration dataset $\{D_t\}$, the sensitivity of the i -th parameter θ_t^i with respect to a data sample $d_k^t \in D_t$ is defined as the change in the loss function:

$$S_{i,k}^t = [\mathcal{L}(\theta_t) - \mathcal{L}(\theta_t - \theta_t^i)]_{d=d_k^t}, \quad (6)$$

where θ_t^i is a vector with only the i -th parameter being non-zero.

For computational efficiency, we approximate this value using a first-order Taylor expansion. This simplifies sensitivity to the product of the parameter and its corresponding

gradient, $g_{i,d_k}^t = \frac{\partial \mathcal{L}(\theta_t)}{\partial \theta_t^i}$, as follows:

$$S_{i,k}^t \approx \left\| \frac{\partial \mathcal{L}(\theta_t)}{\partial \theta_t^i} \cdot \theta_t^i \right\|_{d=d_k^t} \quad (7)$$

$$\approx \|g_{i,d_k}^t \cdot \theta_t^i\|_{d=d_k^t}.$$

A lower sensitivity score indicates that the parameter θ_t^i contributes positively to the model’s performance in the specific domain, as its presence reduces the loss.

Reasoning Preservation Indicator

To prevent the primary drawback of output collapse emerging from previous methods when merging with reasoning models, we introduce a preserving function to indicate important weights in the model merging process. Inspired by Kirkpatrick et al. (2016), we adopt the Bayesian rule where the reasoning model’s parameter distribution serves as a prior for the posterior distribution of the final merged model’s parameters. This approach constrains the weights to remain close to values crucial for reasoning, a detailed derivation is available in Appendix A. Our goal is to find the parameters θ_t that maximize the posterior probability (MAP estimation), which is equivalent to minimizing the negative log-posterior:

$$\theta_{MAP} = \arg \min_{\theta_t} [-\log P(D_t|\theta_t) - \log P(\theta_t|D_r)]. \quad (8)$$

The term $-\log P(\theta_t|D_r)$ acts as a regularization term, discouraging the parameters from deviating significantly from the optimal weights learned on the reasoning task, which we denote as θ_r^* .

However, directly computing the true posterior $P(\theta_t|D_r)$ is intractable for complex neural networks. To address this, we employ the Laplace approximation, which approximates the posterior with a Gaussian distribution centered at the mode θ_r^* : $P(\theta_t|D_r) \approx \mathcal{N}(\theta_t|\theta_r^*, F_r^{-1})$. The precision matrix of this Gaussian is the Fisher Information Matrix (FIM), F_r , which measures the curvature of the log-likelihood landscape. The probability density function is:

$$P(\theta_t|D_r) \approx \frac{|F_r|^{1/2}}{(2\pi)^{k/2}} \exp\left(-\frac{1}{2}(\theta_t - \theta_r^*)^T F_r (\theta_t - \theta_r^*)\right) \quad (9)$$

By taking the natural logarithm and discarding terms that are constant with respect to θ_t , we simplify the expression for optimization purposes. This yields a tractable form for the log-posterior preservation matrix:

$$\log P(\theta_t|D_r) \approx -\frac{1}{2}(\theta_t - \theta_r^*)^T F_r (\theta_t - \theta_r^*). \quad (10)$$

This quadratic term measures how much the updated parameters θ_t have diverged from the reasoning-optimal parameters θ_r^* , weighted by the FIM F_r . A higher value in F_r for a certain parameter indicates its importance for the reasoning task, and thus incurs a larger preservation for any deviation.

To make this computation more feasible, we assume a diagonal FIM. As shown in Equation 5, this simplifies $p_{r,t}^i$ into a sum of per-parameter contributions, where for each parameter θ_t^i , the penalty is $\log P(\theta_t^i|D_r) \approx -\frac{1}{2}F_{r,ii}(\theta_t^i - \theta_{r,i}^*)^2$. The i -th diagonal elements of the FIM, $F_{r,ii}$, can be approximated by the average of the squared gradients over the calibration reasoning dataset $D_r = \{d_k\}_{k=1}^{N_r}$. Combining these steps, we define the final reasoning preservation indicating matrix $p_{r,t}^i$ for each parameter θ_t^i as:

$$p_{r,t}^i = \left\| -\frac{1}{2N_r} \sum_{k=1}^{N_r} (g_{i,d_k}^r)^2 (\theta_t^i - \theta_{r,i}^*)^2 \right\|. \quad (11)$$

Here, g_{i,d_k}^r is the gradient of the loss for sample d_k with respect to parameter θ_r^i . This metrics quantifies how much the new parameter θ_t^i impairs the model's reasoning ability.

Reasoning-preserved Merging

To integrate domain knowledge while preserving core reasoning skills, we propose reasoning-aware merging strategy. We implement this by defining a Constraint metric $C_{i,k}^{r,t}$ for each parameter θ_i to quantify the importance of long CoT capability, combining its Domain Knowledge Sensitivity ($S_{i,k}^t$) and Reasoning Capability Indicator ($p_{r,t}^i$):

$$C_{i,k}^{r,t} = S_{i,k}^t + \lambda_r \cdot p_{r,t}^i. \quad (12)$$

Here, the hyperparameter λ_r balances the trade-off between domain performance and long CoT capability preservation.

Next, we filter parameter updates using a majority vote criterion. An update for parameter θ_t^i is accepted if more data samples in the domain dataset D_t yield a negative conflict score than a non-negative one:

$$\text{Accept update for } \theta_t^i \quad \text{if} \quad N(C_{i,k}^{r,t} < 0) > N(C_{i,k}^{r,t} \geq 0). \quad (13)$$

This condition generates a binary mask $M \in \{0, 1\}^N$, where $M_i = 1$ signifies an accepted update for the corresponding parameter.

Finally, we use this mask to create a filtered domain-specific task vector, $\delta_t^{filtered}$, via an element-wise product with the original task vector $\delta_t = \theta_t - \theta_{pre}$. The final model weights, θ_{merged} , are then obtained by adding the complete reasoning vector δ_r and the weighted sum of these filtered task vectors to the pre-trained weights θ_{pre} :

$$\delta_t^{filtered} = M \odot \delta_t, \quad (14)$$

$$\theta_{merged} = \theta_{pre} + \delta_r + \sum_{t=1}^T \lambda_t \cdot \delta_t^{filtered}, \quad (15)$$

where T is the number of domain-specific tasks and λ_t are scaling coefficients. This approach ensures the model benefits from domain-specific knowledge while robustly maintaining its reasoning abilities.

Experiment

Experimental Setup

Baselines. We compare RCP-Merging with multiple merging baselines: **Average** (Izmailov et al. 2018), **Task Arithmetic** (Ilharco et al. 2023), **TIES-Merging** (Yadav et al. 2023), **DARE-Merging**, **DARE-Merging with TIES** (Yu et al. 2024), **FuseLLM** (Wan et al. 2024), **FuseLLM with AIM**, **DARE TIES with AIM** (Nobari et al. 2025), **Sens-Merging** (Liu et al. 2025), and **CAT-Merging** (Sun et al. 2025a). We utilize mergekit (Goddard et al. 2024b) as merging tools for baseline methods, detailed discussions and recommended hyperparameters are listed in Appendix B.1 and B.2.

Datasets&Metrics. We assess merged model performance through four pillars: (1) Mathematical reasoning (Math) via GSM8k (Cobbe et al. 2021) and AIME2024 (Veeraboina 2023) (Accuracy↑ with CoT); (2) Code generation (Code) evaluated by HumanEval (Chen et al. 2021) and LiveCodeBench (Jain et al. 2024) (Pass@1↑); (3) Medical question answering (BioMedicine) through PubMedQA (Jin et al. 2019) and MedQA (Jin et al. 2020) (Accuracy↑); (4) General knowledge question answering with ARC-C (Clark et al. 2018) and GPQA (Rein et al. 2023) (Accuracy↑). More details of datasets can be found in Appendix B.3.

Models. The experiment involves a set of models built upon the Qwen2.5-7B (Qwen 2024) Base model architecture. The domain-specific model is Meditron3-Qwen2.5-7B (Chen et al. 2023) for BioMedicine, and the Reasoning model is DeepSeek-R1-Distill-Qwen-7B (Guo et al. 2025).

RCP-Merging's Superior Performance

RCP-Merging achieves SOTA average performance on BioMedicine domain, surpassing all existing merging methods and even the original BioMedicine model meanwhile maintaining long CoT capability. The results, summarized in Table 1, demonstrate that RCP-Merging achieves a superior balance between domain-specific expertise and reasoning capabilities. It obtains the highest average score of

Method/Task	Math		Code		BioMedicine		Knowledge		Average
	GSM8K	AIME2024	HumanEval	LiveCodeBench	PubMedQA	MedQA	ARC-C	GPQA	
Base	69.4	0.0	50.6	12.4	32.5	22.9	60.9	7.6	32.0
BioMedicine	81.5	0.0	54.3	2.2	51.0	53.5	74.9	9.6	40.9
Reasoning	86.7	56.7	76.6	29.8	38.0	30.2	76.5	15.2	51.2
Linear	46.4	0.0	32.3	2.8	34.0	20.2	32.6	15.7	23.0
Task Arithmetic	63.5	0.0	21.3	2.8	31.0	39.3	27.7	14.7	25.0
TIES-Merging	40.6	0.0	39.6	1.3	22.5	22.5	25.8	1.0	19.2
DARE Linear	70.2	0.0	49.4	2.5	31.5	37.3	29.5	14.7	29.4
DARE TIES	43.1	0.0	38.4	4.3	24.5	41.4	24.6	8.6	23.1
FuseLLM	41.5	26.7	53.7	5.0	35.0	27.3	57.2	9.6	32.0
DARE TIES & AIM	37.8	0.0	18.9	0.4	22.0	19.3	24.2	2.0	15.6
FuseLLM & AIM	40.3	20.0	26.2	3.8	20.5	25.5	57.8	7.6	25.2
Sens-Merging	79.8	16.7	53.0	21.7	34.0	46.6	59.4	8.1	39.9
CAT-Merging	60.7	10.0	39.0	20.1	39.0	40.4	60.7	8.1	34.8
RCP-Merging	84.3	33.3	71.3	18.4	55.5	54.1	82.5	15.7	49.4

Table 1: Performance comparison of merging Qwen2.5-7B (Base), Meditron3-Qwen2.5-7B (BioMedicine) and DeepSeek-R1-Distill-Qwen-7B (Reasoning) on all datasets across Math, Code, BioMedicine, and Knowledge areas. The best performance among all merging methods on each dataset is highlighted in **bold**.

49.4 across all benchmarks, significantly surpassing all other merging methods. Specifically, in the target BioMedicine domain, RCP-Merging achieves top scores on both PubMedQA and MedQA with scores of 55.5 and 54.1, effectively integrating the knowledge from the BioMedicine model.

Simultaneously, it not only preserves but also enhances the reasoning abilities inherited from the reasoning model, leading to state-of-the-art performance among the merged models in Math with 84.3 on GSM8K, 33.3 on AIME2024 and Code shown as the performance of 71.3 on HumanEval. In contrast, most baseline methods struggle to maintain this balance, sacrificing reasoning for domain performance. This highlights RCP-Merging’s unique effectiveness in creating a truly versatile and capable model. We also utilizing the merged model’s output content length to represent the reasoning thinking process of the merged models, the detailed analysis and results are shown in Table 6 in Appendix. To further analysis the merged model’s reasoning process, we also conduct case study in Appendix.

RCP-Merging’s Output Stability. To address the emerging issue of model stability in merging methods, we measure the “gibberish rate”: the frequency of nonsensical outputs identified by a GPT-4 based evaluator (OpenAI 2023). This metric validates whether benchmark performance is genuine or from output degeneration.

As shown in Figure 3, RCP-Merging demonstrates superior stability, achieving the lowest average gibberish rate of 14.3%, with rates of 0% on PubMedQA and 5.8% on MedQA. In stark contrast, baseline methods like TIES and DARE TIES & AIM suffer from output collapse, showing high gibberish rates of 82.3% and 79.5% respectively. This confirms RCP-Merging’s robust performance stems from a genuine integration of capabilities. Moreover, we further evaluate the stability of RCP-Merging’s output using Distinct-N and PPL metrics, results are shown in Table 5 in Appendix.

Different Domain-specific Task

To verify the generalizability of our method across different domains, we conduct experiments where specific domain is shifted from BioMedicine to finance. In this setup, we merge WiroAI-Finance-Qwen-7B as Finance model with the same Base and Reasoning models.

RCP-Merging demonstrates top performance across all evaluated categories, including GSM8k in Math, HumanEval in Code, ARC-C in Knowledge, and ConvFinQA in Finance (Cheng, Huang, and Wei 2024). Shown in Table 2, results demonstrate that RCP-Merging achieves the highest average score of 72.2, decisively outperforming all baseline methods on Finance domain. As the performance shown in the table, RCP-Merging demonstrates the best performance across four benchmarks, this further verifies the scalability of RCP-Merging in different fields. Results show RCP-Merging balances domain-specific performance and long CoT capability across multiple domains.

Method/Task	Math	Code	Finance	Knowledge	Average
Base	69.4	50.6	50.3	60.9	57.8
Finance	50.2	1.2	58.7	47.9	39.5
Reasoning	86.7	76.8	36.2	76.5	69.1
Linear	16.6	32.3	34.0	27.7	27.7
Task Arithmetic	8.4	39.6	17.4	43.3	27.2
TIES-Merging	7.2	21.3	18.8	42.0	22.3
DARE Linear	8.4	49.4	17.7	43.1	29.7
DARE TIES	7.6	38.4	18.4	43.6	27.0
FuseLLM	7.4	53.7	18.4	42.7	30.6
DARE TIES & AIM	6.4	18.9	19.7	46.5	22.9
FuseLLM & AIM	5.3	26.2	20.4	47.1	24.8
Sens-Merging	60.7	53.7	4.2	25.8	36.1
CAT-Merging	60.7	39.0	10.1	24.8	33.7
RCP-Merging	82.0	71.3	59.2	76.4	72.2

Table 2: Performance comparison of merging Qwen2.5-7B (Base), WiroAI-Finance-Qwen-7B (Finance) and DeepSeek-R1-Distill-Qwen-7B (Reasoning) on four datasets across Math, Code, Finance and Knowledge areas. The best performance among all merging methods on each dataset is highlighted in **bold**.

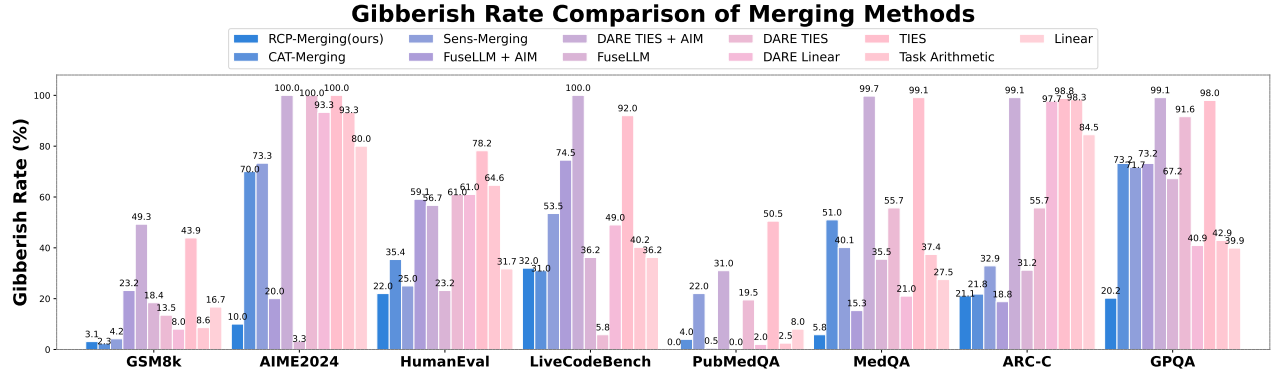


Figure 3: Gibberish rate comparison of merging Qwen2.5-7B (Base), Meditron3-Qwen2.5-7B (BioMedicine) and DeepSeek-R1-Distill-Qwen-7B (Reasoning) on all datasets. The gibberish rate is the percentage of gibberish or nonsensical content generated for a given benchmark, and a lower value means higher-quality content.

Different Model Architecture.

RCP-Merging demonstrates consistent performance across different architectures. We have verified this by conducting experiments on the Llama3.1-8B based models, which is distinct from our primary setup. In this alternative configuration, we used Llama3.1-8B (Grattafiori et al. 2024) as the Base model, Llama3-OpenBioLLM-8B (Ankit Pal 2024) as the BioMedicine model, and DeepSeek-R1-Distill-Llama-8B (Guo et al. 2025) as the Reasoning model.

We use GSM8k, HumanEval, ARC-C and PubMedQA to indicate the performance of different merge methods on Math, Code, Knowledge and BioMedicine domain. As the results in Table 3, RCP-Merging achieves the best average score of 68.3 among all merging techniques. Although FuseLLM with AIM shows a slightly better score in the specific BioMedicine domain, RCP-Merging has the best overall capability. We evaluate our method on different model size, with results shown in Table 7 in Appendix.

Method/Task	Math	Code	BioMedicine	Knowledge	Average
Base	60.9	42.7	55.0	60.7	54.8
BioMedicine	39.4	37.8	58.0	56.0	47.8
Reasoning	68.8	89.6	51.5	84.0	73.5
Linear	3.2	37.2	31.0	59.0	32.6
Task Arithmetic	55.3	48.2	23.0	45.9	43.1
TIES-Merging	47.5	40.2	53.5	62.2	50.9
DARE Linear	58.3	40.2	23.0	45.9	41.9
DARE TIES	45.6	47.6	32.5	22.2	37.0
FuseLLM	48.8	61.0	55.5	53.3	54.7
DARE TIES & AIM	38.1	49.4	13.0	26.0	31.6
FuseLLM & AIM	56.1	59.8	57.5	59.3	58.2
Sens-Merging	65.7	46.3	55.5	65.5	58.3
CAT-Merging	62.5	55.5	54.0	64.3	59.1
RCP-Merging	67.2	73.2	57.0	75.8	68.3

Table 3: Performance comparison of merging Llama-3.1-8B (Base), Llama3-OpenBioLLM-8B (BioMedicine) and DeepSeek-R1-Distill-Llama-8B (Reasoning) on four datasets across Math, Code, BioMedicine, and Knowledge areas. The best performance under certain dataset is highlighted in **bold**.

Ablation Study

This section performs an ablation study to evaluate the effectiveness of the parameter-specific trimming techniques in RCP-Merging, including the pruning of Knowledge Sensitivity and Reasoning Preservation. As results shown in Table 4, excluding Domain Sensitivity (w/o Domain Sensitivity) causes the average score to drop significantly from 68.3 to 48.7. The effect is even more severe when removing the Reasoning Preservation (w/o Reasoning Preservation), which plunges the average score to 41.4. These results underscore that both trimming strategies are indispensable.

Method/Task	Math	Code	BioMedicine	Knowledge	Average
Base	60.9	42.7	55.0	60.7	54.8
BioMedicine	39.4	37.8	58.0	56.0	47.8
Reasoning	68.8	89.6	51.5	84.0	73.5
w/o Domain Sensitivity	58.4	56.1	33.0	47.4	48.7
w/o Reasoning Preservation	57.1	37.2	30.5	40.9	41.4
RCP-Merging	67.2	73.2	57.0	75.8	68.3

Table 4: Ablation Study. Performance comparison when merging Qwen2.5-7B (Base), Meditron3-Qwen2.5-7B (BioMedicine) and DeepSeek-R1-Distill-Qwen-7B (Reasoning) on four datasets across Math, Code, BioMedicine, and Knowledge areas. Performance The best performance under certain dataset is highlighted in **bold**.

Conclusion

We propose a novel model merging framework, RCP-Merging, which effectively integrates domain-specific models with long-chain-of-thought reasoning models by treating reasoning ability as a prior. Our method applies a reasoning capability penalty to preserve core reasoning parameters while selectively merging essential domain-specific weights. Notably, RCP-Merging enhances performance in the BioMedicine and Finance domains by 9.5% and 9.2% respectively, compared to state-of-the-art methods. Our approach creates powerful, unified models that excel in both domain-specific knowledge and general long-chain-of-thought reasoning, effectively addressing the challenge of balancing domain performance with reasoning capability.

References

- Alexandrov, A.; Raychev, V.; Müller, M. N.; Zhang, C.; Vechev, M.; and Toutanova, K. 2024. Mitigating Catastrophic Forgetting in Language Transfer via Model Merging. In Al-Onaizan, Y.; Bansal, M.; and Chen, Y.-N., eds., *Findings of the Association for Computational Linguistics: EMNLP 2024*, 17167–17186. Miami, Florida, USA: Association for Computational Linguistics.
- Ankit Pal, M. S. 2024. OpenBioLLMs: Advancing Open-Source Large Language Models for Healthcare and Life Sciences. <https://huggingface.co/aaditya/OpenBioLLM-Llama3-70B>.
- Chen, H.; Dong, Y.; Wei, Z.; Huang, Y.; Zhang, Y.; Su, H.; and Zhu, J. 2025. Understanding Pre-training and Fine-tuning from Loss Landscape Perspectives. *arXiv preprint arXiv:2505.17646*.
- Chen, M.; Tworek, J.; Jun, H.; Yuan, Q.; Pinto, H. P. D. O.; Kaplan, J.; Edwards, H.; Burda, Y.; Joseph, N.; Brockman, G.; et al. 2021. Evaluating Large Language Models Trained on Code. *CoRR*, abs/2107.03374.
- Chen, Z.; Hernández-Cano, A.; Romanou, A.; Bonnet, A.; Matoba, K.; Salvi, F.; Pagliardini, M.; Fan, S.; Köpf, A.; Mohtashami, A.; Sallinen, A.; Sakhaeirad, A.; Swamy, V.; Krawczuk, I.; Bayazit, D.; Marmet, A.; Montariol, S.; Hartley, M.; Jaggi, M.; and Bosselut, A. 2023. MEDITRON-70B: Scaling Medical Pretraining for Large Language Models. *CoRR*, abs/2311.16079.
- Cheng, D.; Huang, S.; and Wei, F. 2024. Adapting Large Language Models via Reading Comprehension. In *The Twelfth International Conference on Learning Representations*.
- Clark, P.; Cowhey, I.; Etzioni, O.; Khot, T.; Sabharwal, A.; Schoenick, C.; and Tafford, O. 2018. Think you have Solved Question Answering? Try ARC, the AI2 Reasoning Challenge. *CoRR*, abs/1803.05457.
- 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. *CoRR*, abs/2110.14168.
- Cui, H.; Shamsi, Z.; Cheon, G.; Ma, X.; Li, S.; Tikhanovskaya, M.; Norgaard, P. C.; Mudur, N.; Plomecka, M. B.; Raccuglia, P.; Bahri, Y.; Albert, V. V.; Srinivasan, P.; Pan, H.; Faist, P.; Rohr, B. A.; Statt, M. J.; Morris, D.; Purves, D.; Kleeman, E.; Alcantara, R.; Abraham, M.; Mohammad, M.; VanLee, E. P.; Jiang, C.; Dorfman, E.; Kim, E.-A.; Brenner, M.; Ponda, S. S.; and Venugopalan, S. 2025. CURIE: Evaluating LLMs on Multitask Scientific Long-Context Understanding and Reasoning. In *The Thirteenth International Conference on Learning Representations*.
- Dong, Z.; Li, J.; Jiang, J.; Xu, M.; Zhao, W. X.; Wang, B.; and Chen, W. 2025. LongReD: Mitigating Short-Text Degradation of Long-Context Large Language Models via Restoration Distillation. *CoRR*, abs/2502.07365.
- Du, G.; Lee, J.; Li, J.; Jiang, R.; Guo, Y.; Yu, S.; Liu, H.; Goh, S. K.; Tang, H.; He, D.; and Zhang, M. 2024. Parameter Competition Balancing for Model Merging. In Globerson, A.; Mackey, L.; Belgrave, D.; Fan, A.; Paquet, U.; Tomczak, J. M.; and Zhang, C., eds., *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.
- Fisher, R. A. 1925. Theory of statistical estimation. In *Mathematical proceedings of the Cambridge philosophical society*, volume 22, 700–725. Cambridge University Press.
- Goddard, C.; Siriwardhana, S.; Ehghaghi, M.; Meyers, L.; Karpukhin, V.; Benedict, B.; McQuade, M.; and Solawetz, J. 2024a. Arcee’s MergeKit: A Toolkit for Merging Large Language Models. In Dernoncourt, F.; Preoctiuc-Pietro, D.; and Shimorina, A., eds., *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, 477–485. Miami, Florida, US: Association for Computational Linguistics.
- Goddard, C.; Siriwardhana, S.; Ehghaghi, M.; Meyers, L.; Karpukhin, V.; Benedict, B.; McQuade, M.; and Solawetz, J. 2024b. Arcee’s MergeKit: A Toolkit for Merging Large Language Models. In Dernoncourt, F.; Preoctiuc-Pietro, D.; and Shimorina, A., eds., *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, 477–485. Miami, Florida, US: Association for Computational Linguistics.
- Grattafiori, A.; Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Vaughan, A.; et al. 2024. The Llama 3 Herd of Models. *CoRR*, abs/2407.21783.
- Guo, D.; Yang, D.; Zhang, H.; Song, J.; Zhang, R.; Xu, R.; Zhu, Q.; Ma, S.; Wang, P.; Bi, X.; et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.
- Hao, S.; Sukhbaatar, S.; Su, D.; Li, X.; Hu, Z.; Weston, J. E.; and Tian, Y. 2025. Training Large Language Model to Reason in a Continuous Latent Space.
- Hu, Y.; Huang, Q.; Tao, M.; Zhang, C.; and Feng, Y. 2024. Can Perplexity Reflect Large Language Model’s Ability in Long Text Understanding? In *The Second Tiny Papers Track at ICLR 2024*.
- Ilharco, G.; Ribeiro, M. T.; Wortsman, M.; Schmidt, L.; Hashizirzi, H.; and Farhadi, A. 2023. Editing models with task arithmetic. In *The Eleventh International Conference on Learning Representations*.
- Izmailov, P.; Podoprikin, D.; Garipov, T.; Vetrov, D. P.; and Wilson, A. G. 2018. Averaging Weights Leads to Wider Optima and Better Generalization. In Globerson, A.; and Silva, R., eds., *Proceedings of the Thirty-Fourth Conference on Uncertainty in Artificial Intelligence, UAI 2018, Monterey, California, USA, August 6-10, 2018*, 876–885. AUAI Press.
- Jaech, A.; Kalai, A.; Lerer, A.; Richardson, A.; El-Kishky, A.; Low, A.; Helyar, A.; Madry, A.; Beutel, A.; Carney, A.; et al. 2024. Openai o1 system card. *arXiv preprint arXiv:2412.16720*.
- Jain, N.; Han, K.; Gu, A.; Li, W.-D.; Yan, F.; Zhang, T.; Wang, S. I.; Solar-Lezama, A.; Sen, K.; and Stoica, I. 2024. LiveCodeBench: Holistic and Contamination Free Evaluation of Large Language Models for Code. *arXiv preprint arXiv:2403.07974*.

- Jin, D.; Pan, E.; Oufattole, N.; Weng, W.; Fang, H.; and Szolovits, P. 2020. What Disease does this Patient Have? A Large-scale Open Domain Question Answering Dataset from Medical Exams. *CoRR*, abs/2009.13081.
- Jin, Q.; Dhingra, B.; Liu, Z.; Cohen, W. W.; and Lu, X. 2019. PubMedQA: A Dataset for Biomedical Research Question Answering. In Inui, K.; Jiang, J.; Ng, V.; and Wan, X., eds., *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, 2567–2577. Association for Computational Linguistics.
- Jin, X.; Ren, X.; Preotiuc-Pietro, D.; and Cheng, P. 2023. Dataless Knowledge Fusion by Merging Weights of Language Models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.
- Kaplan, J.; McCandlish, S.; Henighan, T.; Brown, T. B.; Chess, B.; Child, R.; Gray, S.; Radford, A.; Wu, J.; and Amodei, D. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.
- Kirkpatrick, J.; Pascanu, R.; Rabinowitz, N. C.; Veness, J.; Desjardins, G.; Rusu, A. A.; Milan, K.; Quan, J.; Ramalho, T.; Grabska-Barwinska, A.; Hassabis, D.; Clopath, C.; Kumaran, D.; and Hadsell, R. 2016. Overcoming catastrophic forgetting in neural networks. *CoRR*, abs/1612.00796.
- Li, J.; Galley, M.; Brockett, C.; Gao, J.; and Dolan, B. 2015. A diversity-promoting objective function for neural conversation models. *arXiv preprint arXiv:1510.03055*.
- Li, W.; Peng, Y.; Zhang, M.; Ding, L.; Hu, H.; and Shen, L. 2023. Deep Model Fusion: A Survey. *CoRR*, abs/2309.15698.
- Li, Y.; Yan, Y.; Tong, Z.; Wang, Y.; Yang, Y.; Bai, M.; Pu, D.; Xie, J.; Liu, C.; Li, B.; Liu, M.; and Shu, K. 2025. Efficient fine-tuning of small-parameter large language models for biomedical bilingual multi-task applications. *Appl. Soft Comput.*, 175: 113084.
- Liu, S.; Wu, H.; He, B.; Han, X.; Yuan, M.; and Song, L. 2025. Sens-Merging: Sensitivity-Guided Parameter Balancing for Merging Large Language Models. *CoRR*, abs/2502.12420.
- Liu, Y.; Liu, Y.; Chen, X.; Chen, P.-Y.; Zan, D.; Kan, M.-Y.; and Ho, T.-Y. 2024. The devil is in the neurons: Interpreting and mitigating social biases in language models. In *The Twelfth International Conference on Learning Representations*.
- Lu, J.; Pang, Z.; Xiao, M.; Zhu, Y.; Xia, R.; and Zhang, J. 2024. Merge, Ensemble, and Cooperate! A Survey on Collaborative Strategies in the Era of Large Language Models. *CoRR*, abs/2407.06089.
- Ma, Q.; Liu, D.; Qian, C.; Zhang, L.; and Shao, J. 2025. LED-Merging: Mitigating Safety-Utility Conflicts in Model Merging with Location-Election-Disjoint. *CoRR*.
- Maini, P.; Mozer, M. C.; Sedghi, H.; Lipton, Z. C.; Kolter, J. Z.; and Zhang, C. 2023. Can Neural Network Memorization Be Localized? In *ICML*.
- Matena, M.; and Raffel, C. 2022. Merging Models with Fisher-Weighted Averaging. In Koyejo, S.; Mohamed, S.; Agarwal, A.; Belgrave, D.; Cho, K.; and Oh, A., eds., *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.
- Michel, P.; Levy, O.; and Neubig, G. 2019. Are sixteen heads really better than one? *Advances in neural information processing systems*, 32.
- Nobari, A. H.; Alimohammadi, K.; ArjomandBigdeli, A.; Srivastava, A.; Ahmed, F.; and Azizan, N. 2025. Activation-Informed Merging of Large Language Models. *CoRR*, abs/2502.02421.
- OpenAI. 2023. GPT-4 Technical Report. *CoRR*, abs/2303.08774.
- OpenAI. 2025. Openai o3 and o4-mini system card.
- Qwen, T. 2024. Qwen2.5 Technical Report. *CoRR*, abs/2412.15115.
- Rein, D.; Hou, B. L.; Stickland, A. C.; Petty, J.; Pang, R. Y.; Dirani, J.; Michael, J.; and Bowman, S. R. 2023. GPQA: A Graduate-Level Google-Proof Q&A Benchmark. *arXiv preprint arXiv:2311.12022*.
- Ruan, W.; Yang, T.; Zhou, Y.; Liu, T.; and Lu, J. 2025. From Task-Specific Models to Unified Systems: A Review of Model Merging Approaches. *CoRR*, abs/2503.08998.
- Shrikumar, A.; Greenside, P.; and Kundaje, A. 2017. Learning Important Features Through Propagating Activation Differences. In *ICML*.
- Springenberg, J.; Dosovitskiy, A.; Brox, T.; and Riedmiller, M. 2015. Striving for Simplicity: The All Convolutional Net. In *ICLR (workshop track)*.
- Sun, W.; Li, Q.; Geng, Y.-a.; and Li, B. 2025a. Cat merging: A training-free approach for resolving conflicts in model merging. *arXiv preprint arXiv:2505.06977*.
- Sun, W.; Li, Q.; Wang, W.; ao Geng, Y.; and Li, B. 2025b. Task Arithmetic in Trust Region: A Training-Free Model Merging Approach to Navigate Knowledge Conflicts. *ArXiv*, abs/2501.15065.
- Sundararajan, M.; Taly, A.; and Yan, Q. 2017. Axiomatic Attribution for Deep Networks. In Precup, D.; and Teh, Y. W., eds., *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, 3319–3328. PMLR.
- Tang, X.; Wang, X.; Lv, Z.; Min, Y.; Zhao, W. X.; Hu, B.; Liu, Z.; and Zhang, Z. 2025. Unlocking General Long Chain-of-Thought Reasoning Capabilities of Large Language Models via Representation Engineering. *CoRR*, abs/2503.11314.
- Thapa, R.; Wu, Q.; Wu, K.; Zhang, H.; Zhang, A.; Wu, E.; Ye, H.; Bedi, S.; Aresh, N.; Boen, J.; Reddy, S.; Athiwaratkun, B.; Song, S. L.; and Zou, J. 2025. Disentangling Reasoning and Knowledge in Medical Large Language Models. *ArXiv*, abs/2505.11462.
- Veeraboina, H. 2023. AIME Problem Set 1983-2024.

- Wan, F.; Huang, X.; Cai, D.; Quan, X.; Bi, W.; and Shi, S. 2024. Knowledge Fusion of Large Language Models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.
- Wang, L.; Li, L.; Dai, D.; Chen, D.; Zhou, H.; Meng, F.; Zhou, J.; and Sun, X. 2023. Label Words are Anchors: An Information Flow Perspective for Understanding In-Context Learning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 9840–9855. Association for Computational Linguistics.
- 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 Chaudhuri, K.; Jegelka, S.; Song, L.; Szepesvári, C.; Niu, G.; and Sabato, S., eds., *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, 23965–23998. PMLR.
- xAI. 2025. Grok 3.5: Advanced Reasoning AI Model by xAI. <https://grok.x.ai/>. Accessed: 2025-05-15.
- Yadav, P.; Tam, D.; Choshen, L.; Raffel, C.; and Bansal, M. 2023. TIES-Merging: Resolving Interference When Merging Models. In *Thirty-seventh Conference on Neural Information Processing Systems*.
- Yang, E.; Shen, L.; Guo, G.; Wang, X.; Cao, X.; Zhang, J.; and Tao, D. 2024a. Model merging in llms, mllms, and beyond: Methods, theories, applications and opportunities. *arXiv preprint arXiv:2408.07666*.
- 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 International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 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. *arXiv:2311.03099*.
- Zeng, Z.; Cheng, Q.; Yin, Z.; Zhou, Y.; and Qiu, X. 2025. Revisiting the Test-Time Scaling of o1-like Models: Do they Truly Possess Test-Time Scaling Capabilities? *CoRR*, abs/2502.12215.
- Zhang, Y.; Khalifa, M.; Zhang, L.; Liu, X.; Lee, A.; Zhang, X. F.; Bayat, F. F.; and Wang, L. 2025. Logit Arithmetic Elicits Long Reasoning Capabilities Without Training. *arXiv preprint arXiv:2507.12759*.

Appendix

A: Bayesian Framework

This appendix provides a detailed derivation of the Bayesian framework used to formulate the Reasoning Capability Penalty. The central goal is to find an optimal set of parameters for a domain-specific model, denoted as θ_t , by leveraging information from two sources: the new domain-specific dataset, D_t , and the pre-existing knowledge from a model trained on a reasoning dataset, D_r .

Our objective is to estimate the posterior probability distribution $P(\theta_t|D_t, D_r)$. This term represents the probability of the parameters θ_t being optimal after we have observed both the domain data D_t and the reasoning data D_r . Maximizing this posterior probability allows us to find the most plausible parameter values.

The derivation begins with the product rule of probability, which allows us to express the joint probability $P(\theta_t, D_t|D_r)$ in two equivalent ways. First, by factoring out θ_t :

$$P(\theta_t, D_t|D_r) = P(D_t|\theta_t, D_r)P(\theta_t|D_r). \quad (16)$$

Alternatively, by factoring out D_t :

$$P(\theta_t, D_t|D_r) = P(\theta_t|D_t, D_r)P(D_t|D_r). \quad (17)$$

By equating these two expressions, we can solve for our target posterior distribution, $P(\theta_t|D_t, D_r)$:

$$P(\theta_t|D_t, D_r) = \frac{P(D_t|\theta_t, D_r)P(\theta_t|D_r)}{P(D_t|D_r)}. \quad (18)$$

In this formulation, the term $P(\theta_t|D_r)$ plays the crucial role of the prior distribution. It encapsulates our prior belief about the parameters θ_t before encountering the new domain data D_t . In our method, we define this prior as the posterior distribution of the parameters obtained after training on the reasoning dataset D_r . Thus, its function is to act as a regularizer, ensuring that the final parameters do not stray far from the values established as important for reasoning.

The term $P(D_t|\theta_t, D_r)$ is the likelihood, which measures how probable the new domain data D_t is for a given set of parameters θ_t . We apply a standard conditional independence assumption, stating that the generation of new data D_t depends only on the parameters θ_t , not on the old data D_r . This simplifies the likelihood to $P(D_t|\theta_t)$.

Finally, the denominator $P(D_t|D_r)$ is the marginal likelihood or evidence. It serves as a normalization constant to ensure the posterior is a valid probability distribution. Since it does not depend on the parameters θ_t that we are optimizing, it can be disregarded when our goal is to maximize the posterior.

Considering these points, for the purpose of optimization, the posterior probability is proportional to the product of the likelihood and the prior. This leads to the final relationship used in our methodology:

$$P(\theta_t|D_t, D_r) \propto P(D_t|\theta_t)P(\theta_t|D_r). \quad (19)$$

This proportionality forms the theoretical foundation for MAP (Maximum A Posteriori) estimation:

$$\theta_{MAP} = \arg \max_{\theta_t} [\log P(D_t|\theta_t) + \log P(\theta_t|D_r)], \quad (20)$$

where maximizing the posterior is equivalent to minimizing the negative of its logarithm, as shown in Equation 8.

B: Experiment Details

B.1 Model Merging Baselines

This section provides detailed descriptions of the model merging baselines used in our experiments. Each method is briefly explained, highlighting its core idea and relevant formulation.

- **Linear Averaging** (Izmailov et al. 2018): This basic method merges models by directly averaging their corresponding parameters.
- **Task Arithmetic** (Ilharco et al. 2023): Task Arithmetic combines task-specific knowledge by adding or subtracting parameter vectors. It computes a task vector as the difference between a fine-tuned model and its base model, then scales and adds this vector to another model.
- **TIES-Merging** (Yadav et al. 2023): TIES-Merging addresses parameter redundancy by identifying and merging significant parameters. It involves pruning, re-scaling, and merging parameter differences to combat interference.
- **DARE-Merging** (Yu et al. 2024): DARE (Drop and Restore) aims to mitigate catastrophic forgetting during merging by selectively dropping and then restoring parameters. It introduces a dropout mechanism on the parameter differences before merging.
- **FuseLLM** (Wan et al. 2024): FuseLLM proposes a method to merge large language models by aligning their activation spaces. It focuses on combining representations learned by different models rather than directly manipulating parameters.
- **AIM** (Nobari et al. 2025): AIM (Activation Informed Merging) provides a more sophisticated way to align and combine the activation patterns of different models for improved merged performance.
- **Sens-Merging** (Liu et al. 2025): Sens-Merging focuses on the sensitivity of model parameters to specific tasks. It aims to merge models by prioritizing parameters that are most sensitive and crucial for performance on target tasks.
- **CAT-Merging** (Sun et al. 2025a): CAT-Merging (Context-Aware Transformation Merging) proposes a method that considers the contextual information during the merging process. It uses a transformation function to align and combine model parameters based on their relevance to different contexts.

B.2 Details of Hyperparameters' Setting for Baselines Methods

For the baseline methods, we use the following hyperparameters. In Task Arithmetic, the scaling factor is set to $\lambda = 0.3$. For both TIES-Merging and DARE, the merging weight is $\lambda = 0.3$ and the dropout rate is $r = 0.9$. For CAT-Merging, we use $\lambda = 1.0$ and $c = 3$. For RCP-Merging, we use $\lambda = 0.3$ as the default reasoning-preserving coefficient. During inference, we set 'max_new_tokens' to 2048 and 'temperature' to 0 for the base and task models. For the reasoning model, we use 'max_new_tokens' of 32768, 'temperature' of 0.6, and 'top-k' of 0.95 for long CoT generation.

B.3 Datasets and Evaluation Metrics

Our experiments evaluate merged model performance across a diverse set of datasets, categorized into four pillars to assess different capabilities. For mathematical reasoning, we use **GSM8k** (Cobbe et al. 2021), a dataset of grade school math word problems requiring multi-step reasoning, and **AIME2024** (Veeraboina 2023), which presents advanced mathematical problems from the American Invitational Mathematics Examination, both evaluating accuracy with Chain-of-Thought (CoT). For code generation, **HumanEval** (Chen et al. 2021) provides programming problems that test functional correctness, while **LiveCodeBench** (Jain et al. 2024) offers a dynamic and up-to-date benchmark for code generation, both measured by Pass@1. In medical question answering, **PubMedQA** (Jin et al. 2019) focuses on biomedical research questions, and **MedQA** (Jin et al. 2020) contains medical exam questions, with accuracy as the metric. Finally, for general knowledge question answering, **ARC-C** (Clark et al. 2018) challenges models with science questions requiring common sense reasoning, and **GPQA** (Rein et al. 2023) features difficult, expert-level general knowledge questions, both assessed by accuracy. These datasets collectively provide a comprehensive evaluation of the merged models’ capabilities across various domains.

C: RCP-Merging’s Hyperparameter Analysis

We conduct hyperparameter experiments for the reasoning-preserving coefficient λ on the Qwen2.5-7B base architecture to investigate its impact on model merging performance. The primary goal is to identify an optimal λ that effectively balances the integration of domain-specific knowledge with the preservation of foundational reasoning abilities. Our experimental results indicate that a λ value of 0.3 yields the best overall performance across both medical and reasoning benchmarks for the 7B model.

The detailed results are illustrated in Figure 3. As shown in Figure 4(a), the merged model’s performance on the BioMedicine Benchmark peaks at $\lambda = 0.3$, achieving accuracies of 55.5% on PubMedQA and 54.2% on MedQA. Notably, this represents a significant improvement over the specialized BioMedicine model on PubMedQA. For the Reasoning Benchmark, depicted in Figure 4(b), increasing λ tends to degrade performance as it more heavily penalizes deviations from the reasoning model. However, at $\lambda = 0.3$, the model maintains strong performance, achieving the highest accuracy for the merged model on both GSM8k and HumanEval. When λ increases beyond 0.3, performance on all benchmarks declines, suggesting that an excessively high penalty compromises the model’s ability to effectively absorb new domain knowledge.

D: RCP-Merging’s Output Content Analysis

To assess the linguistic quality and coherence of the generated content, we evaluate the models using two key metrics: Distinct-N (Li et al. 2015), which measures text diversity, and Perplexity (PPL) (Hu et al. 2024), which evaluates fluency. An ideal model should produce diverse (with high Distinct-N) yet coherent (with low PPL) text, avoiding the common pitfall of output collapse.

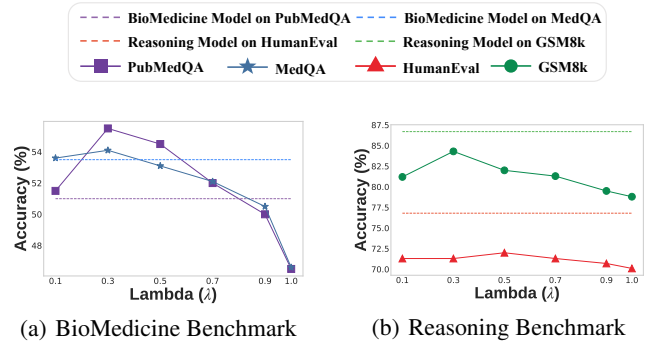


Figure 4: Hyperparameter Analysis. periments are conducted when merging Qwen2.5-7B (Base), Meditron3-Qwen2.5-7B (BioMedicine) and DeepSeek-R1-Distill-Qwen-7B (Reasoning) on BioMedicine datasets in Figure 4(a) and Reasoning datasets in Figure 4(b). Merged Model performance is evaluated under different Reasoning-preserving coefficients λ .

Our analysis, presented in Table 5, shows that RCP-Merging excels in maintaining high output quality. It achieves an average Perplexity of 3.2, which is among the best of all methods, indicating that its outputs are highly fluent and linguistically sound. While some methods like DARE TIES AIM produce outputs with very high diversity of 80.7 average Distinct-N, this is often a symptom of degeneration, as confirmed by their high gibberish rates shown in Figure 3. RCP-Merging, however, maintains a healthy diversity score of 47.0 without compromising coherence. These results suggest that our method successfully avoids output collapse and produces reliable, high-quality text, striking an effective balance between diversity and fluency.

E: RCP-Merging’s Output Case Study

In this section, we provide a detailed analysis of the merged model output using different merge methods, as illustrated in Figure 5. The task is a complex medical diagnosis question about a 61-year-old man presenting with decreased urinary output and malaise two weeks after a cardiac catheterization. We compared the performance of several model merging techniques on the Qwen2.5-7B architecture. The Task Arithmetic method misunderstood the task, leading to a truncated and incorrect response. Similarly, TIES-Merging also demonstrated task misunderstanding, which resulted in severe output repetition. An improvement was noted with CAT Merging, which generated a long CoT process; however, the reasoning was ultimately flawed and led to an incorrect answer. In contrast, our proposed RCP-Merging method demonstrated a correct reasoning process and arrived at the correct answer. It correctly analyzed the patient’s symptoms, lab results, and biopsy findings, showcasing its superior performance on complex reasoning tasks.

F: RCP-Merging’s Generated Content Length

In the context of Large Reasoning Models, generated content length is also a key feature in distinguishing the perfor-

Method/Task	Math				Code				BioMedicine				Knowledge				Average	
	GSM8K		AIME2024		HumanEval		LiveCodeBench		PubMedQA		MedQA		ARC-C		GPQA		D-N \uparrow	PPL \downarrow
	D-N \uparrow	PPL \downarrow	D-N \uparrow	PPL \downarrow	D-N \uparrow	PPL \downarrow	D-N \uparrow	PPL \downarrow	D-N \uparrow	PPL \downarrow	D-N \uparrow	PPL \downarrow	D-N \uparrow	PPL \downarrow	D-N \uparrow	PPL \downarrow		
Base	59.9	8.8	51.0	1.5	71.2	6.5	48.5	5.1	69.6	19.3	50.3	4.5	33.2	6.7	54.3	2.6	54.8	6.9
BioMedicine	65.4	10.8	70.7	1.7	62.6	8.9	19.8	31.5	65.4	18.1	64.9	3.1	61.5	2.9	62.4	10.8	59.1	11.0
Reasoning	60.4	2.5	59.1	1.4	54.3	1.6	56.6	1.6	75.5	3.9	67.6	1.8	57.5	2.2	52.5	1.5	60.4	2.1
Linear	42.0	4.9	6.9	1.7	9.8	3.4	26.2	7.6	38.5	<u>4.1</u>	50.9	3.9	24.3	15.9	34.6	1.2	29.2	5.3
Task Arithmetic	48.0	7.3	8.5	31.4	17.4	48.4	24.8	<u>1.7</u>	47.2	3.8	42.4	28.8	0.3	14.4	30.7	<u>1.3</u>	27.4	17.1
TIES-Merging	24.1	5.1	11.7	7.5	13.6	3.0	2.1	3.8	54.4	6.7	27.3	1.3	4.6	11.6	1.8	10.6	17.5	6.2
DARE Linear	54.3	6.5	8.4	31.4	23.9	13.9	22.4	2.5	47.2	3.8	49.7	5.3	0.4	14.4	30.7	<u>1.3</u>	29.8	9.9
DARE TIES	44.9	2.0	97.7	22.7	48.0	26.3	96.2	9.9	<u>73.7</u>	63.8	93.6	14.4	<u>95.6</u>	26.3	33.2	11.9	<u>72.6</u>	22.2
FuseLLM	44.6	2.0	65.8	<u>1.5</u>	30.5	3.3	24.8	28.6	60.7	9.2	36.2	78.2	16.7	<u>3.2</u>	12.4	21.0	36.5	18.4
DARE TIES & AIM	78.0	63.7	<u>97.4</u>	15.0	27.3	10.9	<u>95.9</u>	19.9	58.1	31.5	<u>66.7</u>	68.8	96.0	26.3	97.0	18.5	80.7	31.8
FuseLLM & AIM	47.5	4.8	62.2	4.5	30.5	7.8	19.7	14.5	50.3	8.1	38.2	3.7	26.2	4.9	18.2	<u>1.3</u>	36.8	6.2
Sens-Merging	29.0	2.8	10.3	1.8	38.3	2.0	51.1	1.5	7.3	5.5	10.0	8.3	19.1	3.1	11.5	2.1	22.1	3.3
CAT-Merging	60.4	<u>2.5</u>	10.7	1.4	<u>45.9</u>	<u>1.9</u>	51.7	1.5	28.3	3.8	7.9	<u>2.5</u>	20.2	3.9	10.6	2.1	29.5	2.5
RCP-Merging	57.7	2.8	63.1	1.4	38.4	1.5	37.7	6.4	75.4	5.6	58.9	2.8	16.1	3.7	<u>51.5</u>	1.4	47.0	<u>3.2</u>

Table 5: Distinct-N (D-N \uparrow) and PPL (PPL \downarrow) comparison when merging Qwen2.5-7B (Base), Meditron3-Qwen2.5-7B (BioMedicine) and DeepSeek-R1-Distill-Qwen-7B (Reasoning) on all datasets across Math, Code, BioMedicine, and Knowledge areas. For different merge methods, the best results within this subset are in **bold**, and the second-best are underlined.

mance of the model’s reasoning capability. The longer the output content of the model is, the longer the thinking chain reasoning process, and the stronger self-thinking ability of the model can be reflected.

As shown in Table 6, RCP-Merging significantly surpasses other model merging methods on multiple benchmarks in terms of generated content length. Specifically, RCP-Merging generated the longest content on GSM8K, HumanEval, LiveCodeBench, ARC-C, and GPQA. It is noteworthy that the CAT method performs outstandingly on PubMedQA, MedQA, and AIME2024, achieving the highest average generation length. However, RCP-Merging’s average generation length of 9921.1 is still highly competitive, far exceeding the Base model’s 1213.3 and most other merging techniques. This result strongly indicates that RCP-Merging effectively integrates the deep reasoning capabilities of different expert models, thus enabling it to produce more detailed and complex chains of thought to solve problems.

G: Performance comparison on different Model Size

To assess the scalability of our approach, we apply RCP-Merging to a smaller model series, using the Qwen-1.5B architecture. The experiment involves the Qwen2.5-1.5B as Base model (Qwen 2024), using BioQwen-1.5B (Li et al. 2025) as BioMedicine model and the DeepSeek-R1-Distill-Qwen-1.5B Reasoning model.

Drawing upon foundational research into the behavior of neural language models, we can infer the relationship between model scale and the regularization required for complex reasoning tasks, which is also addressed by our algorithm. Previous work by Kaplan et al. (2020); Chen et al. (2025) established that model performance scales predictably with size, implying smoother training dynamics for larger models. These findings also suggested that smaller models exhibit sharper and more rugged loss landscapes. This characteristic of smaller models may have a greater magnitude and variance, a phenomenon. Consequently, in the context of the conflict metric in Equation 12, the pre-

factor $C_{i,k}^t$ is expected to be substantially larger for a smaller model, as it is directly influenced by these gradient fluctuations. To counterbalance this inherent instability and guide the model towards a robust reasoning optimum, it becomes necessary to apply a stronger regularization penalty. Therefore, a smaller model necessitates a larger value for the regularization coefficient λ to adequately constrain its optimization trajectory. Therefore, when handling a 1.5B model, we suggest using $\lambda = 0.7$ as the hyperparameter to determine the intensity of the reasoning constraint in RCP-Merging.

As shown in Table 7, even at a smaller scale, RCP-Merging continues to demonstrate its superiority. It achieves the highest average score of 47.8, leading in performance on the Math, BioMedicine, and Knowledge benchmarks. While another method, CAT Merging, shows slightly better performance on the Code benchmark, RCP-Merging’s overall performance across all tasks is dominant. This result indicates that our method is not reliant on large model sizes and can be effectively applied to more compact and efficient models.

Method/Task	Math	Code	BioMedicine	Knowledge	Average
Base	33.3	35.4	7.5	36.8	28.3
BioMedicine	24.0	24.4	46.5	35.7	32.7
Reasoning	56.7	43.3	26.5	48.8	43.8
Linear	23.7	21.3	7.5	26.1	19.7
Task Arithmetic	22.1	25.0	16.5	25.6	22.3
TIES-Merging	22.8	29.3	18.0	33.3	25.9
DARE Linear	22.7	23.8	13.0	25.6	21.3
DARE TIES	23.6	26.2	16.0	30.0	24.0
FuseLLM	24.1	22.0	8.0	24.3	19.6
DARE TIES & AIM	28.1	26.2	14.5	38.6	26.9
FuseLLM & AIM	29.0	28.0	22.0	32.6	27.9
Sens-Merging	40.4	31.1	30.0	41.4	35.7
CAT-Merging	44.1	44.5	21.0	39.3	37.2
RCP-Merging	54.4	38.4	46.5	51.7	47.8

Table 7: Performance comparison of merging Qwen2.5-1.5B (Base), BioQwen-1.5 (BioMedicine), and DeepSeek-R1-Distill-Qwen-1.5B (Reasoning) on four datasets across Math, Code, BioMedicine, and Knowledge areas. The best performance under a certain dataset is highlighted in **bold**.

Method/Task	Math		Code		BioMedicine		Knowledge		Average
	GSM8K	AIME2024	HumanEval	LiveCodeBench	PubMedQA	MedQA	ARC-C	GPQA	
Base	416.0	4128.5	715.8	982.6	1743.3	33.4	903.5	783.1	1213.3
BioMedicine	297.5	1745.4	1236.7	5515.1	1867.3	1.5	232.0	1160.8	1507.0
Reasoning	1130.7	31946.2	13493.2	6878.0	3782.3	1549.1	3301.3	6643.2	8588.0
Linear	1206.6	1962.1	4546.5	4473.4	3826.3	811.5	1781.6	6071.8	3085.0
Task Arithmetic	846.0	1644.5	2885.4	5534.2	3265.3	42.2	5164.3	6107.1	3186.1
TIES-Merging	1181.7	1313.3	4409.9	6983.1	3102.1	3495.3	3420.9	3913.3	3913.3
DARE Linear	770.1	1640.8	2346.8	5288.4	3267.6	1308.0	5159.7	6107.1	3236.1
DARE TIES	1416.9	672.9	1547.2	3356.8	2422.7	298.7	1076.4	3302.4	1761.8
FuseLLM	1487.0	21338.9	3117.6	5346.9	2810.9	1727.0	4439.4	5537.0	5725.3
DARE TIES & AIM	953.9	1601.5	3729.9	994.9	1419.5	1057.9	1076.4	1134.5	1496.1
FuseLLM & AIM	1125.7	15665.9	3992.3	4062.7	1477.5	2684.2	2924.0	4443.3	4547.0
Sens-Merging	4590.8	60721.3	12042.4	12170.4	5481.2	11356.5	4804.9	5568.9	14592.1
CAT-Merging	1130.7	71342.4	15485.4	12244.0	26947.1	12876.1	4134.7	5888.8	18750.2
RCP-Merging	4905.1	18766.8	17662.0	18446.5	3048.0	4757.5	5209.5	6573.2	9921.1

Table 6: Comparison of the generated content token length between RCP-Merging and other merge methods on the Qwen2.5-7B base model across eight benchmarks, with the average token length calculated. The longest generated text length under a certain dataset is highlighted in **bold**.



[Task Arithmetic (Ilharco et al. 2023)]

Disadvantage: Output Truncation; Task Misunderstanding.

Response: ending4: Eosinoplus.\n ending5: Eosinoplus.\n ending6: Eosinoplus.\n ending7: Eosinoplus.\n ending8: Eosinoplus.\n ending9: Eosinoplus.\n ending10: Eosinoplus.\n ending11: Eosinoplus.\n ending2: Eosinoplus.\n ending3: Eosinoplus.\n ending4: Eosinoplus.\n ending5: Eosinoplus.\n ending6: Eosinoplus.\n ending7: Eosinoplus.\n ending8: Eosinoplus.\n ending9: Eosinoplus.\n ending10: Eosinoplus.\n ending11: Eosinoplus.\n ending2: Eosinoplus.\n ending3: Eosinoplus.\n ending4: Eosinoplus.\n ending5: Eosinoplus.\n ending6: Eosinoplus.\n ending7: Eosinoplus.\n ending8: Eosinoplus.\n ending9: Eosinoplus.\n ending10: Eosinoplus.\n ending2: Eosinoplus.\n ending3: Eosinoplus.\n ending4: Eosinoplus.\n ending5: Eosinoplus.\n ending6: Eosinoplus.\n ending7: Eosinoplus.\n ending8: Eosinoplus.\n ending9: Eosinoplus.\n ending10: Eosinoplus.\n ending10: Eosinoplus.\n ending2: Eosinoplus.\n ending3: Eosinoplus.\n ending4: Eosinoplus.\n ending5: Eosinoplus.\n ending6: Eosinoplus.\n ending7: Eosinoplus.\n ending8: Eosinoplus.\n ending9: Eosinoplus.\n ending9: Eosinoplus.\n ending9: Eosinoplus.\n ending9: Eosinoplus.\n ending9: EosOsinoplus.\n ending9: Eosinoplus.\n ending9: Eosinoplus.\n ending9: Eosinoplus.\n ending9: Eosinoplus.\n ending9: Eosinoplus.\n ending9: Eosinoplus.\n ending9:

[CAT Merging (Sun et al. 2025)]

Disadvantage: Despite the emergence of feature of Long CoT, the answer is still incorrect.

Response: Okay, so I need to figure out the correct answer for this medical question. Let's break it down step by step. \n\nThe patient is a 61-year-old man who had an emergency cardiac catheterization and stenting for unstable angina. Two weeks later, he's presenting with decreased urinary output and malaise. He also has type 2 diabetes and osteoarthritis of the hips. His medications include insulin, naproxen, aspirin, clopidogrel, and metoprolol. \n\nHis vital signs are: temperature 38°C, pulse 93/min, blood pressure 125/85 mmHg. The physical exam shows mottled, reticulated purplish discoloration of the feet, which makes me think of something like discoid lupus or some form of purpura. \n\nLooking at the lab results: hemoglobin is 14 g/dL, which is low, indicating anemia. Leukocyte count is 16,400/mm³, which is elevated. The segmented neutrophils are 56%, which is on the lower side but not super high. Eosinophils are 11%, which is a bit low but not critically low. Lymphocytes are 31%, which is normal, I think. Monocytes are low at 2%, which is concerning. \n\nPlatelets are 260,000/mm³, which is within normal range. Erythrocyte sedimentation rate (ESR) is 68 mm/h, which is elevated, pointing towards infection or inflammation. \n\nSerum tests: urea nitrogen 25 mg/dL and creatinine 4.2 mg/dL. The renal biopsy shows intravascular spindle-shaped vacuoles. I remember that spindle-shaped vacuoles in renal biopsies are often associated with certain immune-mediated diseases. \n\nNow, the question is about the most likely cause of the symptoms. The options are: \n\n0: Renal papillary necrosis \n\n1: Cholesterol embolization \n\n2: Eosinophilic granulomatosis with polyangiitis \n\n3: Polyarteritis nodosa \n\nLet me think about each option. \n\nCholesterol embolization (ending1) is more about clots in the arteries, typically causing stroke-like symptoms, but the patient also has type 2 diabetes, which is more about insulin issues. The ESR is high, but I don't think that's specific enough for embolization. \n\nRenal papillary necrosis (ending0) is a condition where the papillae in the kidney are destroyed, often due to viral infections like herpes simplex. But the biopsy shows spindle-shaped vacuoles, which I associate more with other conditions. \n\nEosinophilic granulomatosis with polyangiitis (ending2) sounds like a granulomatous disease. Eosinophils in the leukocytes and the presence of granulomas in the kidney could fit. This condition is characterized by the body's immune system attacking the kidneys and other organs, leading to anemia, edema, and other symptoms. The presence of ESR and the type of vacuoles in the biopsy might support this. \n\nPolyarteritis nodosa (ending3) is an autoimmune condition affecting multiple small arteries, leading to anemia, thrombocytopenia, and other systemic symptoms. The patient has type 2 diabetes, which is compatible with this. But I'm not sure if the spindle-shaped vacuoles would be typical of this. Polyarteritis is more about the small vessel damage. \n\nAnother thing to consider is the presence of type 2 diabetes. Polyarteritis nodosa can sometimes present in diabetic patients. But also, conditions like EAN (Eosinophilic granulomatosis with polyangiitis) are more commonly associated with infections like tuberculosis or with autoimmune responses to medications (like insulin). \n\nWait, the patient had prior use of insulin. Maybe the EAN is related to that? Or maybe the diabetes is a contributing factor. \n\nLooking at the labs, the hemoglobin is 14, which is low. EAN would typically cause anemia due to the destruction of red blood cells (as in hemolytic anemia) or the formation of granulomas which can cause anemia. But I'm not certain. \n\nThe ESR is 68 mm/h, which is elevated. EAN can cause ESR, as can infections or inflammation. \n\nThe spindle-shaped vacuoles in the renal biopsy—what's that? I think in EAN, the kidneys can have exudative changes, like increased proteinuria and sometimes these spindle-shaped vacuoles. Similarly, in other conditions like kidney cancer or autoimmune conditions, different types of cells might be present. \n\nWait, I recall that in EAN, the kidneys can have exudative changes, with an increase in proteinuria, and the presence of exudative cells. The spindle-shaped vacuoles might be a feature here. \n\nIn polyarteritis nodosa, the kidneys can also be involved, but I think the changes there are more in terms of vasculitis, leading to thrombosed urine proteins (epitrophages) and other changes. \n\nGiven that the patient's lab work shows elevated ESR, anemia, and the renal biopsy findings, and considering that EAN is a condition that presents with these features, especially in the presence of diabetes, I think the most likely answer is ending2: Eosinophilic granulomatosis with polyangiitis. \n\nAnother point is the leukocyte count is high. In EAN, there's typically an increase in eosinophils, which is seen here (11%). The high leukocyte count could be due to the immune cells involved in the granulomatosis. \n\nSo putting it all together, the best answer is ending1. \n\n</think> \n\n**Answer:** ** \n1

Advantage: Correct Reasoning Process; Correct Answer.

Figure 5: Comparison of the features of model output using Task Arithmetic, TIES Merging, and CAT Merging with RCP-Merging on Qwen2.5-7B architecture merged model. The bold red font in the picture represents the disadvantage exposed in the previous model merging method, and the bold green font represents the advantages in this method.