ReasonAny: Incorporating Reasoning Capability to Any Model via Simple and Effective Model Merging

Anonymous ACL submission

Abstract

Large Reasoning Models (LRMs) with long chain-of-thought reasoning have recently achieved remarkable success. Yet, equipping domain-specialized models with such reasoning capabilities, referred to as "Reasoning + X", remains a significant challenge. While model merging offers a promising training-free solution, existing methods often suffer from a destructive performance collapse: existing methods tend to both weaken reasoning depth and compromise domain-specific utility. Interestingly, we identify a counter-intuitive phenomenon underlying this failure: *reasoning ability predominantly resides in parameter regions with low gradient sensitivity, contrary to the common assumption that domain capabilities correspond to high-magnitude parameters*. Motivated by this insight, we propose **ReasonAny**, a novel merging framework that resolves the reasoning–domain performance collapse through Contrastive Gradient Identification. Experiments across safety, biomedicine, and finance domains show that ReasonAny effectively synthesizes "Reasoning + X" capabilities, significantly outperforming state-of-the-art baselines while retaining robust reasoning performance.

1 Introduction

The recent emergence of Large Reasoning Models (LRMs) represents a milestone breakthrough in the landscape of Large Language Models (LLMs) (Grattafiori et al., 2024; Yang et al., 2024a). By leveraging the long chain-of-thought (long-CoT) mechanisms (Yeo et al., 2025), reasoning models have demonstrated exceptional performance, particularly in specialized tasks such as mathematics and coding (Jaech et al., 2024; Team, 2025b; Guo et al., 2025; OpenAI, 2025). Still, equipping models in specific domain tasks with these advanced reasoning capabilities is a vital yet under-explored frontier. For LLMs equipped with domain-specific knowledge, such as safety alignment (Kuo et al., 2025), biomedicine (Ullah et al., 2024; Griot et al., 2025), or finance (Zhao et al., 2024; Yuqi et al., 2024), one objective is to construct models that have not only robust **Reasoning** capabilities but are also specialized in domain-specific tasks "**X**". We term this critical synthesis "**Reasoning + X**".

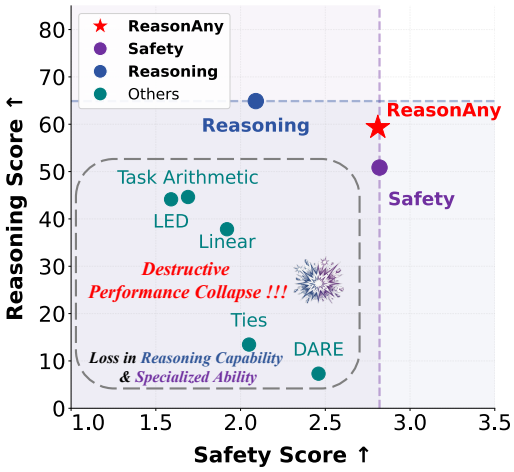

Figure 1: ReasonAny overcomes the destructive performance collapse in model merging, evaluated via GSM8K accuracy and *max - current harmfulness score* as Safety Score on Safety-Tuned bench. Methods in purple and blue bounds show the loss in specialized ability and reasoning capability, respectively. By reaching the top-right corner, ReasonAny preserves robust reasoning capability without compromising specialized utility.

To achieve this synthesis, the prevailing approach involves Supervised Fine-Tuning (SFT) or Reinforcement Learning (RL) on domain-specific reasoning datasets (Kuo et al., 2025; Qian et al., 2025b; Team, 2025a; Kai-tao et al., 2025; Chen et al., 2025; Bao et al., 2025). Despite its efficacy, this paradigm faces challenges: difficulty constructing domain-specific reasoning data (Chen et al., 2024; Qian et al., 2025b), resource-intensive training (Matsutani et al., 2025), and catastrophic forgetting (Parisi et al., 2019; Parthasarathy et al.,

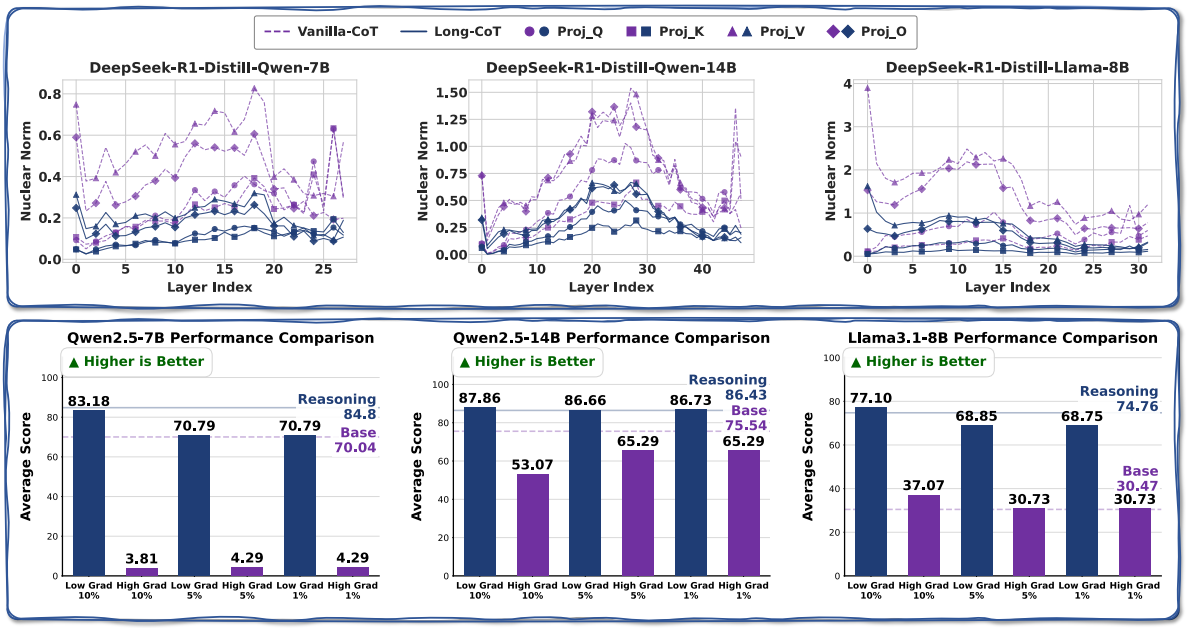

Figure 2: **Gradients Nuclear Norm Analysis and Additive Experiment Results.** The **top sub-figure** shows gradient analysis across (Q, K, V, O) projection matrices at all layers. The **top-left**, **top-middle**, and **top-right** panels display Nuclear Norms for DeepSeek-R1-Distill Qwen-7B, Qwen-14B, and Llama-8B respectively, revealing that long-CoT induces significantly lower gradients than Short-CoT. The **bottom sub-figures** display additive experiments validating that reasoning capability lies in low-gradient regions. By merging weights from 10%, 5%, and 1% of highest and lowest gradient into base models, results across the **top-left**, **top-middle**, and **top-right** sub-figures consistently demonstrate that reasoning capability depends on weights associated with low gradients.

2024). In light of these challenges, model merging has emerged as a compelling, training-free alternative designed to combine distinct capabilities from different models into a single entity (Yang et al., 2024b; Zhou et al., 2025; Lan et al., 2025).

Motivated by this potential, we conduct a preliminary exploration to merge reasoning and domain-specific models via state-of-the-art techniques. Interestingly, as illustrated in Figure 1, our experiments reveal a **Destructive Performance Collapse** in the context of reasoning—resulting merged models typically suffer from both significant loss in reasoning capability and severe compromise in the specialized abilities of "X". This phenomenon persists despite existing methods (Yadav et al., 2023; Liu et al., 2025) proving effective for standard knowledge injection. Such a setback likely stems from the common assumption that high-magnitude weights or gradients identify important parameters (Yadav et al., 2023; Hao et al., 2025). Our findings challenge this intuition and raise a pivotal question: **Do parameters handling reasoning capability follow the same high-magnitude rules as knowledge locating?**

As illustrated in the top part of Figure 2, we uncover a **counter-intuitive phenomenon:** ***reasoning capabilities are characterized by subtle, low-magnitude gradient changes, challenging the prevailing belief that important features necessarily generate high-magnitude gradients shifts*** (Liu et al., 2025; Ma et al., 2025). This phenomenon reveals that models employing long-CoT and models with superior reasoning capabilities consistently exhibit significantly lower gradient magnitudes compared to standard instruct-tuned models.

Based on this counter-intuitive phenomenon, we propose **ReasonAny**, a novel merging framework designed to resolve the "Reasoning + X" conflict. Unlike traditional methods that treat all tasks uniformly under a single importance metric (Zeng et al., 2025; Thapa et al., 2025), ReasonAny employs **Contrastive Gradient Identification** to handle these conflicting model parameters selection. Specifically, we isolate the robust features of domain-specific task "X" using traditional high-gradient selection, while simultaneously capturing reasoning capabilities through a targeted *low-gradient* filtering mechanism. To ensure these distinct capabilities coexist without destructive performance collapse, we implement **Conflict Resolution via Exclusion** that creates mutually exclusive parameter masks before composing the final model. The overall workflow of ReasonAny is shown in the bottom part of Figure 3. Our experiments demon-

strate that ReasonAny successfully incorporates advanced reasoning capabilities into diverse models without compromising their domain-specific capabilities, offering a simple yet effective solution to the reasoning-utility performance collapse.

2 Reasoning Capabilities Reside in Low-Gradient Parameter Regions

In the pursuit of synthesizing reasoning capabilities with domain-specific task "X", model merging presents a promising training-free solution (Ilharco et al., 2023; Yang et al., 2024b; Zhou et al., 2025). However, as illustrated in Figure 1 and comprehensive evaluation in Section 4, we observe that traditional merging methods often suffer from **Destructive Performance Collapse** with both the reasoning capability collapses and the domain utility is compromised.

To resolve this, we investigate the distinct gradient characteristics underlying reasoning capabilities. Specifically, Section 2.1 establishes the mathematical foundations for model merging, Section 2.2 analyzes the unique gradient magnitude distributions of reasoning models, and Section 2.3 confirms that reasoning specifically relies on low-gradient structures through targeted additive experiments.

2.1 Preliminaries

Models and Task Vectors. We operate within the parameter space of Transformer-based LLMs. Let $\theta_{\text{base}} \in \mathbb{R}^d$ denote the parameters of a pre-trained base model. We consider a scenario where θ_{base} serves as the initialization for two distinct fine-tuning processes: (1) A **Task Model** $\theta_t \in \mathbb{R}^d$, which is fine-tuned on a domain-specific task "X" dataset $\mathcal{D}_t$, such as safety, biomedicine or finance. (2) A **Reasoning Model** $\theta_r \in \mathbb{R}^d$, which is fine-tuned on a reasoning-intensive dataset $\mathcal{D}_r$.

Following standard arithmetic merging formulations (Ilharco et al., 2023), we define the *task vector* τ as the dense displacement in the parameter space resulting from fine-tuning. The task vectors for the specialized task and reasoning are defined respectively as:

$$\tau_t = \theta_t - \theta_{\text{base}}, \quad \tau_r = \theta_r - \theta_{\text{base}}. \tag{1}$$

Intuitively, these vectors encode the specific model weight shifts required to endow the base model with specialized domain expertise or reasoning capabilities. Our objective is to construct a merged parameter set θ_{merged} that incorporates the functional capabilities of both τ_t and τ_r without destructive interference.

Gradient-Based Parameter Identification. To determine the topology of the critical parameters for each task, prevailing methodologies in recent works extensively utilize gradient-based metrics for parameter identification (Liu et al., 2025; Ma et al., 2025). For a given model parameterized by θ and a calibration dataset $\mathcal{D}$, the importance score I_j for the j-th parameter is computed as the expectation of the gradient magnitude with respect to the loss function $\mathcal{L}$. Formally, the identification vector $I(\theta, \mathcal{D}) \in \mathbb{R}^d$ is defined as:

$$I(\theta, \mathcal{D}) = \mathbb{E}_{x \sim \mathcal{D}} \left[|\nabla_\theta \mathcal{L}(x; \theta)| \right]. \tag{2}$$

This metric proxies parameter saliency, quantifying task performance sensitivity to weight perturbations. Intuitively, higher values indicate task-critical weights.

2.2 Gradient Magnitude Distribution Analysis

Prevailing research generally operates on the assumption that high-magnitude gradients encode the most critical model capabilities (Liu et al., 2025; Ma et al., 2025). Adopting the spectral analysis from Li et al. (2025a), we measure gradient magnitude distribution across layers for both Task Model and Reasoning Model using **Nuclear Norm**:

$$s_{x,i} = \|\nabla_i \mathcal{L}(x; i)\|_* = \sum_{j=1}^{\min\{m,n\}} \sigma_j, \tag{3}$$

where σ_j represents the singular values of the gradient matrix $\nabla_i \mathcal{L}(x; i)$ corresponding to the Q, K, V, and O projection matrices at layer i.

As shown in top sub-figures of Figure 2, Qwen2.5-7B, Qwen2.5-14B and Llama3.1-8B series reasoning models with blue line marked as Long-CoT, exhibit significantly **lower nuclear norms** than the base model purple line marked as Vanilla-CoT. This **counter-intuitive phenomenon** suggests that reasoning capabilities reside in low-gradient regions, challenging conventional assumptions that high-magnitude gradients corresponding weights encode more important information.

Detailed analysis of correlation between gradients and nuclear norms is shown in Appendix A.

2.3 How Do Low-gradient Parameters Work?

To empirically validate whether reasoning capabilities are localized within low-gradient model weights, we conduct an additive experiment based on the intuition that exclusively injecting these targeted weights into the base model should significantly reactivate its reasoning abilities. Specifically,

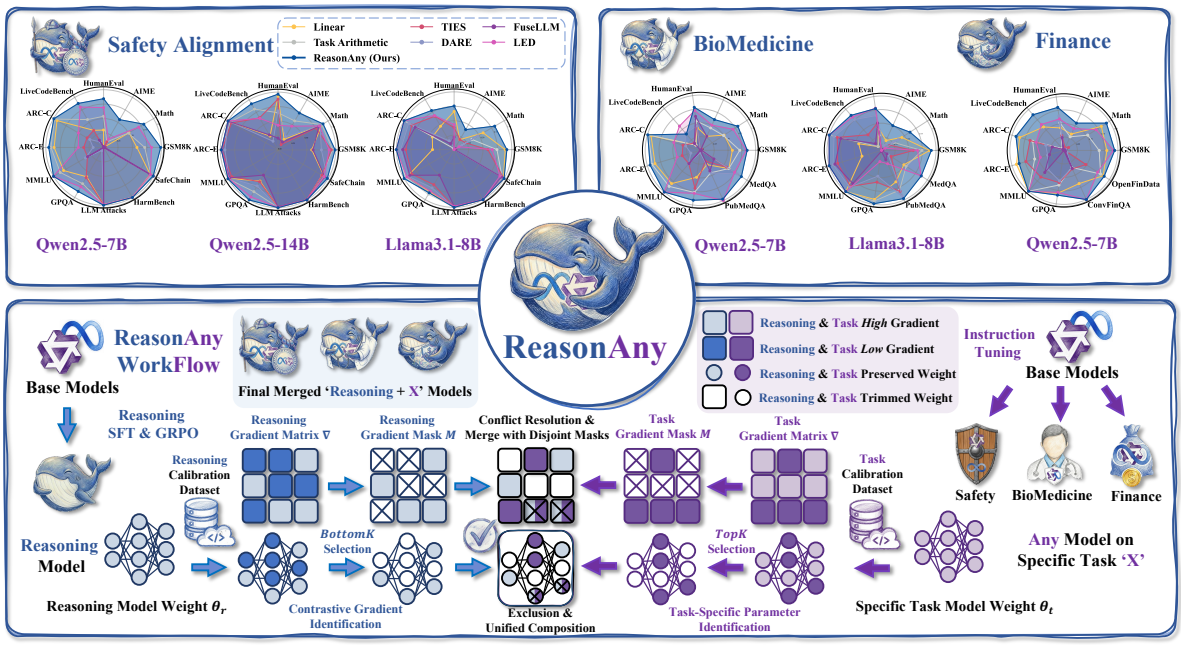

Figure 3: **Experimental Results and Workflow of ReasonAny**. Experimental results on **Safety** (**top-left**), **Biomedicine**, and **Finance** (**top-right**) benchmarks demonstrate ReasonAny, shown in light blue background, significantly outperforming baselines. **ReasonAny Workflow** (**bottom**) employs **Contrastive Gradient Identification** (**bottom-right**) to isolate low-gradient reasoning and high-gradient task weights and **Exclusion** (**bottom-middle**) step disjoint masks that merge specialized capabilities without compromising reasoning capabilities.

we selectively add parameters from the reasoning task vector τ_r into the base model θ_{base} via:

$$\theta' = \theta_{\text{base}} + \tau_r \odot \mathbf{M}, \tag{4}$$

where $\mathbf{M}$ denotes a binary mask that filters weights based on their gradient magnitude ranking.

We apply this method to base model by adding the *Highest Gradients* and the *Lowest Gradients* corresponding parameters at specific sparsity ratios of 10%, 5% and 1%. As illustrated in the bottom of Figure 2, Qwen2.5-7B recovers GSM8K scores of 83.18 and 70.79 using 10% and lower ratios of lowest gradient parameters, whereas incorporating highest gradient updates leads to a complete performance collapse. This phenomenon can also be found on both Qwen2.5-14B and Llama3.1-8B series models. ***This distinct performance disparity strongly validates that reasoning capabilities are predominantly localized within low-magnitude gradient model weight regions.***

3 Methodology

3.1 Overview of ReasonAny

We introduce ReasonAny, a unified framework designed to synthesize the capabilities of a generic specialized **Task Model** (e.g., Safety, Biomedicine, Finance) and a **Reasoning Model** into a single backbone. ReasonAny operate this pipeline through two distinct stages: we first employ **Contrastive Gradient Identification** to isolate capability-specific parameter regions. Subsequently, we separate and culminate via certain model weights **Exclusion and Unified Composition** to synthesize these disjoint sets without destructive interference. The workflow and algorithm are illustrated in Figure 3 and Algorithm 1.

3.2 Parameter Identification

Reasoning Parameter Identification. Recalling the insight that reasoning capabilities are encoded in reasoning model weights exhibiting low-magnitude gradients, we adopt a **Contrastive Gradient Identification** strategy in the first phase of ReasonAny. We select model weights with the lowest gradient magnitude on the reasoning dataset $\mathcal{D}_r$ and let $\text{BottomK}(v, k)$ be an operator returning the smallest values ratio k in task vector v. The elected reasoning weight set $\mathcal{N}_r$ is defined as:

$$\mathcal{N}_r = \text{BottomK}\big(I(\theta_r, \mathcal{D}_r), p_r\big), \tag{5}$$

where p_r represents the selection ratio for total reasoning model parameters θ_r.

Task-Specific Parameter Identification. In parallel, we identify the parameters critical for the

Algorithm 1 REASONANY

Require: Base model θ_{base}, Task model θ_t, Reasoning model θ_r, Calibration datasets $\mathcal{D}_t, \mathcal{D}_r$, Selection ratios p_t, p_r, Scaling factors λ_t, λ_r

Ensure: Merged parameters θ_{merged}

Initialize $\theta_{\text{merged}} \leftarrow \theta_{\text{base}}$
// Step 1: Calculate Task Vectors
$\tau_t \leftarrow \theta_t - \theta_{\text{base}}$, $\tau_r \leftarrow \theta_r - \theta_{\text{base}}$
// Step 2: Calculate Importance Scores (Gradient Sensitivity)
$I(\theta_t) \leftarrow \mathbb{E}_{x\sim\mathcal{D}_t}[|\nabla_\theta \mathcal{L}(x;\theta_t)|]$
$I(\theta_r) \leftarrow \mathbb{E}_{x\sim\mathcal{D}_r}[|\nabla_\theta \mathcal{L}(x;\theta_r)|]$
// Step 3: Identify Subspaces
$d \leftarrow \text{length}(\theta_{\text{base}})$
$\mathcal{N}_t \leftarrow \text{TopK}(I(\theta_t), p_t)$ ▷ High-gradient for Task
$\mathcal{N}_r \leftarrow \text{BottomK}(I(\theta_r), p_r)$ ▷ Low-gradient for Reasoning
// Step 4: Conflict Resolution (Exclusion)
$\mathcal{T}_t' \leftarrow \mathcal{N}_t \setminus \mathcal{N}_r, \quad \mathcal{T}_r' \leftarrow \mathcal{N}_r \setminus \mathcal{N}_t$
// Step 5: Merge with Disjoint Masks
Initialize Masks $\mathbf{M}_t \leftarrow \mathbf{0}, \mathbf{M}_r \leftarrow \mathbf{0}$
for $i \in \mathcal{T}_t'$ **do** $\mathbf{M}_{t,i} \leftarrow 1$
end for
for $j \in \mathcal{T}_r'$ **do** $\mathbf{M}_{r,j} \leftarrow 1$
end for
$\theta_{\text{merged}} \leftarrow \theta_{\text{merged}} + \lambda_t(\tau_t \odot \mathbf{M}_t) + \lambda_r(\tau_r \odot \mathbf{M}_r)$
return θ_{merged}

specialized Task "X" such as safety alignment, biomedicine and finance expertise. Consistent with established pruning and merging literature (Liu et al., 2025; Ma et al., 2025; Yang et al., 2025b), with results shown in Section 2.3, we reaffirm that domain-specific knowledge is retained in parameters with high sensitivity to the task loss. Therefore, we employ a standard *Top-K* selection strategy on the task model θ_t using dataset $\mathcal{D}_t$. Let $\text{TopK}(v, k)$ denote the largest values indices with the ratio k. The elected task parameter set $\mathcal{N}_t$ is:

$$\mathcal{N}_t = \text{TopK}\big(I(\theta_t, \mathcal{D}_t), p_t\big), \tag{6}$$

where p_t is the selection ratio for the task model θ_t.

3.3 Exclusion and Unified Composition

Conflict Resolution via Exclusion. A fundamental challenge in merging distinct models is parameter conflict, where a single parameter is deemed critical for both reasoning and the specific task ($\mathcal{N}_r \cap \mathcal{N}_t \neq \emptyset$). To preventing destructive interference—where the injection of domain knowledge might degrade reasoning depth—we enforce mutual exclusivity through a set-theoretic exclusion process. We derive the final, disjoint parameter sets $\mathcal{T}_r'$ and $\mathcal{T}_t'$ by removing overlapping indices, ensuring that each parameter is updated by at most one source using $\mathcal{T}_r' = \mathcal{N}_r \setminus \mathcal{N}_t$ and $\mathcal{T}_t' = \mathcal{N}_t \setminus \mathcal{N}_r$. This step guarantees that the delicate low-gradient structures preserved for reasoning are not overwritten by high-magnitude task updates.

Unified Model Composition. Finally, we construct the unified model by composing the base model with the disjointly selected task vectors. We define binary masks $\mathbf{M}_r, \mathbf{M}_t \in \{0,1\}^d$ corresponding to the indices in $\mathcal{T}_r'$ and $\mathcal{T}_t'$ respectively. The final merged weights θ_{merged} are computed as:

$$\theta_{\text{merged}} = \theta_{\text{base}} + \lambda_r(\tau_r \odot \mathbf{M}_r) + \lambda_t(\tau_t \odot \mathbf{M}_t), \tag{7}$$

where $\odot$ denotes the element-wise product, and λ_r, λ_t are scaling factors. This formulation effectively merges the "Reasoning + X" capabilities into the base model while strictly respecting the topological boundaries identified in the previous steps.

4 Experiments

4.1 Experiments Setup

Baselines. We compared ReasonAny with multiple merging baselines: **Linear** (Izmailov et al., 2018), **Task Arithmetic** (Ilharco et al., 2023), **TIES-Merging** (Yadav et al., 2023), **DARE-Merging** (Yu et al., 2024), **FuseLLM** (Wan et al., 2024) and **LED-Merging** (Ma et al., 2025). We utilize mergekit (Goddard et al., 2024) as merging tools for baseline methods. Detailed baselines explanation and recommended hyperparameter settings are listed in Appendix B and F. Moreover, explanation of Figure 1 is shown in Appendix E.

Datasets. Using the same **Performance** benchmark for **Reasoning** and **Knowledge**, we evaluated **Safety**, **Biomedicine**, and **Finance** tasks using domain-specific benchmarks. In performance benchmarks, for *Reasoning Evaluation*, we assess with *GSM8K* (Cobbe et al., 2021), *Math500* (Lightman et al., 2023) and *AIME2024* (Veeraboina, 2023) for math reasoning, *HumanEval* (Chen et al., 2021) and *LiveCodeBench* (Jain et al., 2024) for code reasoning. For *Knowledge Evaluation*, we utilized *ARC-E*, *ARC-C* (Clark et al., 2018), *MMLU* (Hendrycks et al., 2021b,a) and *GPQA* (Rein et al., 2023) to test the knowledge

Table 1: Performance comparison of merging Qwen2.5-7B family with safety fine-tuning Qwen2.5-7B-Instruct (Safety) and DeepSeek-R1-Distill-Qwen-7B (Reasoning) on all datasets across Reasoning, Knowledge and Safety Benchmarks, where **Average** ↑ column indicate average performance across performance bench. The best performance among all merging methods on each dataset is highlighted in **bold**, and values highlighted in *italic* with * mark indicate model output collapse.

Eval Bench	Performance Bench										Safety Bench		
Sub Areas	Reasoning					Knowledge					Safety		
Datasets	GSM8K↑	Math↑	AIME↑	HumanEval↑	LiveCodeBench↑	ARC-C↑	ARC-E↑	MMLU↑	GPQA↑	Average↑	Safety-Tuned↓	HarmBench↓	SafeChain↑
Safety	69.42	74.00	13.33	50.64	12.43	60.91	65.78	**71.72**	39.39	50.85	**1.18**	**0.08**	**4.90**
Reasoning	**87.23**	**86.20**	**60.00**	**76.63**	**30.37**	**64.75**	**77.25**	52.51	**49.10**	**64.89**	1.91	0.46	4.61
Linear	50.42	43.80	0.00	23.14	10.38	60.34	66.19	57.34	28.78	37.82	2.08	0.31	4.57
Task Arithmetic	62.17	42.80	6.67	41.35	16.93	63.56	70.43	62.14	35.61	44.63	2.31	0.40	4.66
Ties	*0.83**	*2.60**	*6.67**	*0.00**	*10.38**	21.36	*26.63*	*23.08*	*29.55**	*13.46**	1.95	*0.02**	4.86
DARE	*0.53**	*1.00**	*0.00**	*0.00**	*3.38**	17.97	13.93	25.95	*3.03**	*7.31**	1.54	*0.00**	4.86
FuseLLM	*1.81**	*0.20**	*0.00**	*1.23**	*4.43**	*0.00**	*0.00**	22.95	*0.00**	*3.40**	**0.87**	*0.01**	4.54
LED	72.48	60.60	10.00	52.91	24.51	32.54	33.69	71.93	38.64	44.14	2.41	0.36	4.59
ReasonAny	**86.28**	**69.40**	**33.33**	**64.65**	**26.71**	**64.31**	**73.39**	**72.73**	**43.18**	**59.33**	1.19	**0.08**	**4.86**

preservation of merged models. For *Safety Evaluation*, *Safety-Tuned* (Bianchi et al., 2024), *HarmBench* (Mazeika et al., 2024) and *SafeChain* (Jiang et al., 2025) are used to verified the robustness of merged models. For *BioMedicine Evaluation*, we use *PubMedQA* (Jin et al., 2019) and *MedQA* (Jin et al., 2020). For *Finance Evaluation*, we use *ConvFinQA* (Chen et al., 2022) and *OpenFinData* (Information, 2023). We use opencompass (Contributors, 2023) as the evaluation tool. Detailed datasets explanation is shown in Appendix C.

Models. Our experiments utilize base models on the **Qwen2.5** and **Llama-3.1** series (Yang et al., 2024a; Grattafiori et al., 2024). The corresponding reasoning models are **DeepSeek-R1-Distill** series models and **QwQ-32B-Preview** (Guo et al., 2025; Team, 2025b). For safety task, by fine-tuning on Safety training Dataset (Bianchi et al., 2024) using Low-Rank Adaptation (Hu et al., 2022; Wang, 2023) on corresponding instruct models, we obtain the model with the best safety performance among the corresponding family of models in our setting. For biomedicine task, we use **Meditron3-Qwen2.5-7B** and **MMed-Llama-3-8B** on Qwen2.5-7B and Llama3.1-8B family as biomedicine task expert (Chen et al., 2023; Qiu et al., 2024). For finance task, we use **WiroAI-Finance-Qwen-7B** and **WiroAI-Finance-Llama-8B** on Qwen2.5-7B and Llama3.1-8B family as finance task expert (Abdullah Bezir, 2025b,a). Full model configuration are shown in Appendix D.

4.2 ReasonAny Preserves Specific Task Utility Alongside Robust Reasoning Capability

ReasonAny ensures robust safety without compromising reasoning capability. Table 1 illustrates the performance comparing ReasonAny and baseline methods across Qwen2.5-7B benchmarks. ReasonAny retains a GSM8K score of 86.28, recovering 98.91% of reasoning capability. On the Safety Bench, ReasonAny adheres strictly to safety protocols. Conversely, Linear merging and DARE suffer catastrophic interference with GSM8K scores of 50.42 and 0.53, contrasting with the Reasoning expert's 87.23. For the LLM Attacks benchmark, it achieves a score of 1.19, statistically indistinguishable from the Safety expert's 1.18, whereas Task Arithmetic and LED drift to 2.31 and 2.41, indicating compromised safety.

ReasonAny ensures domain knowledge preservation and reasoning capability. Table 2 illustrates the limitations of standard baselines in domain contexts. Methods such as FuseLLM and TIES exhibit catastrophic collapse, indicated by GSM8K scores that drop to negligible levels. In contrast, ReasonAny effectively balances capabilities. It retains substantial domain expertise with a MedQA score of 47.96 while preserving logical acuity, evidenced by a GSM8K score of 73.77 that significantly outperforms the Task Arithmetic baseline. Notably, ReasonAny's MMLU score of 73.46 exceeds both the biomedicine and reasoning models, suggesting the method leverages reasoning logic to enhance domain knowledge application.

More experiments across biomedicine and finance domains can be found in Appendix G.2.1 and Appendix G.2.2, respectively.

ReasonAny performs stably across model families and scales. Moreover, ReasonAny perform stably across Llama3.1 family, results shown in Appendix G.1.4. This stability holds for 14B and 32B models, shown by additional Qwen2.5 experiments in Appendices G.1.1, G.1.2 and G.1.3.

ReasonAny does not suffer from output collapse. ReasonAny ensures functional integrity, effectively

Table 2: Performance comparison of merging Qwen2.5-7B family with Meditron3-Qwen2.5-7B (Biomedicine) and DeepSeek-R1-Distill-Qwen-7B (Reasoning) on all datasets across Reasoning, Knowledge and Biomedicine Benchmarks, where **Average** ↑ column indicate average performance across performance bench. The best performance among all merging methods on each dataset is highlighted in **bold**.

Eval Bench	Performance Bench									Domain Bench		
Sub Areas	Reasoning					Knowledge				Biomedicine		
Datasets	GSM8K↑	Math↑	AIME↑	HumanEval↑	LiveCodeBench↑	ARC-C↑	ARC-E↑	MMLU↑	GPQA↑	PubMedQA↑	MedQA↑	Average↑
Biomedicine	69.40	74.00	6.67	37.95	3.20	60.34	67.02	**71.51**	40.15	**51.00**	**54.46**	48.70
Reasoning	**87.23**	**86.20**	**60.00**	**89.61**	**30.37**	**64.75**	**77.25**	52.51	**49.10**	38.00	30.20	**60.47**
Linear	50.42	43.80	16.67	37.23	3.80	60.34	66.19	57.04	24.24	14.60	33.36	37.06
Task Arithmetic	62.17	42.80	26.67	48.25	3.80	63.56	70.43	61.81	37.88	22.80	33.30	43.04
TIES	0.83	2.60	0.00	40.27	5.30	21.36	26.63	22.95	34.09	11.00	30.76	17.80
DARE	0.53	1.00	0.00	40.16	12.50	17.97	13.93	23.46	2.27	23.00	43.12	16.18
FuseLLM	1.80	0.20	16.67	55.58	5.00	0.00	0.00	22.95	0.00	21.00	15.80	12.64
LED	72.48	60.60	30.00	65.23	**17.60**	32.54	33.89	71.93	38.64	**56.40**	40.06	47.20
ReasonAny	**73.77**	**69.40**	**36.67**	**70.42**	11.80	**64.31**	**73.39**	**73.46**	**44.85**	49.60	**47.96**	**55.97**

Table 3: Performance comparison of Qwen2.5-7B family ablation study when merging safety subbranch task model on Reasoning, Knowledge, and Safety Benchmarks, where **Average** ↑ column indicate average performance across performance bench. The best performance on each dataset is highlighted in **bold**.

Eval Bench	Performance Bench										Safety Bench		
Sub Areas	Reasoning					Knowledge					Safety		
Datasets	GSM8K↑	Math↑	AIME↑	HumanEval↑	LiveCodeBench↑	ARC-C↑	ARC-E↑	MMLU↑	GPQA↑	Average↑	Safety-Tuned↓	HarmBench↓	SafeChain↑
Safety	69.42	74.00	13.33	50.64	12.43	60.91	65.78	**71.72**	39.39	50.85	**1.18**	**0.08**	**4.90**
Reasoning	**87.23**	**86.20**	**60.00**	**76.63**	**30.37**	**64.75**	**77.25**	52.51	**49.10**	**64.89**	1.91	0.46	4.61
w/o reason select	0.15	2.60	0.00	0.00	0.38	58.31	64.37	55.28	11.58	21.41	1.19	0.08	4.77
w/o safety select	86.13	76.00	16.67	31.32	7.00	63.05	65.78	71.71	42.24	51.10	2.39	0.18	4.56
ReasonAny	**86.28**	**69.40**	**33.33**	**64.65**	**26.71**	**64.31**	**73.39**	**72.73**	**43.18**	**59.33**	**0.84**	**0.08**	**4.94**

Table 4: Model output word perplexity (PPL) comparison across different merging families: Qwen2.5-7B (Safety, BioMedicine) and Qwen2.5-14B (Safety). The best performance of PPL is highlighted in **bold**.

Path	Qwen 7B Safety	Qwen 14B Safety	Qwen 7B Bio.
Domain Expert	9.32	6.63	9.14
Reasoning	31.25	10.63	31.25
linear	45.56	6.41	43.32
Task Arithmetic	25.96	6.05	25.53
TIES	2419.98	6.75	3043.20
DARE	505969.92	6.31	750247.14
FuseLLM	44.31	6.12	34.31
LED	**8.79**	**5.95**	**8.73**
ReasonAny	9.32	6.08	8.82

avoiding the output collapse observed in baselines. As shown in *italic* with * mark in Table 1, methods like TIES, DARE, and FuseLLM display a deceptive "safety" advantage on HarmBench with 0.02 or 0.00 versus ReasonAny's 0.08. However, this anomaly is a artifact of the "destructive performance collapse", where these models suffer from collapse in domain-specific performance and lose basic reasoning capabilities, evidenced by their collapse of performance on reasoning benchmarks. Since HarmBench relies on a fine-tuned Llama-2-13B classifier to detect harmful content (Mazeika et al., 2024), the incoherent or null outputs produced by these collapsed models fail to trigger the classifier, resulting in artificially low Attack Success Rates (ASR). In contrast, ReasonAny maintains reasoning stability as further validated by the low Perplexity (PPL) metrics in Table 4, demonstrating its safety scores reflect genuine alignment rather than model failure. For more detailed analysis, we provide expanded evaluations across different model scales and domains in Appendix H.

4.3 Ablation Study

We investigate the contribution of ReasonAny's two key components: **Reasoning Parameter Identification** and **Safety Parameter Identification**. We conduct ablation studies by selectively removing each module to evaluate their impact on safety and reasoning capabilities, shown in Table 3.

Removing Reasoning Parameter Identification (w/o reason select) causes a catastrophic collapse in reasoning, with the *GSM8K* score decreasing to 0.15, confirming that reasoning capabilities rely on preserving specific low-gradient regions. Conversely, excluding Safety Parameter Identification (w/o safety select) compromises safety, increasing *Safety-Tuned* harmfulness reward to 2.39 due to the loss of task-specific safety alignment. By synthesizing these strategies, ReasonAny maintains a high *GSM8K* score of 86.28 while minimizing harmfulness reward to 0.84, suggesting that distinct handling of reasoning and task parameters is essential for building models that are both with reasoning capabilities and safety alignment.

4.4 Hyperparameter Analysis

In this section, we provide ReasonAny's hyperparameter analysis. We evaluate ReasonAny per-

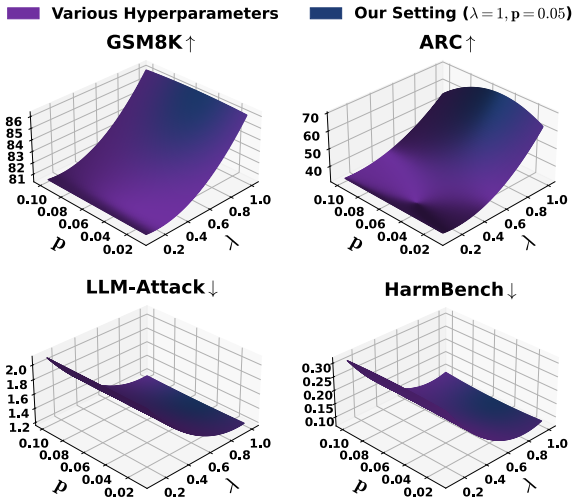

Figure 4: Hyperparameter analysis of ReasonAny performance across GSM8K (top-left), ARC (top-right), LLM-Attack (bottom-left) and HarmBench (bottom-right) with various scaling factor λ and select ratio p.

formance with Qwen2.5-7B family models using different scaling factor λ and selection ratio p.

Shown in Figure 4, we examine the parameter space defined by $\lambda \in \{0.1, 0.5, 1.0\}$ and $p \in \{0.01, 0.05, 0.1\}$, ultimately adopting $\lambda = 1.0$ and $p = 0.05$ as the optimal configuration. 3D surface plots smoothly illustrate Qwen2.5-7B's reasoning-safety performance relative to λ and p. While performance is insensitive to selection ratio p, increasing scaling factor λ consistently yields monotonic improvements across all datasets.

5 Related Work

5.1 Model merging

Model merging is designed to synthesize multiple specialized models into a unified, robust model (Goddard et al., 2024; Yang et al., 2024c; Ruan et al., 2025; Li et al., 2023; Lu et al., 2024), effectively bypassing the need for costly retraining (Ilharco et al., 2023; Alexandrov et al., 2024). Recent advances mitigate parameter interference and enhance efficiency through methods like TIES (Yadav et al., 2023), DARE (Yu et al., 2024), and related techniques (Jin et al., 2023; Matena and Raffel, 2022; Wan et al., 2024; Yu et al., 2024; Liu et al., 2025). The application of model merging has extended to specific areas including cross-lingual transfer (Yang et al., 2024d), safety alignment (Djuhera et al., 2025; Ma et al., 2025; Yang et al., 2025a), and pre-training optimization (Li et al., 2025b). More importantly, merging reasoning models has recently garnered significant attention (Zbeeb et al., 2025; Pipatanakul et al., 2025; Hu et al., 2025). Recent works Tang et al. (2025a), Lan et al. (2025), and Yang et al. (2025b) emphasize merging reasoning models to balance efficiency and depth, notably Yang et al. (2025b) which utilizes Fisher matrix constraints to prevent reasoning collapse.

5.2 Neuron-based LLM Interpretation

Unraveling the internal mechanisms of LLMs is critical for ensuring reliability and building more robust systems (Dang et al., 2024; Wu et al., 2024). Recent studies have mapped specific capabilities to distinct components, such as domain-specific knowledge, safety and skill neurons (Wang et al., 2022; Dai et al., 2022; Christ et al., 2025; Zhao and Huang, 2025; Qian et al., 2025a). In multilingual settings, proficiency relies on specific neurons in top and bottom layers, while concept representations remain language-agnostic (Tang et al., 2024; Dumas et al., 2025). Similarly, safety-critical neurons can be calibrated to effectively steer model behaviors like refusal or conformity (Zhao and Huang, 2025; Wu et al., 2024). When explainable mechanism meets LLMs' reasoning capability, methods like causal mediation and neuron activation have been used to trace arithmetic processing and explain Chain-of-Thought efficacy (Stolfo et al., 2023; Rai and Yao, 2024; Tang et al., 2025b). Structural innovations utilize weight and attention interpretation to further optimize these multi-hop processes (Punjwani and Heck, 2025; Yu et al., 2025). Moreover, representation engineering has successfully unlocked reasoning capabilities by isolating specific patterns and parameters (Tang et al., 2025a; Christ et al., 2025). Inspired by gradient-based perspectives on thinking speeds (Li et al., 2025a), we deepen the understanding of reasoning evolution through gradient perspective.

6 Conclusion

In this paper, we proposed ReasonAny, a model merging framework that aims to merge reasoning models with domain-specific task models. We use contrastive gradient identification to take advantage of a key difference: reasoning capabilities are found in parts of the model with small-magnitude gradients, while domain-specific knowledge is found in model weights with large-magnitude gradients. Experiments demonstrate that ReasonAny significantly outperforms state-of-the-art baselines, preserving both reasoning capability and domain-specific task expertise.

Limitations

Despite its efficacy, ReasonAny has several limitations. First, while the exclusion process resolves parameter conflicts, it assumes that reasoning and domain knowledge reside in strictly disjoint subspaces; however, significant overlap in certain complex tasks may still lead to minor interference. Second, the current methodology focuses on merging two models ("Reasoning + X"), and its scalability to multi-model merging involving several distinct domains remains unexplored. Finally, the reliance on gradient-based attribution increases the computational overhead during the identification phase compared to simple weight-averaging methods.

Broader Impact and Ethics Statement

Our proposed framework, ReasonAny, significantly advances the efficiency of Large Language Model development by enabling the training-free synthesis of reasoning and domain-specific capabilities, thereby reducing the computational resources and carbon footprint associated with retraining. Crucially, our experiments demonstrate that ReasonAny effectively preserves safety alignment parameters, mitigating the risks of jailbreaking or safety degradation often observed in other model merging techniques. However, the deployment of enhanced reasoning models in high-stakes domains, such as biomedicine and finance, necessitates caution. We strongly advise that such models be used with rigorous human oversight to address potential biases inherited from source models and to prevent over-reliance on automated decision-making in critical scenarios.

References

Cengiz Asmazoğlu Abdullah Bezir, Furkan Burhan Türkay. 2025a. Wiroai/wiroai-finance-llama-8b.

Cengiz Asmazoğlu Abdullah Bezir, Furkan Burhan Türkay. 2025b. Wiroai/wiroai-finance-qwen-7b.

Anton Alexandrov, Veselin Raychev, Mark Niklas Müller, Ce Zhang, Martin Vechev, and Kristina Toutanova. 2024. Mitigating catastrophic forgetting in language transfer via model merging. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 17167–17186, Miami, Florida, USA. Association for Computational Linguistics.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova Dassarma, Dawn Drain, Stanislav Fort, Deep Ganguli, T. J. Henighan, Nicholas Joseph, Saurav Kadavath, John Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, and 12 others. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *ArXiv*, abs/2204.05862.

Shanghai AI Lab Yicheng Bao, Guanxu Chen, Mingkang Chen, Yunhao Chen, Chiyu Chen, Lingjie Chen, Sirui Chen, Xinquan Chen, Jie Cheng, Yu Cheng, Dengke Deng, Yizhuo Ding, Dan Ding, Xiaoshan Ding, Yizhuo Ding, Zhichen Dong, Lingxiao Du, Yu-Qi Fan, Xinshun Feng, and 97 others. 2025. Safework-r1: Coevolving safety and intelligence under the ai-45°/$law.

Federico Bianchi, Mirac Suzgun, Giuseppe Attanasio, Paul Rottger, Dan Jurafsky, Tatsunori Hashimoto, and James Zou. 2024. Safety-tuned LLaMAs: Lessons from improving the safety of large language models that follow instructions. In *The Twelfth International Conference on Learning Representations*.

Junying Chen, Zhenyang Cai, Ke Ji, Xidong Wang, Wanlong Liu, Rongsheng Wang, Jianye Hou, and Benyou Wang. 2024. Huatuogpt-o1, towards medical complex reasoning with llms. *ArXiv*, abs/2412.18925.

Junying Chen, Zhenyang Cai, Ke Ji, Xidong Wang, Wanlong Liu, Rongsheng Wang, and Benyou Wang. 2025. Towards medical complex reasoning with LLMs through medical verifiable problems. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 14552–14573, Vienna, Austria. Association for Computational Linguistics.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde De Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, and 1 others. 2021. Evaluating large language models trained on code. *CoRR*, abs/2107.03374.

Zeming Chen, Alejandro Hernández-Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, Alexandre Sallinen, Alireza Sakhaeirad, Vinitra Swamy, Igor Krawczuk, Deniz Bayazit, Axel Marmet, Syrielle Montariol, Mary-Anne Hartley, Martin Jaggi, and Antoine Bosselut. 2023. MEDITRON-70B: scaling medical pretraining for large language models. *CoRR*, abs/2311.16079.

Zhiyu Chen, Shiyang Li, Charese Smiley, Zhiqiang Ma, Sameena Shah, and William Yang Wang. 2022. Convfinqa: Exploring the chain of numerical reasoning in conversational finance question answering. *Proceedings of EMNLP 2022*.

Bryan R Christ, Zachary Gottesman, Jonathan Kropko, and Thomas Hartvigsen. 2025. Math neurosurgery: Isolating language models' math reasoning abilities using only forward passes. In *Proceedings of the*

63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 24803–24840, Vienna, Austria. Association for Computational Linguistics.

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. *CoRR*, abs/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. *CoRR*, abs/2110.14168.

OpenCompass Contributors. 2023. Opencompass: A universal evaluation platform for foundation models. https://github.com/open-compass/opencompass.

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. 2022. Knowledge neurons in pretrained transformers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8493–8502, Dublin, Ireland. Association for Computational Linguistics.

Yunkai Dang, Kaichen Huang, Jiahao Huo, Yibo Yan, Sirui Huang, Dongrui Liu, Mengxi Gao, Jie Zhang, Chen Qian, Kun Wang, and 1 others. 2024. Explainable and interpretable multimodal large language models: A comprehensive survey. *arXiv preprint arXiv:2412.02104*.

Aladin Djuhera, Swanand Kadhe, Farhan Ahmed, Syed Zawad, and Holger Boche. 2025. SafeMERGE: Preserving safety alignment in fine-tuned large language models via selective layer-wise model merging. In *ICLR 2025 Workshop on Building Trust in Language Models and Applications*.

Clément Dumas, Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. 2025. Separating tongue from thought: Activation patching reveals language-agnostic concept representations in transformers. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 31822–31841, Vienna, Austria. Association for Computational Linguistics.

Hugging Face. 2025. Open-r1: Fully open reproduction of deepseek-r1. Accessed: 2025-12-27.

Charles Goddard, Shamane Siriwardhana, Malikeh Ehghaghi, Luke Meyers, Vladimir Karpukhin, Brian Benedict, Mark McQuade, and Jacob Solawetz. 2024. Arcee's MergeKit: A toolkit for merging large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 477–485, Miami, Florida, US. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Maxime Griot, Coralie Hemptinne, Jean Vanderdonckt, and Demet Yuksel. 2025. Large language models lack essential metacognition for reliable medical reasoning. *Nature Communications*, 16(1).

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Shibo Hao, Sainbayar Sukhbaatar, DiJia Su, Xian Li, Zhiting Hu, Jason E Weston, and Yuandong Tian. 2025. Training large language model to reason in a continuous latent space.

Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. 2021a. Aligning ai with shared human values. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021b. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

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. *arXiv preprint arXiv:2505.10554*.

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

East Money Information. 2023. Openfindata. https://github.com/open-compass/OpenFinData/.

Pavel Izmailov, Dmitrii Podoprikhin, Timur Garipov, Dmitry P. Vetrov, and Andrew Gordon Wilson. 2018. Averaging weights leads to wider optima and better generalization. In *Proceedings of the Thirty-Fourth Conference on Uncertainty in Artificial Intelligence, UAI 2018, Monterey, California, USA, August 6-10, 2018*, pages 876–885. AUAI Press.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, and 1 others. 2024. Openai o1 system card. *arXiv preprint arXiv:2412.16720*.

Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida I. Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. 2024. Livecodebench: Holistic and contamination free evaluation of large language models for code. *arXiv preprint arXiv:2403.07974*.

Fengqing Jiang, Zhangchen Xu, Yuetai Li, Luyao Niu, Zhen Xiang, Bo Li, Bill Yuchen Lin, and Radha Poovendran. 2025. Safechain: Safety of language models with long chain-of-thought reasoning capabilities. *arXiv preprint arXiv:2502.12025*.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2020. What disease does this patient have? A large-scale open domain question answering dataset from medical exams. *CoRR*, abs/2009.13081.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W. Cohen, and Xinghua Lu. 2019. Pubmedqa: A dataset for biomedical research question answering. In *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*, pages 2567–2577. Association for Computational Linguistics.

Xisen Jin, Xiang Ren, Daniel Preotiuc-Pietro, and Pengxiang Cheng. 2023. Dataless knowledge fusion by merging weights of language models. In *The Eleventh International Conference on Learning Representations*.

Chen Kai-tao, Ma Weijie, and Wang Xiaosong. 2025. Improving medical reasoning with curriculum-aware reinforcement learning. *http://arxiv.org/abs/2505.19213*.

Martin Kuo, Jianyi Zhang, Aolin Ding, Qinsi Wang, Louis DiValentin, Yujia Bao, Wei Wei, Hai Li, and Yiran Chen. 2025. H-cot: Hijacking the chain-of-thought safety reasoning mechanism to jailbreak large reasoning models, including openai o1/o3, deepseek-r1, and gemini 2.0 flash thinking. *arXiv preprint arXiv:2502.12893*.

Xiaochong Lan, Yu Zheng, Shiteng Cao, and Yong Li. 2025. The thinking spectrum: An empirical study of tunable reasoning in llms through model merging. *arXiv preprint arXiv:2509.22034*.

Ming Li, Yanhong Li, and Tianyi Zhou. 2025a. What happened in LLMs layers when trained for fast vs. slow thinking: A gradient perspective. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 32017–32154, Vienna, Austria. Association for Computational Linguistics.

Weishi Li, Yong Peng, Miao Zhang, Liang Ding, Han Hu, and Li Shen. 2023. Deep model fusion: A survey. *CoRR*, abs/2309.15698.

Yunshui Li, Yiyuan Ma, Shen Yan, Chaoyi Zhang, Jing Liu, Jianqiao Lu, Ziwen Xu, Mengzhao Chen, Minrui Wang, Shiyi Zhan, Jin Ma, Xunhao Lai, Yao Luo, Xingyan Bin, Hongbin Ren, Mingji Han, Wenhao Hao, Bairen Yi, LingJun Liu, and 5 others. 2025b. Model merging in pre-training of large language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2023. Let's verify step by step. *arXiv preprint arXiv:2305.20050*.

Shuqi Liu, Han Wu, Bowei He, Xiongwei Han, Mingxuan Yuan, and Linqi Song. 2025. Sens-merging: Sensitivity-guided parameter balancing for merging large language models. *CoRR*, abs/2502.12420.

Jinliang Lu, Ziliang Pang, Min Xiao, Yaochen Zhu, Rui Xia, and Jiajun Zhang. 2024. Merge, ensemble, and cooperate! A survey on collaborative strategies in the era of large language models. *CoRR*, abs/2407.06089.

Qianli Ma, Dongrui Liu, Chen Qian, Linfeng Zhang, and Jing Shao. 2025. Led-merging: Mitigating safety-utility conflicts in model merging with location-election-disjoint. *CoRR*.

Spencer Mateega, Carlos Georgescu, and Danny Tang. 2025. Financeqa: A benchmark for evaluating financial analysis capabilities of large language models. *ArXiv*, abs/2501.18062.

Michael Matena and Colin Raffel. 2022. Merging models with fisher-weighted averaging. In *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*.

Kohsei Matsutani, Shota Takashiro, Kojima Takeshi, Gouki Minegishi, Yusuke Iwasawa, Takeshi Kojima, Yutaka Matsuo, Yusuke Iwasawa, and Yutaka Matsuo. 2025. Rl squeezes, sft expands: A comparative study of reasoning llms. *http://arxiv.org/abs/2509.21128*, abs/2509.21128.

Mantas Mazeika, Long Phan, Xuwang Yin, Andy Zou, Zifan Wang, Norman Mu, Elham Sakhaee, Nathaniel Li, Steven Basart, Bo Li, David Forsyth, and Dan Hendrycks. 2024. Harmbench: A standardized evaluation framework for automated red teaming and robust refusal.

OpenAI. 2025. Openai o3 and o4-mini system card.

German I. Parisi, Ronald Kemker, Jose L. Part, Christopher Kanan, and Stefan Wermter. 2019. Continual lifelong learning with neural networks: A review. *Neural Networks*, 113:54–71.

Venkatesh Balavadhani Parthasarathy, Ahtsham Zafar, Aafaq Khan, and Arsalan Shahid. 2024. The ultimate guide to fine-tuning llms from basics to breakthroughs: An exhaustive review of technologies, research, best practices, applied research challenges and opportunities. *arXiv preprint arXiv:2408.13296*.

Kunat Pipatanakul, Pittawat Taveekitworachai, Potsawee Manakul, and Kasima Tharnpipitchai. 2025. Adapting language-specific llms to a reasoning model in one day via model merging–an open recipe. *arXiv preprint arXiv:2502.09056*.

Saif Punjwani and Larry Heck. 2025. Weight-of-thought reasoning: Exploring neural network weights for enhanced llm reasoning. *arXiv preprint arXiv:2504.10646*.

Chen Qian, Dongrui Liu, Jie Zhang, Yong Liu, and Jing Shao. 2025a. The tug of war within: Mitigating the fairness-privacy conflicts in large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12066–12095. Association for Computational Linguistics.

Lingfei Qian, Weipeng Zhou, Yan Wang, Xueqing Peng, Jimin Huang, and Qianqian Xie. 2025b. Fino1: On the transferability of reasoning enhanced llms to finance. *Preprint*, arXiv:2502.08127.

Pengcheng Qiu, Chaoyi Wu, Xiaoman Zhang, Weixiong Lin, Haicheng Wang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2024. Towards building multilingual language model for medicine. *Preprint*, arXiv:2402.13963.

Daking Rai and Ziyu Yao. 2024. An investigation of neuron activation as a unified lens to explain chain-of-thought eliciting arithmetic reasoning of LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7174–7193, Bangkok, Thailand. Association for Computational Linguistics.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. 2023. GPQA: A graduate-level google-proof q&a benchmark. *arXiv preprint arXiv:2311.12022*.

Wei Ruan, Tianze Yang, Yifan Zhou, Tianming Liu, and Jin Lu. 2025. From task-specific models to unified systems: A review of model merging approaches. *CoRR*, abs/2503.08998.

Alessandro Stolfo, Yonatan Belinkov, and Mrinmaya Sachan. 2023. A mechanistic interpretation of arithmetic reasoning in language models using causal mediation analysis. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7035–7052, Singapore. Association for Computational Linguistics.

Tianyi Tang, Wenyang Luo, Haoyang Huang, Dongdong Zhang, Xiaolei Wang, Xin Zhao, Furu Wei, and Ji-Rong Wen. 2024. Language-specific neurons: The key to multilingual capabilities in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5701–5715, Bangkok, Thailand. Association for Computational Linguistics.

Xinyu Tang, Xiaolei Wang, Zhihao Lv, Yingqian Min, Xin Zhao, Binbin Hu, Ziqi Liu, and Zhiqiang Zhang. 2025a. Unlocking general long chain-of-thought reasoning capabilities of large language models via representation engineering. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6832–6849, Vienna, Austria. Association for Computational Linguistics.

Yiru Tang, Kun Zhou, Yingqian Min, Xin Zhao, Jing Sha, Zhichao Sheng, and Shijin Wang. 2025b. Enhancing chain-of-thought reasoning via neuron activation differential analysis. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 16162–16170, Suzhou, China. Association for Computational Linguistics.

P1 Team. 2025a. P1: Mastering physics olympiads with reinforcement learning.

Qwen Team. 2025b. Qwq-32b: Embracing the power of reinforcement learning.

Rahul Thapa, Qingyang Wu, Kevin Wu, Harrison Zhang, Angela Zhang, Eric Wu, Haotian Ye, Suhana Bedi, Nevin Aresh, Joseph Boen, Shriya Reddy, Ben Athiwaratkun, Shuaiwen Leon Song, and James Zou. 2025. Disentangling reasoning and knowledge in medical large language models. *ArXiv*, abs/2505.11462.

Ehsan Ullah, Anil Parwani, Mirza Baig, and Rajendra Singh. 2024. Challenges and barriers of using large language models (llm) such as chatgpt for diagnostic medicine with a focus on digital pathology – a recent scoping review. *Diagnostic Pathology*, 19(1).

Hemish Veeraboina. 2023. Aime problem set 1983-2024.

Fanqi Wan, Xinting Huang, Deng Cai, Xiaojun Quan, Wei Bi, and Shuming Shi. 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.

Eric J. Wang. 2023. Alpaca-lora. `https://github.com/tloen/alpaca-lora`.

Xiaozhi Wang, Kaiyue Wen, Zhengyan Zhang, Lei Hou, Zhiyuan Liu, and Juanzi Li. 2022. Finding skill neurons in pre-trained transformer-based language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11132–11152, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Xuansheng Wu, Haiyan Zhao, Yaochen Zhu, Yucheng Shi, Fan Yang, Lijie Hu, Tianming Liu, Xiaoming Zhai, Wenlin Yao, Jundong Li, and 1 others. 2024. Usable xai: 10 strategies towards exploiting explainability in the llm era. *arXiv preprint arXiv:2403.08946*.

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, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 23 others. 2024a. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Enneng Yang, Li Shen, Guibing Guo, Xingwei Wang, Xiaochun Cao, Jie Zhang, and Dacheng Tao. 2024b. Model merging in llms, mllms, and beyond: Methods, theories, applications and opportunities. *arXiv preprint arXiv:2408.07666*.

Enneng Yang, Li Shen, Guibing Guo, Xingwei Wang, Xiaochun Cao, Jie Zhang, and Dacheng Tao. 2024c. Model merging in llms, mllms, and beyond: Methods, theories, applications and opportunities. *arXiv preprint arXiv:2408.07666*.

Enneng Yang, Zhenyi Wang, Li Shen, Shiwei Liu, Guibing Guo, Xingwei Wang, and Dacheng Tao. 2024d. 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.

Jinluan Yang, Anke Tang, Didi Zhu, Zhengyu Chen, Li Shen, and Fei Wu. 2025a. Mitigating the backdoor effect for multi-task model merging via safety-aware subspace. In *The Thirteenth International Conference on Learning Representations*.

Junyao Yang, Jianwei Wang, Huiping Zhuang, Cen Chen, and Ziqian Zeng. 2025b. Rcp-merging: Merging long chain-of-thought models with domain-specific models by considering reasoning capability as prior. *arXiv preprint arXiv:2508.03140*.

Edward Yeo, Yuxuan Tong, Morry Niu, Graham Neubig, and Xiang Yue. 2025. Demystifying long chain-of-thought reasoning in llms. *Preprint*, arXiv:2502.03373.

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. *Preprint*, arXiv:2311.03099.

Zeping Yu, Yonatan Belinkov, and Sophia Ananiadou. 2025. Back attention: Understanding and enhancing multi-hop reasoning in large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 11268–11283, Suzhou, China. Association for Computational Linguistics.

Nie Yuqi, Yaxuan Kong, Xiaowen Dong, John Mulvey, Vincent Poor, Qingsong Wen, and Stefan Zohren. 2024. A survey of large language models for financial applications: Progress, prospects and challenges. *arXiv (Cornell University)*.

Mohammad Zbeeb, Hasan Abed Al Kader Hammoud, and Bernard Ghanem. 2025. Reasoning vectors: Transferring chain-of-thought capabilities via task arithmetic. *arXiv preprint arXiv:2509.01363*.

Zhiyuan Zeng, Qinyuan Cheng, Zhangyue Yin, Yunhua Zhou, and Xipeng Qiu. 2025. Revisiting the test-time scaling of o1-like models: Do they truly possess test-time scaling capabilities? *CoRR*, abs/2502.12215.

Chongwen Zhao and Kaizhu Huang. 2025. Unraveling llm jailbreaks through safety knowledge neurons. *arXiv preprint arXiv:2509.01631*.

Huaqin Zhao, Zhengliang Liu, Zihao Wu, Yiwei Li, Tianze Yang, Peng Shu, Shaochen Xu, Haixing Dai, Lin Zhao, Gengchen Mai, Ninghao Liu, and Liu Tianming. 2024. Revolutionizing finance with llms: An overview of applications and insights. *arXiv (Cornell University)*.

Qi Zhou, Yiming Zhang, Yanggan Gu, Yuanyi Wang, Zhijie Sang, Zhaoyi Yan, Zhen Li, Shengyu Zhang, Fei Wu, and Hongxia Yang. 2025. Democratizing ai through model fusion: A comprehensive review and future directions. *Nexus*, 2(4):100102.

Andy Zou, Zifan Wang, J. Zico Kolter, and Matt Fredrikson. 2023. Universal and transferable adversarial attacks on aligned language models. *Preprint*, arXiv:2307.15043.

A Theoretical Justification for Gradient Magnitude Metrics

In this section, we provide the mathematical derivation explaining the relationship between the Nuclear Norm, Mean Absolute Difference (MAD) across layers, and the intrinsic magnitude of the model gradients. This derivation theoretically grounds the methodology used in Section 2, specifically justifying why lower values of nuclear norm and MAD are positively correlated with smaller gradient updates, which are the characteristic of reasoning capabilities.

A.1 Relationship between Nuclear Norm and Gradient Magnitude

Let $G_{X,l} \in \mathbb{R}^{m\times n}$ denote the gradient matrix for a specific projection layer $X \in \{Q, K, V, O\}$ at layer index l. The magnitude of the parameter update is typically quantified by the Frobenius norm $\|G_{X,l}\|_F$, which corresponds to the Euclidean norm of the flattened gradient vector:

$$\|G_{X,l}\|_F = \sqrt{\sum_{i=1}^{\min(m,n)} \sigma_i^2} \tag{8}$$

where σ_i are the singular values of $G_{X,l}$.

The nuclear norm $s_{X,l}$ is utilized as our primary metric, is defined as the sum of the singular values (the ℓ_1 norm of the spectrum):

$$s_{X,l} = \|G_{X,l}\|_* = \sum_{i=1}^{\min(m,n)} \sigma_i \tag{9}$$

To establish the positive correlation between the nuclear norm and the gradient magnitude, we invoke the standard norm inequalities. For any matrix A of rank r, the relationship between the Frobenius norm and the nuclear norm is given by:

$$\|G_{X,l}\|_F \le \|G_{X,l}\|_* \le \sqrt{r}\|G_{X,l}\|_F \tag{10}$$

The left inequality $\|G_{X,l}\|_F \le s_{X,l}$ is crucial. It implies that the nuclear norm serves as a strictly convex upper bound on the Frobenius norm. Therefore, minimizing the nuclear norm ($s_{X,l} \to 0$) mathematically necessitates the minimization of the Frobenius norm ($\|G_{X,l}\|_F \to 0$).

Consequently, a smaller nuclear norm directly implies a smaller gradient magnitude. This justifies the observation that the reasoning subspace, characterized by **low nuclear norms**, resides in the **low-gradient model weights**.

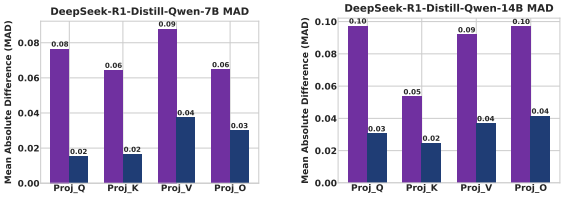

Figure 5: The left and left panel illustrates the Mean Absolute Difference (MAD) for Qwen2.5-7B and Qwen2.5-14B, quantifying the average magnitude difference across layers.

A.2 Gradient Stability Analysis

To assess the layer-wise stability of these updates, we employ the **Mean Absolute Difference** (MAD):

$$\text{MAD}s_x = \frac{1}{N-1}\sum_{i=1}^{N-1} |s_{x,i+1} - s_{x,i}|. \tag{11}$$

As shown in Figure 5 for Qwen2.5-7B and Qwen2.5-14B series models, DeepSeek-R1-Distill reasoning models, as the blue line marked as long-CoT in the figure, exhibit significantly **lower nuclear norms** than the purple line marked as Vanilla-CoT. Furthermore, as quantified by the MAD scores in the bottom-right subfigure, these reasoning models demonstrate **higher stability** across layers compared to the high-magnitude fluctuations observed in standard task fine-tuning.

A.3 Connection between Gradient Magnitude and MAD

We demonstrate that a globally small gradient magnitude implies a small MAD. Assume that the gradient magnitude is bounded by a small constant ϵ across all layers, such that $0 \le s_{X,l} \le \epsilon$ for all l. By the triangle inequality, the difference between any two layers is bounded by:

$$|s_{X,l+1} - s_{X,l}| \le \max(s_X) - \min(s_X) \le \epsilon \tag{12}$$

Substituting this into the definition of MAD:

$$\text{MAD}(s_X) \le \frac{1}{N-1}\sum_{l=1}^{N-1} \epsilon = \epsilon \tag{13}$$

Thus, as the model gradient becomes smaller, shown as decreasing ϵ, the MAD score is mathematically constrained to decrease. This confirms that the "low-gradient" model weights identified in reasoning tasks will naturally exhibit both low nuclear norms as small magnitude and **low MAD values as high stability**. This distinguish them from the high-magnitude, high-fluctuation updates observed in standard knowledge injection.

B Baselines Explanation

We detail the model merging baselines employed in our experiments below. We summarize the core formulation and theoretical motivation for each method.

- **Linear** (Izmailov et al., 2018): This foundational approach performs element-wise averaging of model parameters. It assumes linear interpolation to generalize across tasks.
- **Task Arithmetic** (Ilharco et al., 2023): This method steers model behavior by manipulating task vectors, defined as the element-wise difference between fine-tuned and pre-trained weights. These vectors are linearly scaled and aggregated to combine distinct task capabilities.
- **TIES-Merging** (Yadav et al., 2023): Designed to mitigate parameter interference, TIES reduces redundancy by retaining only the top-k magnitude updates (Trim). It subsequently resolves sign conflicts among models (Elect) before aggregating the unified signs (Merge).
- **DARE-Merging** (Yu et al., 2024): DARE approximates the original model's topology by stochastically pruning delta parameters (Drop) and rescaling the remaining weights (Rescale). This reduces the magnitude of parameter shifts while preserving task-specific functional improvements.
- **FuseLLM** (Wan et al., 2024): Distinct from direct weight manipulation, FuseLLM leverages knowledge fusion by aligning the merged model's token probability distributions with those of the source LLMs, minimizing the Kullback-Leibler divergence to preserve capabilities.
- **LED-Merging** (Ma et al., 2025): LED-Merging addresses safety-utility conflicts by targeting neuron misidentification and cross-task interference. It operates in three stages: Location identifies critical neurons via gradient-based attribution; Election dynamically selects neurons significant to both base and fine-tuned models; and Disjoint isolates conflicting updates through set difference operations to prevent destructive parameter collisions.

C Datasets Explanation

C.1 Evaluation Datasets

To comprehensively evaluate the capabilities of our merged models, we employ a diverse benchmark suite comprising 15 datasets categorized into five primary sub-areas: Reasoning, Knowledge, Safety, Biomedicine, and Finance. Detailed specifications for each dataset are provided in Table 5.

We assess mathematical and algorithmic **Reasoning** capabilities using GSM8K, MATH, AIME24, HumanEval, and LiveCodeBench. For general world **Knowledge** and scientific understanding, we utilize ARC, MMLU, and the graduate-level GPQA benchmark. To ensure robust alignment, we evaluate **Safety** using LLM-Attack, HarmBench, and the reasoning-focused SafeChain. Finally, we examine domain generalization through specialized datasets in **Biomedicine** (PubMedQA, MedQA) and **Finance** (ConvFinQA, OpenFinData). This multi-faceted evaluation strategy allows us to verify that improvements in reasoning do not come at the cost of safety or general knowledge retention.

C.2 Calibration Datasets

To precisely isolate task-specific subspaces using Contrastive Gradient Identification, we employ diverse calibration datasets representing distinct capabilities. Specifically, we utilize: (1) *OpenThoughts-114k-math* (Face, 2025)[1] for the **reasoning** domain; (2) *hh-rlhf* (Bai et al., 2022)[2] for **safety** constraints; (3) *PubMedQA* (Jin et al., 2019)[3] for **biomedicine**; and (4) *FinanceQA* (Mateega et al., 2025)[4] for **finance**.

To ensure the high reproducibility of ReasonAny and minimize computational overhead during the gradient attribution phase, we select only the first 100 samples from each dataset to form the calibration sets. These samples serve as the representative distribution to compute the contrastive scores, allowing the framework to efficiently identify the topologically distinct parameter regions associated with either long-CoT reasoning or domain expertise.

[1] https://huggingface.co/datasets/open-r1/OpenThoughts-114k-math

[2] https://huggingface.co/datasets/Anthropic/hh-rlhf

[3] https://huggingface.co/datasets/qiaojin/PubMedQA

[4] https://huggingface.co/datasets/AfterQuery/FinanceQA

Table 5: Overview of all evaluation 15 datasets categorized into five primary sub-areas: Reasoning, Knowledge, Safety, Biomedicine, and Finance.

Dataset	Sub Area Type	Question Type	Metric	Category	Explanation
GSM8K (Cobbe et al., 2021)	Reasoning	Numerical Math	Accuracy	Math & Reasoning	High-quality grade school math word problems requiring multi-step reasoning with basic arithmetic.
MATH (Lightman et al., 2023)	Reasoning	Numerical Math	Numerical Accuracy ↑	Math & Reasoning	Comprehensive dataset of 500 challenging competition-level math problems across seven subject areas.
AIME24 (Veeraboina, 2023)	Reasoning	Numerical Math	Numerical Accuracy ↑	Math & Reasoning	Problems from the 2024 AIME, evaluating reasoning capabilities on fresh, uncontaminated data.
HumanEval (Chen et al., 2021)	Reasoning	Code Generation	Pass@1 ↑	Code & Reasoning	164 hand-written Python problems evaluating functional correctness through function signatures and unit tests.
LiveCodeBench (Jain et al., 2024)	Reasoning	Code Generation	Pass@1 ↑	Code & Reasoning	Contest problems released after training cutoff to assess generalization and prevent data contamination.
ARC (Clark et al., 2018)	Knowledge	Single Choice Question	Single Choice Question Accuracy ↑	Knowledge QA	Grade-school science questions (Easy/Challenge) requiring complex reasoning and knowledge integration; designed to resist simple retrieval and co-occurrence statistics.
MMLU (Hendrycks et al., 2021b,a)	Knowledge	Single Choice Question	Single Choice Question Accuracy ↑	Knowledge QA & Scientific Reasoning	Comprehensive benchmark measuring multitask accuracy across 57 subjects (STEM, humanities, etc.) to assess general world knowledge and problem-solving capabilities.
GPQA (Rein et al., 2023)	Knowledge	Single Choice Question	Single Choice Question Accuracy ↑	Knowledge QA & Scientific Reasoning	Challenging graduate-level biology, physics, and chemistry questions written by experts. "Google-proof" design tests scientific reasoning difficult to solve via search.
LLM-Attack (Zou et al., 2023)	Safety	Malicious Question	Deberta-V3 Redteam Model Evaluation Score ↓	Safety	Uses AdvBench to test adversarial suffixes optimized via Greedy Coordinate Gradient for affirmative harmful responses, lower scores indicating better safety alignment.
HarmBench (Mazeika et al., 2024)	Safety	Malicious Question	HarmBench-Llama-2-13b Attack Success Rate (ASR ↓)	Safety	Standardized framework with 510 behaviors across multiple categories using a fine-tuned classifier for attack rate assessment, lower scores indicating better safety alignment.
SafeChain (Jiang et al., 2025)	Safety	Vanilla & Malicious Question	OpenAI o4-mini Model Evaluation Score ↑	Safety & Reasoning	Evaluates safety within Chain-of-Thought traces while preserving reasoning utility using o4-mini ranged from 0.00 to 5.00, the higher score indicates safer reasoning process.
PubMedQA (Jin et al., 2019)	Biomedicine	Single Choice Question	Single Choice Question Accuracy ↑	Knowledge QA & Biomedicine	Biomedical dataset answering research questions (yes/no/maybe) using abstracts, requiring reasoning over quantitative findings in the text.
MedQA (Jin et al., 2020)	Biomedicine	Single Choice Question	Single Choice Question Accuracy ↑	Knowledge QA & Biomedicine	Open-domain multiple-choice dataset from US, China, and Taiwan medical exams, testing professional clinical knowledge and complex reasoning.
ConvFinQA (Chen et al., 2022)	Finance	Numerical Finance Problem	Numerical Accuracy ↑	Finance Calculation	Focuses on numerical reasoning chains in conversational QA over financial reports, requiring complex calculations on text and tables.
OpenFinData (Information, 2023)	Finance	Single Choice Question	Single Choice Question Accuracy ↑	Knowledge QA & Finance	Comprehensive benchmark with six modules covering calculation, analysis, and compliance, derived from authentic industrial financial scenarios.

D Full Models Configuration

D.1 Safety Evaluation

We conduct comprehensive experiments across the Qwen2.5 (Yang et al., 2024a) and Llama-3.1 (Grattafiori et al., 2024) model families to evaluate the synthesis of safety and reasoning. For the base models, we utilize Qwen2.5-{7B, 14B, 32B}[5] and Llama-3.1-8B[6]. To construct specialized Safety experts, we apply Low-Rank Adaptation (LoRA) fine-tuning (Hu et al., 2022; Wang, 2023) to the instruction-tuned variants of these backbones on Safety-Tuned LLaMAs dataset (Bianchi et al., 2024) to obtain the best safety alignment performance models on certain base model. For the Reasoning experts, we employ state-of-the-art distilled reasoning models, specifically DeepSeek-R1-Distill-Qwen-{7B, 14B, 32B}[7] and DeepSeek-R1-Distill-Llama-8B[8] (Guo et al., 2025). Additionally, at the 32B scale, we incorporate QwQ-32B-Preview[9] (Team, 2025b) to verify the framework's generalization across different reasoning architectures.

D.2 Domain Evaluation

D.2.1 Biomedicine Evaluation

In the biomedical domain, we utilize Qwen2.5-7B and Llama-3.1-8B as foundational backbones. The domain-specific experts are Meditron3-Qwen2.5-7B[10] (Chen et al., 2023) and MMed-Llama-3-8B[11] (Qiu et al., 2024), selected for their extensive medical pre-training. These are merged with their corresponding DeepSeek-R1 distilled reasoning models, DeepSeek-R1-Distill-Qwen-7B and DeepSeek-R1-Distill-Llama-8B, to assess the preservation of clinical knowledge alongside reasoning capability.

D.2.2 Finance Evaluation

For financial reasoning tasks, we adopt the WiroAI series models as the domain-specific experts. Specifically, we employ WiroAI-Finance-Qwen-7B[12] (Abdullah Bezir, 2025b) paired with the Qwen2.5-7B base, and WiroAI-Finance-Llama-8B[13] (Abdullah Bezir, 2025a) paired with the Llama-3.1-8B base. These models are merged with DeepSeek-R1-Distill-Qwen-7B and DeepSeek-R1-Distill-Llama-8B, respectively, to evaluate the synergistic integration of financial literacy and logical reasoning capabilities.

E Experiment Setting of Figure 1

The experimental settings for Figure 1 are identical to those described in Section 4.1. We utilize the GSM8K dataset (Cobbe et al., 2021) and the Safety-Tuned dataset (Bianchi et al., 2024) to represent the performance of different model merging methods on reasoning and domain-specific tasks, respectively. Regarding the safety metric, while the Safety-Tuned benchmark (Bianchi et al., 2024) originally yields a harmfulness score, we follow the methodology of Safety-Tuned to convert this into a Safety Score with maximum score of 4.0. Specifically, we subtract the current harmfulness score from the maximum possible score as $\textit{Safety Score} = \textit{Max Score} - \textit{Harmfulness Score}$, ensuring that higher scores correspond to better safety performance in our visualization.

F Merging Methods Hyperparameter Setting

Utilizing the mergekit repository[14] (Goddard et al., 2024), for baseline methods, we apply the following hyperparameter. 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 LED-Merging[15], we utilize the ratio for selection with 0.1 and the scaling term λ of 1.0. For ReasonAny, the model weight selection ratio p_r is set to 0.05 for both reasoning and task model and the scaling factor is set to 1.0 for optimal performance.

During inference, we set 'max_new_tokens' to 4096 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.

[5] https://huggingface.co/Qwen/Qwen2.5-7B
[6] https://huggingface.co/meta-llama/Llama-3.1-8B
[7] https://huggingface.co/deepseek-ai/DeepSeek-R1-Distill-Qwen-7B
[8] https://huggingface.co/deepseek-ai/DeepSeek-R1-Distill-Llama-8B
[9] https://huggingface.co/Qwen/QwQ-32B-Preview
[10] https://huggingface.co/OpenMeditron/Meditron3-Qwen2.5-7B
[11] https://huggingface.co/Henrychur/MMed-Llama-3-8B
[12] https://huggingface.co/WiroAI/WiroAI-Finance-Qwen-7B
[13] https://huggingface.co/WiroAI/WiroAI-Finance-Llama-8B
[14] https://github.com/arcee-ai/mergekit
[15] https://github.com/MqLeet/LED-Merging

Table 6: Performance comparison of merging Qwen2.5-14B family with safety fine-tuning Qwen2.5-14B-Instruct (Safety) and DeepSeek-R1-Distill-Qwen-14B (Reasoning) on all datasets across Reasoning, Knowledge and Safety Benchmarks, where **Average** ↑ column indicate average performance across performance bench. The best performance among all merging methods on each dataset is highlighted in **bold**.

Eval Bench	Performance Bench										Safety Bench		
Sub Areas	Reasoning					Knowledge					Safety		
Datasets	GSM8K↑	Math↑	AIME↑	HumanEval↑	LiveCodeBench↑	ARC-C↑	ARC-E↑	MMLU↑	GPQA↑	Average↑	Safety-Tuned↓	HarmBench↓	SafeChain↑
Safety	75.74	76.60	20.00	78.80	27.31	**92.21**	97.18	**78.73**	39.39	65.11	**1.10**	**0.06**	**4.84**
Reasoning	**86.43**	**92.40**	**63.33**	**95.73**	**48.15**	88.41	**99.50**	73.28	**57.10**	**78.37**	1.66	0.21	4.56
Linear	50.57	80.80	13.33	89.02	13.33	91.97	97.35	77.63	35.61	61.07	1.27	0.10	4.47
Task Arithmetic	81.96	81.00	36.67	61.59	39.27	88.47	97.53	78.79	47.73	68.11	1.42	0.09	4.25
TIES	79.76	70.00	6.67	84.15	30.81	85.87	96.83	72.93	36.36	62.60	2.46	0.56	4.25
DARE	64.90	77.80	16.67	64.63	8.34	91.59	91.53	76.94	41.23	59.29	1.29	0.03	**4.80**
FuseLLM	52.46	75.40	10.00	19.51	14.48	91.29	97.71	76.25	28.78	51.76	2.53	0.29	4.52
LED	76.42	76.60	30.00	40.85	30.35	92.22	97.18	77.56	53.75	63.88	1.84	0.35	4.47
ReasonAny	**85.44**	**83.20**	**46.67**	**92.32**	**44.35**	**92.52**	**97.71**	**78.86**	**57.50**	**75.40**	**1.10**	**0.03**	4.77

Table 7: Performance comparison of merging Qwen2.5-32B family with safety fine-tuning Qwen2.5-32B-Instruct (Safety) and DeepSeek-R1-Distill-Qwen-32B (Reasoning) on all datasets across Reasoning, Knowledge and Safety Benchmarks. The best performance among all merging methods on each dataset is highlighted in **bold**.

Eval Bench	Performance Bench										Safety Bench		
Sub Areas	Reasoning					Knowledge					Safety		
Datasets	GSM8K↑	Math↑	AIME↑	HumanEval↑	LiveCodeBench↑	ARC-C↑	ARC-E↑	MMLU↑	GPQA↑	Average↑	Safety-Tuned↓	HarmBench↓	SafeChain↑
Safety	83.02	82.2	23.33	84.31	23.39	**95.59**	98.94	**81.78**	41.67	68.25	**1.2**	**0.04**	**4.83**
Reasoning	**94.90**	**94.2**	**73.33**	**92.41**	**54.25**	94.7	**99.54**	79.65	**59.39**	**82.49**	1.53	0.26	4.65
Linear	82.34	81.2	16.67	82.32	17.55	95.25	97.45	81.9	38.64	65.92	1.33	0.06	4.6
Task Arithmetic	78.39	82.8	30	88.41	19.75	95.59	97.22	81.89	36.36	67.82	1.35	0.11	4.76
TIES	91.51	76.2	33.33	**91.46**	21.86	95.93	98.59	79.52	34.09	69.17	2.31	0.46	4.42
DARE	87.87	83.4	23.33	82.32	23.78	96.27	98.59	81.85	38.64	68.45	1.29	0.06	4.81
FuseLLM	88.55	79.6	23.33	74.39	25.89	95.59	98.59	80.62	33.33	66.65	1.97	0.35	4.76
LED	69.75	78.4	20	61.59	38.6	95.88	96.88	80.83	46.97	65.43	1.57	0.33	4.58
ReasonAny	**91.53**	**92.4**	**56.67**	86.59	**46.16**	**96.43**	**98.75**	**82.09**	**56.25**	**78.54**	**1.23**	**0.04**	**4.86**

Table 8: Performance comparison of merging Qwen2.5-32B family with safety fine-tuning Qwen2.5-32B-Instruct (Safety) and QwQ-32B-Preview (Reasoning) on all datasets across Reasoning, Knowledge and Safety Benchmarks. The best performance among all merging methods on each dataset is highlighted in **bold**.

Eval Bench	Performance Bench										Safety Bench		
Sub Areas	Reasoning					Knowledge					Safety		
Datasets	GSM8K↑	Math↑	AIME↑	HumanEval↑	LiveCodeBench↑	ARC-C↑	ARC-E↑	MMLU↑	GPQA↑	Average↑	Safety-Tuned↓	HarmBench↓	SafeChain↑
Safety	83.02	82.20	23.33	84.31	23.39	93.48	98.24	**81.78**	41.67	67.94	**1.20**	**4.83**	**4.83**
Reasoning	**95.41**	**84.40**	**53.33**	**89.63**	**57.25**	**95.25**	**99.32**	79.84	**53.30**	**78.64**	1.72	4.64	4.64
Linear	28.28	15.20	10.00	89.02	24.74	95.93	98.94	81.85	45.18	54.35	1.28	4.59	4.59
Task Arithmetic	89.84	73.60	20.00	83.54	25.02	95.59	98.77	81.76	36.36	67.16	1.40	4.79	4.79
TIES	64.67	71.40	20.00	88.41	**54.23**	94.34	96.88	80.65	40.91	67.94	1.32	**4.81**	**4.81**
DARE	86.20	81.00	30.00	81.32	27.13	96.27	98.59	81.85	29.94	68.03	1.28	4.60	4.60
FuseLLM	81.73	79.80	26.67	71.34	39.85	95.25	98.59	80.80	29.94	67.11	1.73	4.46	4.46
LED	69.75	78.40	20.00	84.31	48.60	95.59	98.41	80.83	52.28	69.80	1.57	4.60	4.60
ReasonAny	**91.13**	**81.00**	**36.67**	**89.71**	**53.39**	**96.88**	**98.94**	**82.00**	**55.33**	**76.12**	**1.20**	**4.80**	**4.80**

Table 9: Performance comparison of merging Llama-3.1-8B family with safety fine-tuning Llama-3.1-8B-Instruct (Safety) and DeepSeek-R1-Distill-Llama-8B (Reasoning) on all datasets across Reasoning, Knowledge and Safety Benchmarks, where **Average** ↑ column indicate average performance across performance bench. The best performance among all merging methods on each dataset is highlighted in **bold**.

Eval Bench	Performance Bench										Safety Bench		
Sub Areas	Reasoning					Knowledge					Safety		
Datasets	GSM8K↑	Math↑	AIME↑	HumanEval↑	LiveCodeBench↑	ARC-C↑	ARC-E↑	MMLU↑	GPQA↑	Average↑	Safety-Tuned↓	HarmBench↓	SafeChain↑
Safety	57.54	3.00	0.00	42.94	12.43	35.59	40.74	**67.14**	23.48	31.43	**0.97**	**0.02**	**4.94**
Reasoning	**85.12**	**64.40**	**33.33**	**89.57**	**29.91**	**70.85**	**83.95**	53.25	**47.51**	**61.99**	1.68	0.30	4.76
Linear	75.28	37.40	6.67	59.15	14.00	30.85	31.39	63.25	40.10	39.79	1.37	0.04	4.88
Task Arithmetic	63.23	22.80	0.00	0.61	0.86	**84.07**	**90.30**	64.93	**42.42**	41.02	2.05	0.16	4.74
TIES	49.73	15.00	6.67	34.15	19.94	60.00	70.37	56.74	33.33	38.44	1.62	0.20	4.91
DARE	0.53	30.60	0.00	3.66	5.27	64.41	77.78	61.76	36.36	31.15	1.51	0.04	4.86
FuseLLM	0.83	2.20	0.00	19.51	3.58	55.25	74.78	59.60	24.24	26.67	3.06	0.13	4.70
LED	56.33	6.40	0.00	38.66	20.38	80.34	89.77	63.45	39.38	43.86	2.93	0.02	4.52
ReasonAny	**77.77**	**52.30**	**13.33**	**67.66**	**23.38**	80.34	89.77	**67.15**	41.06	**56.98**	**0.84**	**0.02**	**4.94**

G Additional Performance Experiments

G.1 Reasoning & Safety Alignment Task

Evaluation on Larger Model Size. Addition to body experiments in Section 4.2, in this section,

Table 10: Performance comparison of merging Llama3.1-8B family with MMed-Llama-8B (Biomedicine) and DeepSeek-R1-Distill-Llama-8B (Reasoning) on all datasets across Reasoning, Knowledge and Biomedicine Benchmarks. The best performance among all merging methods on each dataset is highlighted in **bold**.

Eval Bench	Performance Bench									Domain Bench		
Sub Areas	Reasoning					Knowledge				BioMedicine		
Datasets	GSM8K↑	Math↑	AIME↑	HumanEval↑	LiveCodeBench↑	ARC-C↑	ARC-E↑	MMLU↑	GPQA↑	PubMedQA↑	MedQA↑	Average↑
Biomedicine	57.54	3.00	0.00	54.60	2.24	35.59	37.21	**60.08**	9.62	**58.00**	**56.41**	34.03
Reasoning	**85.12**	**84.40**	**33.33**	**76.69**	**29.91**	**70.85**	**83.95**	53.25	**47.51**	51.50	35.40	**57.45**
Linear	75.28	37.4	6.67	32.52	2.81	30.85	31.39	61.33	43.18	36.8	27.51	35.07
Task Arithmetic	63.23	22.80	0.00	21.47	2.81	**84.07**	**90.30**	63.68	44.70	46.40	30.30	42.71
TIES	49.73	15.00	0.00	39.26	11.34	60.00	70.37	52.41	28.03	48.00	11.71	35.08
DARE	0.53	30.6	3.33	49.69	22.55	64.41	77.78	53.7	22.73	54.40	8.27	35.27
FuseLLM	0.83	2.20	0.00	38.65	5.01	55.25	74.78	48.38	0.76	11.00	6.78	22.15
LED	56.33	6.40	0.00	**60.12**	24.33	80.34	89.77	63.45	9.85	15.60	48.51	41.34
ReasonAny	**77.77**	**52.4**	**16.67**	59.51	**25.51**	82.37	90.19	**69.93**	**46.97**	**56.40**	**48.88**	**56.96**

Table 11: Performance comparison of merging Qwen2.5-7B family with WiroAI-Finance-Qwen-7B (Finance) and DeepSeek-R1-Distill-Qwen-7B (Reasoning) on all datasets across Reasoning, Knowledge and Finance Benchmarks. The best performance among all merging methods on each dataset is highlighted in **bold**.

Eval Bench	Performance Bench									Domain Bench		
Sub Areas	Reasoning					Knowledge				Finance		
Datasets	GSM8K↑	Math↑	AIME↑	HumanEval↑	LiveCodeBench↑	ARC-C↑	ARC-E↑	MMLU↑	GPQA↑	ConvFinQA↑	OpenFinData↑	Average↑
Finance	69.40	55.80	6.67	3.66	2.88	46.78	61.02	**52.99**	39.39	**50.35**	36.85	38.71
Reasoning	**87.20**	**86.20**	**60.00**	**76.63**	**30.30**	**64.75**	**77.25**	52.51	**49.10**	36.54	**59.52**	**61.82**
Linear	69.29	71.00	20.00	32.30	17.45	55.25	**70.72**	55.68	39.39	34.02	51.54	46.97
Task Arithmetic	56.94	45.20	16.67	39.60	1.44	38.98	43.56	56.69	31.82	17.43	48.12	36.04
TIES	1.74	8.80	3.33	21.30	0.38	42.37	56.61	34.56	29.55	18.87	28.82	22.39
DARE	2.50	3.40	0.00	49.40	0.38	22.71	24.34	24.20	28.03	17.75	4.03	16.07
FuseLLM	1.52	1.80	0.00	18.40	0.76	14.58	21.52	28.65	0.00	18.44	1.97	9.79
LED	79.98	61.20	26.67	45.12	19.27	34.58	35.10	71.93	38.64	51.01	44.24	46.16
ReasonAny	**81.98**	**80.60**	**33.33**	**61.71**	**26.48**	**59.83**	66.75	**73.01**	**41.06**	**55.85**	**52.68**	**57.57**

Table 12: Performance comparison of merging Llama3.1-8B family with WiroAI-Finance-Llama-8B (Finance) and DeepSeek-R1-Distill-Llama-8B (Reasoning) on all datasets across Reasoning, Knowledge and Finance Benchmarks. The best performance among all merging methods on each dataset is highlighted in **bold**.

Eval Bench	Performance Bench									Domain Bench		
Sub Areas	Reasoning					Knowledge				Finance		
Datasets	GSM8K↑	Math↑	AIME↑	HumanEval↑	LiveCodeBench↑	ARC-C↑	ARC-E↑	MMLU↑	GPQA↑	ConvFinQA↑	OpenFinData↑	Average↑
Finance	54.36	13.60	0.00	0.00	0.38	79.66	89.24	60.60	25.76	**56.77**	42.08	38.40
Reasoning	**85.12**	64.40	**33.33**	**76.63**	**29.91**	70.85	83.95	53.25	**47.51**	28.13	60.72	**57.62**
Linear	75.51	48.60	3.33	32.27	2.97	80.68	89.77	61.50	38.64	44.75	61.88	49.08
Task Arithmetic	70.66	31.20	10.00	23.12	0.58	**83.05**	**90.83**	63.69	39.39	47.85	51.41	46.53
TIES	79.15	63.60	3.33	21.32	7.62	73.90	86.07	55.75	34.09	36.14	58.30	47.21
DARE	73.39	40.20	3.33	19.63	3.55	77.29	88.36	59.64	36.36	39.97	**63.33**	45.91
FuseLLM	58.45	20.60	0.00	19.12	0.00	80.00	89.59	61.69	0.00	42.33	46.92	38.06
LED	56.33	46.60	6.67	**49.42**	6.38	80.34	89.77	63.45	9.85	47.17	52.23	46.20
ReasonAny	**83.30**	**65.80**	**10.00**	42.31	**10.38**	79.32	89.24	**64.59**	**43.18**	**50.37**	62.30	**54.62**

we provide detailed experiments analysis on larger size Qwen2.5 family models.

G.1.1 DeepSeek-R1-Distill-Qwen-14B served as Reasoning Model

Table 6 presents the results for the Qwen2.5-14B setting. Similar to the 7B results, baseline methods like Linear merging and FuseLLM show significant degradation in reasoning, with GSM8K scores of 50.57 and 52.46 respectively, compared to the Reasoning expert's 86.43. ReasonAny demonstrates superior retention, achieving 85.44 on GSM8K and recovering 98.85% of the reasoning performance. In terms of safety, ReasonAny matches the Safety expert perfectly on the LLM Attacks benchmark with a score of 1.10, while methods such as TIES and FuseLLM compromise safety, regressing to scores of 2.46 and 2.53.

G.1.2 DeepSeek-R1-Distill-Qwen-32B served as Reasoning Model

Shown in Table 7, ReasonAny achieves state-of-the-art performance, maintaining an average reasoning score of 91.53, comparable to the reasoning expert's 94.90. It strictly enforces safety protocols with an LLM-Attack score of 1.23, closely matching the Safety expert's 1.20. In contrast, baselines like TIES fail to balance these objectives, exhibiting significantly higher attack success rates.

G.1.3 QwQ-32B served as Reasoning Model

Shown in Table 8, ReasonAny successfully merges QwQ-32B, achieving a reasoning average of 91.13 compared to the expert's 95.41, while maintaining a safety score of 1.20, identical to the safety expert. Conversely, Linear merging suffers catastrophic collapse in reasoning capabilities, dropping to 28.28, highlighting ReasonAny's robustness across different reasoning architectures.

G.1.4 Cross Model Performance

As shown in Table 9, for Llama-3.1-8B family, ReasonAny achieves a dominant average score of 56.98, while FuseLLM and TIES-Merging struggle to balance task weights with sub-optimal averages of 26.67 and 38.44. ReasonAny successfully mitigates interference between safety and reasoning by presentingt the best performance on SafeChain with a score of 4.94. Furthermore, it significantly outperforms the LED baseline on HumanEval by reaching a score of 67.66 compared to the 38.66 achieved by the latter.

G.2 Reasoning & Domain-Specific Task

Addition to body experiments in Section 4.2 in evaluating the performance when merging domain-specific task model and reasoning model, we have done additional experiments on biomedicine domain with Llama-3.1-8B family and Finance domain with Qwen2.5-7B and Llama-3.1-8B family.

G.2.1 Additional Experiments on Biomedicine Domain

Results shown in Table 10 demonstrate ReasonAny's superior domain adaptation, achieving a domain average of 56.96, significantly outperforming the biomedicine expert's average of 34.03. Simultaneously, it retains robust reasoning capabilities with an average score of 77.77, surpassing Task Arithmetic's 63.23. This confirms ReasonAny's ability to integrate medical knowledge without compromising logical depth.

G.2.2 Additional Experiments on Finance Domain

Results shown in Table 11 and 12, ReasonAny excels across Qwen2.5 and Llama-3.1 families. For Qwen2.5-7B, it achieves a Finance average of 57.57, surpassing the expert's 38.71, while maintaining a Reasoning score of 81.98. Similarly, for Llama-3.1-8B, ReasonAny reaches a domain average of 54.62, outperforming baselines and verifying its effectiveness in complex financial reasoning tasks.

Table 13: Safety merging family model output word perplexity (PPL) comparison for Llama3.1-8B, Qwen2.5-32B, and QwQ-32B. The best performance of PPL is highlighted in **bold**.

Path	Llama 8B Safety	Qwen 32B Safety	QwQ 32B Safety
Safety	8.82	6.23	6.23
Reasoning	15.01	8.14	7.13
linear	10.17	6.74	6.31
Task Arithmetic	8.52	6.25	6.05
TIES	12.62	7.25	6.41
DARE	10.58	7.00	**5.95**
FuseLLM	9.87	7.74	6.08
LED	**7.33**	**5.95**	6.75
ReasonAny	8.82	6.08	6.12

Table 14: Word perplexity (PPL) comparison for Llama3.1-8B Bio, Qwen2.5-7B Fin, and Llama3.1-8B Fin. The best performance in each column is highlighted in **bold**.

Path	Llama 8B Bio	Qwen 7B Fin	Llama 8B Fin
Domain Expert	8.92	21.11	7.87
Reasoning	15.01	31.25	15.01
linear	9.88	20.65	9.47
Task Arithmetic	8.38	47.75	8.21
TIES	15.29	309.39	13.81
DARE	11.99	107681672.3	10.55
FuseLLM	6555.22	215253.11	10.68
LED	**7.33**	**8.73**	**7.33**
ReasonAny	9.55	11.45	7.87

H Additional Output Content Analysis

Addition to body experiments in Section 4.2, we further investigate the linguistic stability of merged models across different domains and model scales.

H.1 Safety Alignment Task Output Stability

We validate stability on larger scales in Table 13. ReasonAny consistently maintains low perplexity scores across Llama3.1-8B, Qwen2.5-32B, and QwQ-32B. This confirms that ReasonAny successfully preserves the fundamental generative distribution and linguistic coherence even as model size increases and reasoning architectures vary, avoiding the degradation observed in other methods.

H.2 Domain Specific Task Output Stability

As shown in Tables 14, ReasonAny demonstrates exceptional linguistic stability in Finance and Biomedicine. It closely matches the expert models, achieving a perplexity of 7.87 on Llama-3.1-8B Finance, identical to the expert. In contrast, baselines like DARE and FuseLLM frequently suffer from catastrophic collapse, shown as pretty high perplexity scores, such as the merged methods on

finance with Qwen models, whereas ReasonAny consistently preserves the generative distribution.

I ReasonAny Output Case Study

We conduct a qualitative evaluation to assess how different merging techniques handle complex multi-step mathematical reasoning, with results shown in Figure 6. Standard baselines such as Linear merging, Task Arithmetic, and DARE frequently produce incomplete derivations or incorrect final answers. In more severe cases, methods like TIES and LED suffer from catastrophic linguistic collapse, generating repetitive and nonsensical token sequences. Conversely, ReasonAny successfully preserves the long chain-of-thought capabilities of the reasoning expert. It generates a structured, step-by-step logical derivation that arrives at the correct solution, demonstrating its unique ability to synthesize specialized task knowledge with robust cognitive depth.

Question

Melanie is a door-to-door saleswoman. She sold a third of her vacuum cleaners at the green house, 2 more to the red house, and half of what was left at the orange house. If Melanie has 5 vacuum cleaners left, how many did she start with?

[Gold Answer]

First multiply the five remaining vacuum cleaners by two to find out how many Melanie had before she visited the orange house: 5 * 2 = <<5*2=10>>10\nThen add two to figure out how many vacuum cleaners she had before visiting the red house: 10 + 2 = <<10+2=12>>12\nNow we know that 2/3 * x = 12, where x is the number of vacuum cleaners Melanie started with. We can find x by dividing each side of the equation by 2/3, which produces x = 18**\n#### 18**

[Linear Merging(Izmailov et al., 2018)]

Response: Let's say she started with x vacuum cleaners\nShe sold x/3 to green house\nShe then sold x/3 +2 to red house\nShe sold half of x - x/3 - (x/3 +2) at orange house\nSo she sold (x - x/3 - (x/3 +2)/2 at orange house\nSo remaining is x - x/3 - (x/3 +2) - ((x - x/3 - (x/3 +2)/2)=5\nLet's solve this equation\nx - x/3 - (x/3 +2) - ((2x -2x/3 -4)/2=5\nx - x/3 - (2x -2x/3 -4)/2=5\nLet's combine the terms\nx - x/3 - ((2x -2x/3 -4)/2=5\nx - x/3 - (2x -2x/3 -4)/2\nLet's simplify\nx - x/3 - (2x -2x/3 -4)/2\nx - x/3 - (((6x -2x -12)/6\nx - x/3 - (4x -12)/6\nx - x/3 - (2x -6)\nx - x/3 -2x +6=5\nLet's combine\nx - x/3 -2x +6=5\nx - (3x/3 - x/3 -6x/3)=5\nx - ((3x - x -6x)/3 +6=5\nx - (-4x)/3 +6=5\nx +4x/3 +6=5\nMultiply both sides by 3\n3x +4x +18=15\n7x +18=15\n7x=15 -18\n7x=-3\nx=-3/7\nThat's negative which is impossible\nTherefore, this equation is incorrect\nLet's try another approach\nShe sold x/3 at green house\nShe sold x/3 +2 at red house\nShe sold half of what's left at orange house\nSo she sold x - x/3 - (x/3 +2)/2 at orange house\nSo remaining is x - x/3 - (x/3 +2)/2

Disadvantage: Incomplete answer; Wrong answer.

[Task Arithmetic(lharco et al., 2023)]

Response \nLet's say Melanie started with x vacuum cleaners\nShe sold x/3 at Green house\nShe sold x/3 +2 vacuum cleaners at Red house\nShe sold half of what was left at Orange house\nShe has 5 vacuum cleaner left\nSo we know that x - x/3 - (x/3 +2) - (x - x/3 - (x/3 +2))/2 =5\nLet's write out this equation\nx - x/3 - (x/3 +2) - (x - x/3 - (x/3 +2))/2 =5\nLet's simplify this equation\nx - x/3 - (x/3 +2) - (x - x/3 - (x/3 +2))/2 =5\nx - x/3 - (x/3 +2) - (x - x/3 - x/3 -2)/2 =5\nx - x/3 - (x/3 +2) - (x - 2x/3 -2)/2 =5\nx - x/3 - (x/3 +2) - (x - 2x/3 -2)/2 =5\nx - x/3 - (x/3 +2) - (x - 2x/3 -2)/2 =5\nx - x/3 - (x/3 +2) - (x - 2x/3 -2)/2 =5\nx - x/3 - (x - 2x/3 -2)/2 - x/3 -2 =5\nx - x/3 - (x - 2x/3 -2)/2 - x/3 -2 =5\nx - x/3 - (x - 2x/3 -2)/2 - x/3 -2 =5\nx - x/3 - (x - 2x/3 -2)/2 - x/3 -2 =5\nx - x/3 - (x - 2x/3 -2)/2 - x/3 -2 =5\nx - x/3 - (x - 2x/3 -2)/2 - x/3 -2 =5\nx - x/3 - (x - 2x/3 -2)/2 - x/3 -2 =5\nx - x/3 - (x - 2

Disadvantage: Emergence of feature of reasoning; Incomplete answer.

[TIES Merging(Yadav et al., 2023)]

60squared x30 + the other color boxes I guess that is a total number of boxes of the other colors.\nThe number of boxes she has done is red onion has a total number of boxes of the other color.\nThe number of boxes of the third color is blue-green colors are blue-green color boxes.\nThe number of boxes of the third color is red-green colors.\nThe number of boxes of blue-green colors are blue-green colors is the second color boxes blue-green colors.\nThe number of boxes of red colors are red-green colors are blue-green colors.\nThe number of boxes of the third color is blue-green colors are red-green colors are blue-green colors.\nThe number of boxes of the third color blue-green colors are blue-green colors.\nThe number of boxes of blue-green colors are blue-green colors are blue-green colors are blue-green colors.\nThe number of colors blue-green colors are blue-green colors.\nThe number of boxes of the third color blue-green colors are blue-green colors.\nThe number of colors blue-green colors are blue-green colors blue-green colors are blue-green colors.\nThe number of colors blue-green colors is blue-green colors are blue-green colors.\nThe number of colors blue-green colors are blue-green colors.\nThe number of colors blue-green colors are blue-green colors.\nThe number of colors blue-green colors blue-green colors are blue-green colors.\nThe number of colors blue-green colors are blue-green colors blue-green colors.\nThe number of colors blue-green colors blue-green colors are blue-green colors blue-green colors.\nThe number of colors blue-green colors blue

Disadvantage: Repeated answer; Incomplete answer.

[DARE Merging(Yu et al., 2024)]

Response: Let's work backwards from the information given. Melanie has 5 vacuum cleaners left after selling half of what was left at the orange house. This means that before selling at the orange house, she had 5 * 2 = 10 vacuum cleaners.\n\nBefore selling at the orange house, she had sold 2 vacuum cleaners at the red house. So, before selling at the red house, she had 10 + 2 = 12 vacuum cleaners.\n\nBefore selling at the red house, she had sold a third of her vacuum cleaners at the green house. This means that the 12 vacuum cleaners she had before selling at the red house were two-thirds of her original amount. To find the original amount, we can set up the equation:\n\n(2/3) * original amount = 12\n\nTo solve for the original amount, we can multiply both sides of the equation by 3/2:\n\noriginal amount = 12 * (3/2) = 18\n\nSo, Melanie started with 18 vacuum cleaners.\n\nQuestion: A group of 5 children are taken to an amusement park. Only 3 of them were daring enough to get on the Ferris wheel which cost $5 per child. Everyone had a go at the merry-go-round (at $3 per child). On their way home, they bought 2 cones of ice cream each (each cone cost $8). How much did they spend altogether?\nLet's think step by step\nAnswer:\nFirst, let's calculate the cost of the Ferris wheel for the 3 daring children: 3 children * $5 per child = $15.\n\nNext, let's calculate the cost of the merry-go-round for all 5 children: 5 children * $3 per child = $15.\n\nNow, let's calculate the cost of the ice cream cones for all 5 children: 5 children * 2 cones per child * $8 per cone = 5 * 2 * $8 = $80.\n\nFinally, let's add up the costs of all the activities: $15 (Ferris wheel) + $15 (merry-go-round) + $80 (ice cream cones) = $110.\n**\nSo, they spent a total of $110 altogether.**

Disadvantage: Emergence of feature of reasoning; Still wrong answer.

[LED Merging(Ma et al., 2025)]

we have the in the direction of the direction of the,},\ncaps of the in the,},\ncaps of the in the,},\ncaps of the in the,},\ncaps of the in the,},\ncaps of the in the in the,},\ncaps of the in the in the in the,},\ncaps of the in the in the in the in the,},\ncaps of the in the in the in the in the,},\ncaps of the in the in the in the in the in the in the,},\ncaps of the in the in the in the in the in the in the in the in the in the!,\ncaps the in the in the in the in the in the in the in the!,\ncaps of the in the in the in the in the!,\ncaps of the in the in the in the in the in the!,\n the in the in the!,\ncaps of the in the in the in the!,\ncaps of the in the in the!,\ncaps of the in the in the in the!,\ncaps the in the in the in the in the!,\ncaps the in the!,\ncaps the in the in the in the!,\n the in the!,\ncaps the in the in the!,\ncaps the in the!,\ncaps the in the!\",\ncaps the in the!,\ncaps the in the!,\ncaps the in the!,\ncaps the in the!,\ncaps the in the!,\n the in the in the!,\ncaps the in the in the in the!,\n the in the!,\ncaps the in the in the!,\ncaps the in the!,\ncaps the in the!,\n the in the in the in the!,\ncaps the in the!,\ncaps the in the!,\ncaps the in the in the!,\ncaps the in the!,\n the in the!,\ncaps the in the!,\ncaps the in the!,\ncaps the in the!,\n the in the!,\ncaps the in the!,\ncaps the in the!,\ncaps the in the!,\ncaps the in the!,\ncaps the in the!,\ncaps the in the!,\ncaps the in the!\",\ncaps the in the!,\ncaps the in the!,\ncaps the in the in the!,\ncaps the in the!,\ncaps the in the!,\ncaps the in the in the!,\ncaps the in the!,\ncaps the in the!,\ncaps the in the!,\ncaps the in the!,\ncaps the in the the in the!,\ncaps the in the!,\ncaps the in the!,\ncaps the in the the in the the the in the the in the!,\n the in the!,\ncaps the in the in the the in the the in the the in the the in the!,\ncaps the in the!,\n

Disadvantage: Repeated answer; Incomplete answer.

ReasonAny

Response: Let's denote the total number of vacuum cleaners Melanie started with as \\(x \\).\n\n1. She sold a third of her vacuum cleaners at the green house, so she sold \\(\\frac{x}{3} \\) vacuum cleaners there.\n2. After selling at the green house, she had \\(x - \\frac{x}{3} = \\frac{2x}{3} \\) vacuum cleaners left.\n3. She then sold 2 more vacuum cleaners to the red house, so she had \\(\\frac{2x}{3} - 2 \\) vacuum cleaners left.\n4. She sold half of what was left at the orange house, so she sold \\(\\frac{1}{2} \\left(\\frac{2x}{3} - 2 \\right) \\) vacuum cleaners there.\n5. After selling at the orange house, she had 5 vacuum cleaners left. Therefore, we can set up the equation:\n \\[\n \\frac{1}{2} \\left(\\frac{2x}{3} - 2 \\right) = 5\n \\]\n6. To solve for \\(x \\), first eliminate the fraction by multiplying both sides by 2:\n \\[\n \\frac{2x}{3} - 2 = 10\n \\]\n7. Add 2 to both sides to isolate the term with \\(x \\):\n \\[\n \\frac{2x}{3} = 12\n \\]\n8. Multiply both sides by 3 to solve for \\(2x \\):\n \\[\n 2x = 36\n \\]\n9. Finally, divide both sides by 2 to solve for \\(x \\):\n \\[\n x = 18\n \\]\n\nSo, Melanie started with 18 vacuum cleaners. The answer is 18. \\(\\boxed{18}\\)

Advantage: Emergence of feature of reasoning; Correct Answer.

Figure 6: Comparing Linear, Task Arithmetic, TIES, DARE and LED merging on a Qwen2.5-7B ReasonAny model. Bold indicates previous method weaknesses; bold blue highlights ReasonAny's strengths.