We thank all reviewers for their constructive comments! Please find our responses below.

*Table 1.* Comparing performance of different model merging methods on Qwen 1.5 4B model. We added LoRA-Merging for Reviewer aCDT, AdaMerging and WeMOE for Reviewer Q1B5 to provide comprehensive comparisons.

| Algorithm/Tasks | GSM8K | TriviaQA | Winogrande | HumanEval | MMLU | All Tasks Average |
|---|---|---|---|---|---|---|
| base | 47.16 | 44.54 | 56.75 | **41.46** | 54.45 | 48.87 |
| Math | **51.00** | 46.95 | 54.62 | 26.83 | 53.54 | 46.79 |
| Code | 43.29 | 46.39 | 54.14 | 43.29 | **54.82** | 48.39 |
| QA | 45.56 | **48.02** | **57.93** | 39.02 | 52.32 | 48.57 |
| all-sft | 48.52 | 47.73 | 55.88 | 39.14 | 53.93 | **49.04** |
| TIES | 47.76 | 46.59 | 54.14 | 44.51 | 54.58 | 49.5 |
| **AdaMerging** | 47.46 | 47.90 | 54.38 | 44.51 | 54.62 | 49.59 |
| PCB-merging | 47.83 | 47.60 | 56.75 | 43.90 | 54.58 | 49.93 |
| Twin-merging | 47.99 | 44.63 | 57.54 | 40.85 | 52.98 | 48.80 |
| ***LoRA-Merging*** | 48.02 | 44.69 | 57.61 | 40.85 | 53.01 | 48.87 |
| BTX | 48.44 | 46.94 | 57.77 | 42.68 | 53.88 | 49.94 |
| **WeMOE** | 47.83 | 47.84 | 53.99 | **45.73** | 54.70 | 49.83 |
| Mediator | **50.94** | **48.20** | **57.85** | **45.12** | **54.87** | **51.40** |

- **Lora-Merging by reviewer Reviewer #aCDT**. Table 1 show that while LoRA merging slightly outperforms Twin-merging, both matrix decomposition approaches introduce additional approximation errors. In contrast, **our merging method achieves better performance by avoiding these decomposition errors and preserving the original parameter values more faithfully.
- **AdaMerging and WeMOE by reviewer #Q1B5**. We also add experiments to compare with WEMoE and Adamerging. The following table shows that Mediator achieved the best performance through global task-level routing and enhanced classification precision that ensures accurate expert selection for non-OOD tasks; Similar performance gaps are also observed when comparing with AdaMerging and PCB-Merging.

### Mediator routing results by reviewer Reviewer #aCDT

The table shows the total number of layers requiring routing for each model architecture:

- Qwen-4B requires routing in 6 layers total:

  - 5 layers in lower positions (layers 0-4)

  - 1 layer in higher position (layer 39)
- Qwen-7B requires routing in 7 layers total:

  - 6 layers in lower positions (layers 0-5 and 9)

  - 1 layer in higher position (layer 27)
- LLaMA-3.2-3B requires routing in 5 layers total:

  - 4 layers in lower positions (layers 0,1,2,5)

  - 1 layer in higher position (layer 31)
- LLaMA-3.1-8B requires routing in 7 layers total:

  - All 7 routing layers are in lower positions (layers 0-6)

  - No routing needed in higher layers

This analysis shows that across different model architectures, the number of layers requiring routing remains relatively consistent (5-7 layers), with most routing concentrated in the lower layers.

*Table 2.* Layer-wise routing strategy for different model architectures

| Models | Lower routing layers | Higher routing layers |
|---|---|---|
| Qwen-4B | 0-4 | 39 |
| Qwen-7B | 0-5, 9 | 27 |
| llama-3.2-3B | 0,1,2,5 | 31 |
| llama-3.1-8B | 0-6 | NA |

### Mediator routing results by reviewer Reviewer aCDT

The detailed layer-wise merging strategy is available in Table 2

### Question: *More insight in COT experiments* (by reviewer #sV1X).

Following the suggestions of reviewer #sV1X, we redraw the Fig 3 as shown in Fig. 1 as follows.

Table 3 and Table 4 show the detailed results of our experiments on these challenging benchmarks and also compares the results between SFT models on COT and non-cot datasets.

Our ablation studies demonstrate the domain-specific benefits of Chain-of-Thought (CoT) prompting:

- Removing Math CoT mainly affects math tasks (GSM8K: -2.56%, GSM-Plus-mini: -2.26%) while having minimal impact on coding performance.
- Similarly, removing Code CoT primarily impacts coding tasks (HumanEval: -0.61%, MBPP: -1.00%) with negligible effect on math reasoning.

On the challenging benchmarks, Mediator maintains strong performance:

- For GSM-Plus-mini, Mediator (36.25%) outperforms the base model by 3.42
- On MBPP, Mediator achieves 34.20%, showing a 1.00% improvement over the base model.

These results validate our method's effectiveness on complex reasoning tasks.

*Table 3.* Results on additional challenging reasoning benchmarks

| Algorithm/Tasks | GSM-Plus-mini | MBPP |
|---|---|---|
| Base | 32.83 | 33.20 |
| Math Expert | 36.71 | 32.8 |
| Code Expert | 34.12 | 34.2 |
| TIES | 34.12 | 34.00 |
| Twin-merging | 34.12 | 34.00 |
| PCB-merging | 34.12 | 34.20 |
| BTX | 35.79 | 34.00 |
| Mediator | **36.25** | **34.20** |

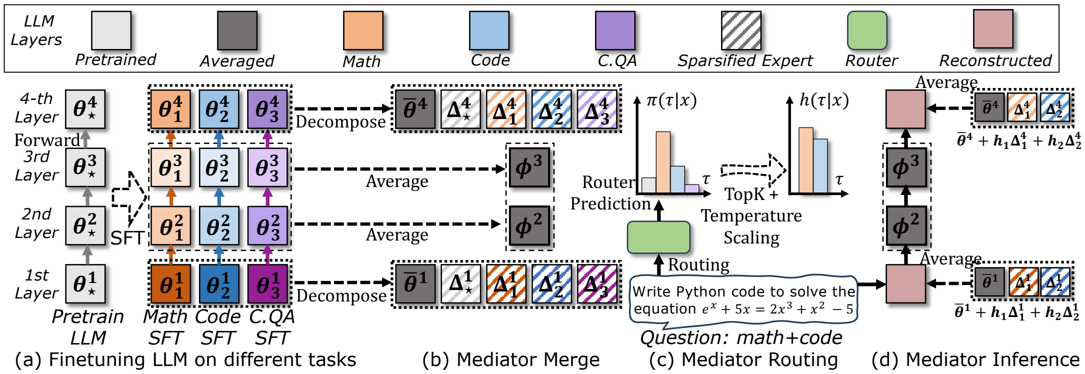

*Figure 1.* The framework of Mediator. For three finetuned LLMs on different tasks, Mediator decomposes layers with large conflicts into expert task arithmetics and sparsifies them. And Mediator averages layers with less conflicts. During inference, Mediator inputs the input question to the router, which output the logits of different experts. Then, top-K experts are chosen and averaged with the base model wit temperature scaling weights. Then, the input question is sent to the final merged model.

*Table 4.* Ablation study w/o CoT (format: w/o CoT/with CoT, ↑ indicates improvement)

| Algorithm/Tasks | GSM8K | HumanEval | GSM-Plus-mini | MBPP |
|---|---|---|---|---|
| w/o Math CoT | 48.44/51.00 (↑2.56) | 27.44/26.83 | 34.45/36.71 (↑2.26) | 33.20/32.8 |
| w/o Code CoT | 46.95/43.29 | 42.68/43.29 (↑0.61) | 33.29/34.12 (↑0.83) | 33.20/34.20 (↑1.00) |