BMIL

Brain and Machine Intelligence Lab.

[ICML 2026]

Post-Hoc Merging is Not Enough: Many-Shot Model Merging with Loss-Gap Balancing

Kyungjin Im*, Miru Kim*, Chanin Eom*, Minhae Kwon

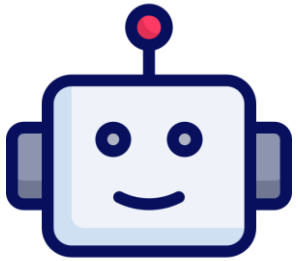
Brain and Machine Intelligence Lab. (BMIL)

https://bmil.skku.edu

Project Page

Toward a Single Multi-Task LLM

Multi-task data



LLM

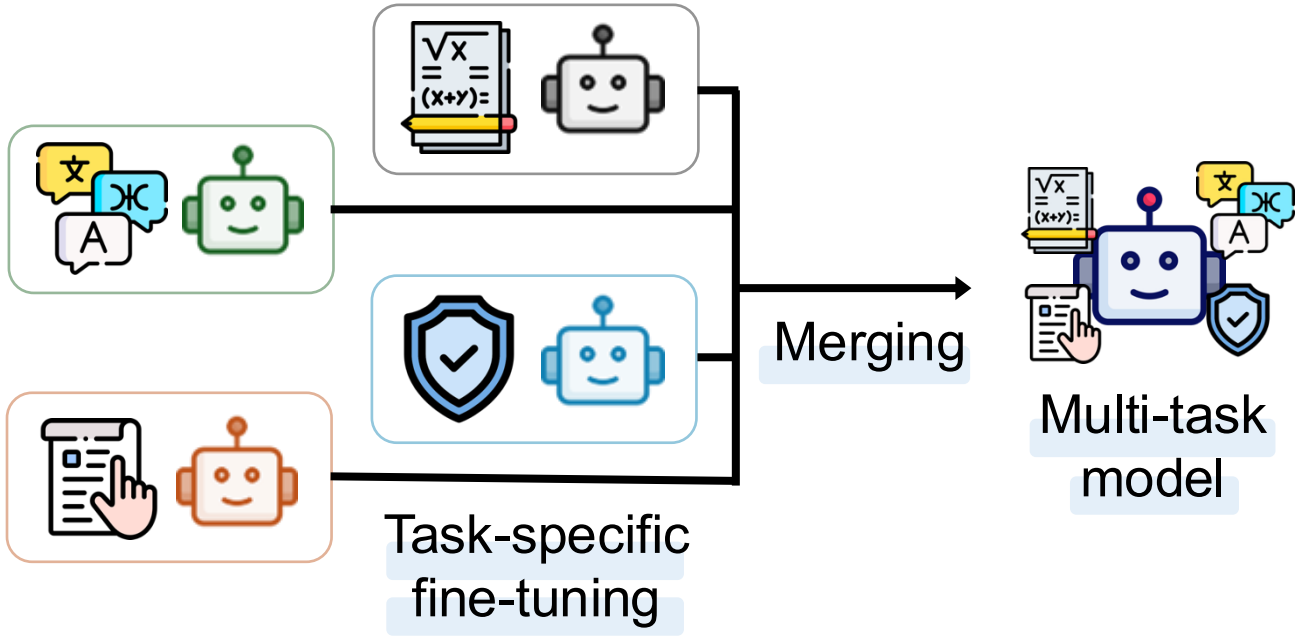
Challenges in Building Multi-Task LLMs

- ✓ Computationally expensive
- ✓ Hard to optimize across different tasks

- Modern LLMs show strong performance across many tasks
- Building one model that performs well on all tasks remains challenging
  - ✓ Joint multi-task post-training is often computationally expensive
  - ✓ Tasks may induce conflicting optimization directions, making joint training unstable

→ A practical alternative is **to reuse task-specific models and combine their knowledge**

Model Merging

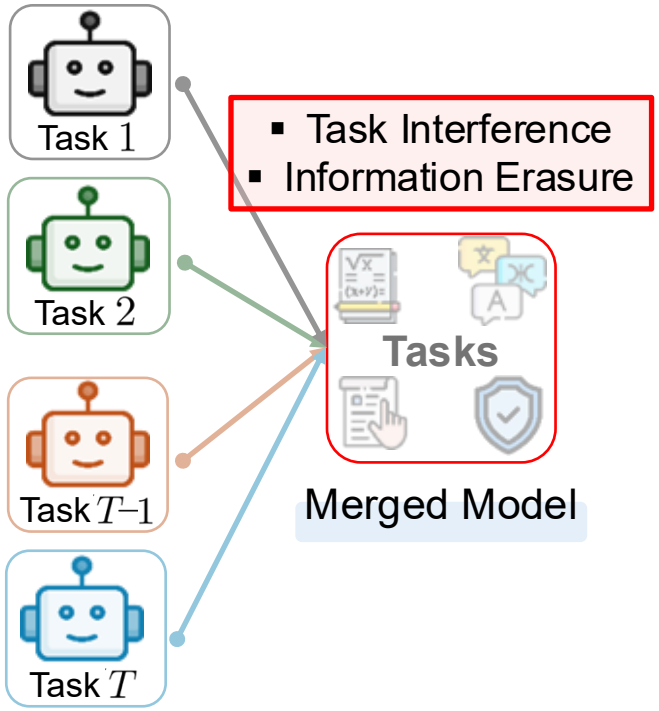


Typical pipeline of merging

Pre-trained model
→ Task-specific fine-tuning
→ Post-hoc one-shot merging
→ Multi-task model

- Model merging constructs a multi-task model by combining multiple task-specific models
- Each task-specific model is fine-tuned from the same pre-trained model
- The task-specific updates are aggregated into one merged model
- This avoids training a single model jointly on all task datasets

Limitations of Post-Hoc Merging



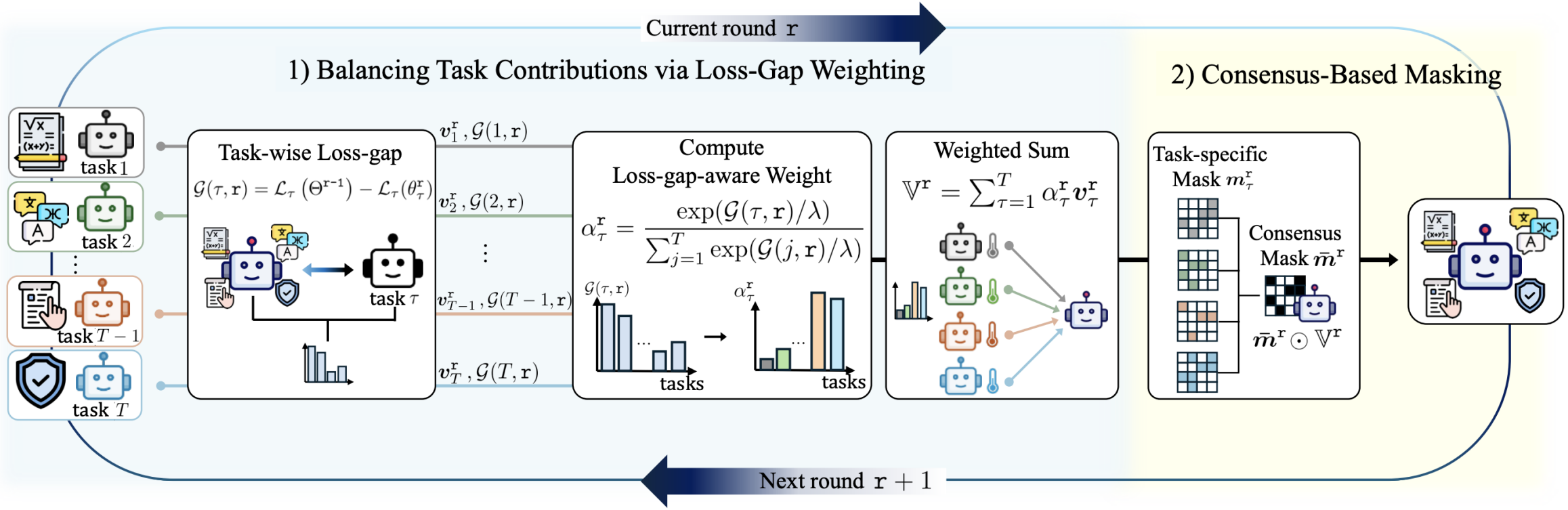
Key research question

Q. How can we improve multi-task capability through model merging without erasing task-specific knowledge?

A. **Gradually integrate task knowledge while reducing destructive interference via many-shot merging !**

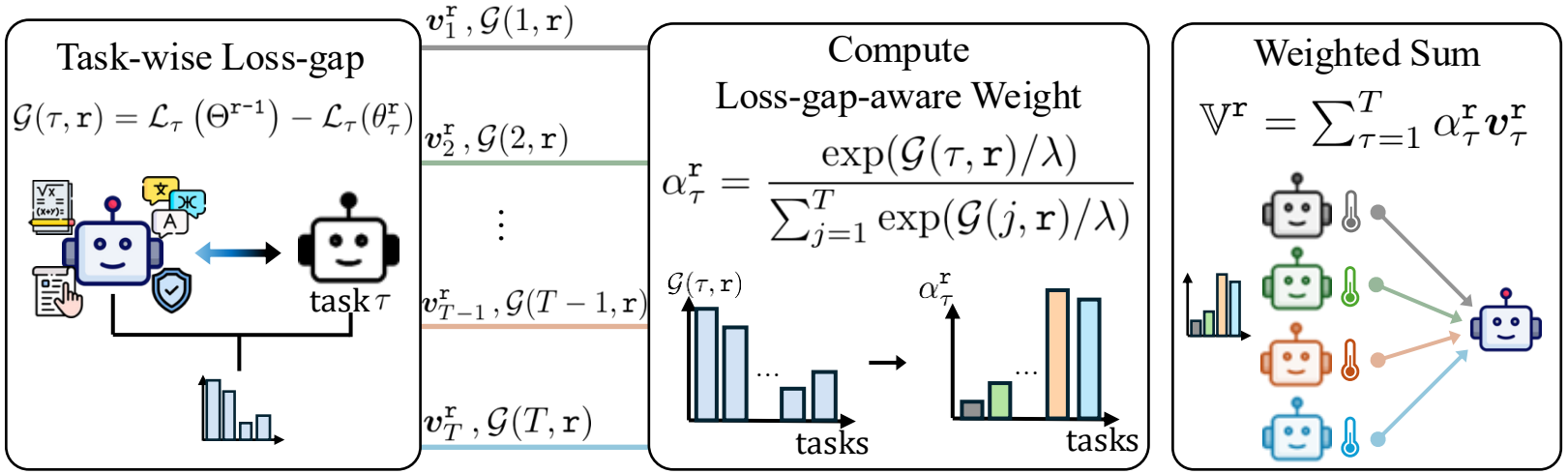
- Most existing methods use post-hoc merging, where task-specific models are merged only once after training
- Post-hoc merging can cause abrupt parameter changes when task updates conflict
- This leads to **information erasure** caused by task interference

Proposed - Overview



- We propose **METIS**, **M**itigating **E**rasure from **T**ask **I**nterference for **S**table many-shot merging, a framework that iteratively merges task-specific models
- METIS use 1**) loss-gap-aware weighting** to recover underrepresented tasks and **2) consensus-based masking** to retain compatible parameter updates

Proposed – 1) Loss-Gap-Aware Weighting

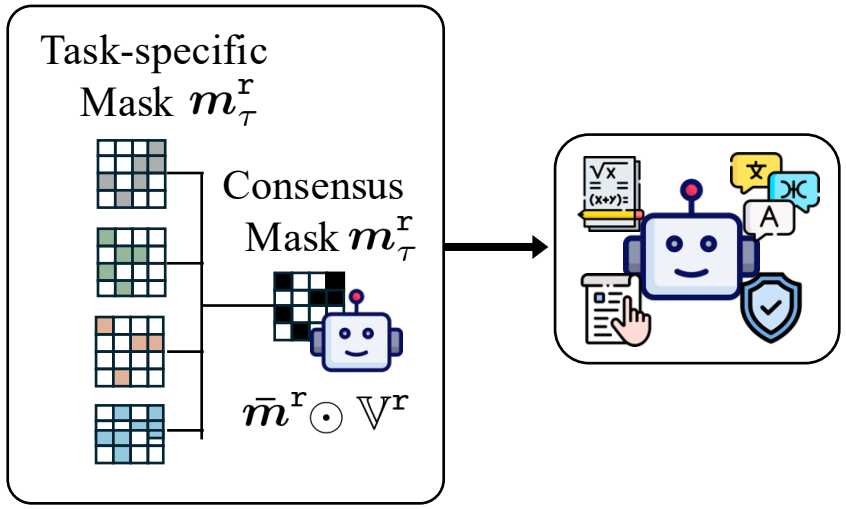


Key Intuition

Tasks with **larger loss gaps** receive **larger weights**, allowing METIS to recover underrepresented task knowledge

- The task-wise loss gap measures **how much task-specific knowledge is degraded** after the previous merging round
- If a task has a large loss gap, it means that task knowledge was erased or underrepresented
- METIS assigns a larger merging weight to such tasks in the next round

Proposed – 2) Consensus-based Masking



Key Intuition

Consensus masking **filters out conflicting updates** and retains parameter changes supported by multiple tasks

- METIS constructs task-specific masks to identify updates aligned with each task

Task vector of τ

$$m_{\tau,i}^r = \mathbb{I}\left(\alpha_\tau^r \left|v_{\tau,i}^r\right| \geq \delta \left|v_i^r - \alpha_\tau^r v_{\tau,i}^r\right|\right) \qquad \bar{m}_i^r = \mathbb{I}\left(\sum_{\tau=1}^{T} m_{\tau,i}^r \geq k\right)$$

Sum of all task vectors

- A consensus mask keeps only the updates supported by multiple tasks
- The final merged model is updated using only consensus-supported parameters

Simulation Setup

- **Models & Baselines**
  - Backbone Models: Gemma-2-2B. Llama-3.2-3B, Llama-3.1-8B, Qwen-3-4B
  - Compared Methods: Task Arithmetic [1], DARE [2], TIES [3], ConsensusTA [4]
- **Tasks & Datasets**
  - **Instruction-following:**
    - Training: TULU-3 / Evaluation: IFEval
  - **Mathematical reasoning:**
    - Training: DART-Math, NuminaMathTIR / Evaluation: GSM8K
  - **Multilingual understanding:**
    - Training: Aya / Evaluation: M-MMLU, M-ARC, M-HellaSwag
  - **Safety:**
    - Training: WildGuardMix, WildJailbreak / Evaluation: XSTest

[1] G. Ilharco, et al., “Editing models with task arithmetic,” in *ICLR*, 2023.
[2] L. Yu, et al., “Language models are super mario: Absorbing abilities from homologous models as a free lunch,” in *ICML,* 2024.
[3] P. Yadav, et al., “TIES-Merging: Resolving interference when merging models,” in *NeurIPS,* 2023.
[4] K. Wang, et al., “Localizing task information for improved model merging and compression,” in *ICML,* 2024.

Simulation Results

[Performance Comparison: Post-Hoc vs. Many-Shot Merging]

(a) Gemma-2-2B

	Method	Avg.	Inst.	Math	Multi.	Safety
Post-Hoc	Task Arithmetic	0.521	0.250	0.080	0.692	0.721
	DARE	0.663	0.375	0.420	0.765	0.885
	TIES	0.726	0.275	0.440	**0.940**	0.820
	ConsensusTA	0.752	0.400	0.600	0.827	1.033
Many-Shot	Task Arithmetic	0.715	0.500	0.520	0.898	0.574
	DARE	0.701	0.350	0.520	0.839	0.820
	TIES	0.776	0.400	0.540	0.840	1.197
	ConsensusTA	0.791	0.450	**0.660**	0.814	1.197
	METIS (Ours)	**0.800**	**0.525**	0.620	0.815	**1.213**

(b) Llama-3.2-3B

	Method	Avg.	Inst.	Math	Multi.	Safety
Post-Hoc	Task Arithmetic	0.706	0.583	0.385	0.715	1.122
	DARE	0.807	0.542	0.615	0.854	1.122
	TIES	0.883	0.375	0.897	**1.082**	0.776
	ConsensusTA	0.942	0.875	0.641	0.958	**1.265**
Many-Shot	Task Arithmetic	0.857	0.375	0.744	1.048	0.878
	DARE	0.914	0.792	0.846	0.955	0.980
	TIES	0.938	0.792	**0.923**	0.924	1.143
	ConsensusTA	0.945	**0.917**	0.872	0.907	1.163
	METIS (Ours)	**1.015**	**0.917**	0.872	1.018	1.245

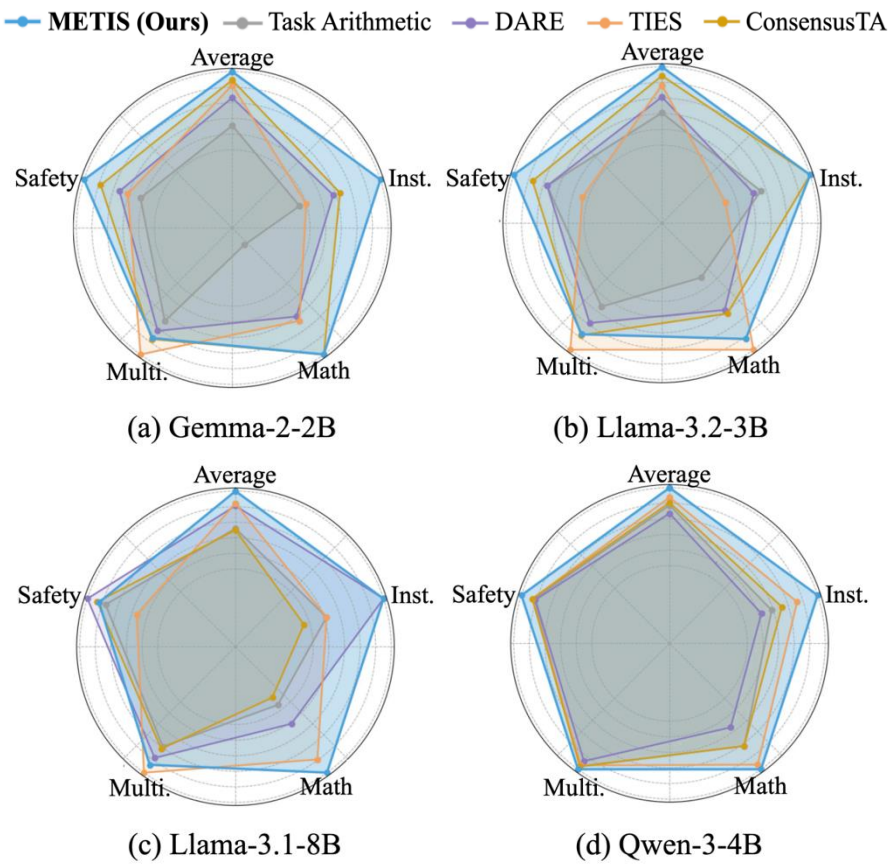
(c) Llama-3.1-8B

	Method	Avg.	Inst.	Math	Multi.	Safety
Post-Hoc	Task Arithmetic	0.704	0.390	0.525	0.796	0.919
	DARE	0.838	**0.634**	0.695	0.885	**1.047**
	TIES	0.852	0.390	1.017	**1.002**	0.698
	ConsensusTA	0.694	0.293	0.458	0.812	0.977
Many-Shot	Task Arithmetic	0.785	0.244	0.966	0.996	0.512
	DARE	0.849	0.439	0.966	0.965	0.791
	TIES	0.902	0.390	1.017	**1.002**	1.000
	ConsensusTA	0.898	0.512	1.085	0.943	0.965
	METIS (Ours)	**0.935**	0.585	**1.136**	0.972	0.977

(d) Qwen-3-4B

	Method	Avg.	Inst.	Math	Multi.	Safety
Post-Hoc	Task Arithmetic	1.048	0.909	0.750	1.036	1.517
	DARE	0.986	0.818	0.614	0.994	1.500
	TIES	1.108	1.136	0.886	1.029	1.534
	ConsensusTA	1.067	1.000	0.750	1.040	1.534
Many-Shot	Task Arithmetic	1.084	0.682	**0.977**	1.069	1.638
	DARE	1.154	1.136	0.886	**1.087**	1.638
	TIES	1.087	1.000	0.818	1.044	1.569
	ConsensusTA	1.128	1.182	0.830	1.052	1.603
	METIS (Ours)	**1.180**	**1.318**	0.920	1.062	**1.655**

[Robustness to Information Erasure]



- METIS achieves **the best average normalized performance** across all backbone models
- METIS improves **robustness to information erasure** by better preserving weak tasks

Simulation Results

[Robustness to Worst-performing Task]

Model (→)	Gemma-2-2B		Llama-3.2-3B	
Method (↓)	**Avg.**	**Worst-task**	**Avg.**	**Worst-task**
Task Arithmetic	0.521	0.080 (-0.44)	0.706	0.385 (-0.32)
DARE	0.663	0.375 (-0.29)	0.807	0.542 (-0.27)
TIES	0.726	0.275 (-0.45)	0.883	0.375 (-0.51)
ConsensusTA	0.752	0.400 (-0.35)	0.942	0.641 (-0.30)
METIS (Ours)	**0.800**	**0.525** (-0.28)	**1.015**	**0.872** (-0.14)

Model (→)	Llama-3.1-8B		Qwen-3-4B	
Method (↓)	**Avg.**	**Worst-task**	**Avg.**	**Worst-task**
Task Arithmetic	0.704	0.390 (-0.31)	1.048	0.750 (-0.30)
DARE	0.838	**0.634** (-0.20)	0.986	0.614 (-0.37)
TIES	0.852	0.390 (-0.46)	1.108	0.886 (-0.22)
ConsensusTA	0.694	0.293 (-0.40)	1.067	0.750 (-0.32)
METIS (Ours)	**0.935**	0.585 (-0.35)	**1.180**	**0.920** (-0.26)

[Robustness to Pre-trained Knowledge Forgetting]

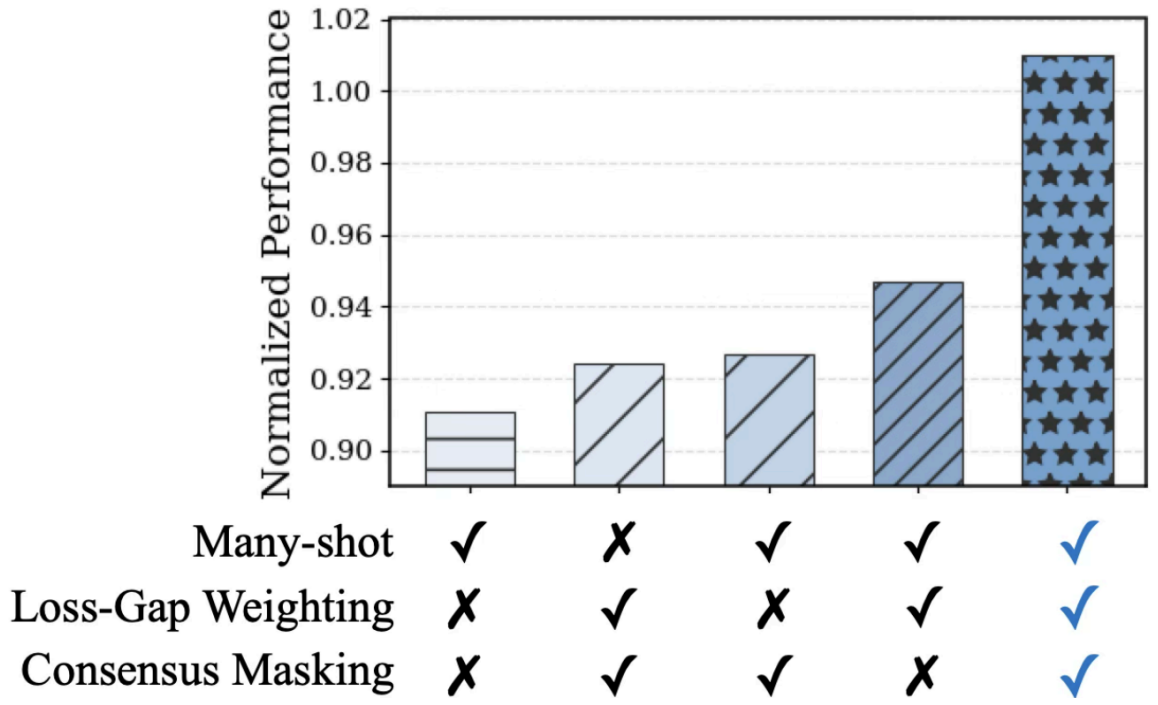
Model (→)	Gemma-2-2B		Llama-3.2-3B	
Method (↓)	CoQA	PubMedQA	CoQA	PubMedQA
Pre-trained	0.695	0.732	0.659	0.750
Task Arithmetic	0.420	0.558	0.548	0.552
DARE	0.561	0.622	0.552	0.556
TIES	0.705	**0.720**	0.662	0.552
ConsensusTA	0.645	0.640	0.583	0.556
METIS (Ours)	**0.714**	0.718	**0.670**	**0.596**

Model (→)	Llama-3.1-8B		Qwen-3-4B	
Method (↓)	CoQA	PubMedQA	CoQA	PubMedQA
Pre-trained	0.697	0.766	0.683	0.746
Task Arithmetic	0.573	0.554	0.473	0.688
DARE	0.687	0.654	0.639	0.708
TIES	0.687	0.574	0.661	**0.728**
ConsensusTA	0.577	0.554	0.521	0.688
METIS (Ours)	**0.698**	**0.728**	**0.684**	**0.728**

- METIS shows **the highest worst-performing task score** and the smallest average–worst gap
- METIS stays closest to the pre-trained model's performance, **reducing knowledge forgetting**

Simulation Results – Ablation Study

- **Many-shot Merging**
  - ✓ Enables stable integration by repeatedly merging task-specific models
- **Loss-Gap Weighting**
  - ✓ Rebalances task contributions based on the task-wise loss gap
- **Consensus Masking**
  - ✓ Filters out potentially conflicting parameter updates
- **Full METIS**
  - ✓ Achieves the best normalized performance, showing that the components are complementary



BMIL

Brain and Machine Intelligence Lab.

[ICML 2026]

ICML International Conference On Machine Learning

SUNGKYUNKWAN UNIVERSITY 1398 성균관대학교

SOONGSIL UNIVERSITY 1897

Post-Hoc Merging is Not Enough: Many-Shot Model Merging with Loss-Gap Balancing

Poster Session 2

July 7th (TUE) 14:00

COEX, Hall A

Project Page