30th ANNIVERSARY EMNLP 2025
Suzhou, China | 中国苏州
November 4-9 | 11月4日-9日

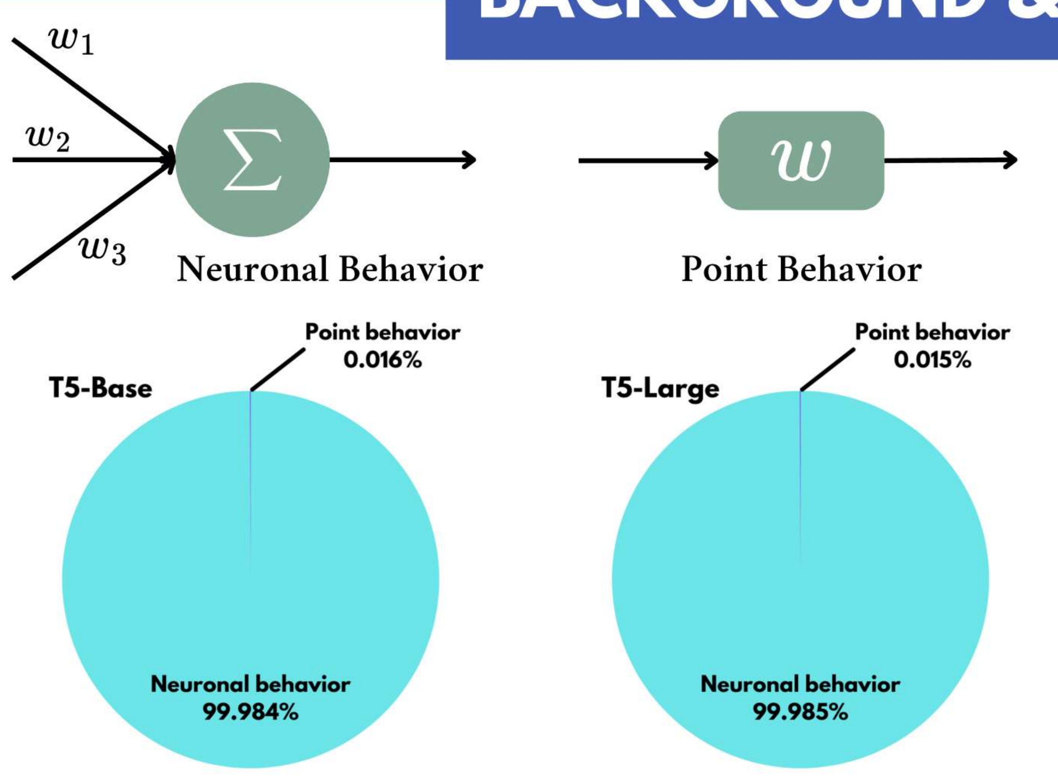
To See a World in a Spark of Neuron: Disentangling Multi-task Interference for Training-free Model Merging

Zitao Fang[1] Guodong Du[2,*] Shuyang Yu[3,†] Yifei Guo[4,†] Yiwei Zhang[1] Yiyao Cao[1] Jing Li[5] Ho-Kin Tang[5] Sim Kuan Goh[1,*]

[1] Xiamen University Malaysia [2] The Hong Kong Polytechnic University
[3] Columbia University [4] Duke University [5] Harbin Institute of Technology (Shenzhen)
[*]Corresponding authors [†]Work conducted while at Xiamen University Malaysia

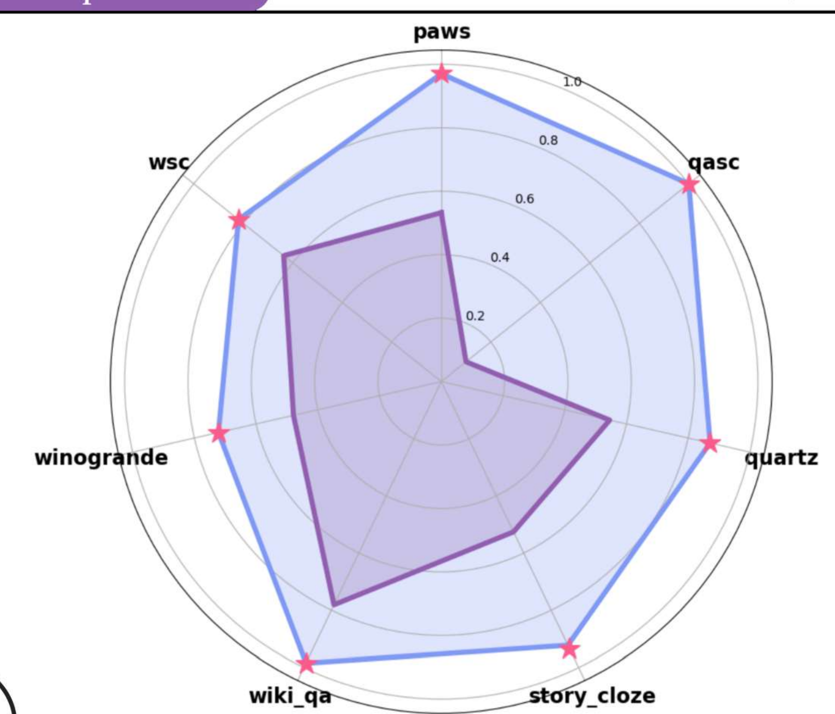
Project Page: https://zzzitaofang.github.io/projects/NeuroMerging/

BACKGROUND & MOTIVATION



- Neuronal behavior dominates (>99.9% parameters);
- But prior methods treat parameters in isolation;
- First merging method considering neuronal mechanisms.

T5-large	In-domain	Out-domain	Total Average
★ Fine-tuned	88.0	53.8	58.7
Keep Orthogonal	**88.0**	**53.9**	**58.8**
Keep Parallel	52.7	51.8	51.9



NLP

Task(→) Method(↓)	T5-Base Acc. (%)	T5-Base Orth. ($\mathcal{O}$)	T5-Base Par. ($\mathcal{P}$)	T5-Large Acc. (%)	T5-Large Orth. ($\mathcal{O}$)	T5-Large Par. ($\mathcal{P}$)
TIES-Merging[NeurIPS2023]	**76.4**	76.2	55.5	79.5	**79.9**	52.5
PCB-Merging[NeurIPS2024]	76.9	**77.1**	54.7	81.0	**81.8**	52.0
NeuroMerging (Ours)	**77.5**	**77.5**	54.6	**82.2**	**82.2**	52.2

CV

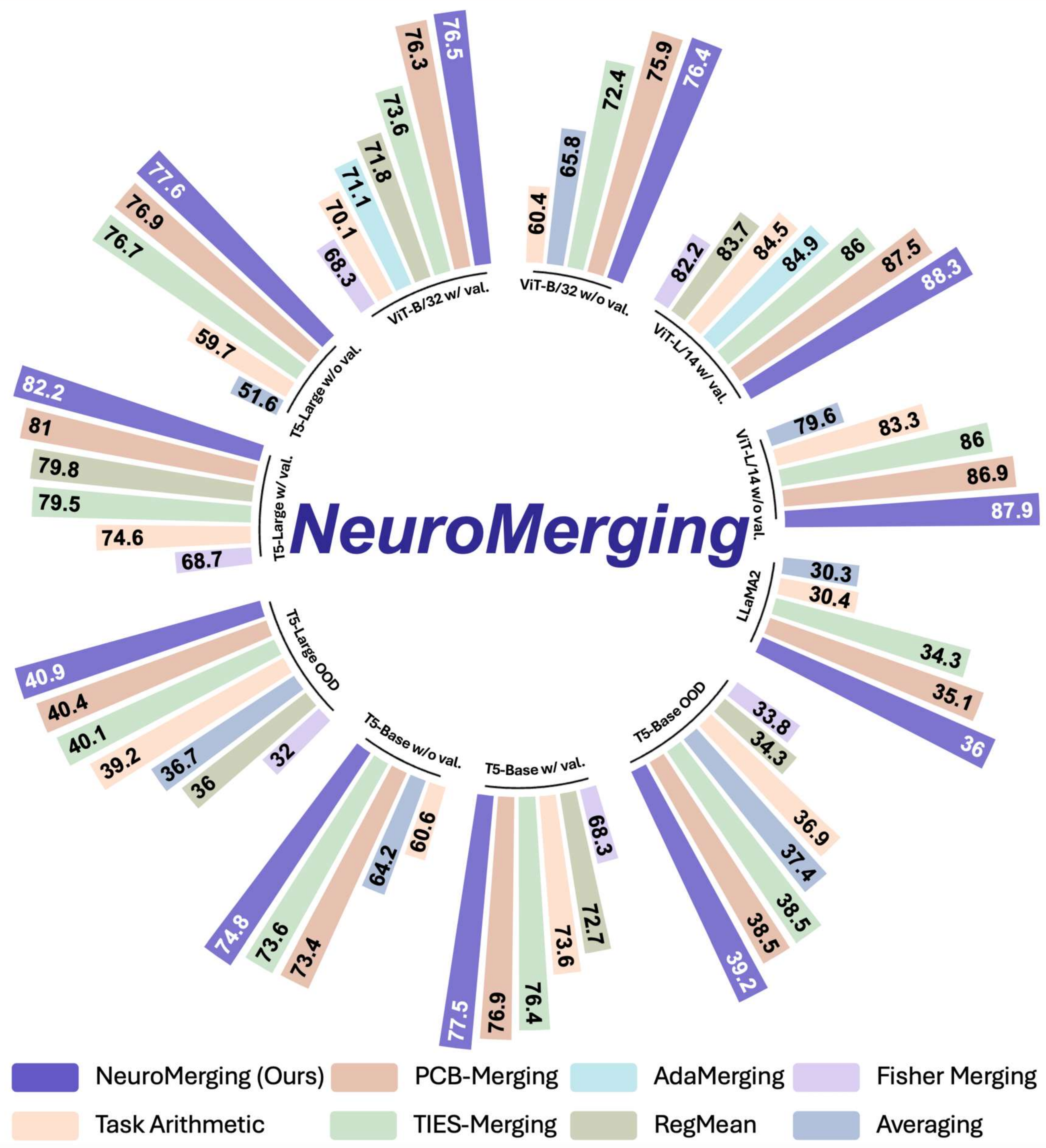
Task(→) Method(↓)	ViT-B/32 Acc. (%)	ViT-B/32 Orth. ($\mathcal{O}$)	ViT-B/32 Par. ($\mathcal{P}$)	ViT-L/14 Acc. (%)	ViT-L/14 Orth. ($\mathcal{O}$)	ViT-L/14 Par. ($\mathcal{P}$)
Layer-wise AdaMerging[ICLR2024]	80.1	**80.4**	50.4	90.8	**91.0**	66.1
Layer-wise AdaMerging++[ICLR2024]	**81.1**	**81.1**	50.3	91.0	**91.2**	65.9
PCB-Merging[NeurIPS2024]	76.3	**76.4**	48.7	87.5	**87.8**	65.0
NeuroMerging (Ours)	76.5	**76.6**	48.7	**88.3**	**88.3**	65.1

▲ We decompose neurons into two complementary subspaces: Orthogonal ($\mathcal{O}$) and Parallel ($\mathcal{P}$);

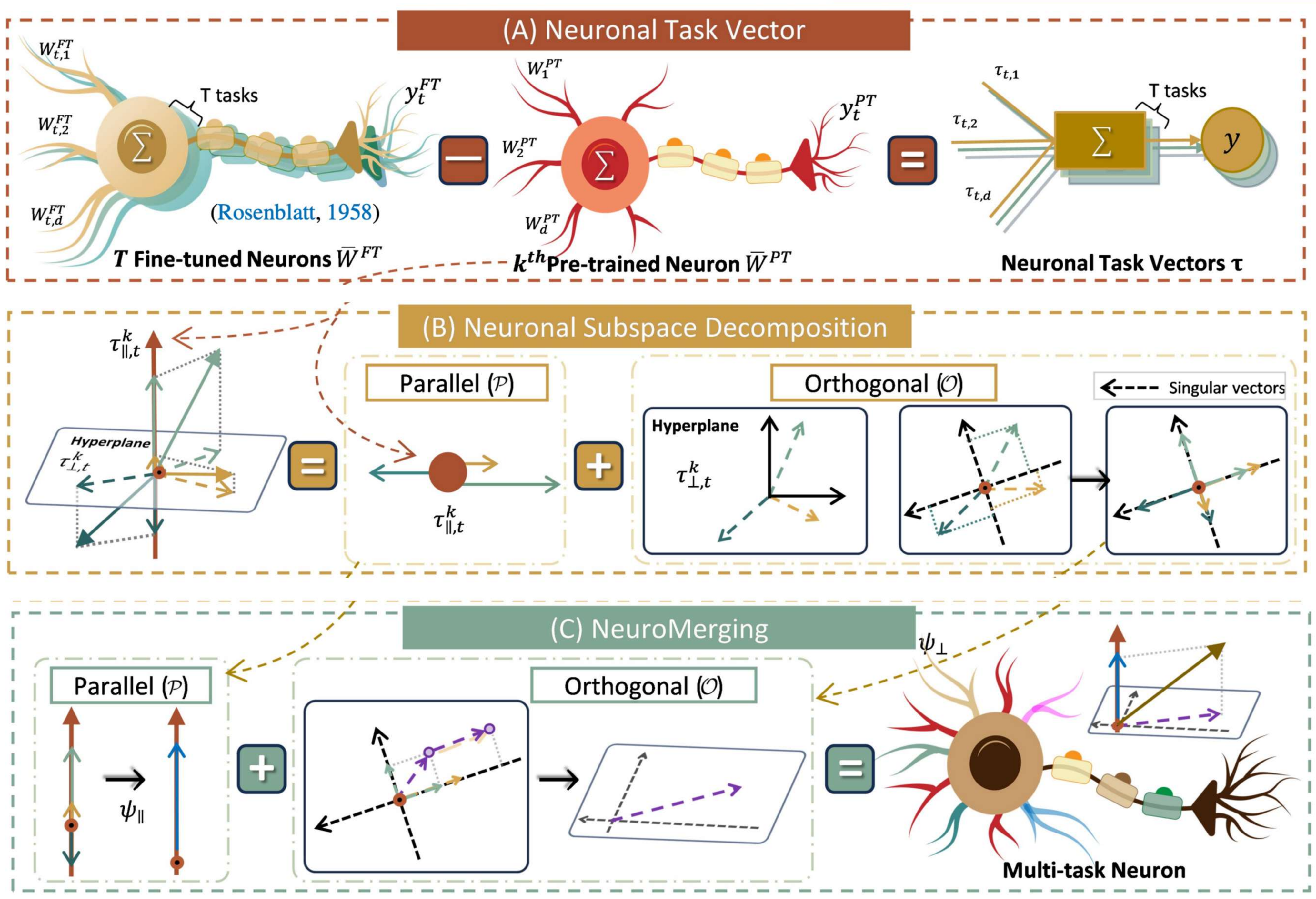
▲ Preserving $\mathcal{O}$ retains fine-tuned performance; Isolating $\mathcal{P}$ reverts to pre-finetuning baseline.

◀ Further exploration across different modalities and merging methods; results remain consistent;

◀ Notably, retaining only the orthogonal subspace $\mathcal{O}$ slightly surpasses the original performance.



METHODOLOGY



Algorithm 1 NeuroMerging

Input: Task-specific models $\tau_1, \tau_2, \dots, \tau_T$, pretrained model θ_0, mask m_t, mask ratio r, proj matrix $\mathbf{P}$

Output: Merged model θ

$\tau_t = m_t \circ \tau_t$ `// Mask task vector based on` r — Retain top-k% of neurons based on their magnitude.

for $k \leftarrow 1$ **to** K **do**

  for $t \leftarrow 1$ **to** T **do**

    ▷ `Create neuronal task vector.` — Neuronal Task Vector

    $\tau_t^k = \mathbf{w}_t^k - \mathbf{w}_0^k$

    ▷ `Decompose neuronal subspaces.` — Neuronal Subspace Decomposition

    $\tau_{\parallel,t}^k = \mathbf{P}\tau_t^k, \quad \tau_{\perp,t}^k = (\mathbf{I}-\mathbf{P})\tau_t^k$

  end

end

▷ `Merge neuronal task vectors.`

for $k \leftarrow 1$ **to** K **do**

  if *Validation data is available* **then**

    Tune λ_1 and λ_2 using the validation dataset

  else

    $\sigma_t = \frac{\|\tau_t^{masked}\|_1}{\|\tau_t\|_1}, \quad \sigma = \max(\sigma_1, .., \sigma_T)$

    $\lambda_1 = 0, \quad \lambda_2 = \frac{1}{1-\sigma}$

  end

  $\overline{\tau}^k = \lambda_1 \psi_\parallel(\tau_{\parallel,1}^k, .., \tau_{\parallel,T}^k) + \lambda_2 \psi_\perp(\tau_{\perp,1}^k, .., \tau_{\perp,T}^k)$

end

▷ `Reconstruct the merged task vector` $\overline{\tau}$ `by combining the` $\overline{\tau}^k$ `for each neuron.` — NeuroMerging

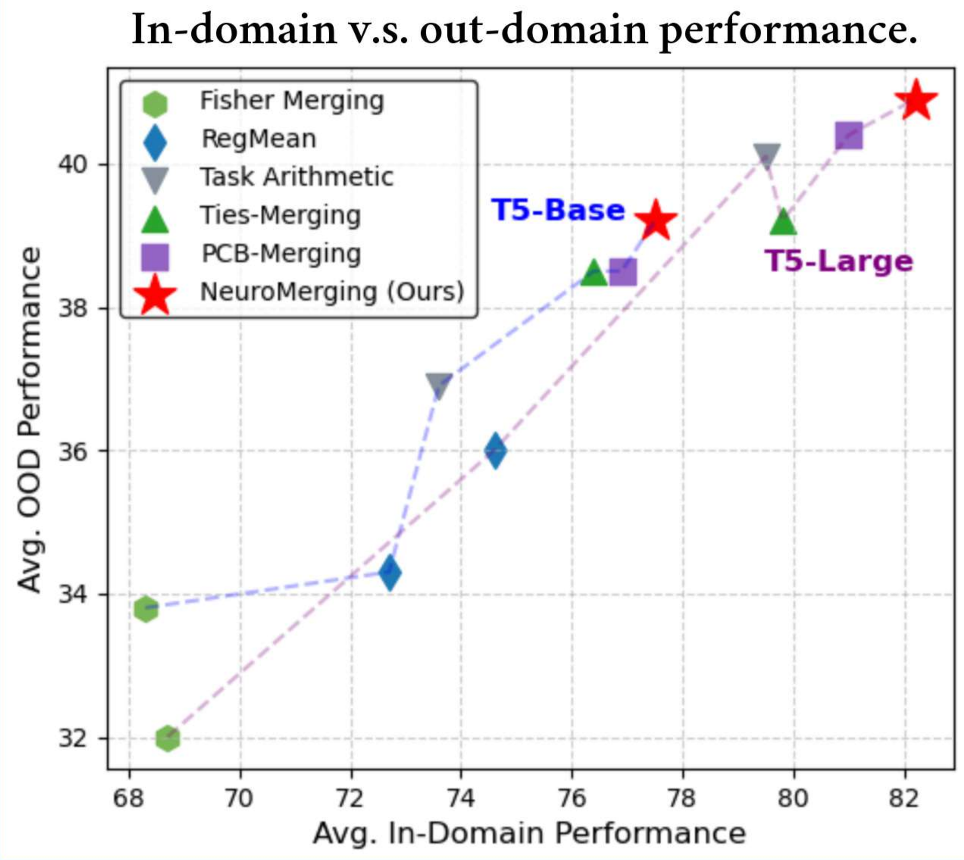
$\bar{\theta} = \theta_0 + \overline{\tau}$ `// final merged model`

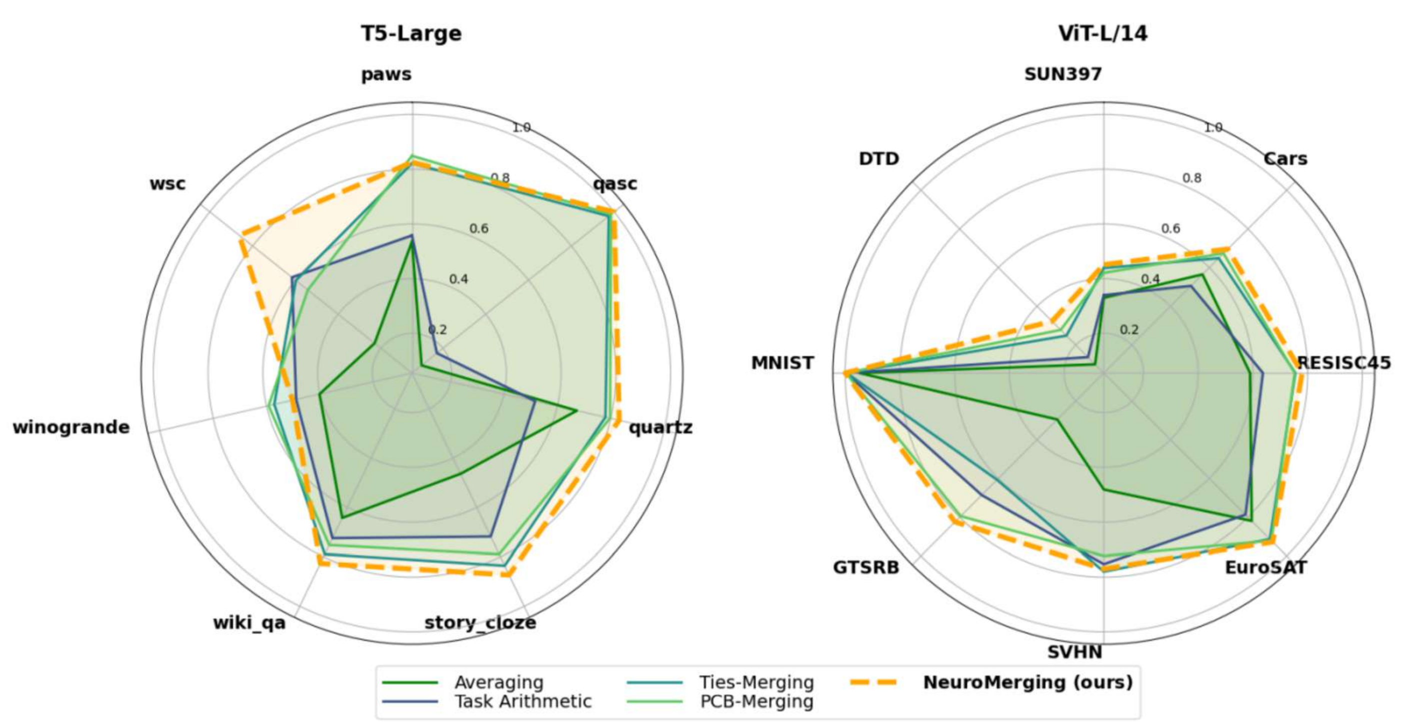
return $\bar{\theta}$

Hyperparameters λ_1 and λ_2
Validation Set ✔ Directly tuning
Validation Set ✖ Estimate based on L1-norm of the neuron weights

RESULTS & ANALYSIS

In-domain v.s. out-domain performance.





Comparison on T5-Large and ViT-L/14 (without validation datasets).

Impacts of λ_1 (parallel) and λ_2 (orthogonal) on T5-Large.