Appendix for TIES-MERGING

Algorithm 1 TIES-MERGING Procedure.

Input: Fine-tuned models $\{\theta_t\}_{t=1}^{n}$, Initialization θ_{init}, k, and λ.
Output: Merged Model θ_m
forall t ***in*** $1, ..., n$ **do**
  ▷ `Create task vectors.`
  $\tau_t = \theta_t - \theta_{\text{init}}$
  ▷ `Step 1: Trim redundant parameters.`
  $\hat{\tau}_t \leftarrow \text{keep\_topk\_reset\_rest\_to\_zero}(\tau_t, k)$
  $\hat{\gamma}_t \leftarrow sgn(\hat{\tau}_t)$
  $\hat{\mu}_t \leftarrow |\hat{\tau}_t|$
end
▷ `Step 2: Elect Final Signs.`
$\gamma_m = sgn(\sum_{t=1}^{n} \hat{\tau}_t)$
▷ `Step 3: Disjoint Merge.`
forall p ***in*** $1, ..., d$ **do**
  $\tau_m^p \leftarrow \text{Mean}_{\{t \in [n] \mid \gamma_t^p = \gamma_m^p\}}(\hat{\tau}_t^p)$
end
▷ `Obtain merged checkpoint`
$\theta_m \leftarrow \theta_{\text{init}} + \lambda * \tau_m$
return θ_m

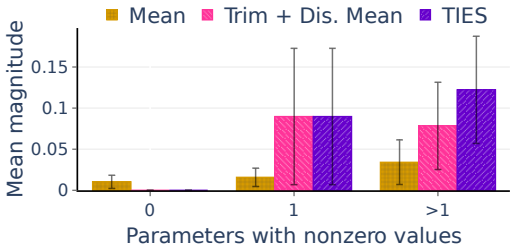
A Additional Results

Method	Estimating Sign			Average
	Multitask	Samples	Init.	
Fine-Tuned	-	-	-	71.4
Multitask	-	-	-	73.1
Averaging [9, 76]	-	-	-	58.0
Task Vectors [29]	-	-	-	63.9
TIES-MERGING	-	-	-	**66.4**
TIES-MERGING	✓	32	scratch	66.5 [+0.1]
	✓	32	mean	67.7 [+1.2]
	✓	All	scratch	**72.0** [+5.6]

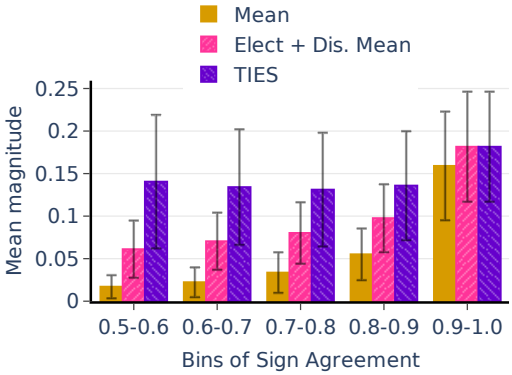
Table 5: **Merging Performance can be improved by estimating the Sign Vector by performing few-shot multitask training.** We use the estimated sign as the elected sign and perform merging.

A.1 Enhancing Performance by Estimating the Multitask Sign Vector.

Considering the findings, we inquire whether it is possible to efficiently acquire multitask sign vectors without extensive multitask training. Our proposal involves utilizing a limited number of validation samples from each task to cheaply train a multitask model and subsequently derive the relevant sign vector. We create two multitask $(IA)^3$ models: one developed from scratch and another initialized using the average of task-specific $(IA)^3$ models intended for merging. We use 32 validation examples from each task to train this model. In Table 4, we notice using the sign vector from the fewshot multitask model initialized with mean yielded a performance increase of 3.8% and 1.3% compared to Task Arithmetic and TIES-MERGING. Interestingly, training fewshot multitask training from scratch did not yield significant improvements over TIES-MERGING without sign estimation. We believe that exploring this area further may result in improved merging techniques.

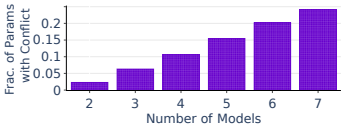


(a) Redundant Parameter Interference for $(IA)^3$ with STD.

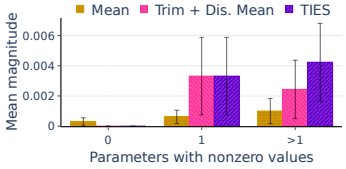


(b) Sign Interference for $(IA)^3$ with STD.

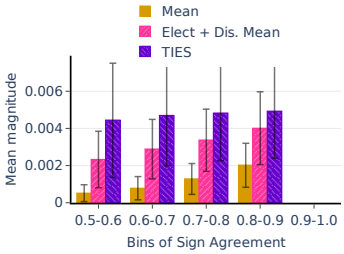
Figure 8: Effect of different types of Merging on the Magnitudes of the Parameters. Here we additionally compare with TIES-MERGING and also provide the standard deviation of parameter values. A high std implies that there is some diversity in magnitude values across different parameters.



(a) Fraction of Parameters with Sign conflicts for T5-Base model versus number of models.



(b) Redundant Parameter Interference for T5-Base with STD.



(c) Sign Interference for T5-Base model with STD.

Figure 9: Plots for T5-Base model.

Method	Validation	Average	rte	cb	winogrande	wic	wsc	copa	h-swag	story cloze	anli-r1	anli-r2	anli-r3
Zeroshot	-	55.3	79.8	46.4	52.8	54.1	45.2	85	36.1	91	39.7	37.6	40.5
Fine-Tuned	-	71.4	82.7	95.8	75.1	71.7	65.3	85.3	44.4	94.9	70.2	46.5	53
Multitask (All, scratch)	-	73.1	88.6	95.8	75.5	61.1	80.6	94.1	42.3	97.6	70.5	49.8	47.7
Multitask (32, scratch)	-	60.9	74.9	79.2	59.3	49.2	63.9	80.9	39.5	91.6	49.4	41.9	40.1
Multitask (32, mean)	-	65.2	79.8	83.3	61.6	54.2	66.7	85.3	41.1	94.4	58.1	46.0	46.5
Averaging	✗	58	81.2	58.3	53.8	55.2	53.5	80.9	40.1	92.5	43.3	39.2	40.2
Task Arithmetic	✗	59.2	76.5	79.2	57.7	51.6	51.4	66.2	31.4	81.5	59.8	47.5	48.2
TIES-MERGING	✗	64.9	81.2	87.5	60.8	59.9	58.3	80.2	42.6	91.1	58.1	46.5	47.4
Fisher Merging	✓	62.2	83.3	83.3	56.7	54.2	58.3	83.1	42.2	94.1	45.9	41.0	42.2
RegMean	✓	58	81.2	58.3	53.8	55.2	53.5	80.9	40.1	92.5	43.3	39.2	40.2
Task Arithmetic	✓	63.9	74.1	83.3	62.8	49.1	49.3	87.5	41.5	95.3	60.8	49.4	50.0
TIES-MERGING	✓	66.4	78.0	83.3	67.9	57.6	59.7	81.7	42.8	90.3	66.9	51.3	51.1

Table 6: Test set performance when merging IA3 models on eleven tasks. Please refer to Section 6 for experimental details.

A.2 Detailed Results for Types of Interference and Their Effect on Merging

In Section 7.1 and Figure 6, we showed the effect of redundant parameters and sign conflicts on parameter magnitudes when comparing simple averaging vs disjoint mean after either trimming or electing and showed that performing these operations helps with the parameter magnitudes. In Figure 8, we additionally compare with TIES-MERGING and show that performing both trimming and electing usually results in higher magnitude and also higher standard deviation in parameter magnitudes. Higher std denotes that all parameter values in the merged model are the same and that there is a significant variation in the magnitude which is in contrast to simple averaging as it decreases the magnitude of not redundant parameters and reduces the magnitude of the influential parameters in the merged model. Similar plots for the T5-base model are provided in Figure 9.

A.3 Breakdown Per Task

We provide the task level for all the in-domain evaluation experiments in the main Table 1. Table 6, 7, 8, 9, and 10 provide the task level results IA3 [41], T5-Base, T5-Large [54], ViT-B/32, and ViT-L/14 [14] respectively. The task level results of the out-of-domain experiments for T5-Base and T5-Large can be found in Table 11, and 12. Lastly, Figure 10, shows the scaling of the T5-Base model as we merge different numbers of tasks.

Method	Vailidation	Average	paws	qasc	quartz	story_cloze	wiki_qa	winogrande	wsc
Zeroshot	-	53.5	49.9	35.8	53.3	48.1	76.2	50	61.1
Fine-tuned	-	82.8	94.3	98.3	80.4	84.7	95.5	64.1	62.5
Multitask	-	83.6	94	97.9	82.5	86.7	95	64.1	65.3
Averaging	✗	65.9	66.4	82.6	60.2	49.5	94.1	50.4	58.3
Task Arithmetic	✗	73.9	73.3	93.5	68.2	76.5	93.7	55.5	56.9
TIES-MERGING	✗	69.7	74	83.3	70.3	64.2	84.7	55.9	55.6
Fisher Merging	✓	68.9	69.3	85.7	63.6	56.4	93.8	50.9	62.5
RegMean	✓	71.2	76.8	96.2	62.5	55	94.8	51.9	61.1
Task Arithmetic	✓	73.2	73.4	93.3	67.1	71.7	94.1	52.9	59.7
TIES-MERGING	✓	73.9	79.3	88.6	71.8	72.9	82.5	61.3	61.1

Table 7: Test set performance when merging T5-base models on seven tasks. Please refer to Section 6 for experimental details.

Method	Validation	Average	paws	qasc	quartz	story_cloze	wiki_qa	winogrande	wsc
Zeroshot	-	51.7	55.4	14.3	54.1	54.1	71	49.3	63.9
Fine-tuned	-	88.8	94.4	98.9	87.8	90.8	96	74.7	79.2
Multitask	-	88.1	94.2	98.5	89.3	92	95.4	73.5	73.6
Averaging	✗	59.6	61.3	82.6	70.5	53.7	63.2	49.7	36.1
Task Arithmetic	✗	73.5	79.2	96.8	80.2	83.6	58.6	60.2	55.6
TIES-MERGING	✗	74.4	80.5	96.2	81.8	78.6	62	61.9	59.7
Fisher Merging	✓	64.6	60.4	81.7	75	60.1	88.6	50	36.1
RegMean	✓	73.2	86	96.9	80.7	78.6	82.6	51.8	36.1
Task Arithmetic	✓	73.3	77.8	96	78.6	86.4	59.1	62.3	52.8
TIES-MERGING	✓	76.9	81.5	96.2	80.1	83.6	64.9	66.5	65.3

Table 8: Test set performance when merging T5-large models on seven tasks. Please refer to Section 6 for experimental details.

Method	Validation	Average	SUN397	Cars	RESISC45	EuroSAT	SVHN	GTSRB	MNIST	DTD
Individual	-	90.5	75.3	77.7	96.1	99.7	97.5	98.7	99.7	79.4
Multitask	-	88.9	74.4	77.9	98.2	98.9	99.5	93.9	72.9	95.8
Averaging	✗	65.8	65.3	63.4	71.4	71.7	64.2	52.8	87.5	50.1
Task Arithmetic	✗	60.4	36.7	41	53.8	64.4	80.6	66	98.1	42.5
TIES-MERGING	✗	72.4	59.8	58.6	70.7	79.7	86.2	72.1	98.3	54.2
Fisher Merging	✓	68.3	68.6	69.2	70.7	66.4	72.9	51.1	87.9	59.9
RegMean	✓	71.8	65.3	63.5	75.6	78.6	78.1	67.4	93.7	52
Task Arithmetic	✓	70.1	63.8	62.1	72	77.6	74.4	65.1	94	52.2
TIES-MERGING	✓	73.6	64.8	62.9	74.3	78.9	83.1	71.4	97.6	56.2

Table 9: Test set performance when merging ViT-B/32 models on eight tasks. Please refer to Section 6 for experimental details.

B Implementation Details

Compute Resources Used and Runtimes. We executed all our experiments on Nvidia A6000 GPUs equipped with 48GB RAM. Single-task $(IA)^3$ models for eleven tasks required 1-2 hours per model, while the multitask vector took around 24 hours on four GPUs. The T5-Base and T5-Large models, based on dataset size, needed between 15 minutes and 2 hours per task, and approximately eight hours for the multitask checkpoint. Vision models ViT-B/32 and ViT-L/14 were utilized, as supplied by Ilharco et al. [29].[2] Merge experiments were efficient, with evaluations consuming less

[2] https://github.com/mlfoundations/task_vectors#checkpoints

Method	Validation	Average	SUN397	Cars	RESISC45	EuroSAT	SVHN	GTSRB	MNIST	DTD
Fine-tuned	-	94.2	82.3	92.4	97.4	100	98.1	99.2	99.7	84.1
Multitask	-	93.5	90.6	84.4	99.2	99.1	99.6	96.3	80.8	97.6
Averaging	✗	79.6	72.1	81.6	82.6	91.9	78.2	70.7	97.1	62.8
Task Arithmetic	✗	83.3	72.5	79.2	84.5	90.6	89.2	86.5	99.1	64.3
TIES-MERGING	✗	86	76.5	85	89.3	95.7	90.3	83.3	99	68.8
Fisher Merging	✓	82.2	69.2	88.6	87.5	93.5	80.6	74.8	93.3	70
RegMean	✓	83.7	73.3	81.8	86.1	97	88	84.2	98.5	60.8
Task Arithmetic	✓	84.5	74.1	82.1	86.7	93.8	87.9	86.8	98.9	65.6
TIES-MERGING	✓	86	76.5	85	89.4	95.9	90.3	83.3	99	68.8

Table 10: Test set performance when merging ViT-L/14 models on eight tasks. Please refer to Section 6 for experimental details.

than 2 minutes for the T5-Base, T5-Large, ViT-B/32, and ViT-L/14 experiments. The assessment of $(IA)^3$ models, due to the necessity of using multiple templates from prompt sources and median result calculations across all templates, required approximately one hour per 11 dataset evaluation.

Model	Average	cosmos_qa	social_iqa	quail	wic	copa	h-swag
PAWS	35.9	18.8	25	24.8	68.8	56.2	21.9
QASC	34.9	15.6	21.9	25.1	75	53.1	18.8
QUARTZ	37.4	31.2	18.8	24.3	71.9	59.4	18.8
Story Cloze	35	6.2	25	25.6	75	65.6	12.5
Wiki QA	24.5	18.8	21.9	24.9	28.1	43.8	9.4
Winogrande	28.3	18.8	25	25.7	34.4	43.8	21.9
WSC	31.7	21.9	21.9	24.6	62.5	46.9	12.5
Pretrained	31.1	21.9	18.8	24.1	65.6	43.8	12.5
Averaging	31.7	21.9	21.9	24.6	68.8	37.5	15.6
Fisher Merging	33.8	15.6	21.9	24.9	65.6	53.1	21.9
Task Arithmetic	31.9	15.6	31.2	25.7	28.1	68.8	21.9
RegMean	34.3	23.1	28.1	24.9	48.4	62.5	18.8
TIES-MERGING	35.3	21.9	25	25.7	50	65.6	23.8

Table 11: Out-of-Distributon performance of T5-Base model checkpoints on six tasks. Please refer to Section 6 for experimental details.

Employed Datasets and Associated Licences. We use the following datasets in the paper with the following licenses. ANLI [47], WiC [49], WSC [37], and Story Cloze [64], QuaRTz [68], Cars [35], GTSRB [67] are under Creative Commons License. Winogrande [60], QASC [33] are under Apache license. COPA [57] is under a BSD-2 Clause license. H-SWAG [80], EuroSAT [24], is under MIT Licence. MNIST [36] is under Gnu General Public License. We could not find the licences of DTD [10], RESISC45 [8], SUN397 [78], SVHN [45], CB [42], RTE [11]), and PAWS [82] but they are publically for research use.

Motivation Experiments Details. For both Figure 3, and 4 in Section 3, we perform experiment on the eleven $(IA)^3$ models used in our PEFT merging experiments (§ 6). For a Figure similar to Fig. 4 demonstrating the fraction of parameters with a sign conflict for T5-base model, please refer to Fig. 9a.

Merging in the absence of the Validation Set. In our investigation into scenarios where a validation set is not available, we have devised a recipe and identified the optimal hyperparameters, employing the PEFT experimental procedure detailed in Section 6. This approach was applied to the eleven task-specific models presented in the same section, utilizing the TIES-MERGING method for tuning the hyperparameters. Our preliminary estimates for the hyperparameters were $k = 20$ and λ close to 1. The hyperparameter search was conducted using the eleven task-specific $(IA)^3$ models, with $k \in \{10, 20, 30\}$, and λ spanning from 0.8 to 3.0, in increments of 0.1. The results of this comprehensive search indicated an optimal value of $k = 20$, with values of $\lambda = 0.9$, $\lambda = 1.0$, and $\lambda = 1.1$ demonstrating equivalent performance. To maintain simplicity in our model, we chose a

Model	Average	cosmos_qa	social_iqa	quail	wic	copa	h-swag
PAWS	32.3	25	28.1	25.6	56.2	46.9	12.5
QASC	33.4	21.9	28.1	25.5	43.8	62.5	18.8
QUARTZ	28.7	25	25	25.1	25	53.1	18.8
Story Cloze	32.1	21.9	34.4	26.8	46.9	53.1	9.4
Wiki QA	27.1	25	28.1	25.2	28.1	46.9	9.4
Winogrande	32.4	31.2	18.8	25.6	43.8	62.5	12.5
WSC	29.7	25	25	25.1	37.5	56.2	9.4
Pretrained	27.6	21.9	21.9	24.9	28.1	56.2	12.5
Averaging	30.4	31.2	25	26.3	31.2	59.4	9.4
Fisher Merging	32	34.4	25	26.1	40.6	56.2	9.4
Task Arithmetic	33.3	21.9	34.4	24.6	40.6	59.4	18.8
RegMean	36	34.4	28.1	25.3	62.5	50	15.6
TIES-MERGING	40.4	31.2	43.8	26.6	59.4	59.4	21.9

Table 12: Out-of-Distributon performance of T5-Large model checkpoints on six tasks. Please refer to Section 6 for experimental details.

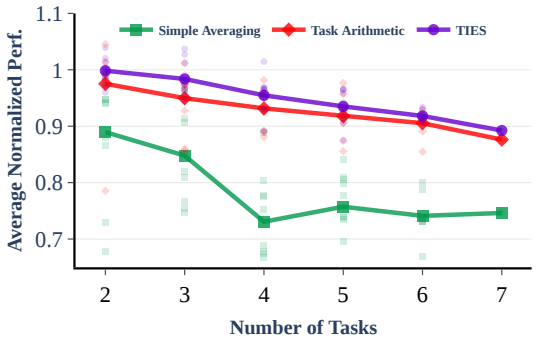


Figure 10: T5-Base with increasing number of task being merged. Average performance when merging a different number of tasks.

λ value of 1. Thus, the final selection of parameters for TIES-MERGING is $k = 20$, signs based on mass, the disjoint mean, and a λ value of 1.

Merging Different Number of Tasks. Here we provide some additional details on the experiments when merging different numbers of tasks. In Fig. 5, we perform the experiment with T5-Large when merging the seven tasks considered in Tab. 1 and described in Sec. 6. The x-axis shows the different number of tasks being merged. Note that when merging T tasks, we have a total of $\binom{7}{T}$ combinations. However, in our experiment, we sample at most 10 distinct combinations for each value of T. A similar plot for the T5-Base model is shown in Fig. 10.

Training Details. In our research, we utilized two variants of the T5 model, specifically the T5-base and T5-large models, which were trained to a maximum of 75,000 steps. An effective training batch size of 1024 was implemented, alongside a learning rate (lr) of 0.0001. We instituted an early stopping mechanism with a patience threshold of 5 to prevent overfitting. During the training process, bfloat16 was adopted to curtail GPU memory expenditure, and the maximum sequence length was set at 128. In contrast, for the PEFT configuration of the $(IA)^3$ approach on the T0-3B model, we modified our parameters. An effective training batch size of 16 was deployed along with an evaluation batch size of 32, while maintaining the learning rate at 0.0001. To accommodate the model's complexity, the early stopping patience was augmented to 10. We do not use any lr scheduler and weight decay for any of our model training.

For the purpose of evaluation, we perform *rank classification*. In this method, the model's log probabilities for all potential label strings are ranked. The model's prediction is deemed accurate if

the choice ranked highest aligns with the correct answer. It should be noted that rank classification evaluation can accommodate both classification tasks and multiple-choice tasks.