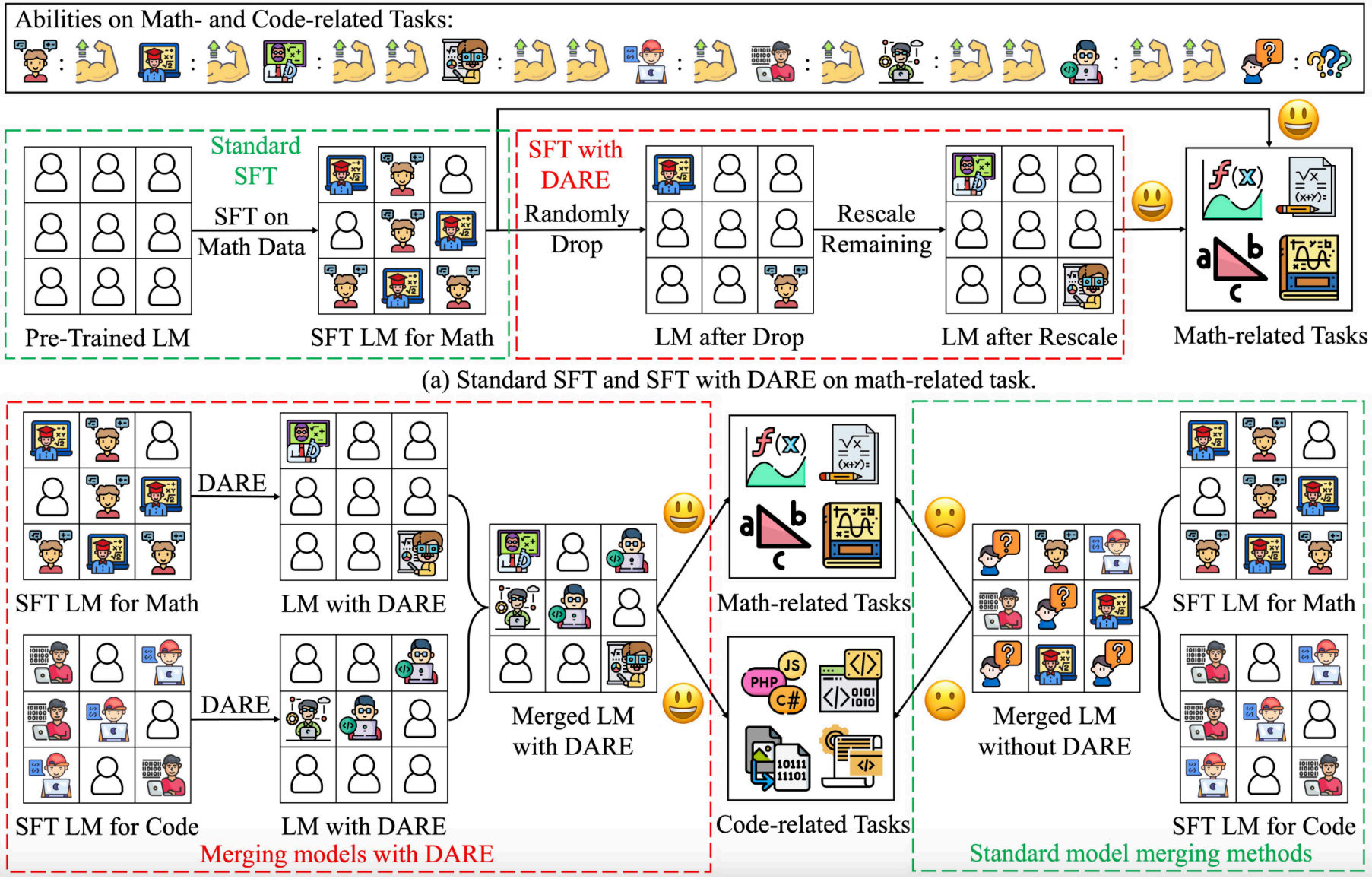
北京大学
PEKING UNIVERSITY

Language Models are Super Mario: Absorbing Abilities from Homologous Models as a Free Lunch

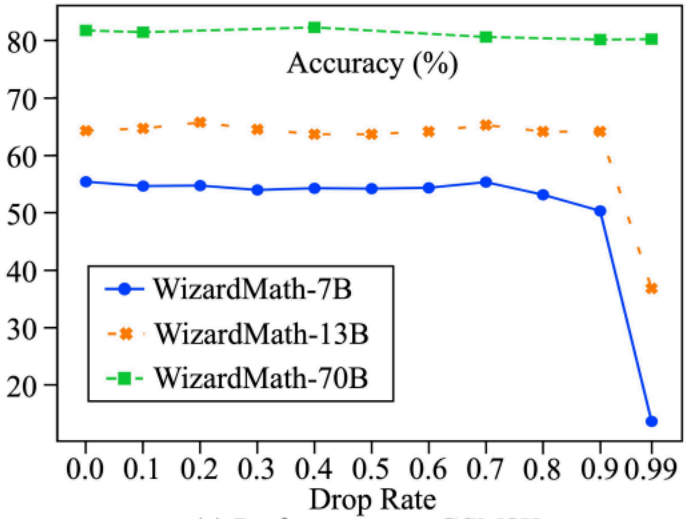
https://arxiv.org/pdf/2311.03099v2.pdf

Jiayu Yao

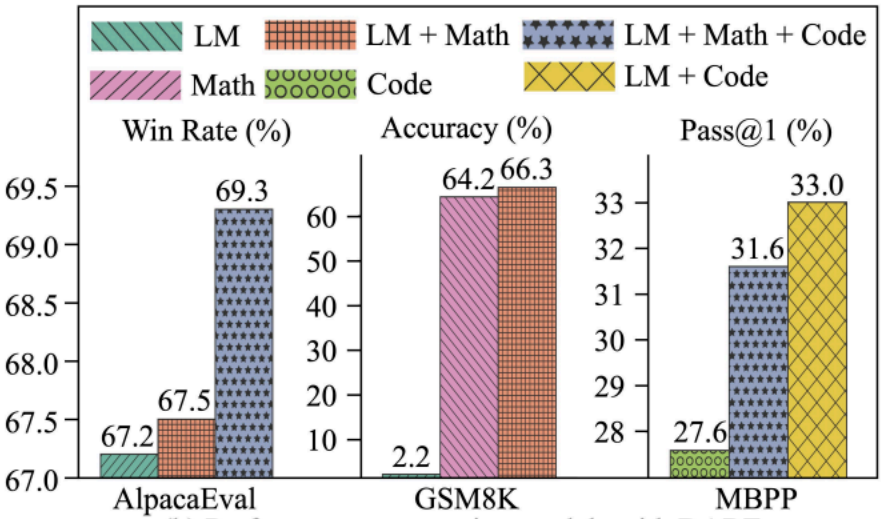
Framework


Abilities on Math- and Code-related Tasks:
Standard SFT
SFT on Math Data
Pre-Trained LM
SFT LM for Math
SFT with DARE
Randomly Drop
LM after Drop
Rescale Remaining
LM after Rescale
Math-related Tasks
(a) Standard SFT and SFT with DARE on math-related task.
DARE
SFT LM for Math
LM with DARE
DARE
SFT LM for Code
LM with DARE
Merging models with DARE
Merged LM with DARE
Math-related Tasks
Code-related Tasks
Merged LM without DARE
SFT LM for Math
SFT LM for Code
Standard model merging methods

DARE



(a) Performance on GSM8K.



(b) Performance on merging models with DARE.

Lottery Ticket Hypothesis & LoRA

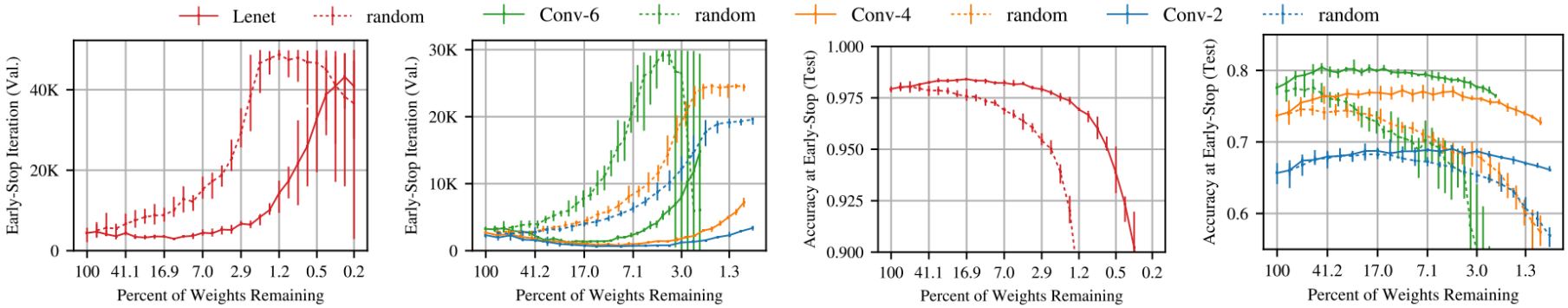


Figure 1: The iteration at which early-stopping would occur (left) and the test accuracy at that iteration (right) of the Lenet architecture for MNIST and the Conv-2, Conv-4, and Conv-6 architectures for CIFAR10 (see Figure 2) when trained starting at various sizes. Dashed lines are randomly sampled sparse networks (average of ten trials). Solid lines are winning tickets (average of five trials).

$$h = W_0 x + \Delta W x = W_0 x + BAx$$

DARE

$$\delta^t = \theta^t_{SFT} - \theta_{PRE}$$

$$m^t = Bernoulli(p)$$

$$\tilde{\delta}^t = \left(1 - m^t\right) \odot \delta^t$$

$$\hat{\delta}^t = \gamma \tilde{\delta}^t$$

$$\gamma = \frac{1}{1-p}$$

Proof

$$\mathbb{E}[h_i] = \mathbb{E}[\sum_{j=1}^{n} (w_{ij} + \Delta w_{ij})\, x_j + (b_i + \Delta b_i)]$$

$$= \sum_{j=1}^{n} x_j \mathbb{E}[w_{ij}] + \mathbb{E}[b_i] + \sum_{j=1}^{n} x_j \mathbb{E}[\Delta w_{ij}] + \mathbb{E}[\Delta b_i]$$

$$= \sum_{j=1}^{n} w_{ij} x_j + b_i + \sum_{j=1}^{n} \Delta w_{ij} x_j + \Delta b_i = h_i^{\text{PRE}} + \Delta h_i$$

$$\mathbb{E}[\hat{h}_i] = \mathbb{E}[\sum_{j=1}^{n} (w_{ij} + \Delta \hat{w}_{ij})\, x_j + (b_i + \Delta \hat{b}_i)]$$

$$= \sum_{j=1}^{n} x_j \mathbb{E}[w_{ij}] + \mathbb{E}[b_i] + \sum_{j=1}^{n} x_j \mathbb{E}[\Delta \hat{w}_{ij}] + \mathbb{E}[\Delta \hat{b}_i]$$

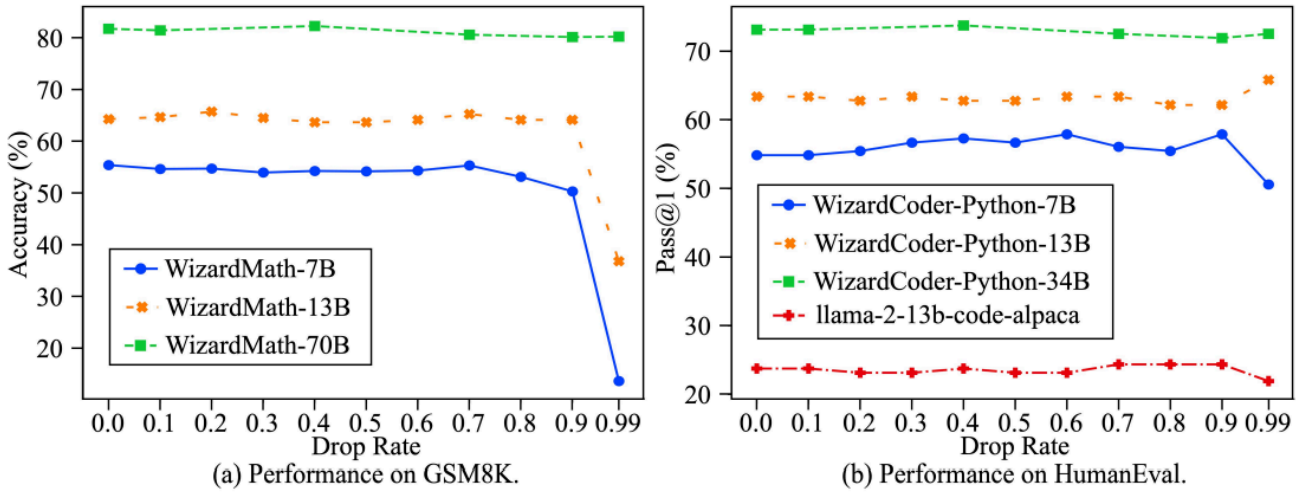
$$= \sum_{j=1}^{n} w_{ij} x_j + b_i + \sum_{j=1}^{n} x_j ((1-p) \cdot \gamma \cdot \Delta w_{ij} + p \cdot 0)$$

$$+((1-p) \cdot \gamma \cdot \Delta b_i + p \cdot 0)$$

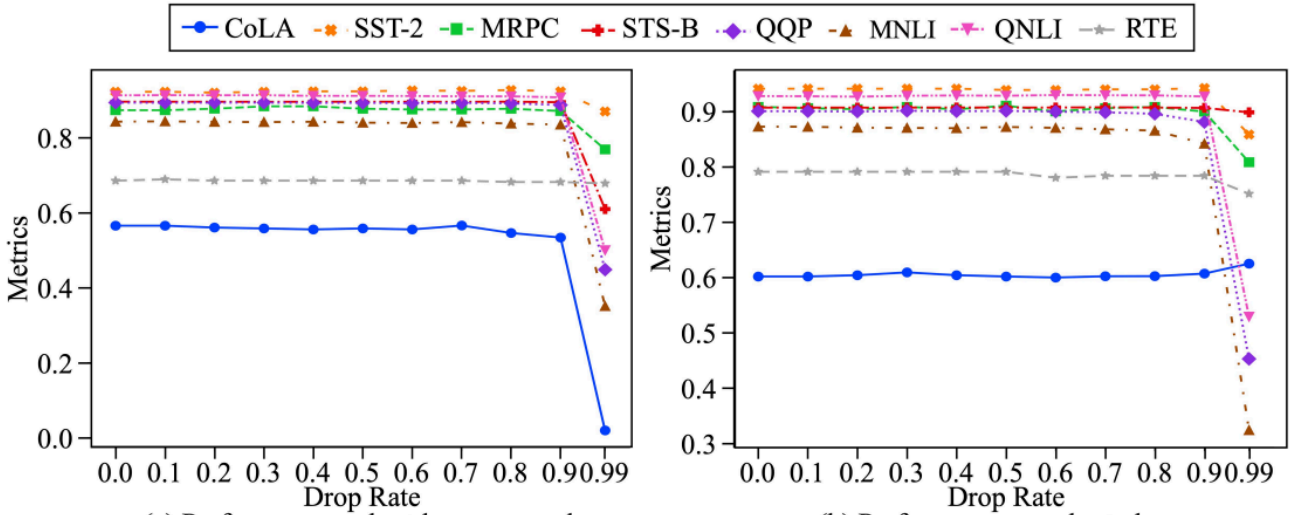
$$= h_i^{\text{PRE}} + (1-p) \cdot \gamma \cdot (\sum_{j=1}^{n} \Delta w_{ij} x_j + \Delta b_i)$$

$$= h_i^{\text{PRE}} + (1-p) \cdot \gamma \cdot \Delta h_i$$

Experiment

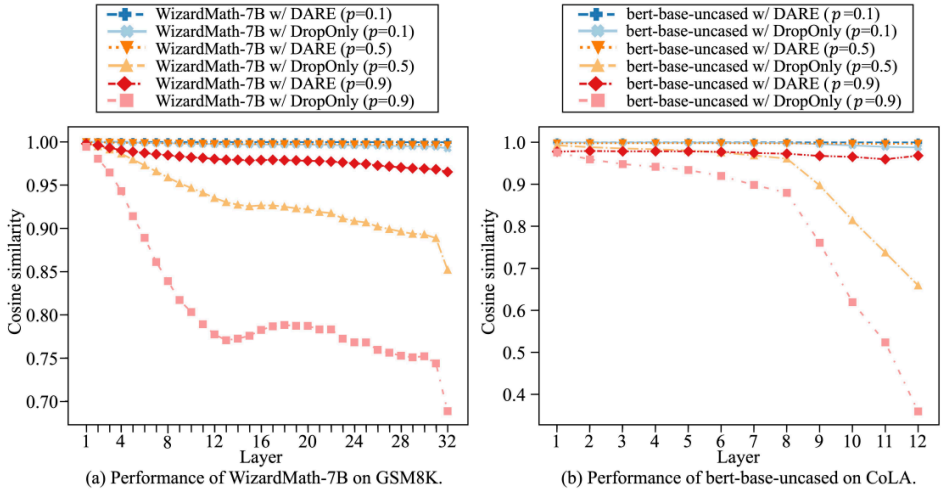


(a) Performance on GSM8K. (b) Performance on HumanEval.

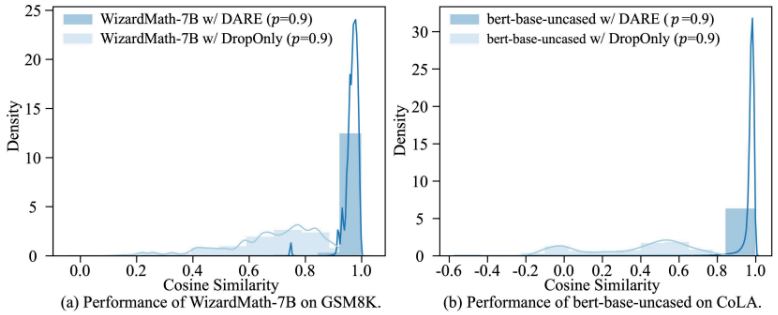


(a) Performance on bert-base-uncased. (b) Performance on roberta-base.

Experiment

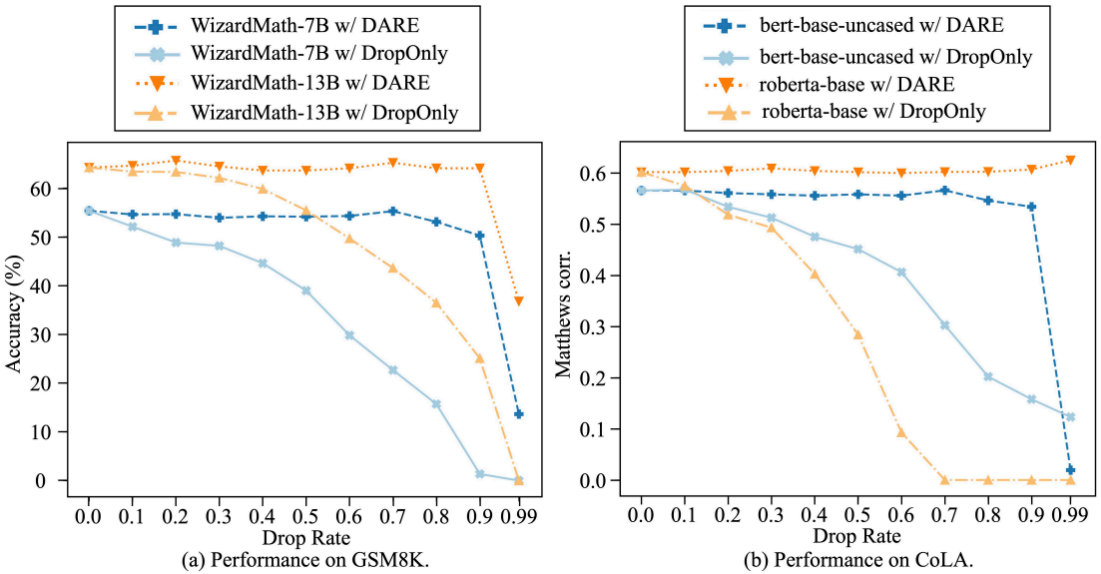


(a) Performance of WizardMath-7B on GSM8K. (b) Performance of bert-base-uncased on CoLA.



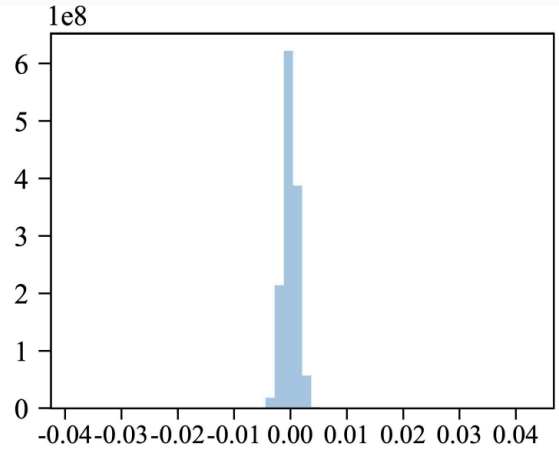
(a) Performance of WizardMath-7B on GSM8K. (b) Performance of bert-base-uncased on CoLA.

Figure 7: Distributions of cosine similarities of the last layer's embeddings between the original LM and LM with DARE or DropOnly.

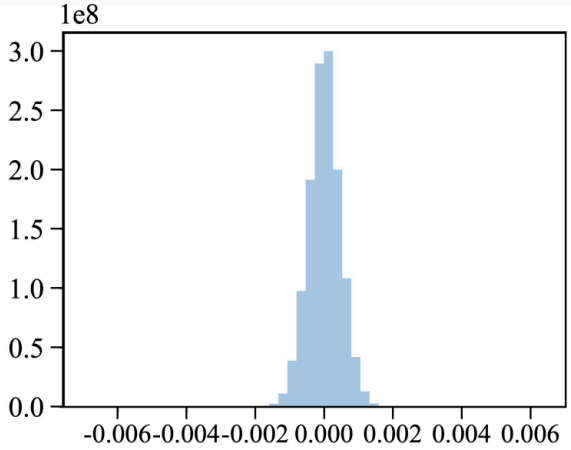


(a) Performance on GSM8K. (b) Performance on CoLA.

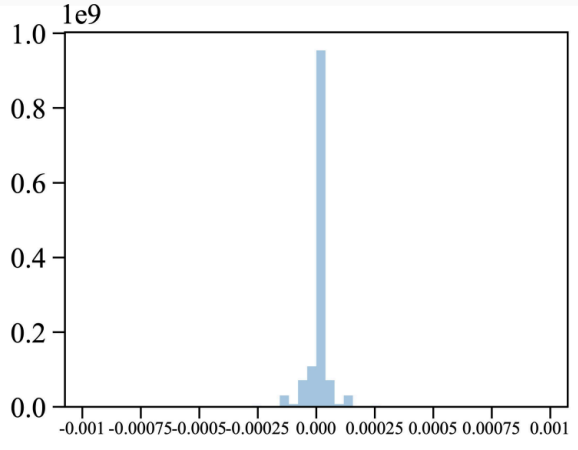
Experiment



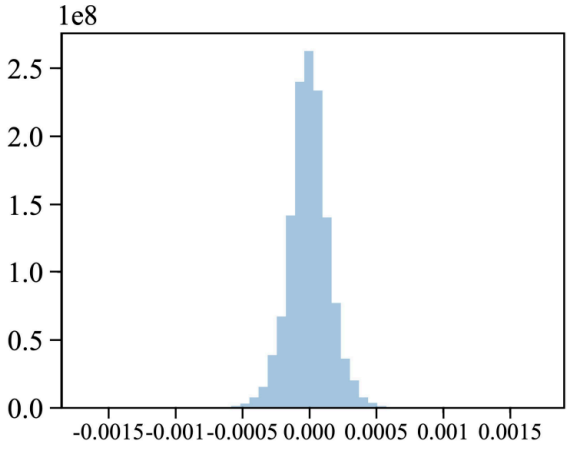
(a) Delta parameter ranges of WizardLM-13B vs. Llama-2-13b.



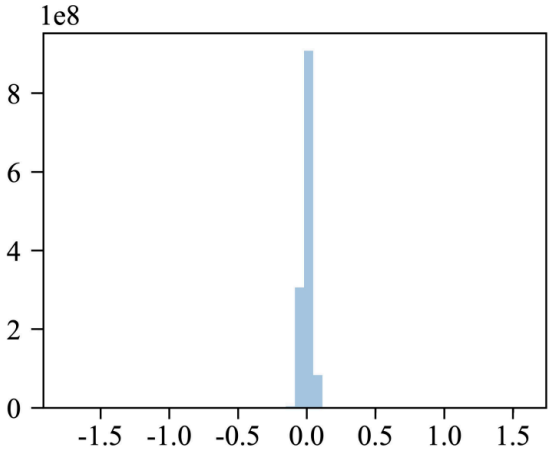
(b) Delta parameter ranges of WizardMath-13B vs. Llama-2-13b.



(c) Delta parameter ranges of llama-2-13b-code-alpaca vs. Llama-2-13b.



(d) Delta parameter ranges of WizardCoder-Python-13B vs. CodeLlama-13b-Python.



(e) Delta parameter ranges of WizardCoder-Python-13B vs. Llama-2-13b.

Experiment

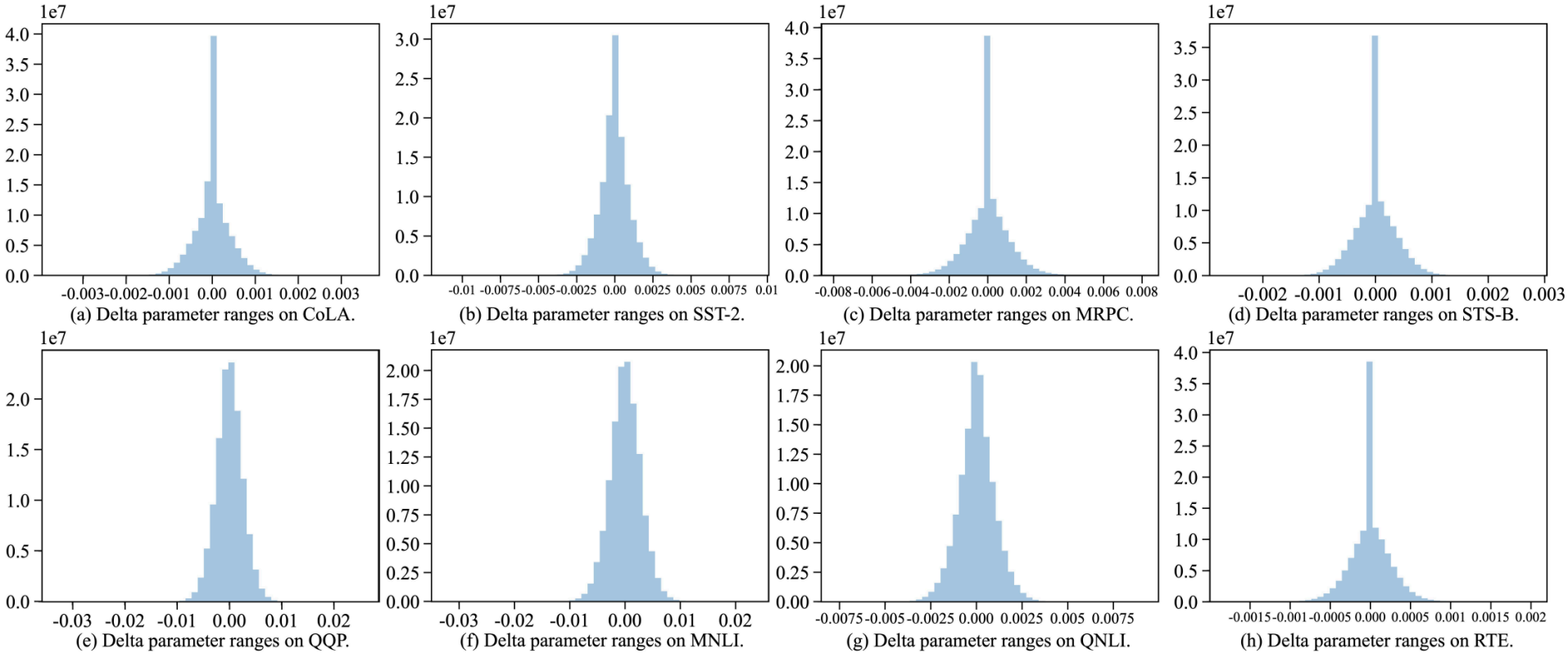


Figure 20: Delta parameter ranges of roberta-base after SFT on GLUE.

Experiment

Table 1: Performance of merging decoder-based WizardLM-13B (LM), WizardMath-13B (Math), and llama-2-13b-code-alpaca (Code). We use single and mixed colors to denote individual and merged models. The best and second-best results are marked in **bold** and underlined fonts.

Merging Methods	Models	Use DARE	AlpacaEval	GSM8K	MBPP
Single Model	LM	No	67.20	2.20	34.00
	Math	No	/	64.22	/
	Code	No	/	/	27.60
Task Arithmetic	LM	No	67.04	**66.34**	30.60
	& Math	Yes	**67.45**	66.26	32.40
	LM	No	**68.07**	/	32.40
	& Code	Yes	67.83	/	**33.00**
	Math	No	/	64.67	8.60
	& Code	Yes	/	**65.05**	9.80
	LM & Math	No	69.03	58.45	29.80
	& Code	Yes	**69.28**	56.48	**31.60**

Experiment

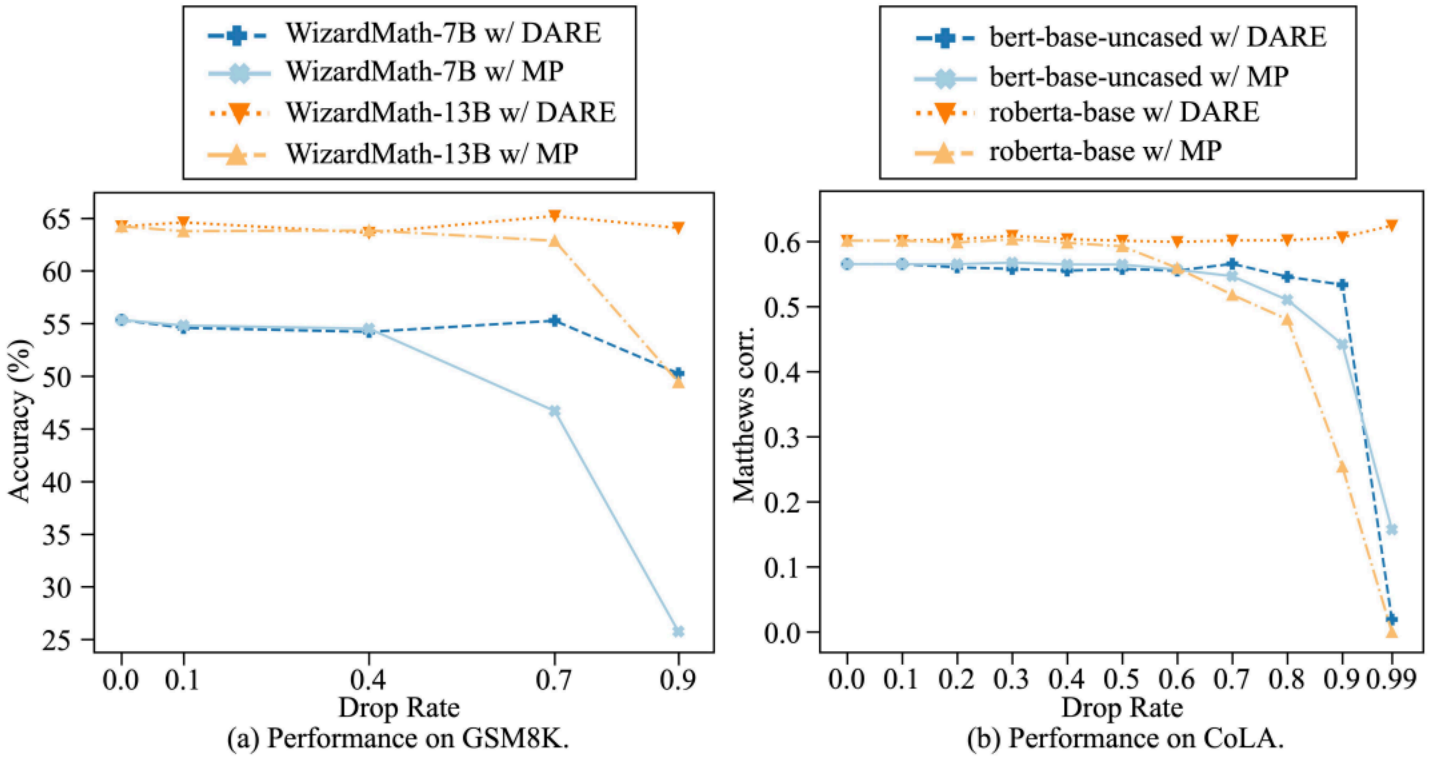


Figure 9: Comparisons between DARE and MP on GSM8K and CoLA on various LMs.

Discussion

- Incremental Learning
- Efficient Tuning
- Subnetwork

Thanks