

A method to efficiently
Adapt an LLM to
A language or domain.

OPTIMER

Optimal Distribution Vector Merging
Is Better than Data Mixing for Continual Pre-Training



Haiyue Song



Vu Huy Hien



Masao Utiyama



Multilingual and Multicultural Communication Laboratory, UCRI, NICT, KYOTO, JAPAN

One Model Does Not Fit Every Language or Domain

■ Thus, governments and companies adapt LLMs to

- **Local Languages:** ASEAN languages, Japanese, ...
- **Domains:** coding, ...



Continual pre-train on
Southeast Asian languages



SEA-LION 4 Model
from **AI Singapore**

Kimi K2.5

CPT on
Codes

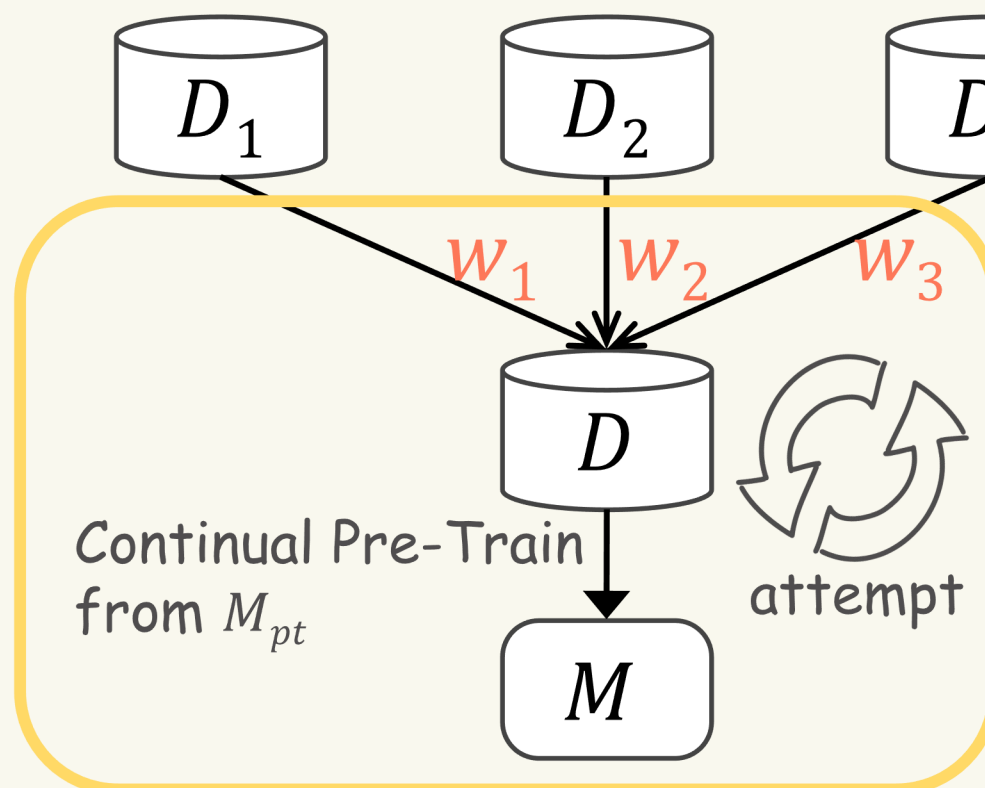
Composer 2

Adapting LLMs is Costly

3

- A widely used method is continual pre-training,
 - i.e. training LLM on target languages / domains data.
- However, Data Mixing is costly because it requires
 - Multiple trials, to search for the optimal **mixing ratio** of training datasets D_i
- Each trial costs ~\$10,000 and ~1 week.
 - e.g., CPT of a 27B model on 10B tokens with 32 H200 GPUs

(a) Data Mix



Fix ratios before training.
Weeks per *attempt*.
Suboptimal? Start over!



Motivation

- Training is expensive,
so we do it only once.
- Adjust the mixing ratio many times,
each in minutes.

WE PROPOSE OPTIMER

■ First,

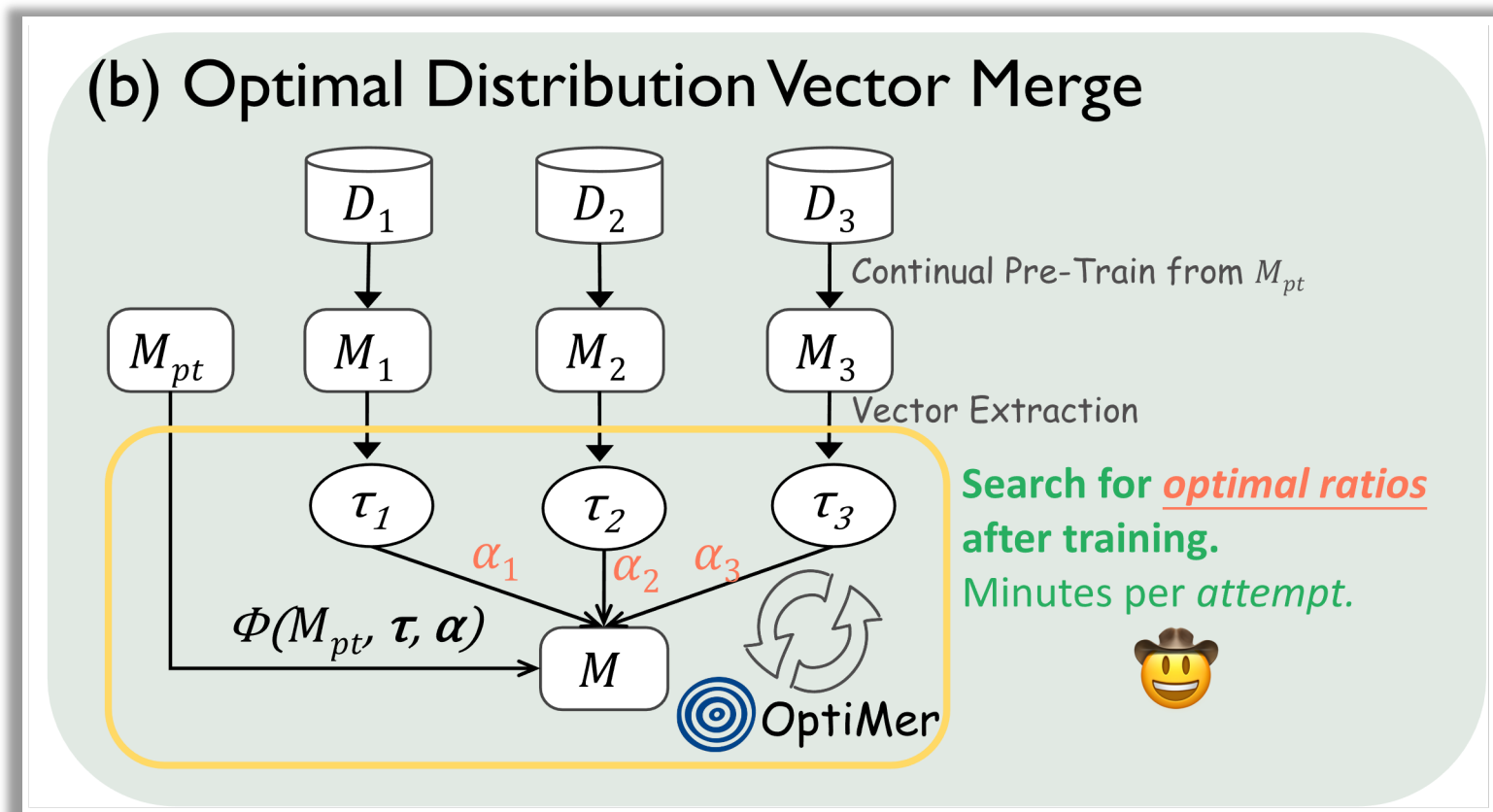
- Train separate models M_i
 - From pre-trained model M_{pt} .
- Train on one dataset, e.g.
 - The target, Japanese $D_1 \rightarrow M_1$
 - English dataset $D_2 \rightarrow M_2$
 - Math dataset $D_3 \rightarrow M_3$

■ Then,

- Extract distribution vectors τ .
 - $\tau_i = M_i - M_{pt}$

■ Final,

- Try multiple times the merge of these vectors τ .
 - Search for the optimal combination through Bayesian optimization using **the Optuna framework**.



Experimental Settings

■ Model

- **Gemma 3 27B**

■ Data

- 1B tokens each: Japanese / English / Chinese / Code / Math

■ Baselines

- DataMix (equal ratio)
- Average merging: Task Arithmetic, TIES, DARE

■ OptiMer

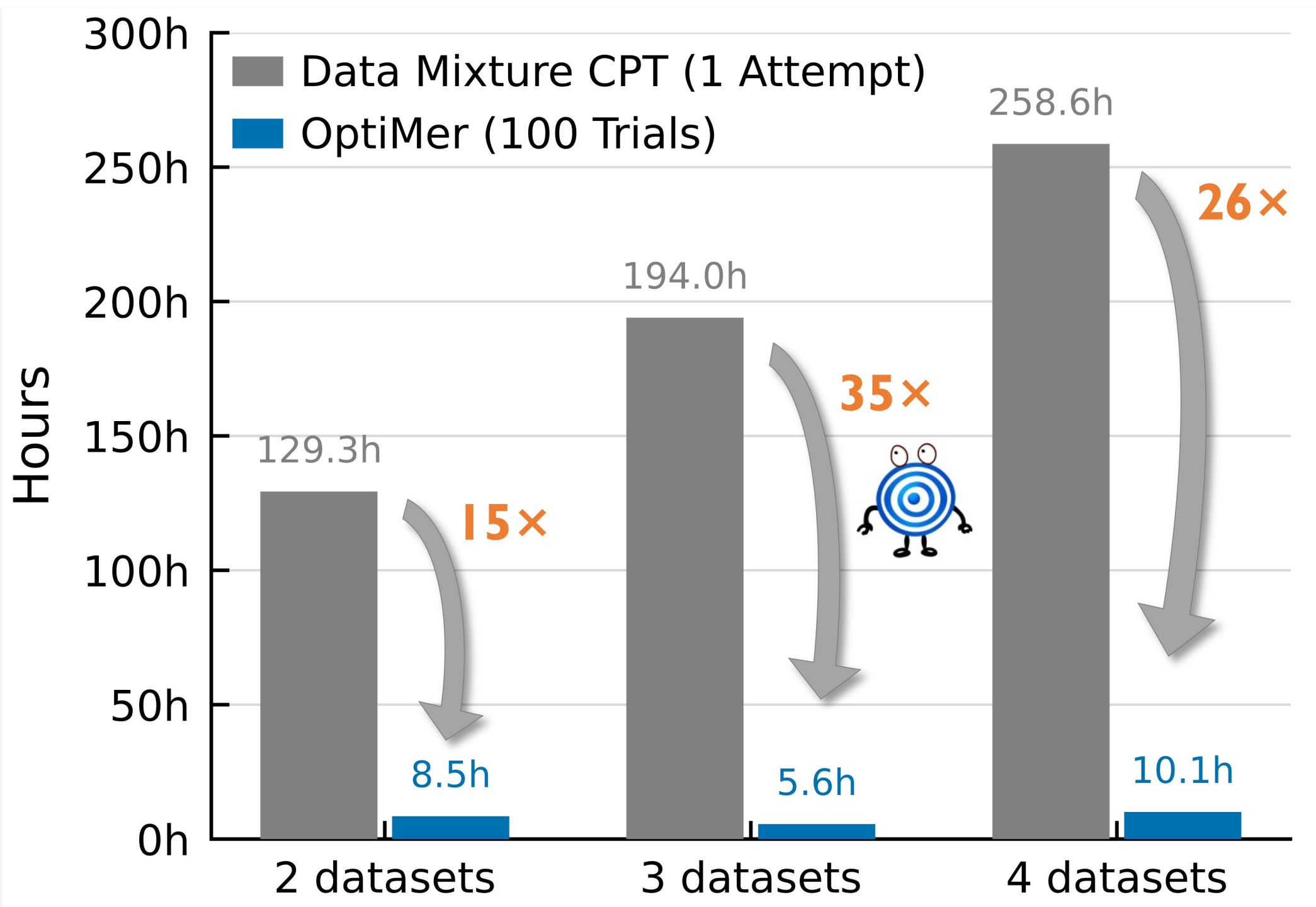
- DARE-Linear + Bayesian Optimization with 100 trials

■ Evaluation

- En / Ja / Zh / Math / Code benchmarks (1-shot)

OPTIMER is Efficient

- After the first round of training,
- Our **ratio optimization** is much more **efficient** than data mixing.



OPTIMER is Effective

Combination	Method	English				Math	Code		Japanese	Chinese	Avg.
		MMLU	ARC-C	HellaSwag	TQA	GSM8K	HE	MBPP	JA LB	C-Eval	
	Gemma 3 27B PT	31.59	27.30	48.41	54.71	0.83	0.00	0.60	39.04	27.41	25.54
	Gemma 3 27B IT	51.03	34.90	55.91	64.87	15.39	2.44	37.80	54.64	37.96	39.44
	<i>Single CPT, merged with IT ($\alpha_{it}=0.6$)</i>										
	CPT _{En}	73.60	69.80	82.82	46.39	79.98	71.95	68.40	43.06	62.56	66.51
	CPT _{Ja}	72.11	67.06	81.98	55.32	83.78	68.29	68.20	72.50	62.48	70.19
	CPT _{Zh}	72.85	68.94	81.70	48.84	82.18	71.95	67.80	57.98	62.85	68.34
	CPT _{Math}	75.07	70.14	82.72	43.57	82.64	57.32	68.00	72.79	63.37	68.40
	CPT _{Code}	71.86	70.05	81.71	51.16	81.96	64.63	68.40	73.39	62.26	69.49
	DataMix $_{\alpha_{it}=0.6}$ (Baseline)	73.17	68.77	81.72	49.57	83.62	50.00	68.80	73.34	61.74	67.86
Ja + Math	DataMixOptiMer ratio (Ours)	72.72	70.05	82.62	50.06	76.19	67.07	64.80	72.85	61.96	68.70
	<i>Task Arithmetic</i> (Ilharco et al., 2023)	75.44	69.88	83.31	41.49	74.45	60.37	1.60	75.11	63.67	60.59
	<i>TIES</i> (Yadav et al., 2023)	71.14	65.70	80.03	30.11	69.52	26.22	0.00	72.40	58.84	52.66
	<i>DARE-Linear</i> (Yu et al., 2024)	75.02	69.71	83.74	42.96	78.17	58.54	65.00	72.37	63.30	67.65
	<i>DARE-TIES</i> (Yu et al., 2024)	74.72	68.94	82.75	39.53	73.77	60.98	0.80	74.39	62.33	59.80
	OPTIMER (Ours)	73.17	68.34	82.50	55.20	84.46	59.76	70.40	72.53	63.45	69.98
	DataMix $_{\alpha_{it}=0.6}$ (Baseline)	71.23	69.20	81.14	43.08	78.62	64.02	65.00	72.38	60.40	67.23
	DataMixOptiMer ratio (Ours)	72.35	68.60	80.22	51.77	74.00	55.49	67.60	74.66	62.26	67.44
Ja + Code	<i>Task Arithmetic</i> (Ilharco et al., 2023)	72.33	70.39	82.44	42.72	78.17	45.12	60.20	74.04	61.96	65.26
	<i>TIES</i> (Yadav et al., 2023)	66.44	66.98	79.69	36.11	70.58	43.29	53.60	70.22	54.75	60.18
	<i>DARE-Linear</i> (Yu et al., 2024)	72.33	70.31	82.50	42.84	79.08	46.34	60.60	74.07	62.18	65.58
	<i>DARE-TIES</i> (Yu et al., 2024)	71.14	69.37	82.00	41.49	76.04	42.68	60.40	73.30	60.25	64.07
	OPTIMER (Ours)	71.41	66.89	81.44	54.22	83.47	64.63	70.20	73.06	61.81	69.68
	DataMix $_{\alpha_{it}=0.6}$ (Baseline)	73.83	69.03	80.54	37.70	79.61	32.32	65.40	72.77	62.18	63.71
	DataMixOptiMer ratio (Ours)	74.34	67.58	81.80	44.80	81.27	66.46	67.20	74.05	60.48	68.66
Ja + Zh + Math	<i>Task Arithmetic</i> (Ilharco et al., 2023)	75.27	69.97	83.18	38.92	73.54	45.73	0.40	73.17	64.41	58.29
	<i>TIES</i> (Yadav et al., 2023)	70.13	66.72	79.65	26.68	65.81	14.63	0.20	69.86	60.03	50.41
	<i>DARE-Linear</i> (Yu et al., 2024)	74.48	69.28	83.65	38.68	76.42	50.00	62.00	70.92	63.89	65.48
	<i>DARE-TIES</i> (Yu et al., 2024)	74.06	69.54	82.59	35.86	70.36	45.12	0.40	72.38	63.60	57.10
	OPTIMER (Ours)	72.94	68.60	82.98	51.77	83.93	67.07	70.20	72.65	63.22	70.37

OPTIMER is Flexible

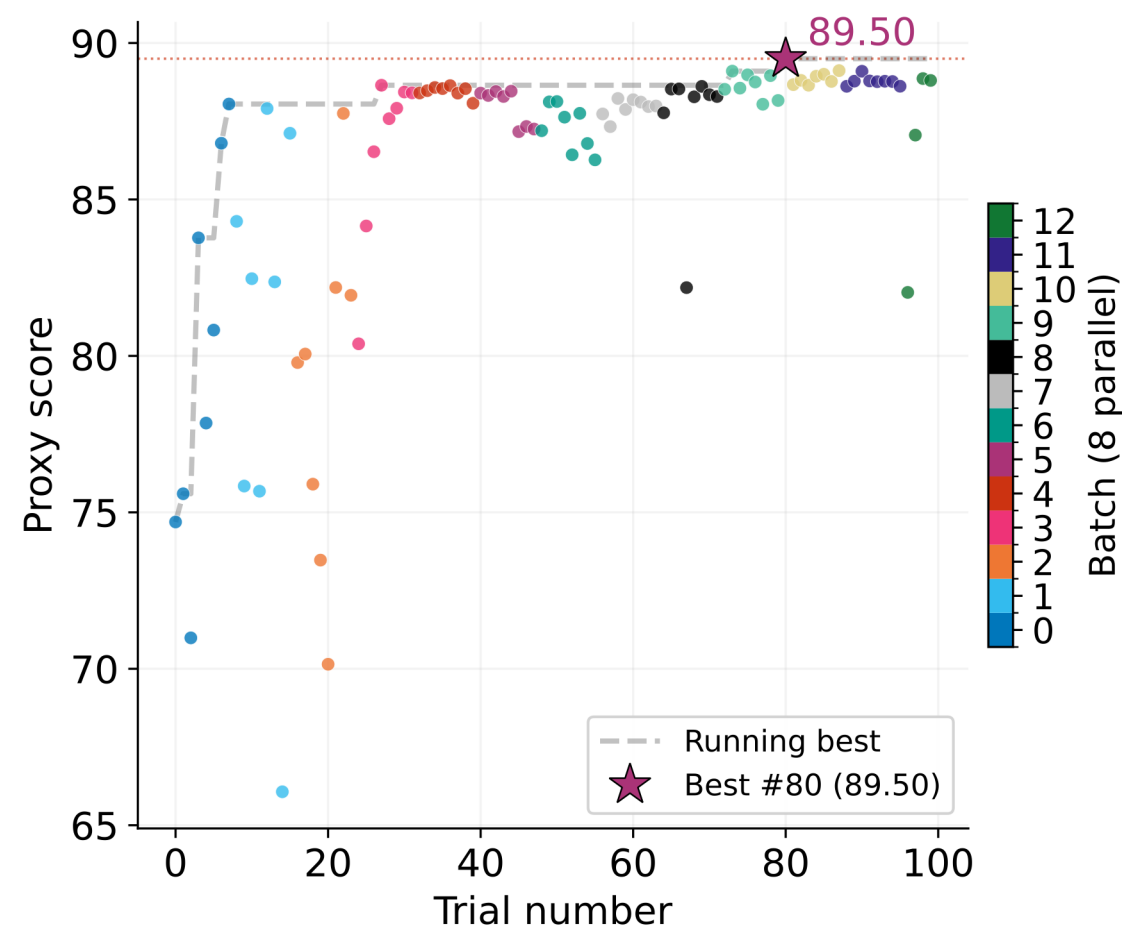
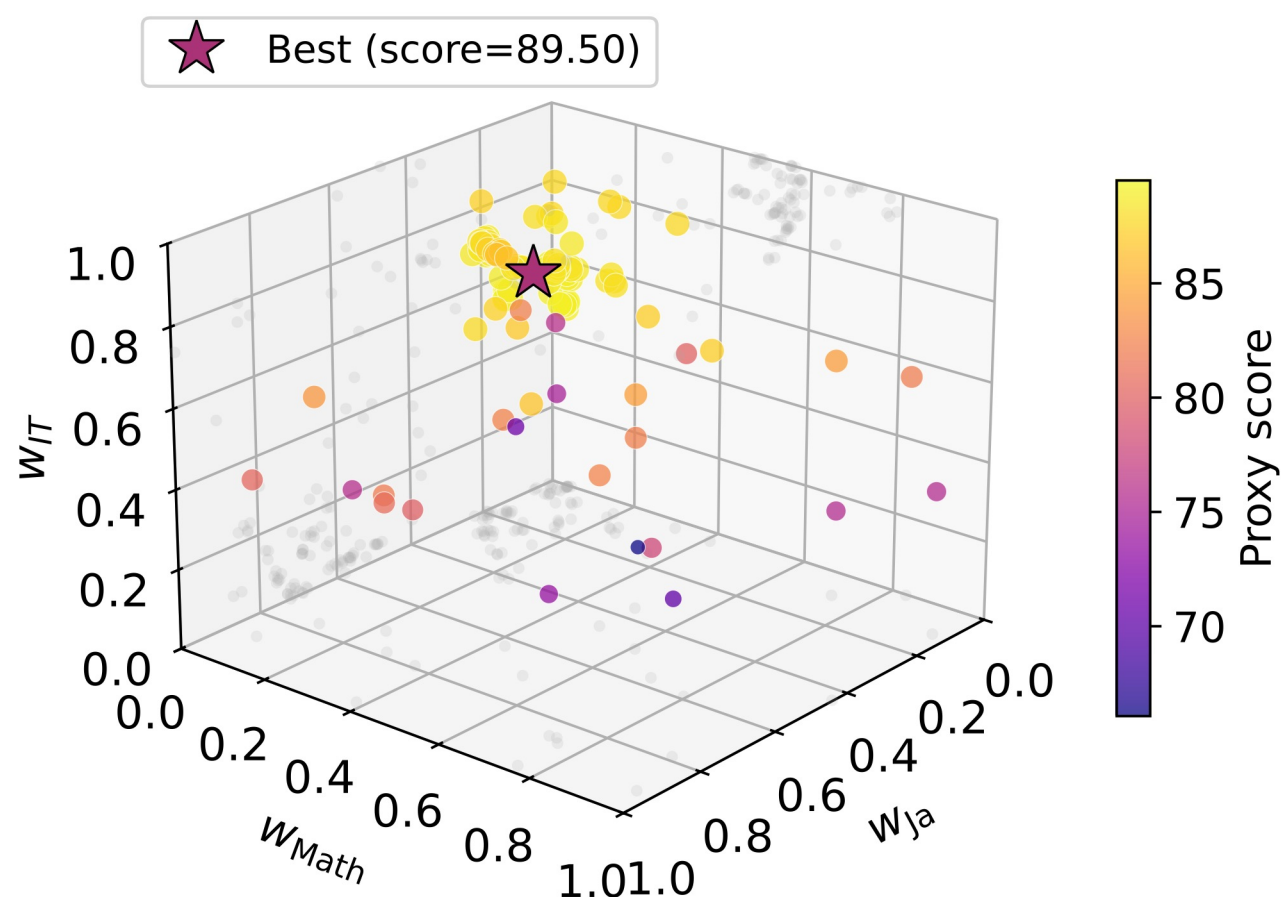
- OptiMer **quickly adapts an LLM to another domain through re-optimization.**
 - Using the already calculated distribution vectors {Ja, Zh, En, Math},
 - We can set another goal and only re-optimize the vector merging ratios, in several **hours**, **without training the whole model again!**

Objective	English				Math	Code		Japanese	Chinese	Avg.
	MMLU	ARC-C	HellaSwag	TQA	GSM8K	HE	MBPP	JA LB	C-Eval	
English	72.87	69.62	82.02	47.74	82.49	44.51	71.40	69.56	62.93	65.35
Math	70.10	63.05	79.16	41.13	82.87	4.88	72.60	68.47	58.99	58.37
Japanese	73.96	68.86	82.73	46.39	84.69	68.29	69.20	73.37	62.85	69.07
Chinese	71.93	69.11	82.37	34.76	72.71	11.59	1.60	40.78	63.60	44.29



OPTIMER's Dynamics

- Good ratio combinations are concentrated (large IT, small CPT weights)
 - -> making grid search impractical
- OptiMer quickly approaches this good region, usually converging within 100 trials.

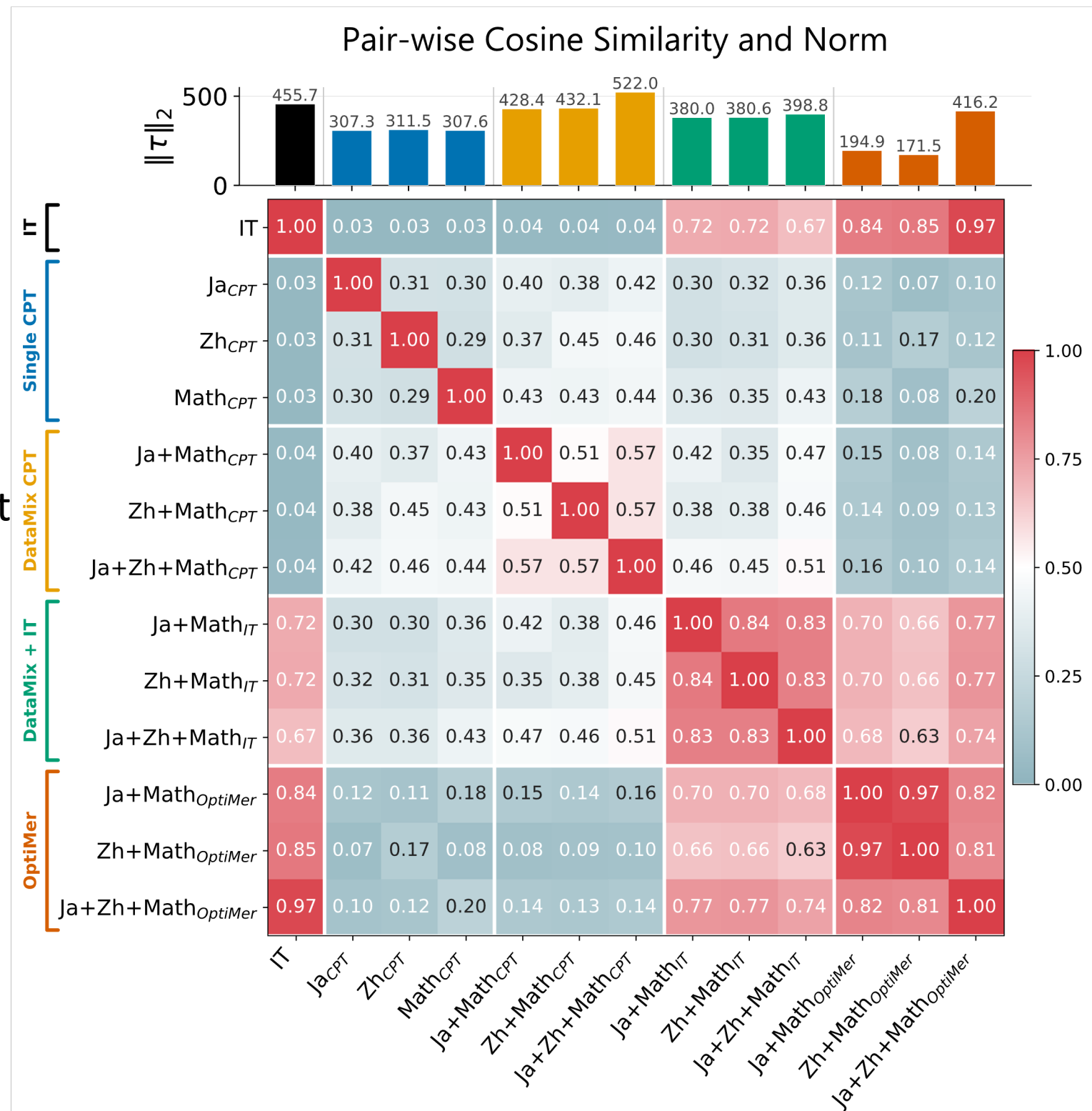


Why Mergeable?

Distribution vectors are nearly **orthogonal**.

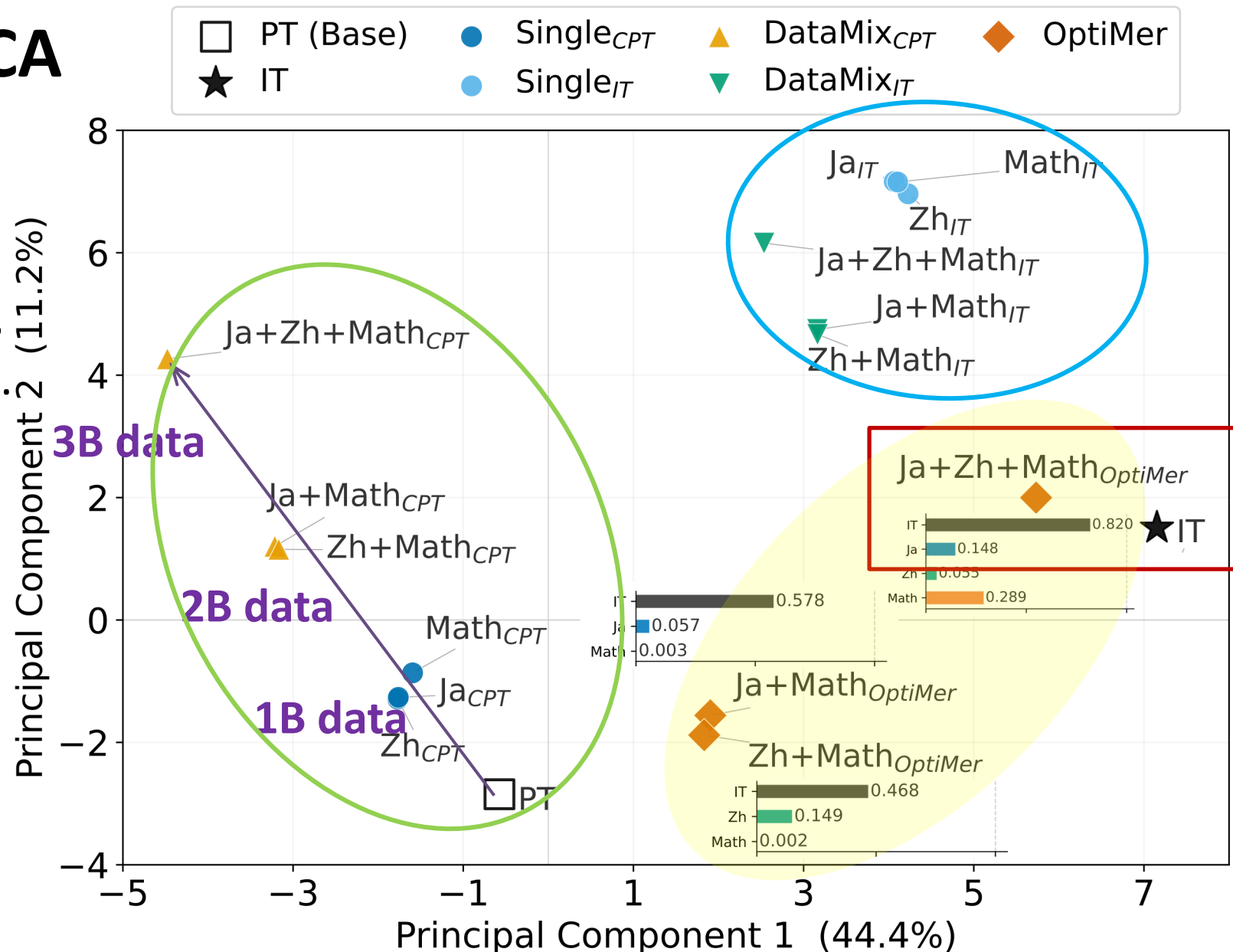
- CPT and IT vectors are nearly **orthogonal** (0.03).
- Different CPT vectors also exhibit **low similarity** (0.29–0.31).

→ Vectors are mergeable with low conflicts.



Distribution Vector PCA

- **CPT** and **IT** form two groups
- Model shifts almost linearly as CPT data (1B $\rightarrow$ 3B) increases.
- OptiMer models form a group
- The **best model** is near **Gemma-IT**



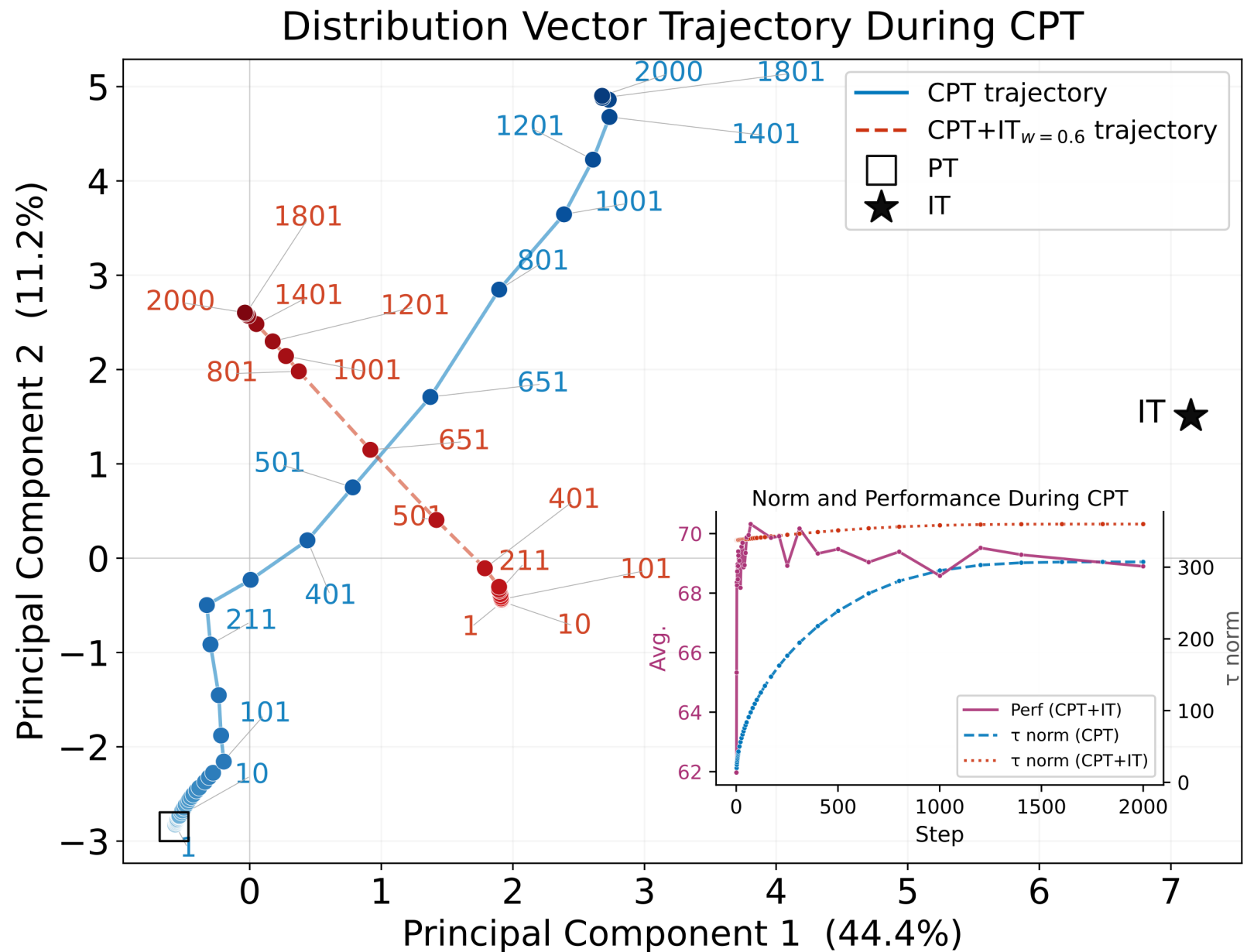
Continual Pre-Training Dynamics

- Vector is **almost linear**

- It suggests that controlling the **vector merging ratio** implies controlling the **training duration**.

- Norm keeps increasing

- Performance $\uparrow$ then $\searrow$



Applications

- We applied **OPTIMER** to a model for ASEAN languages, SEA-LION-v4-27B-PT, to **further enhance its Japanese capability**,
 - using distribution vectors from: Japanese, Chinese, English, ASEAN langs, Math, Code
 - Performance on Japanese tasks is improved from 66.34 to 74.40 (+8.1), while the performance is maintained for ASEAN tasks.

Take-Home Message

- **OPTIMER** adapts open-sourced LLMs
- to Your language and Your domain,
- efficiently, effectively, and flexibly.

Thank you for your attention!

