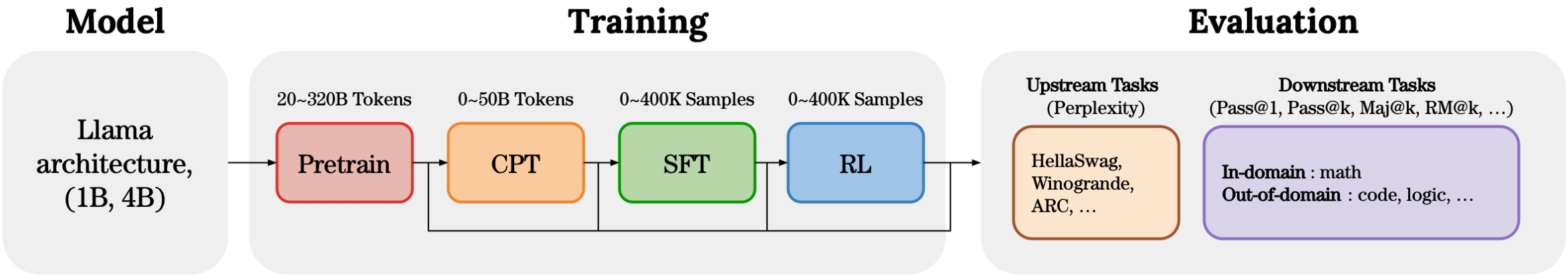
Diffract

Spectral View of LLM Domain Adaptation

Nikita Borodin[1], Maria Krylova[1], Artem Zabolotnyi[1 2], Dmitry Aspisov[1], Egor Shikov[1], Nikita Tyuplyaev[1], Oleg Travkin[1], Roman Alferov[1], Dmitry Vinichenko[1]

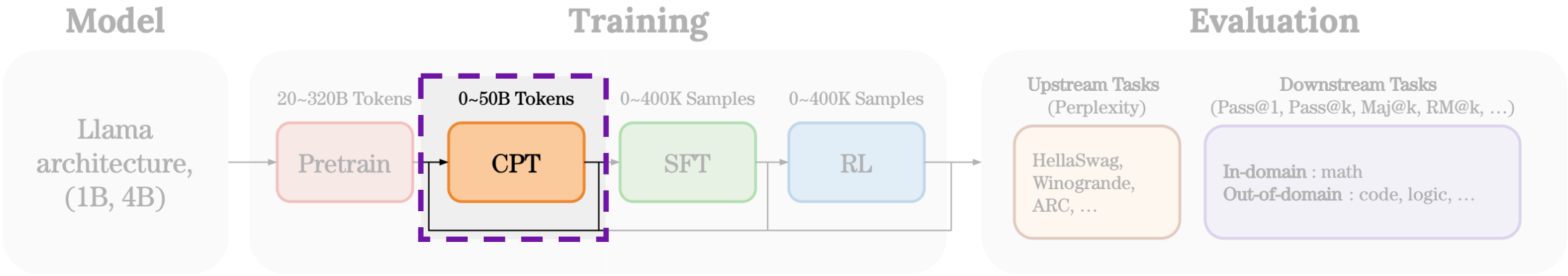
[1] Risk AI Research Lab · [2] Applied AI Institute

Continual pre-training – important stage for LLM domain adaptation



[Qi et al., 2025]

Continual pre-training – important stage for LLM domain adaptation



[Qi et al., 2025]

CENTRAL IN MODERN PIPELINES

Produces domain-specialized models like **DeepSeek-Math** and **LLEMMA**

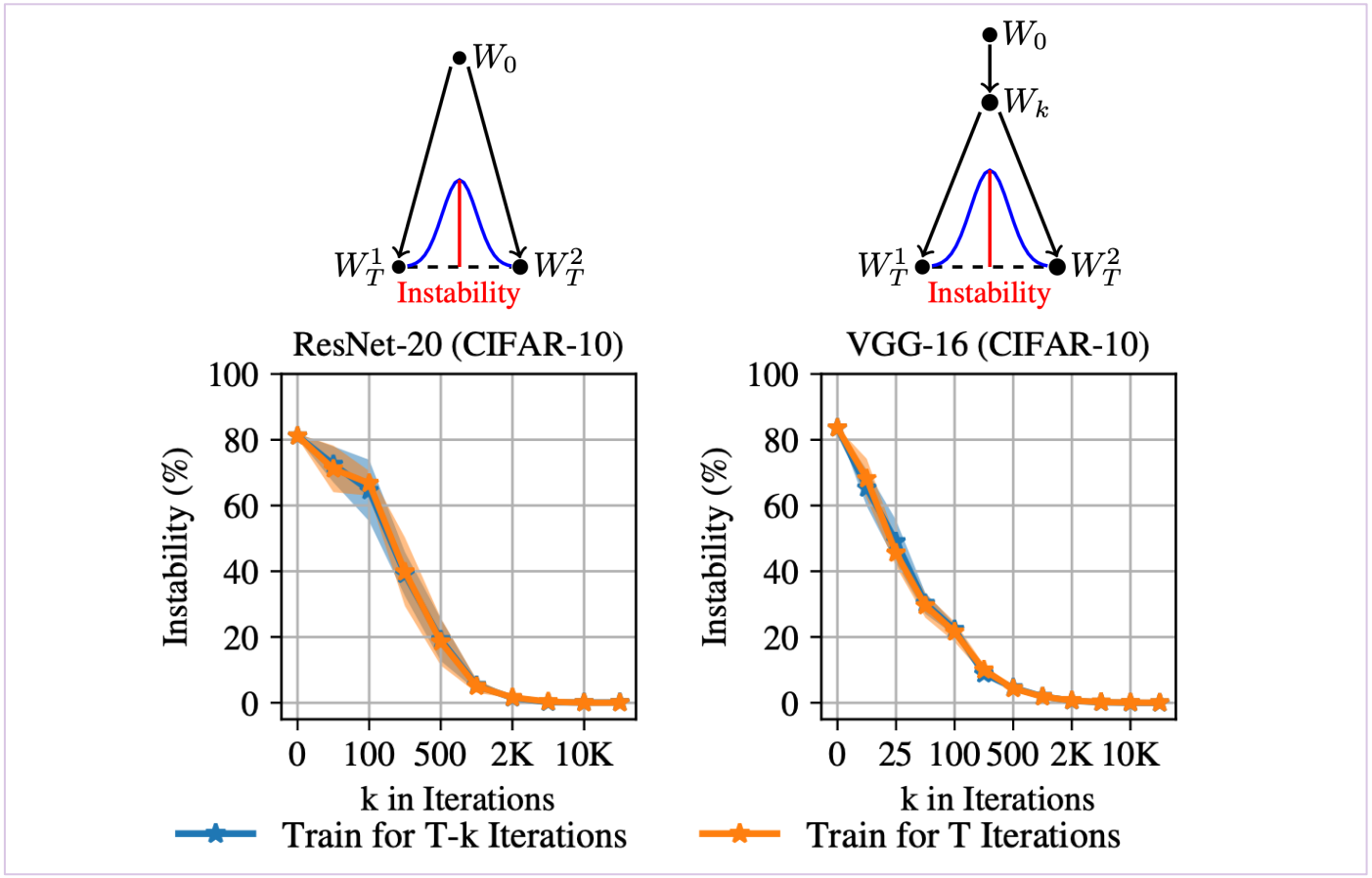
deepseek

LLEMMA

A wide range of phenomena and properties is known for SFT stage

1 Linear mode connectivity

Linear Mode Connectivity

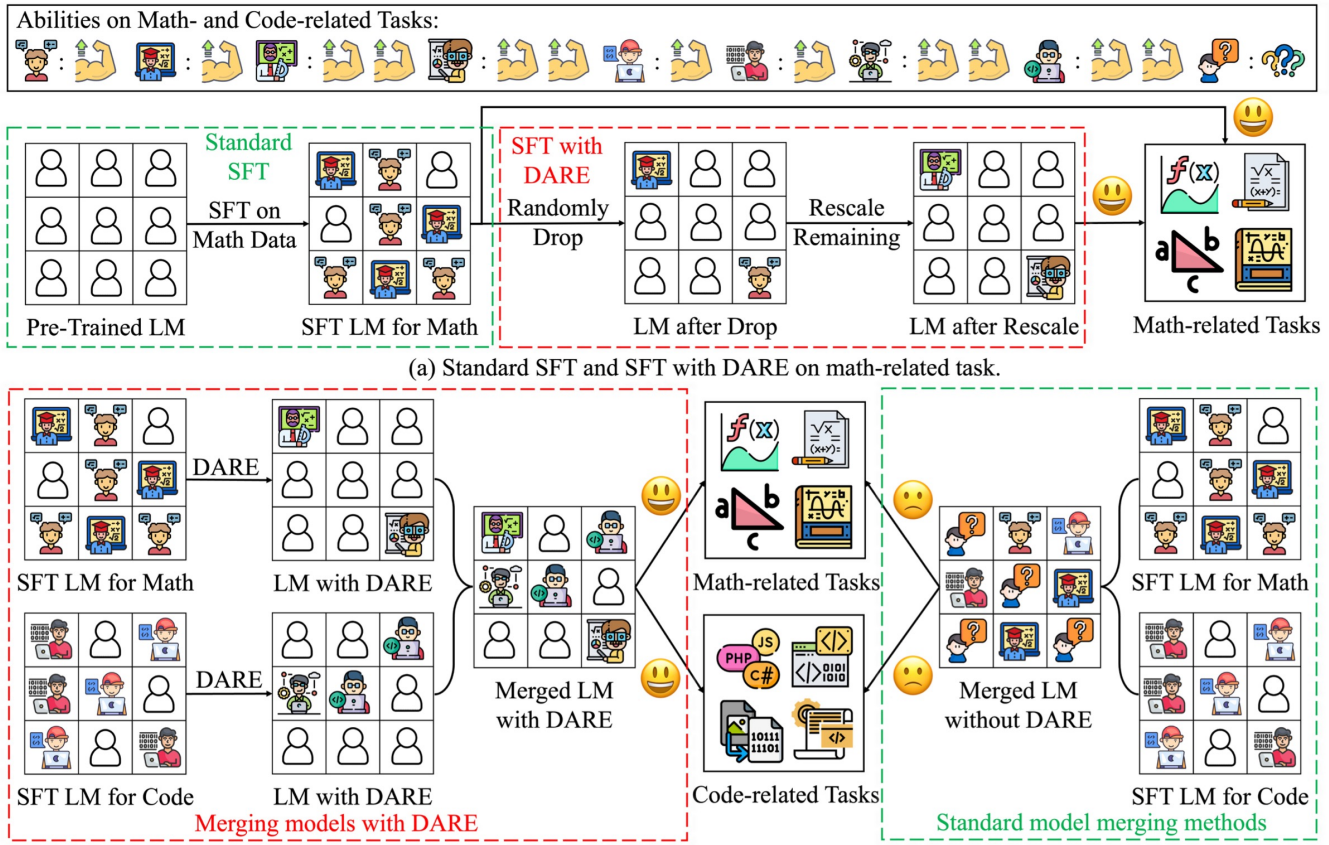


[Frankle et al., 2020]

A wide range of phenomena and properties is known for SFT stage

1. Linear mode connectivity
2. Task arithmetic, ability transfer

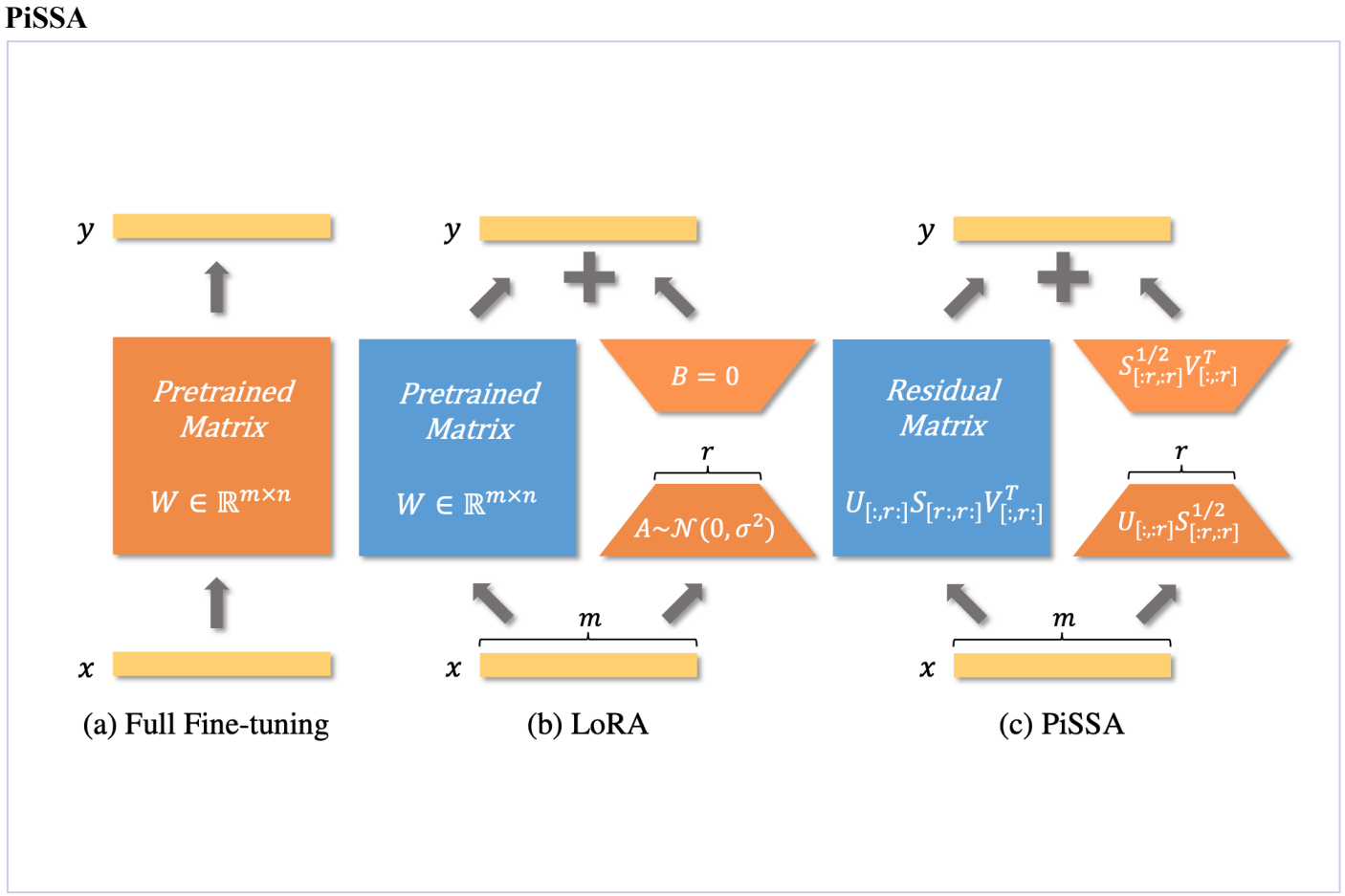
Language Models are Super Mario



[Yu et al., 2024]

A wide range of phenomena and properties is known for SFT stage

1. Linear mode connectivity
2. Task arithmetic, ability transfer
3. Low-rank subspace modification

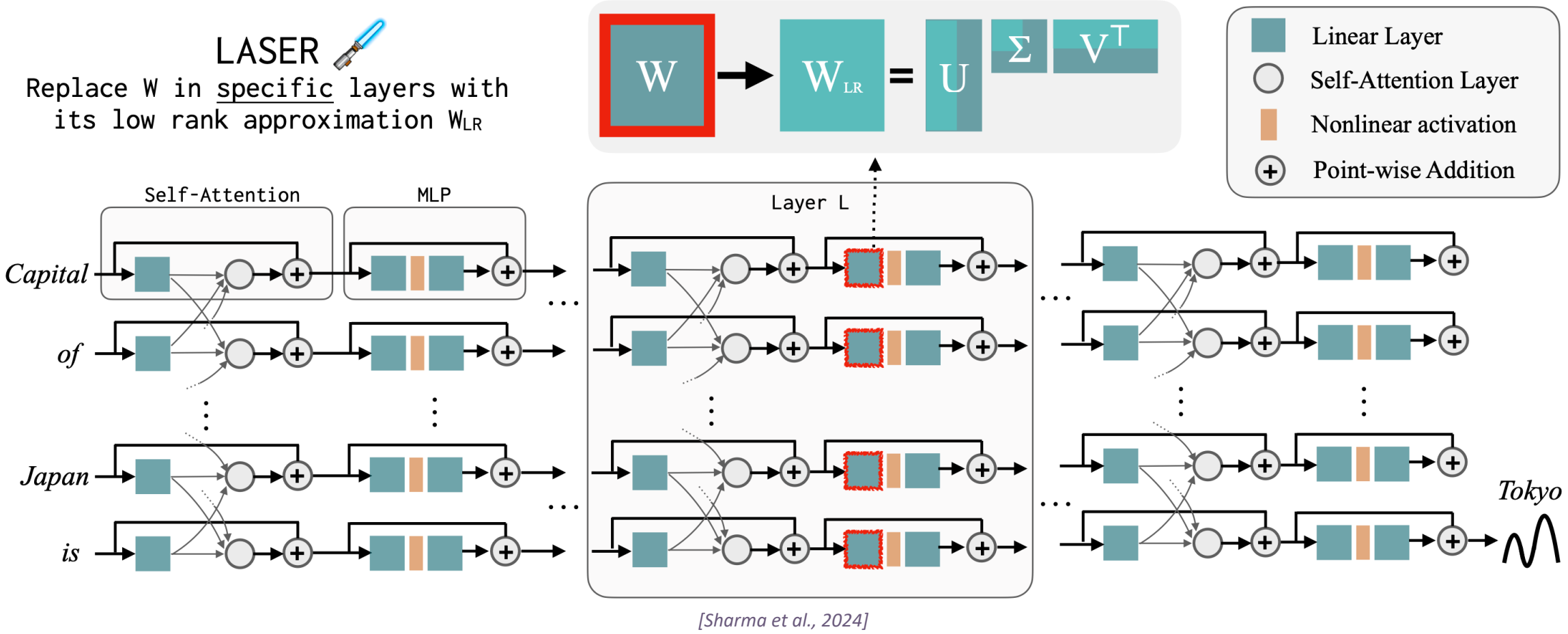


[Meng et al., 2024]

Objective: Investigate the CPT stage and identify its properties

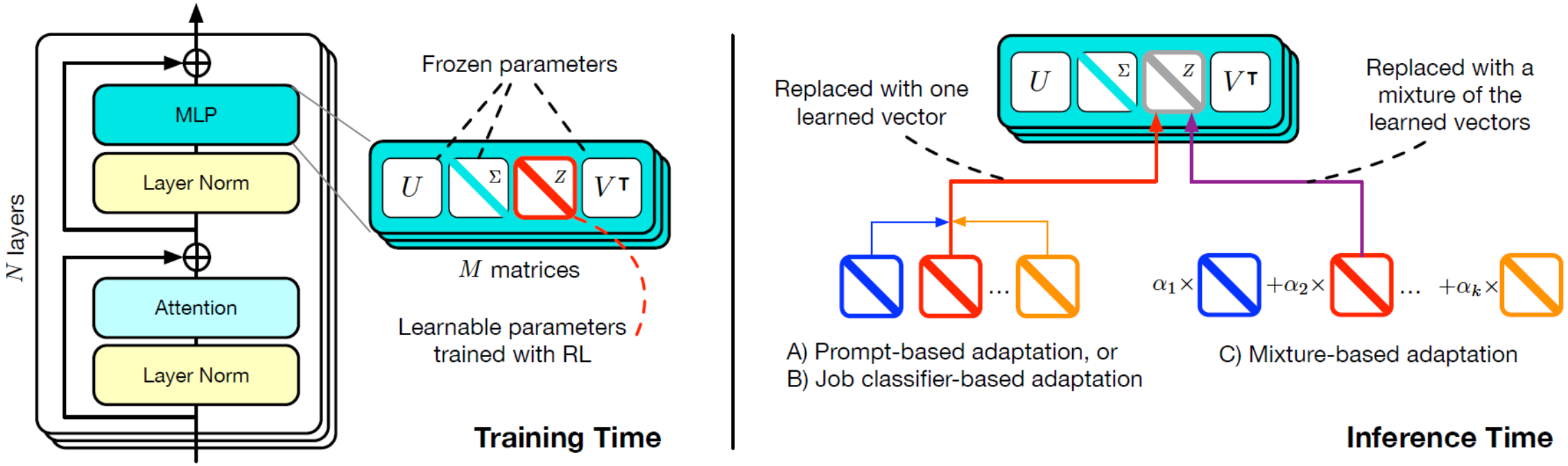
Investigation of weight matrices spectra can lead to new insights

LASER



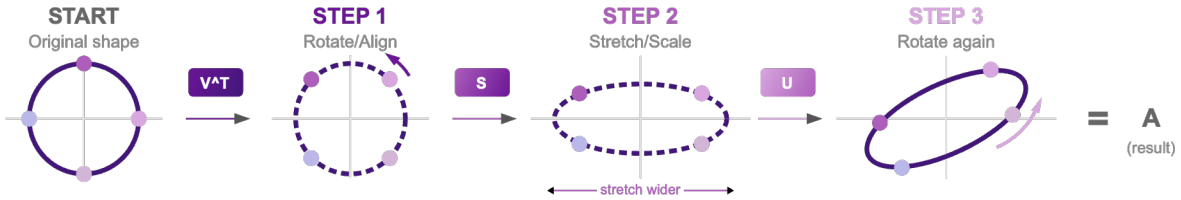
[Sharma et al., 2024]

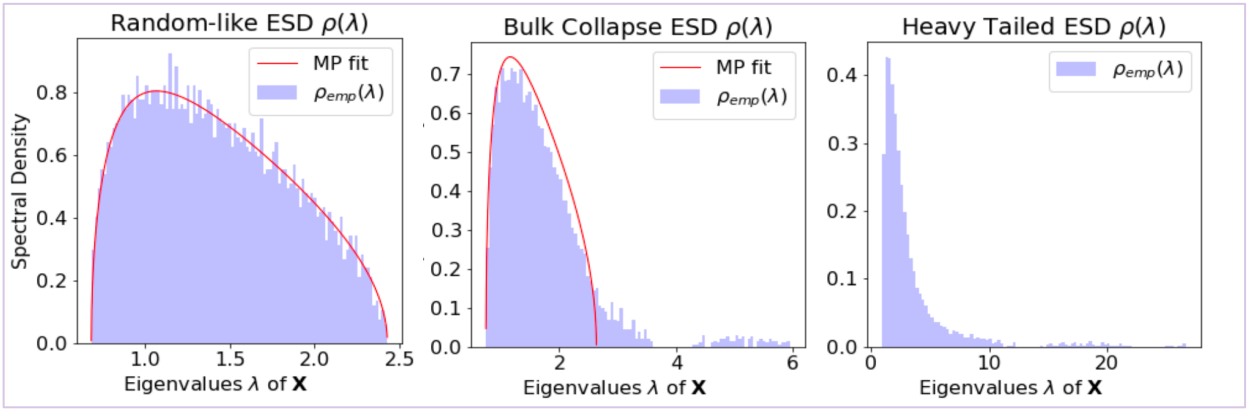
Investigation of weight matrices spectra can lead to new insights Transformer-Squared



[Sun et al., 2025]

Investigation of weight matrices spectra can lead to new insights
Heavy-tailed self-regularization theory





[Martin & Mahoney, 2021]

Marchenko-Pastur distribution for GOE random matrices

$$W_{i,j} \sim \mathcal{N}(0, \mathrm{Var}(\boldsymbol{W}))$$

$$\rho^{\mathrm{MP}}(\lambda; \boldsymbol{X}) = \begin{cases} \dfrac{m}{2\pi n \cdot \mathrm{Var}(\boldsymbol{W})} \dfrac{\sqrt{(\lambda^{\max} - \lambda)(\lambda - \lambda^{\min})}}{\lambda}, & \lambda \in [\lambda^{\min}, \lambda^{\max}], \\ 0, & \text{otherwise}, \end{cases}$$

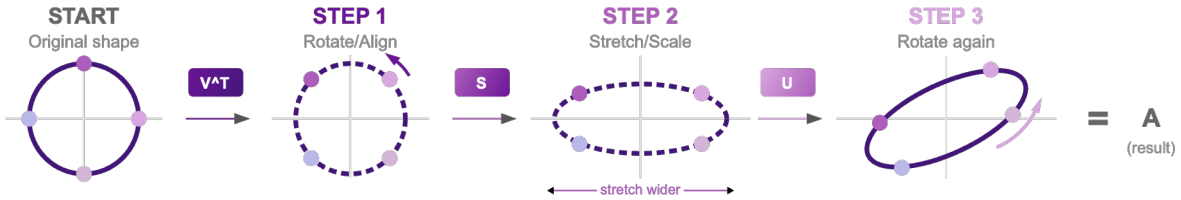
$$\lambda^{\max\,/\,\min}(\boldsymbol{X}) = \mathrm{Var}(\boldsymbol{W})\left(1 \pm \sqrt{n/m}\right)^2$$

Power law tails develop as training progresses

$$\rho^{\mathrm{PL}}(\lambda; \boldsymbol{X}) = c \cdot \lambda^{-\alpha}$$

Hypothesis – power law parameter and other spectral properties correlate with model quality

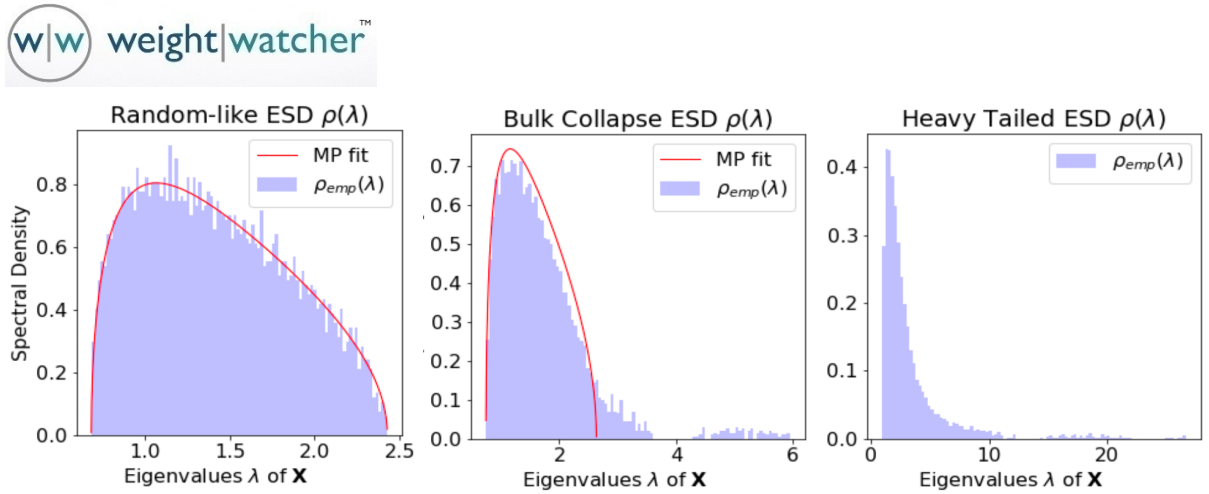
Investigation of weight matrices spectra can lead to new insights
Heavy-tailed self-regularization theory



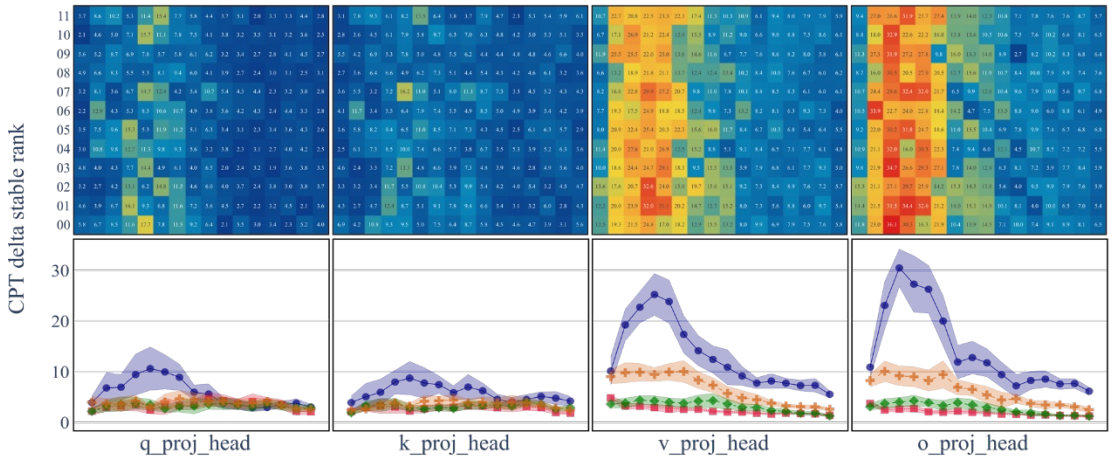
DIFFRACT · OPEN SOURCE PYTHON PACKAGE

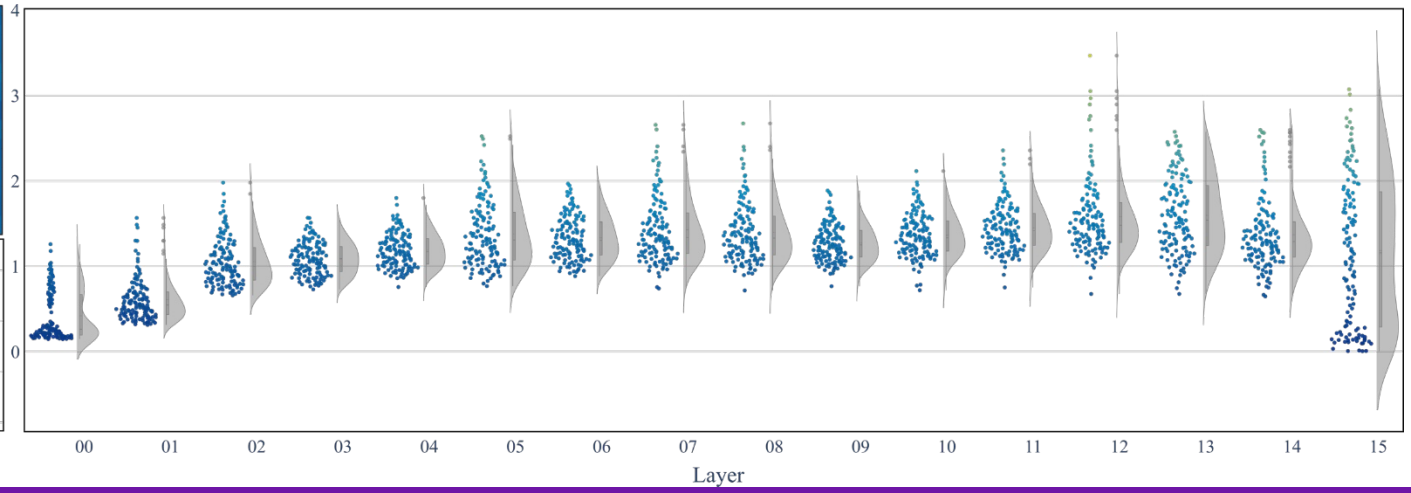
`https://github.com/Risk-AI-Research/diffract`

SVD-based weight analysis for billion-parameter LLMs



[Martin & Mahoney, 2021]





We use Diffract package to investigate CPT stage through the spectral lens

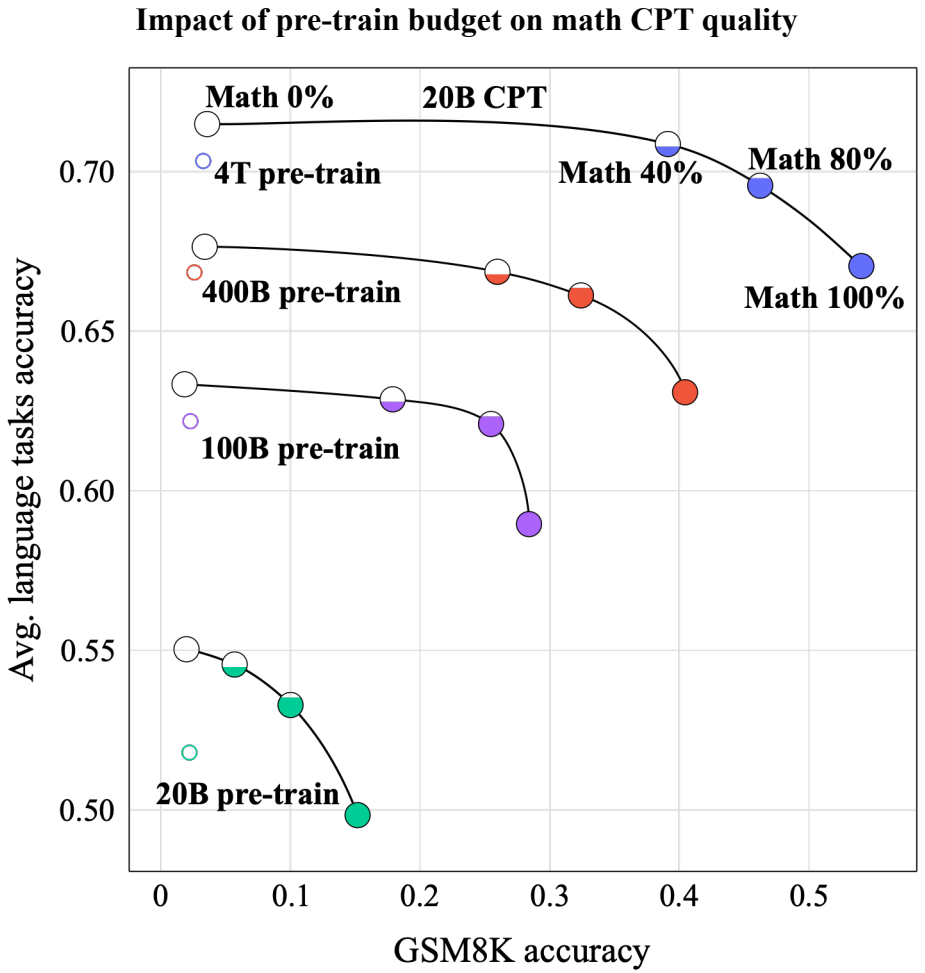
Experimental setup: OLMo 2 pre-train + CPT

MODEL SIZE	PRE-TRAIN BUDGET
1B:	4T, 20 – 400B
7B:	4T
13B:	5T
32B:	6T

CPT DOMAINS

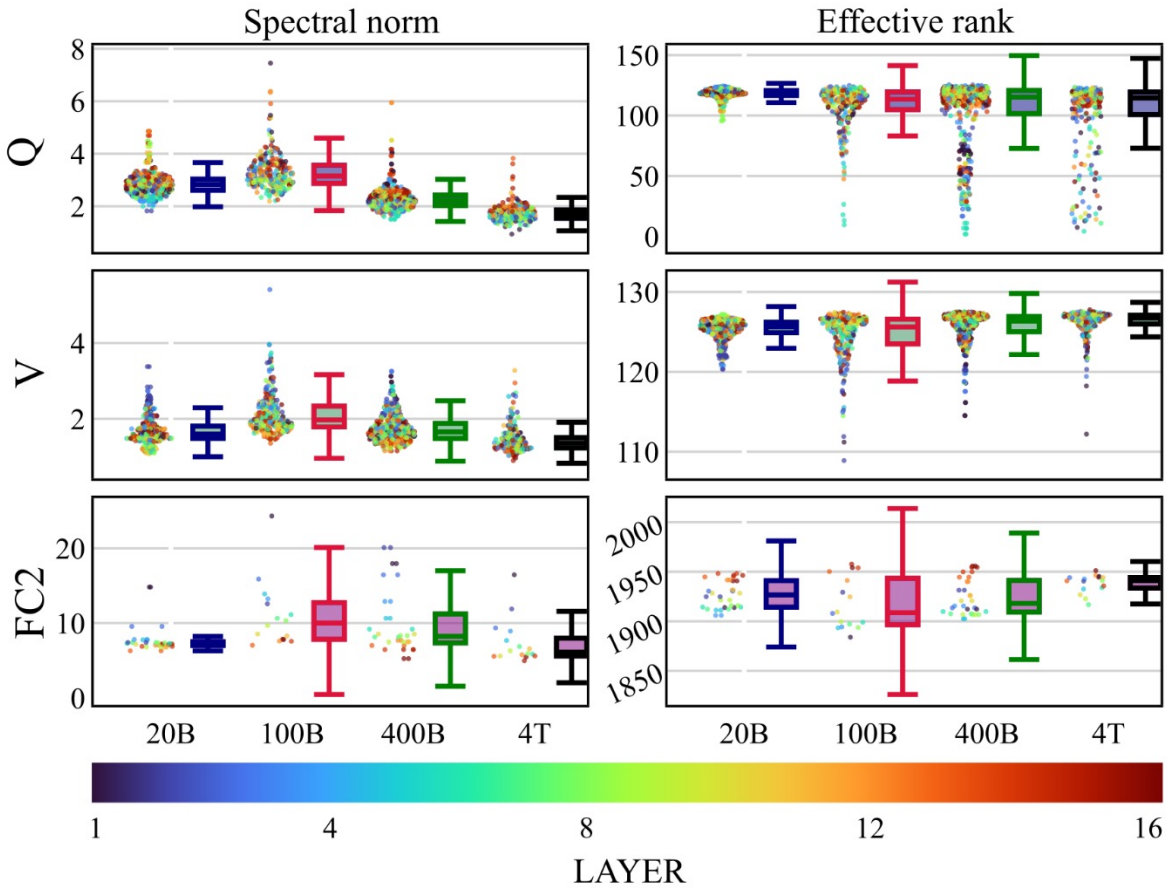
- MATH
- INSTRUCT
- CODE
- TEXT

OLMo-2 CPT checkpoints



Takeaway: longer pre-train enables higher quality after CPT

Spectral evolution along the pre-train stage and CPT spectral properties



1B model
Spectral properties of attention heads:

Frobenius norm $\|\boldsymbol{W}\|_F = \sqrt{\sum_i \sigma_i(\boldsymbol{W})^2}$

Spectral norm $\|\boldsymbol{W}\|_2 = \max_i \sigma_i(\boldsymbol{W})$

Stable rank $R^s(\boldsymbol{W}) = \frac{\|\boldsymbol{W}\|_F^2}{\|\boldsymbol{W}\|_2^2}$

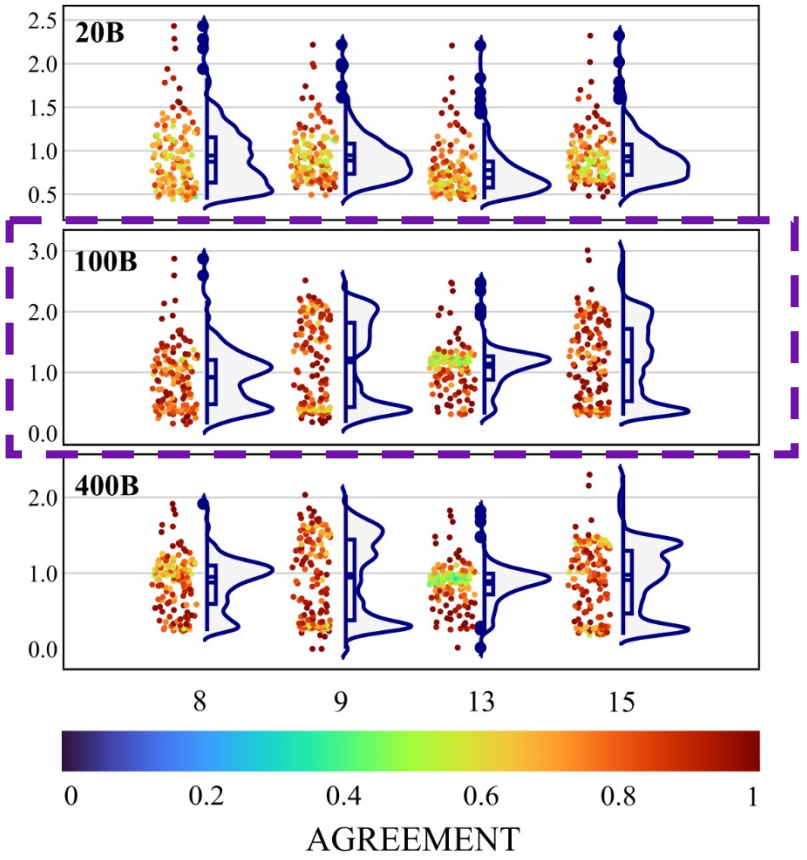
Effective rank

$$R^e(\boldsymbol{W}) = \exp\left(-\sum_{i=1}^{R(\boldsymbol{W})} \frac{\sigma_i(\boldsymbol{W})}{\sum_j \sigma_j(\boldsymbol{W})} \log\left(\frac{\sigma_i(\boldsymbol{W})}{\sum_j \sigma_j(\boldsymbol{W})}\right)\right)$$

Spectral evolution along the pre-train stage and CPT spectral properties

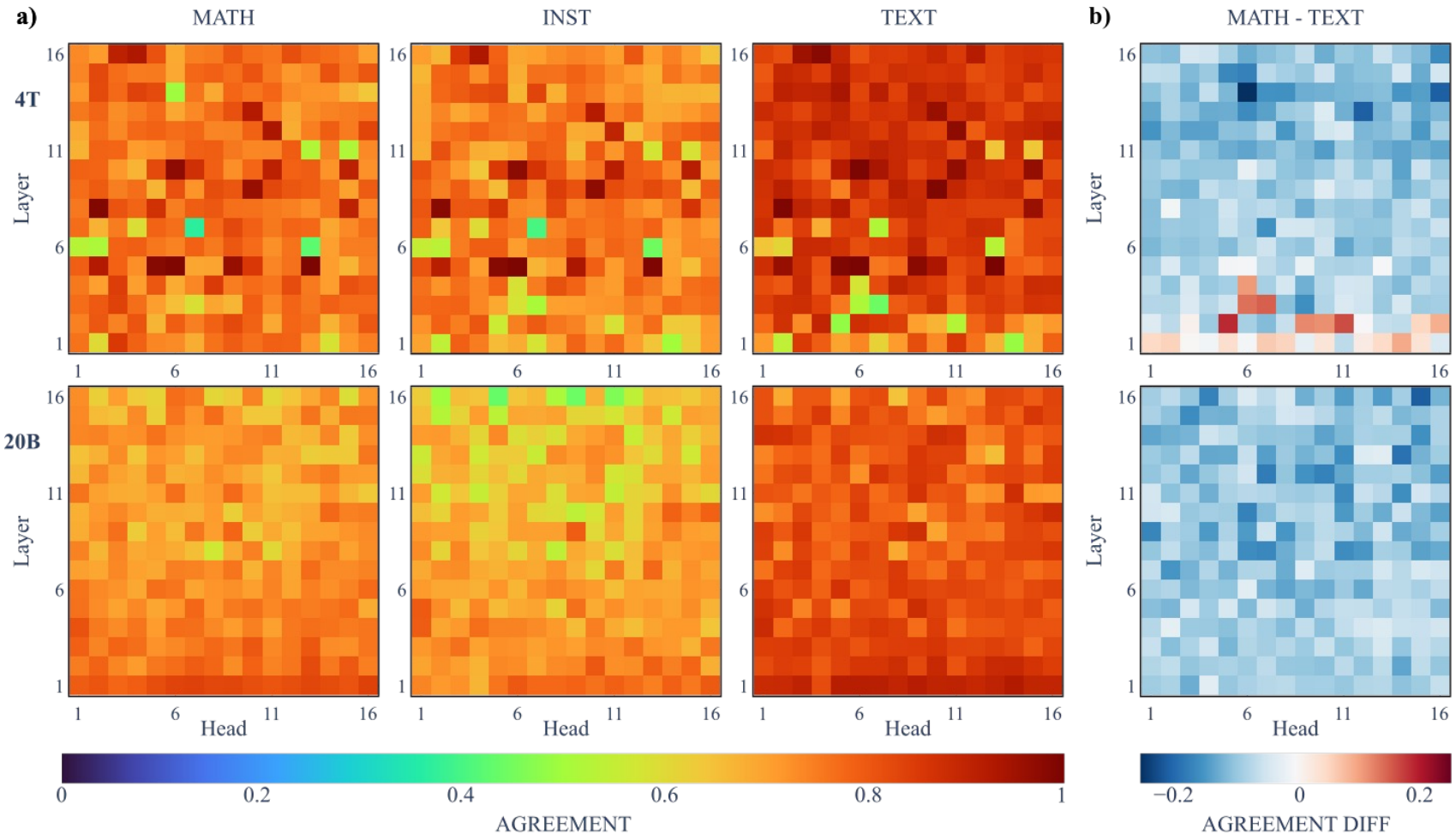
KEY OBSERVATIONS

- Attention spectra develop multi-peak structure at ~100B tokens
- CPT barely changes weights singular spectra
- CPT delta: $\Delta \boldsymbol{W} = \boldsymbol{W}^{\mathrm{domain}} - \boldsymbol{W}^{\mathrm{pre-train}}$, can be expressed via singular vector agreement
- CPT adaptation occurs in singular vectors at spectral peaks

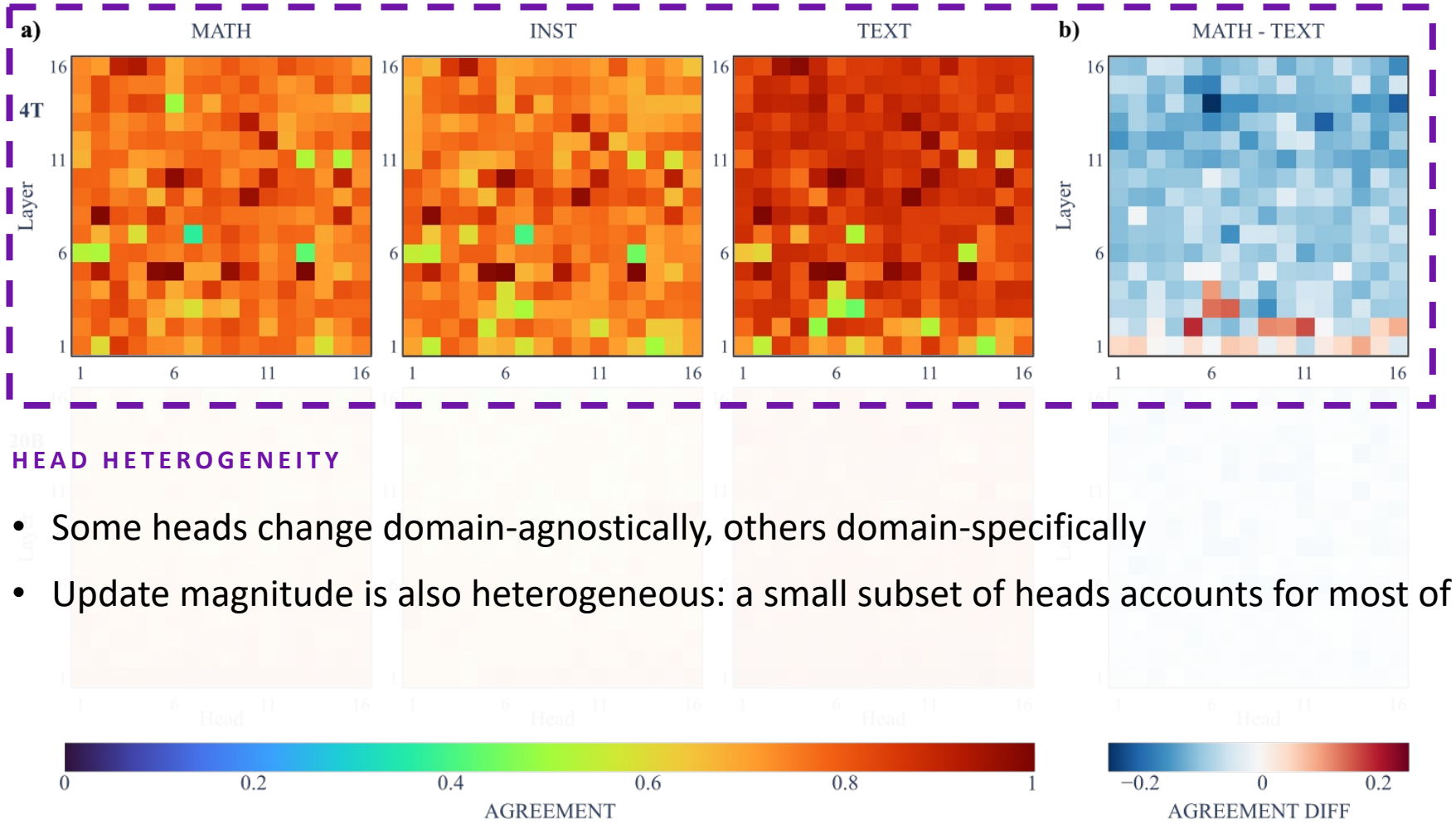


Takeaway: complex spectral shape enables finer CPT delta structure

Structure of CPT delta – head heterogeneity



Structure of CPT delta – head heterogeneity



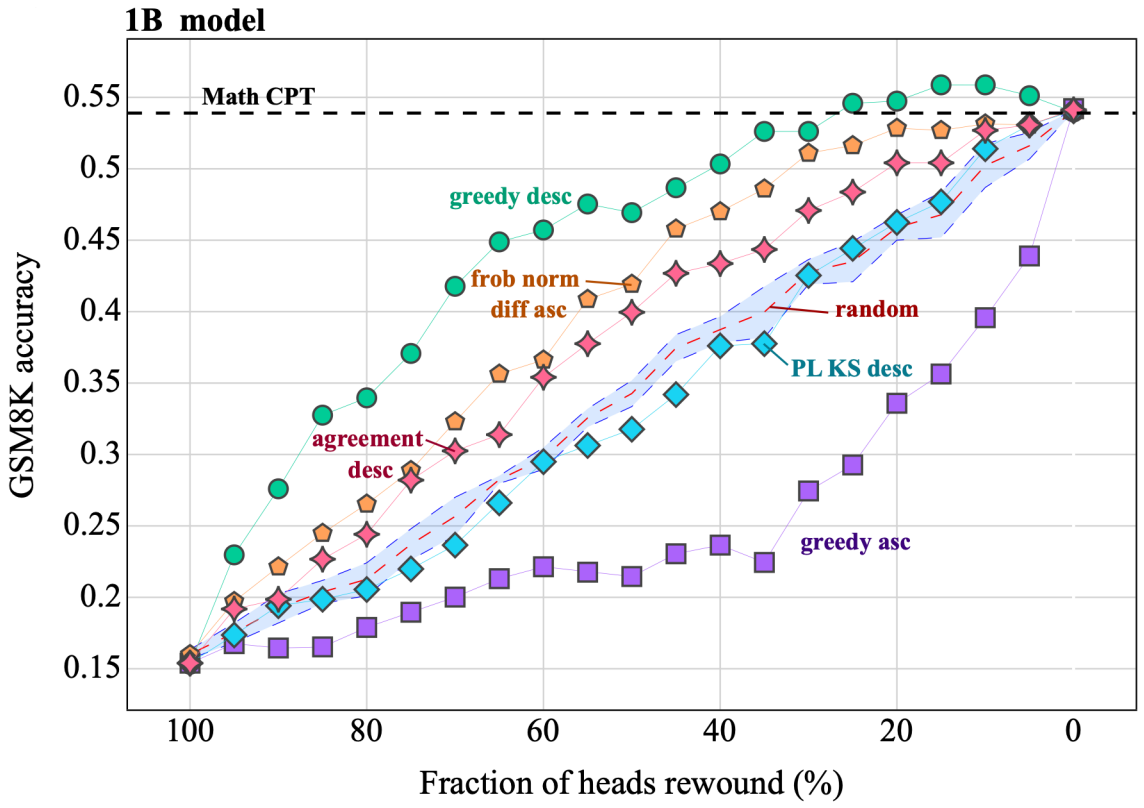
HEAD HETEROGENEITY

- Some heads change domain-agnostically, others domain-specifically
- Update magnitude is also heterogeneous: a small subset of heads accounts for most of the change

Takeaway: CPT changes are concentrated and domain-specific

Structure of CPT delta – head-wise rewind
Effect of head heterogeneity on model quality

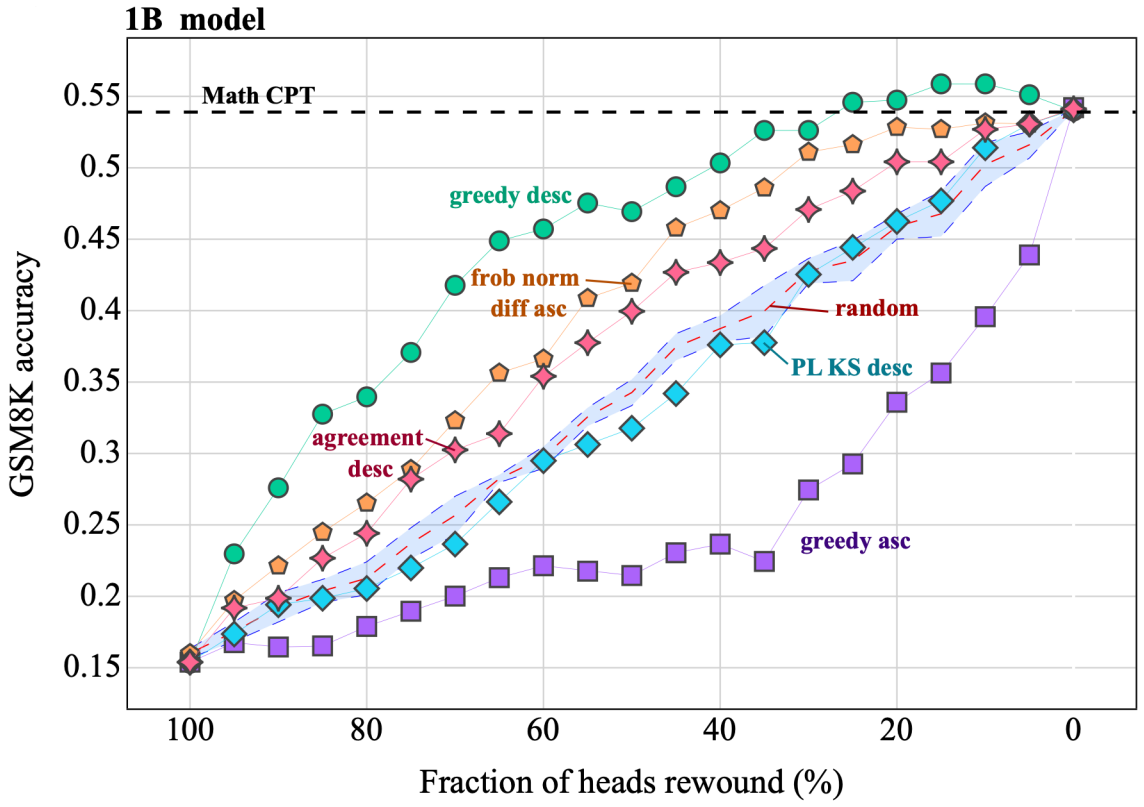
20B math CPT after **4T** pre-train



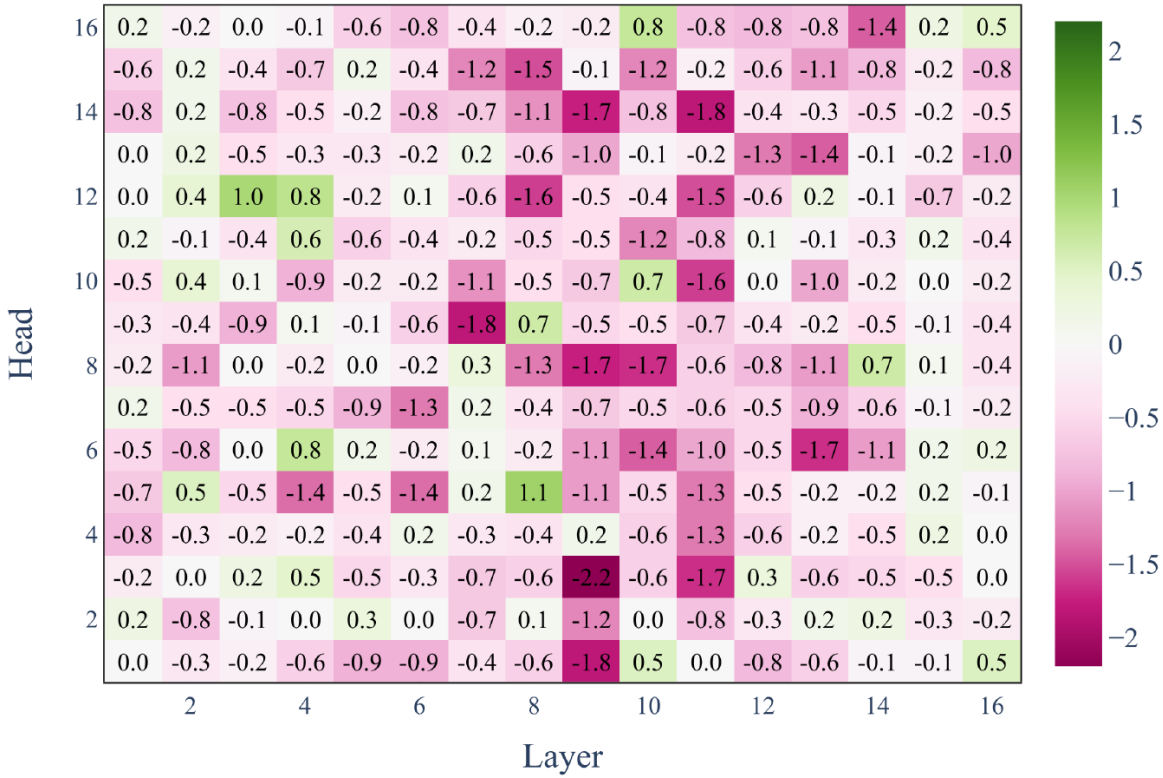
Structure of CPT delta – head-wise rewind

“Greedy” rewind pattern

20B math CPT after **4T** pre-train

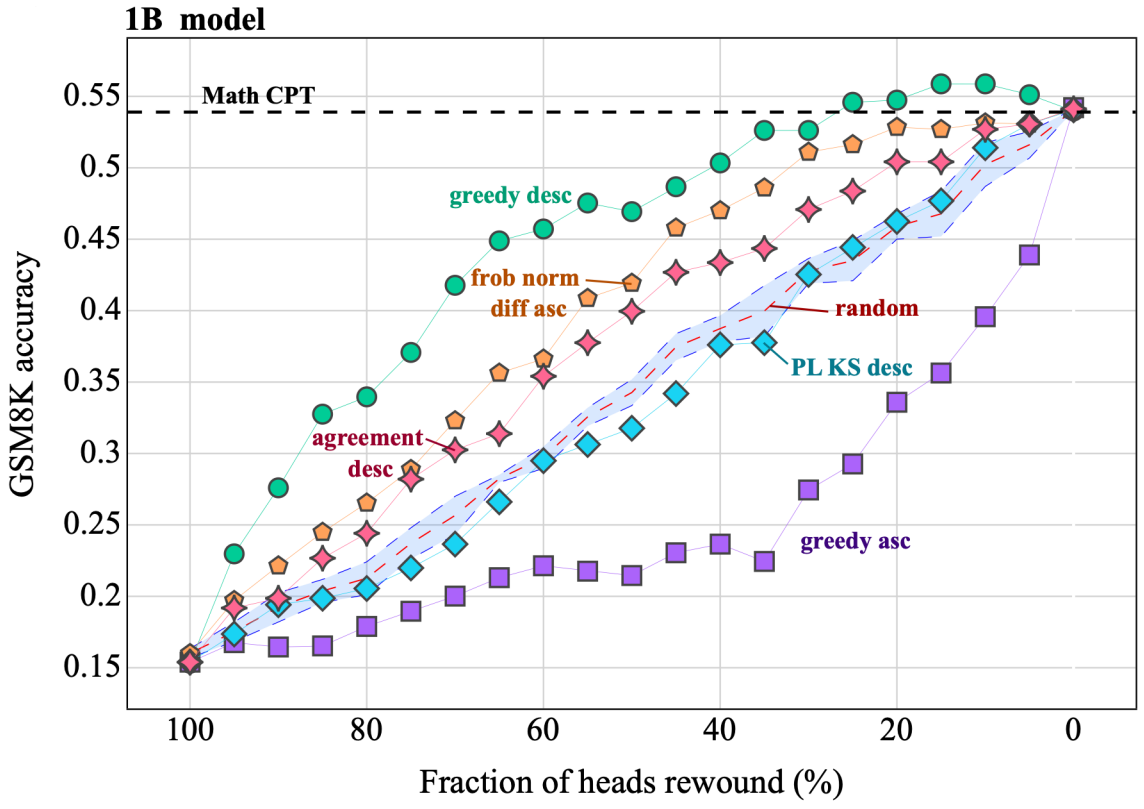


Heads are ranked by the effect of individual rewind on model GSM8K quality

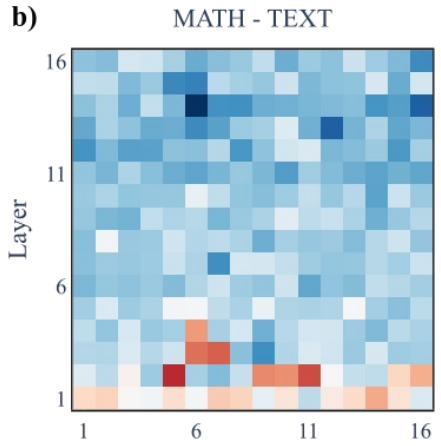


Structure of CPT delta – head-wise rewind
“Reference domain” rewind pattern

20B math CPT after **4T** pre-train



Heads are ranked by difference between math and text CPTs

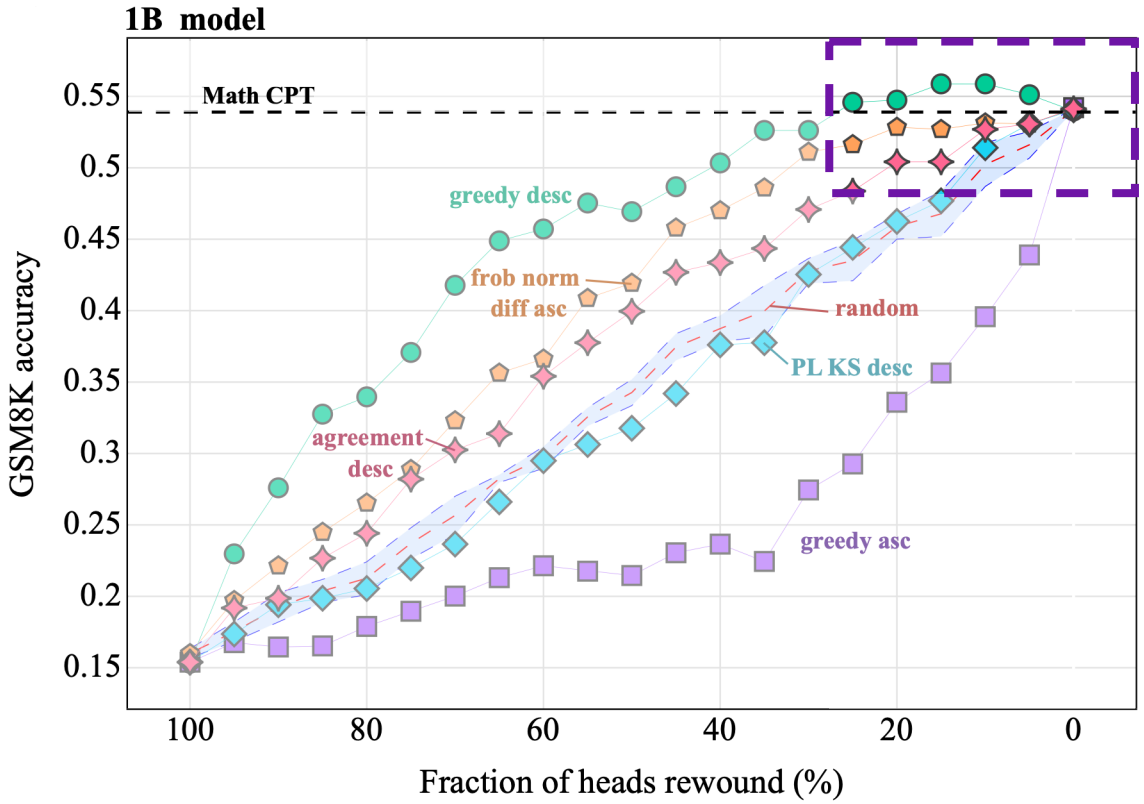


$$s_{l,h} = \text{scale}_{[0,1]}\Big(\|\boldsymbol{W}_{l,h}^{\text{domain}} - \boldsymbol{W}_{l,h}^{\text{pre-train}}\|_F \Big) - \text{scale}_{[0,1]}\Big(\|\boldsymbol{W}_{l,h}^{\text{reference}} - \boldsymbol{W}_{l,h}^{\text{pre-train}}\|_F \Big)$$

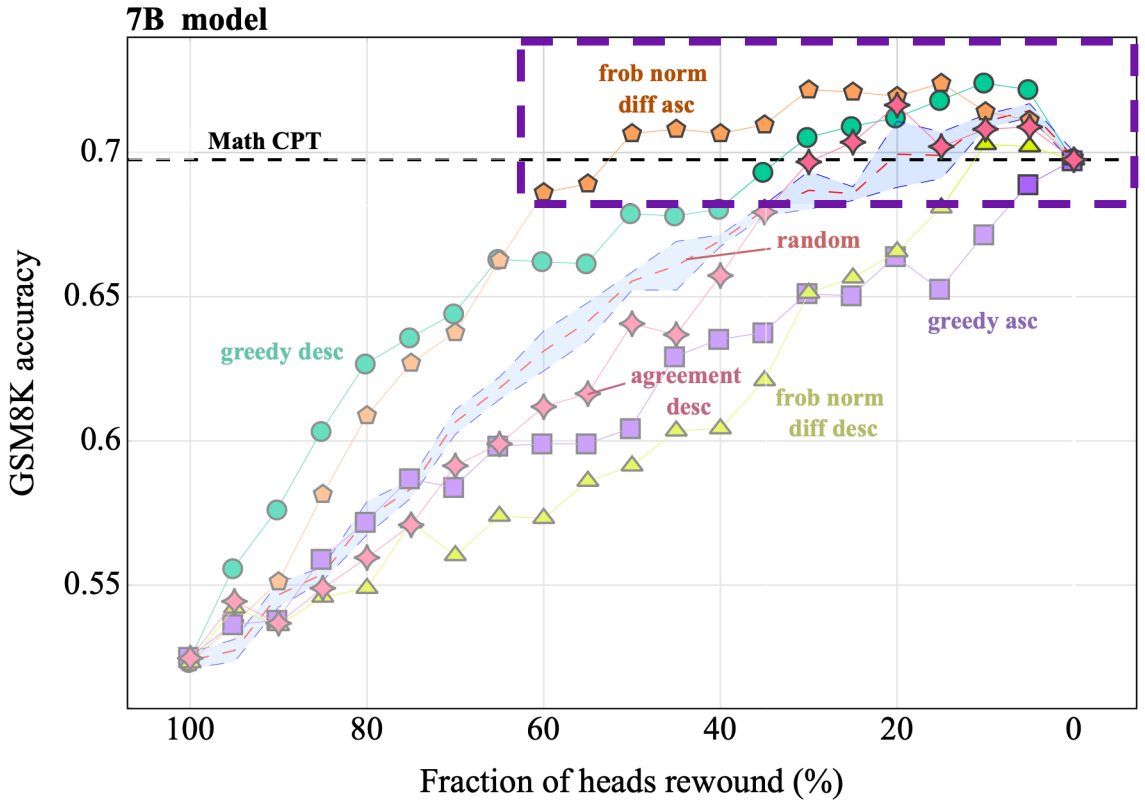
$$\text{scale}_{[0,1]}(\boldsymbol{X}) = (\boldsymbol{X} - X_{\min})/(X_{\max} - X_{\min})$$

Structure of CPT delta – head-wise rewind

- Rewinding 15% heads yields **+2%** GSM8K
- Up to **25%** heads can be rewound without significant quality drop



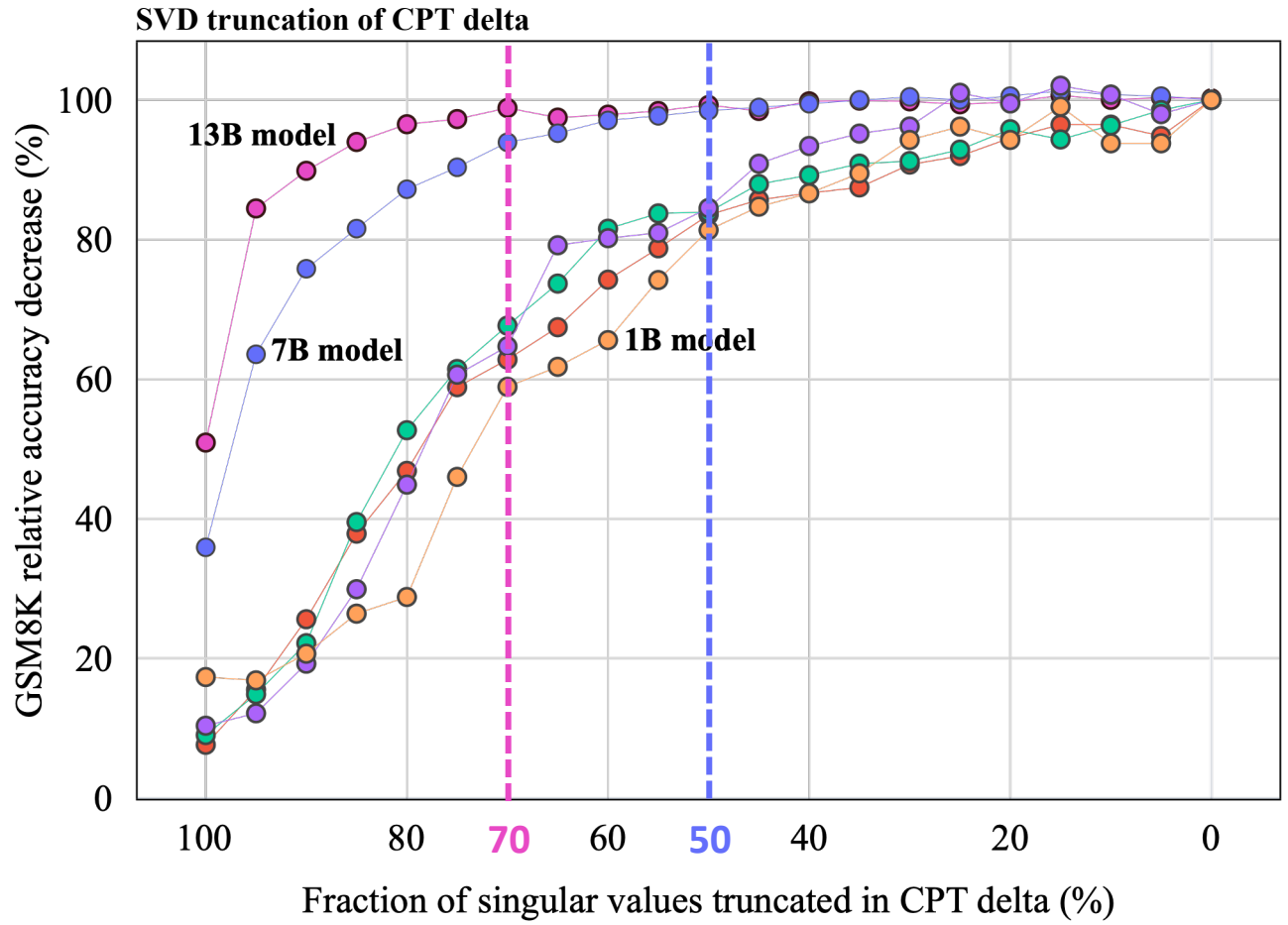
- Rewinding 15% heads yields **+4%** GSM8K
- Up to **60%** heads can be rewound without significant quality drop



Takeaway: head-wise rewind can consistently increase quality for larger models

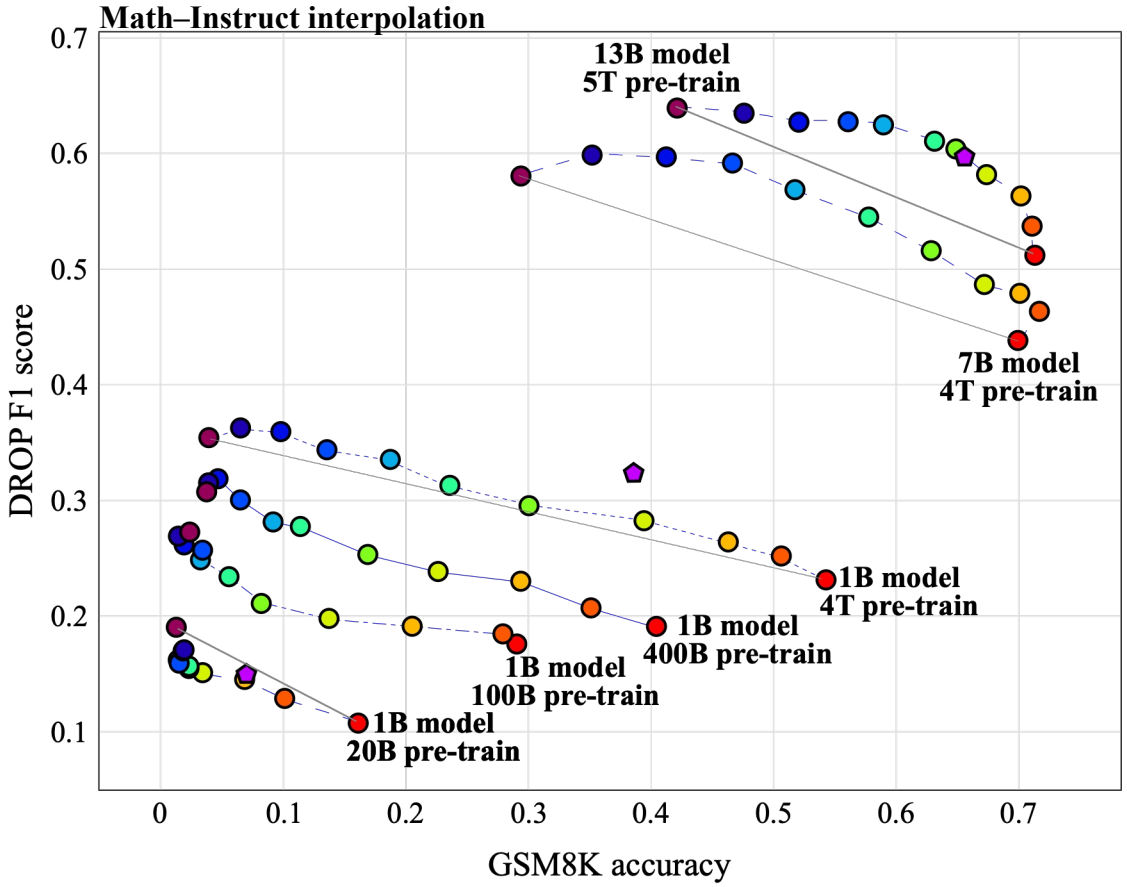
Structure of CPT delta – SVD truncation

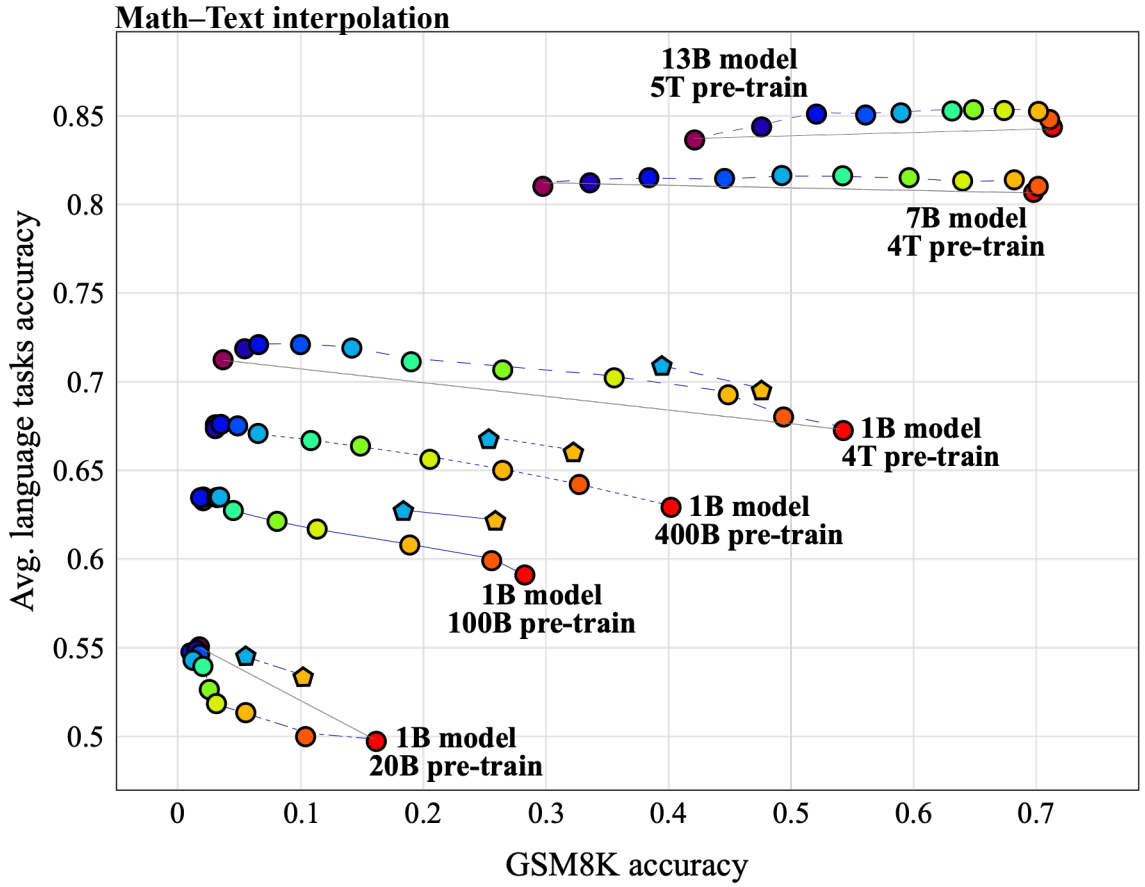
- CPT delta is increasingly redundant with scale
- Larger model ⇒ lower-rank update
- 7B model: **50%** of singular values truncatable
- 13B model: up to **70%** truncatable



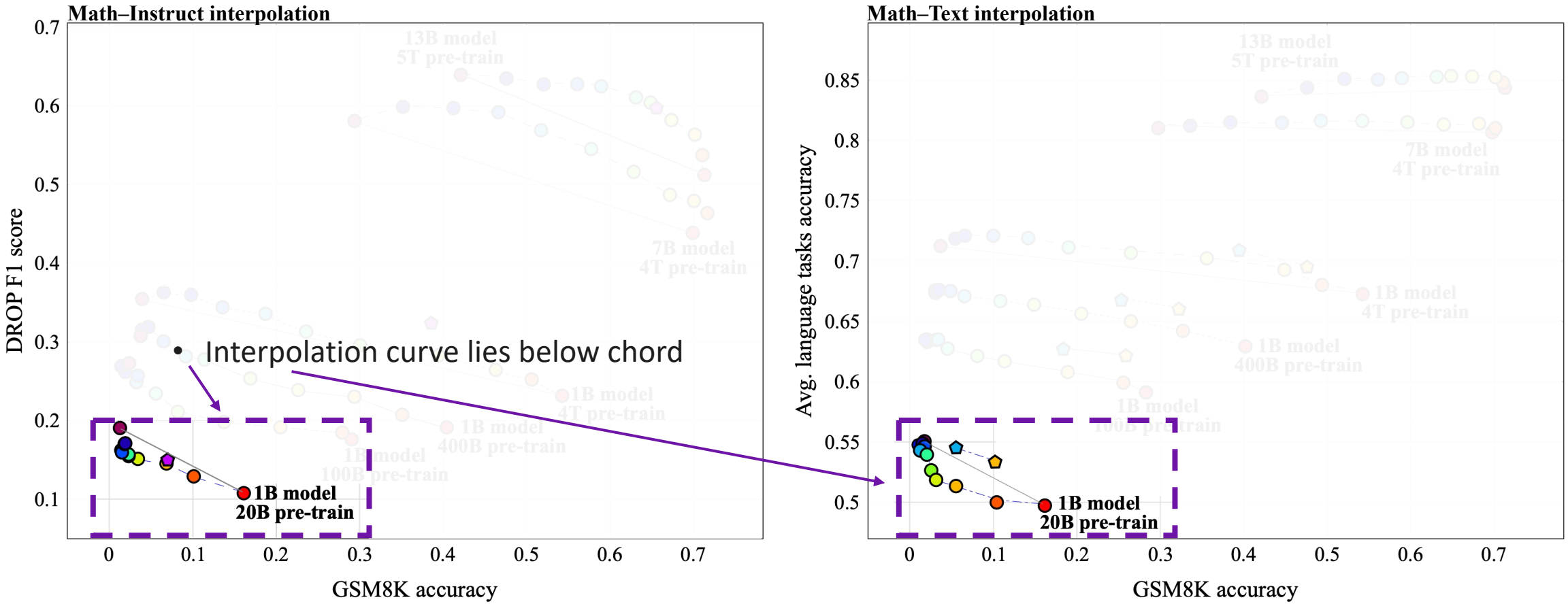
Takeaway: CPT update becomes more compressible for larger models, but not enough for DARE-like methods

Structure of CPT delta – domain connectivity

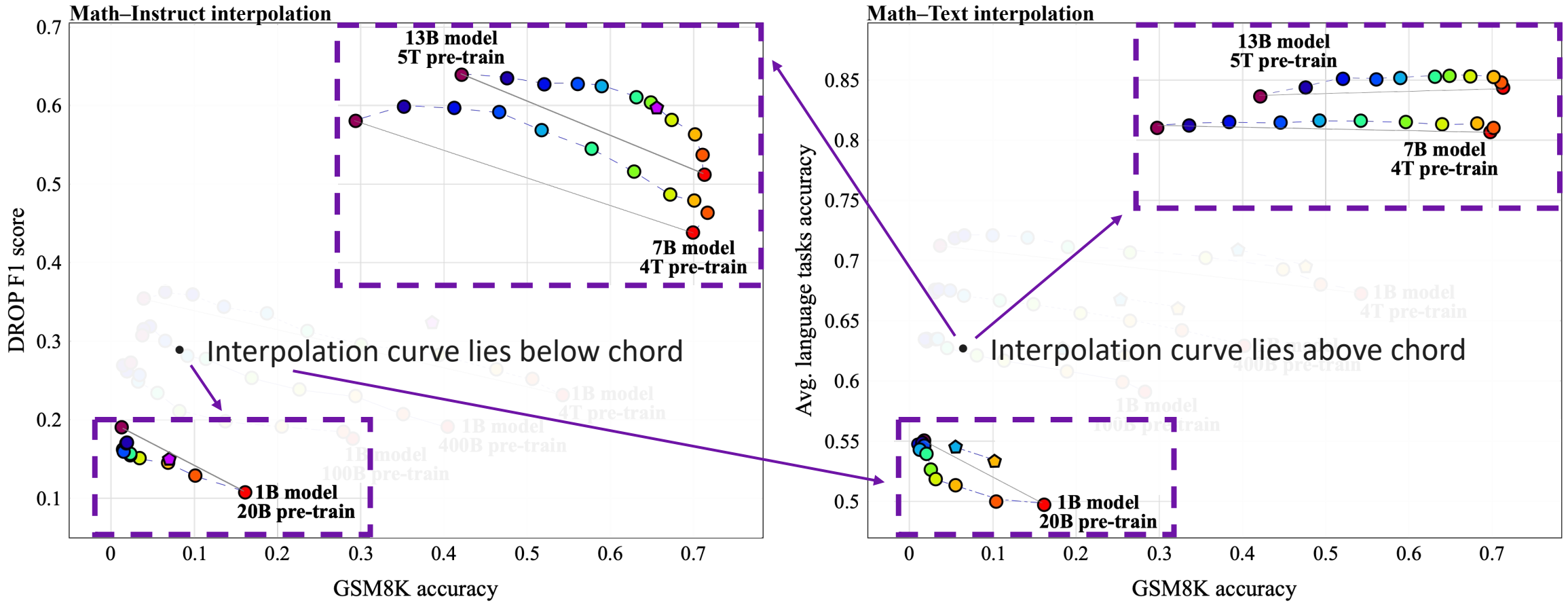




Structure of CPT delta – domain connectivity



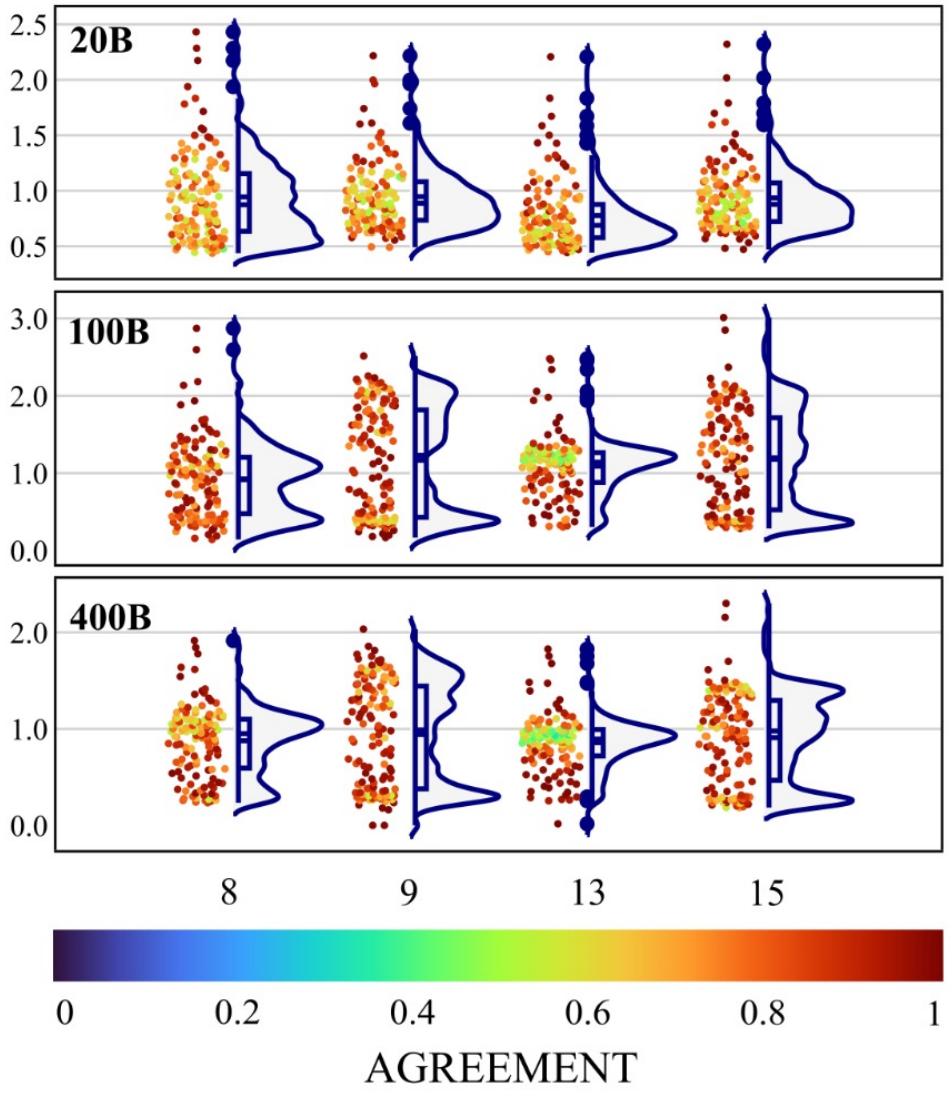
Structure of CPT delta – domain connectivity



Takeaway: larger pre-train budgets enable domain connectivity

Further research directions

1. Theoretical framework for complex spectral shapes
2. Downstream effects of head-wise rewind
3. Spectral fingerprints of SFT and RL fine-tuning
4. Combination with analysis of activations



Thank you for your attention!

DIFFRACT · SPECTRAL VIEW OF LLM DOMAIN ADAPTATION

Nikita Borodin

Maria Krylova

Artem Zabolotnyi

Dmitry Aspisov

Egor Shikov

Nikita Tyuplyaev

Oleg Travkin

Roman Alferov

Dmitry Vinichenko

Code & analysis toolkit

https://github.com/Risk-AI-Research/diffract-training

https://github.com/Risk-AI-Research/diffract

Corresponding author: Dmitry Vinichenko — dmitry.vinichenko@risk-ai-research.tech

ICML International Conference On Machine Learning