MiniPLM: Knowledge Distillation for Pre-Training Language Models

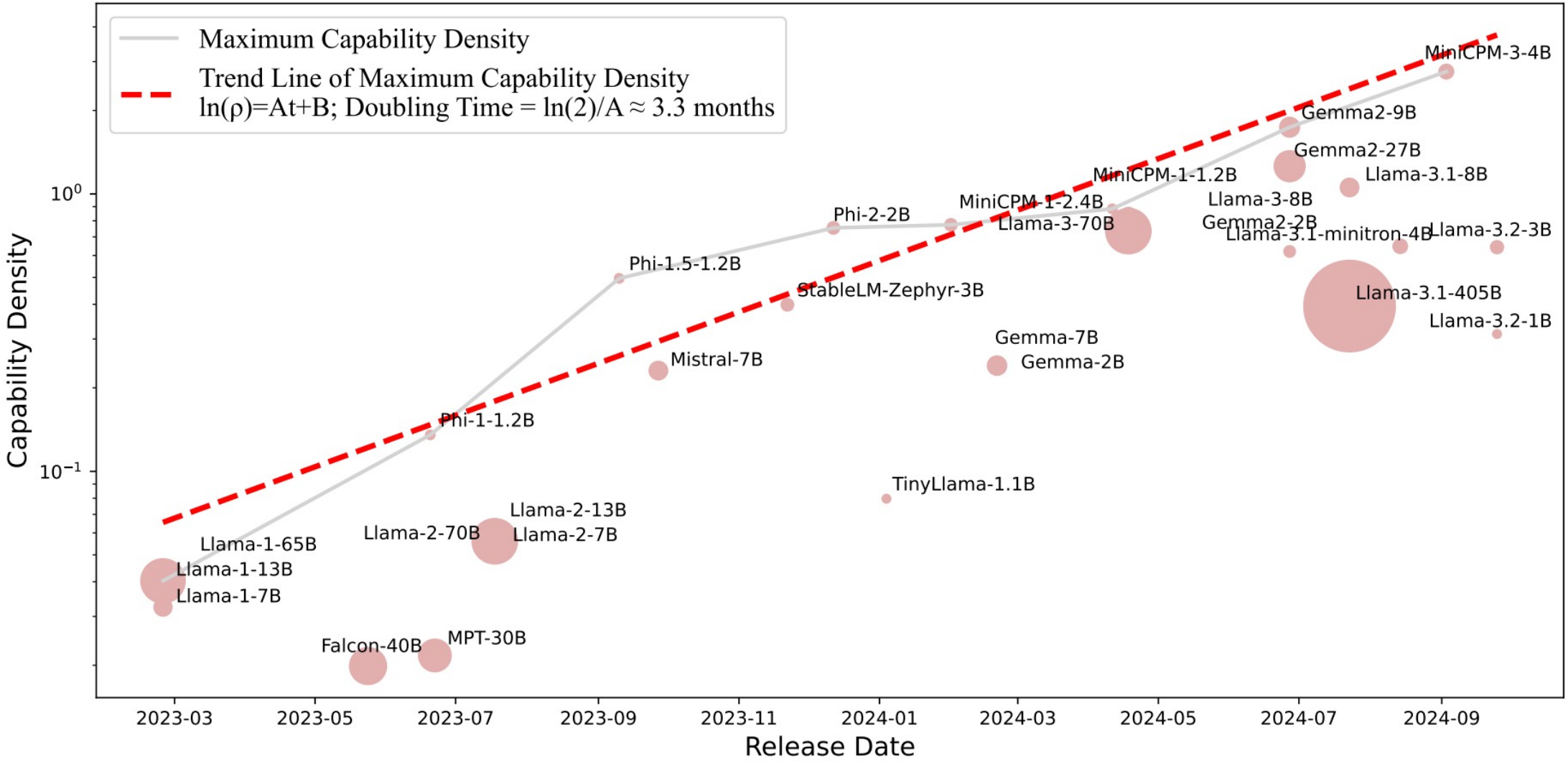
Yuxian Gu[1,2], Hao Zhou[2], Fandong Meng[2], Jie Zhou[2], Minlie Huang[1]

[1]The CoAI Group, Tsinghua University

[2]WeChat AI, Tencent Inc., China

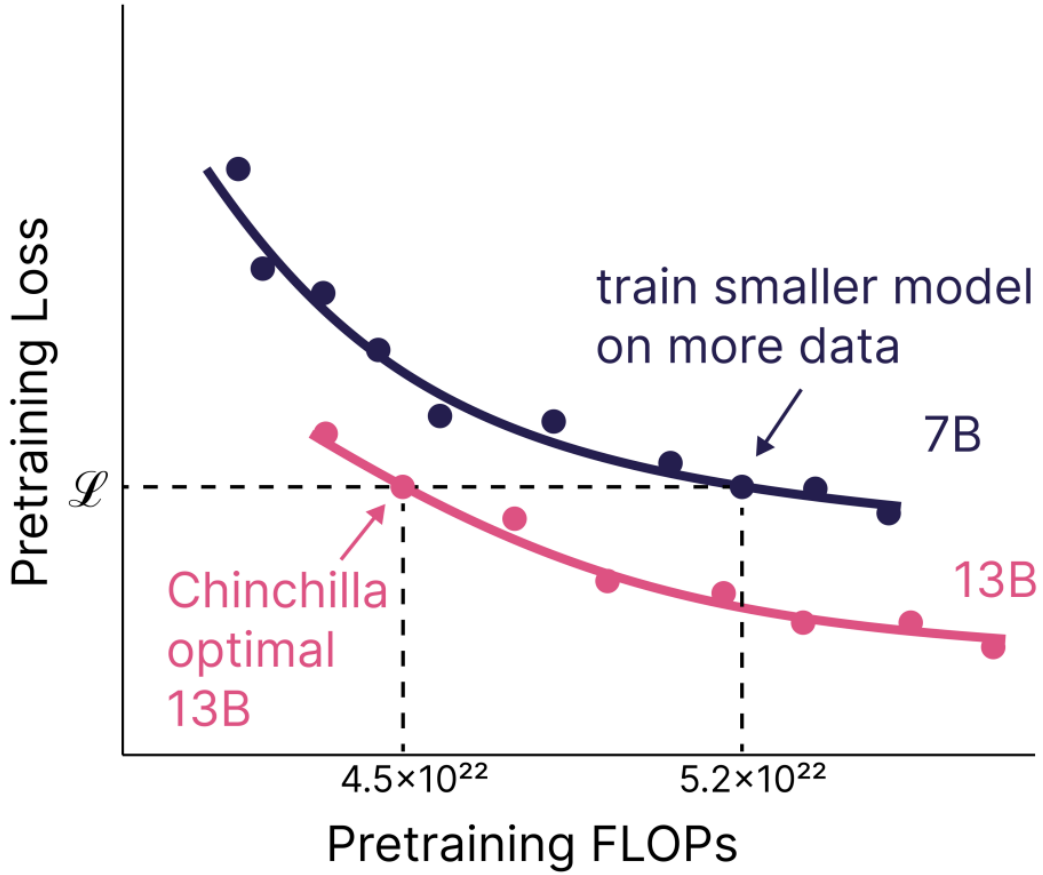
清華大學
Tsinghua University

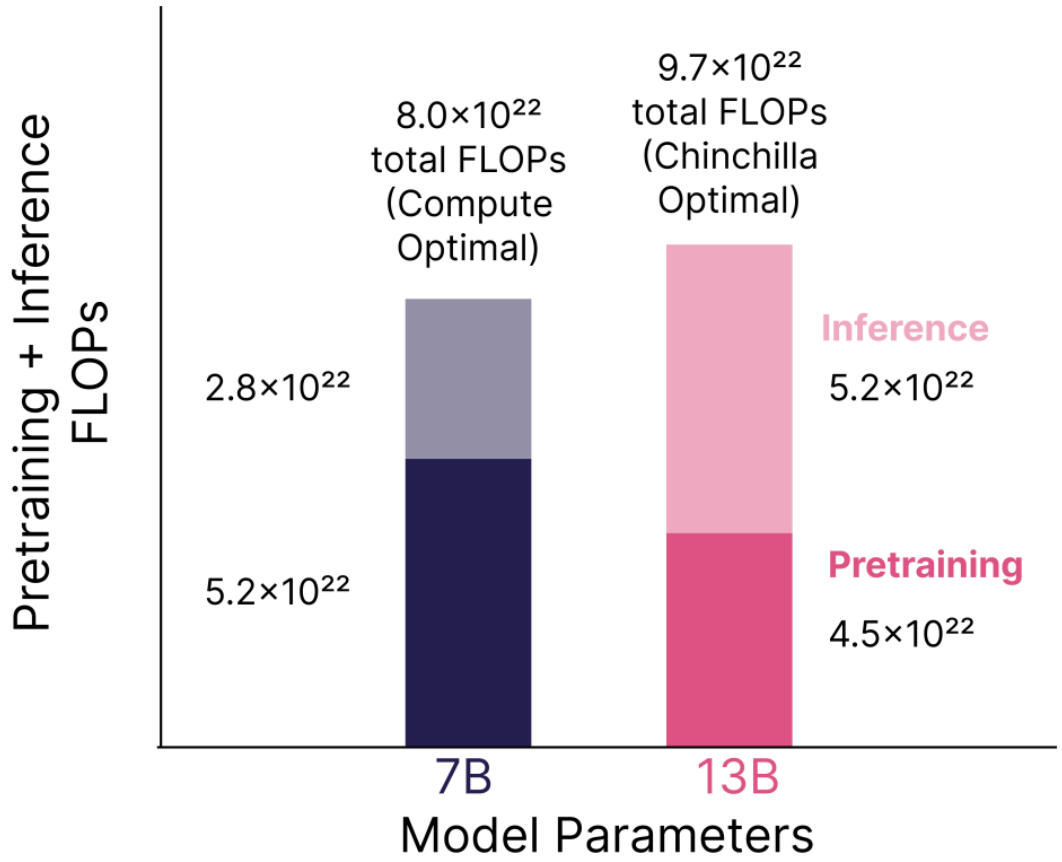
The Trend of Building “Small” Models



Inference Cost Matters for LM Scaling

- When considering inference costs, small models are more close to compute-optimal



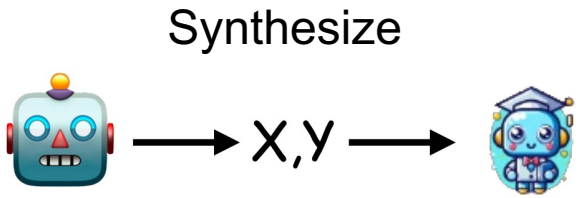


Knowledge Distillation for Small LMs

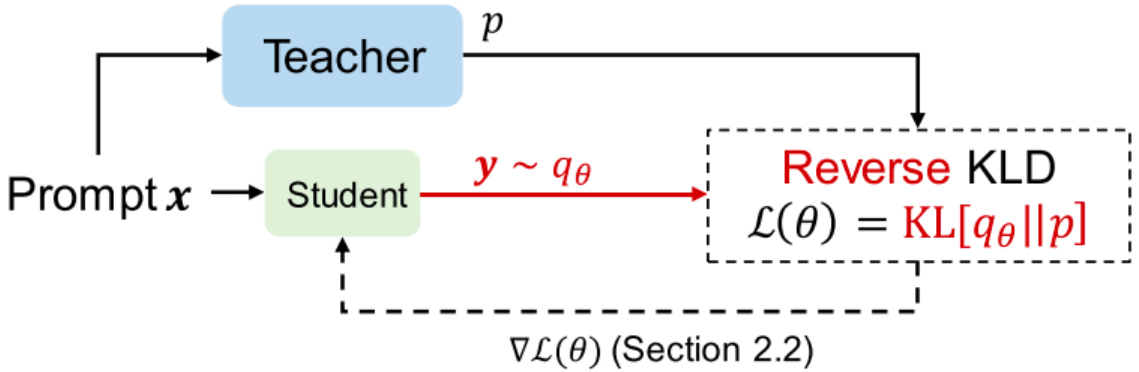
Online KD: Vanilla KD

$$L_{\text{logits}} = \frac{1}{l}\sum_{k=1}^{l} \text{Loss}(p_t^k(x,\tau), p_s^k(x,\tau))$$

Offline KD: Sequence KD



Online KD: MiniLLM

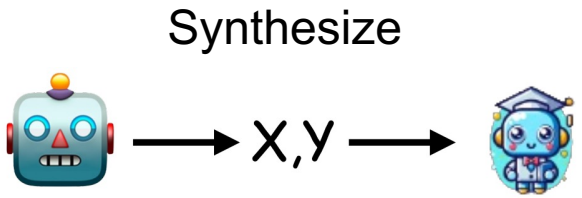


Knowledge Distillation for Small LMs

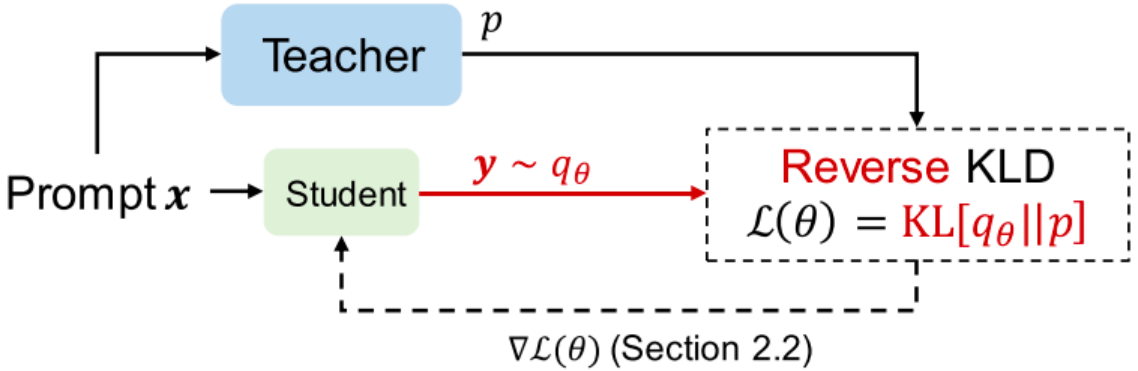
Online KD: Vanilla KD

$$L_{\text{logits}} = \frac{1}{l}\sum_{k=1}^{l} \text{Loss}(p_t^k(x,\tau), p_s^k(x,\tau))$$

Offline KD: Sequence KD



Online KD: MiniLLM

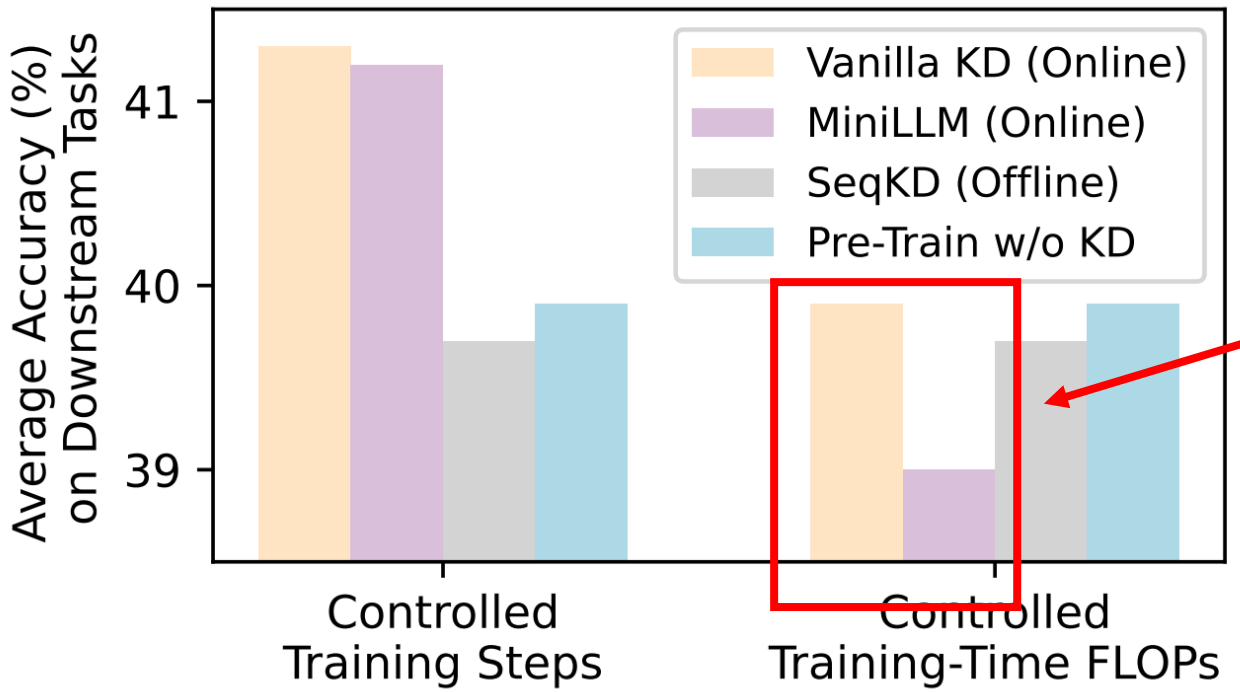


Mostly During SFT

How to do KD for pre-training?

Adapting KD to Pre-Training is Non-Trivial

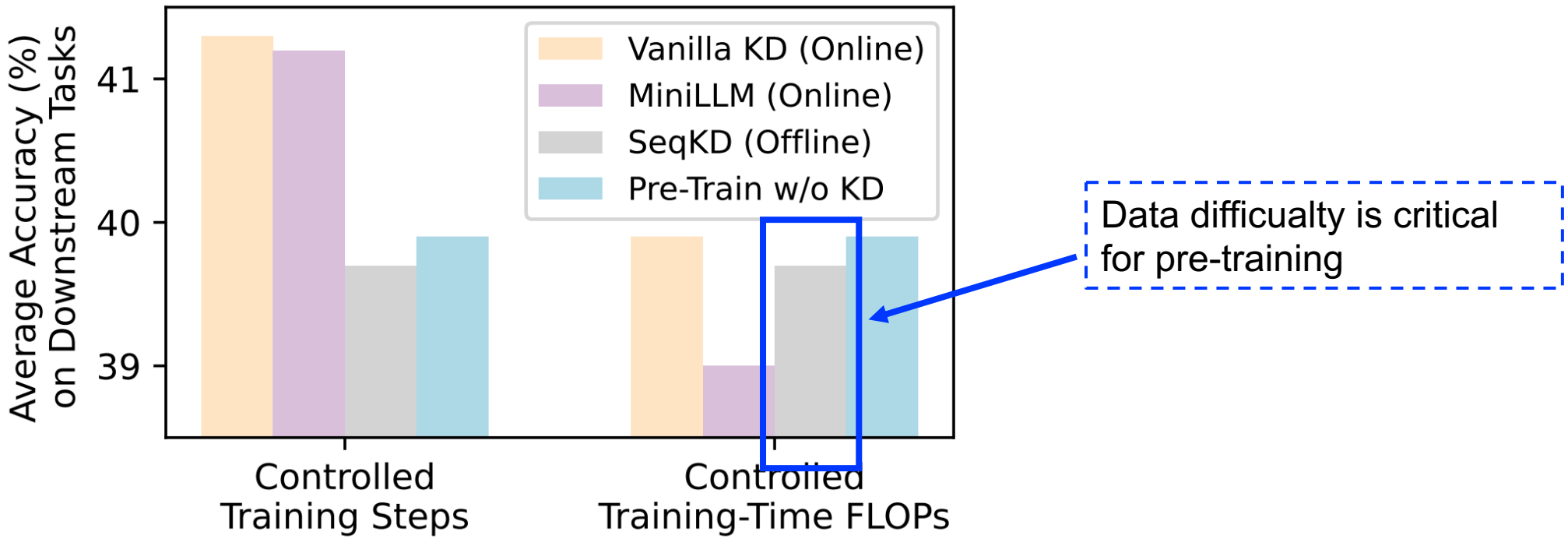
- **Efficiency Issue:** KD introduce additional training-time cost
  - Involving the inference of teacher LM during pre-training



Conventional KD does not show improvement when aligning the training-time computation

Adapting KD to Pre-Training in Non-Trival

- **Effectiveness Issue:** Teacher-generated samples loses difficualty
  - Model generation may “collapse”



Adapting KD to Pre-Training in Non-Trival

- **Flexibility Issue: Cannot KD across model families**
  - Tokenization needs to be aligned

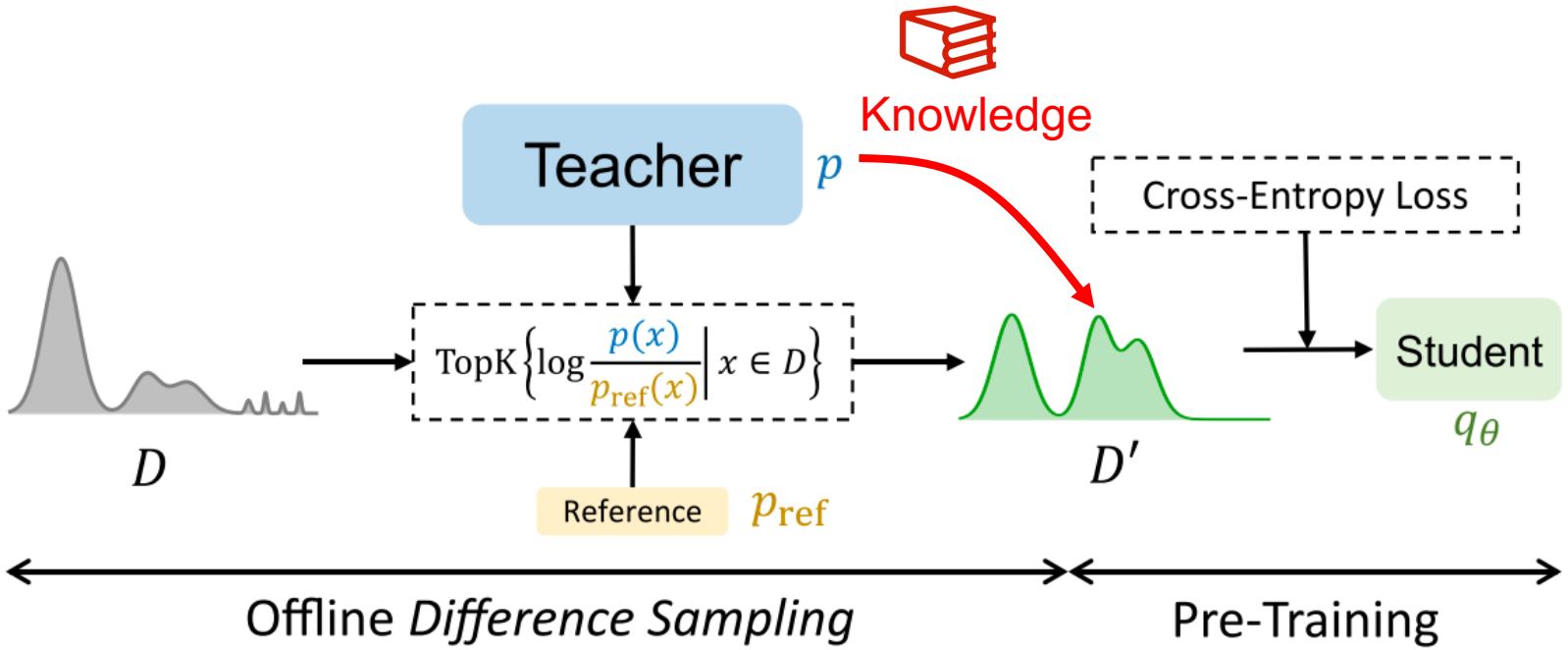
LLaMA

Knowledge Distillation

Mamba

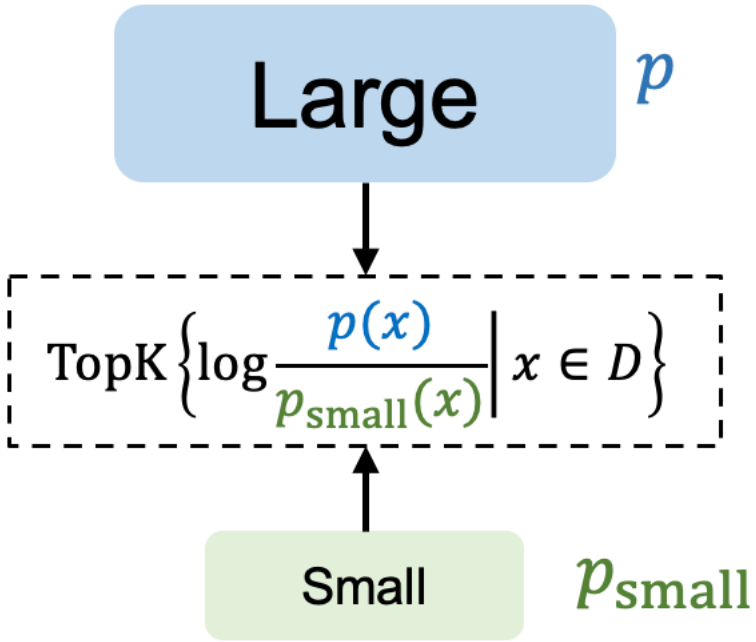
MiniPLM Framework

- Conducting KD ***offline*** with ***Difference Sampling***
  - Difference Sampling: Using the teacher's knowledge to select training data
  - Pre-Training: Pre-training student on the refined data



Difference Sampling: KD via Data Curation

- Distill the knowledge from teacher to student via training data



Select samples that

1. The large model learns well
2. The small model has not learned yet

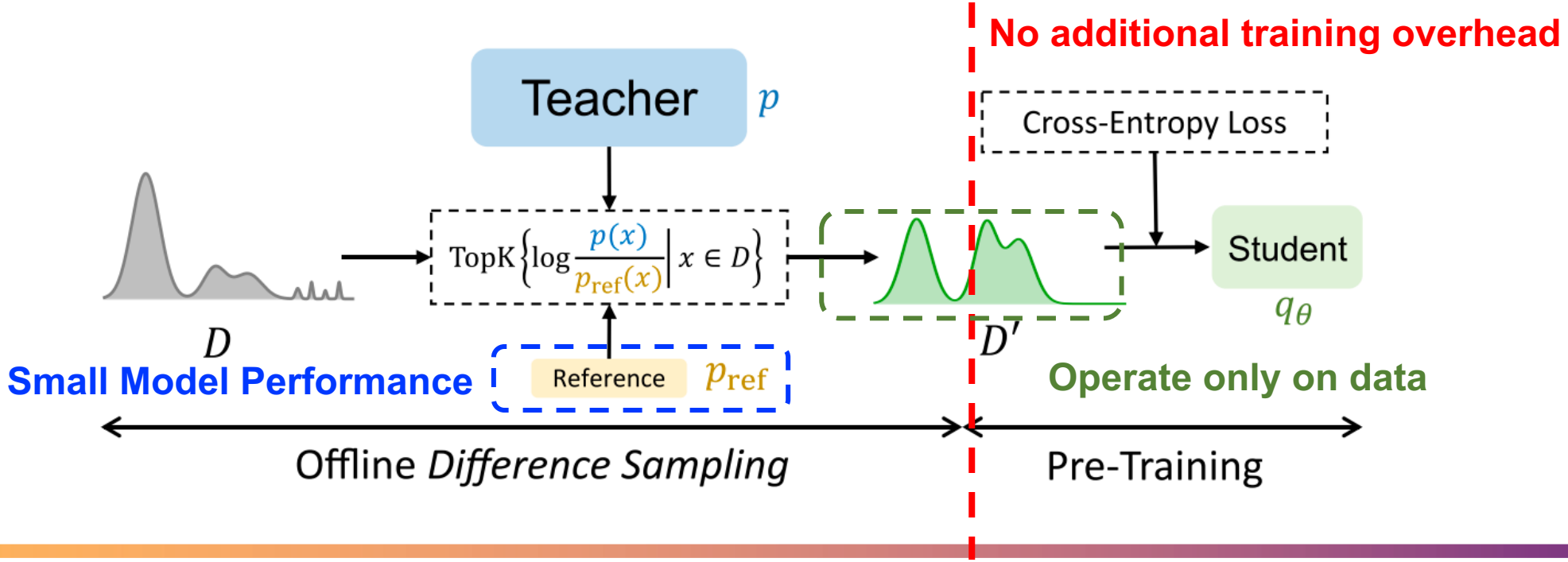
MiniPLM Framework

- Conducting KD ***offline*** with ***Difference Sampling***
  - Inference of the teacher LM once before the student LM training
  - Ensuring the data difficulty by involving small model performance
  - Operating solely on data. No tokenization alignment needed.

Efficiency

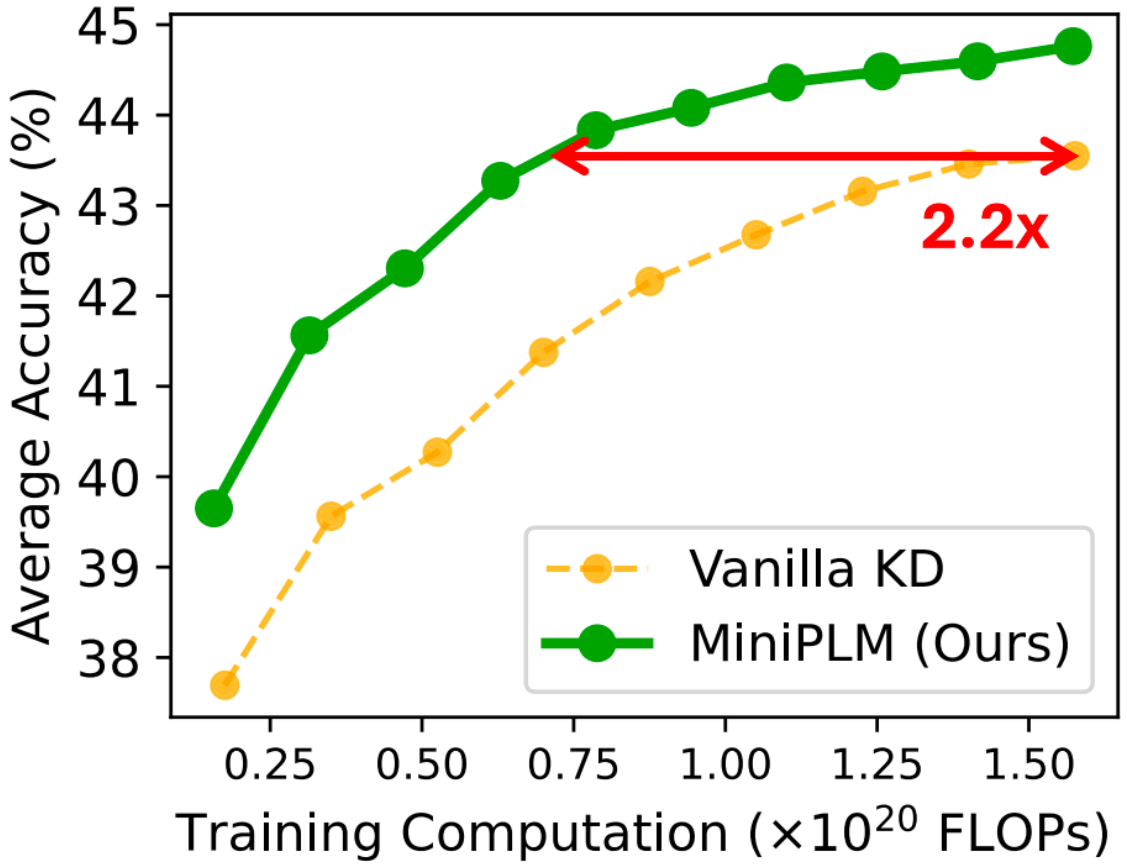
Effectiveness

Flexibility



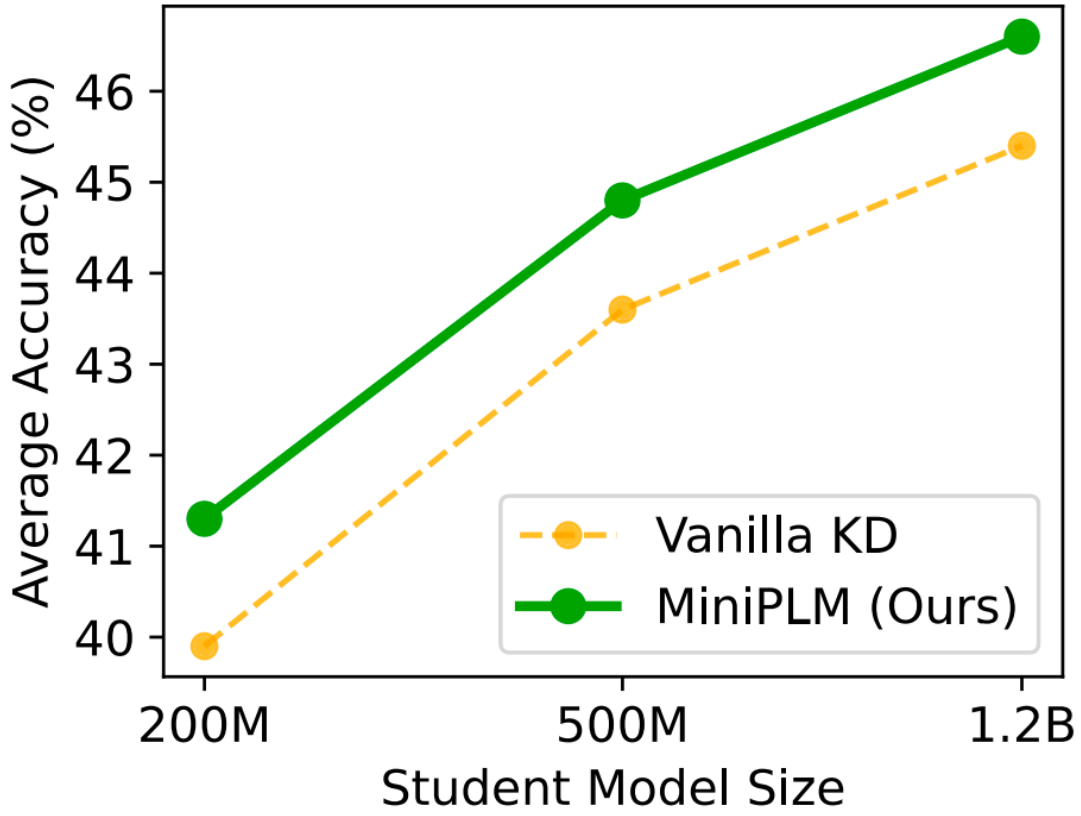
Efficiency

- 2.2x computation saving for pre-training LMs from scratch



Effectiveness

- Constant improvement across model scales



N_{stu}	Method	L_{1T}	L_{10T}
200M	Pre-Train w/o KD	3.35	3.32
	Vanilla KD	3.39	3.35
	MINIPLM	**3.28**	**3.26**
500M	Pre-Train w/o KD	3.12	3.08
	Vanilla KD	3.12	3.07
	MINIPLM	**3.06**	**3.04**
1.2B	Pre-Train w/o KD	2.98	2.94
	Vanilla KD	2.95	2.91
	MINIPLM	**2.92**	**2.88**

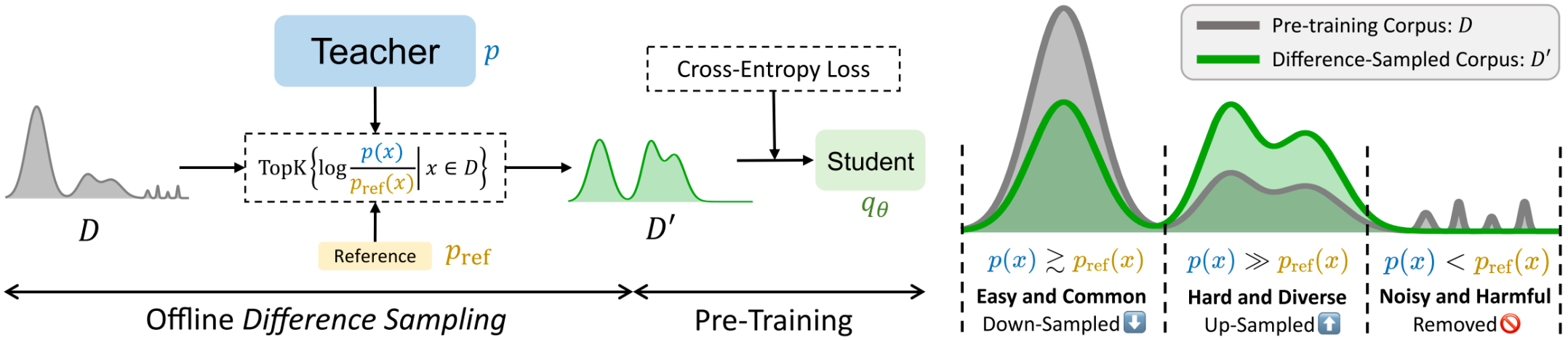
Flexibility

- KD Across Model Families
- Qwen ⟶ Llama/Mamba

	Llama3.1		Mamba	
	Acc.	Loss	Acc.	Loss
Pre-Train w/o KD	41.0	3.52	41.6	3.24
SeqKD	40.8	3.54	41.0	3.27
MINIPLM	**41.8**	**3.43**	**42.6**	**3.15**

Summary

- MiniPLM: A new Paradigm for KD in Pre-Training
  - KD by refining pre-training corpus



✓ Efficiency ✓ Effectiveness ✓ Flexibility

Thanks!

清華大學
Tsinghua University