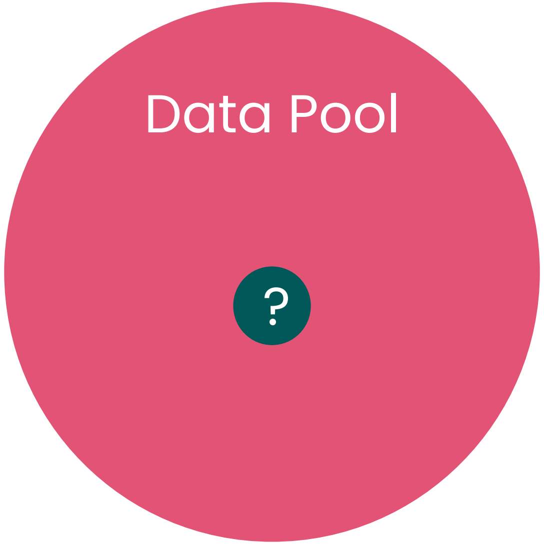
MATES: Model-Aware Data Selection for Efficient Pretraining with Data Influence Models

Zichun Yu, Spandan Das, Chenyan Xiong

Potential of Data Selection in Pretraining

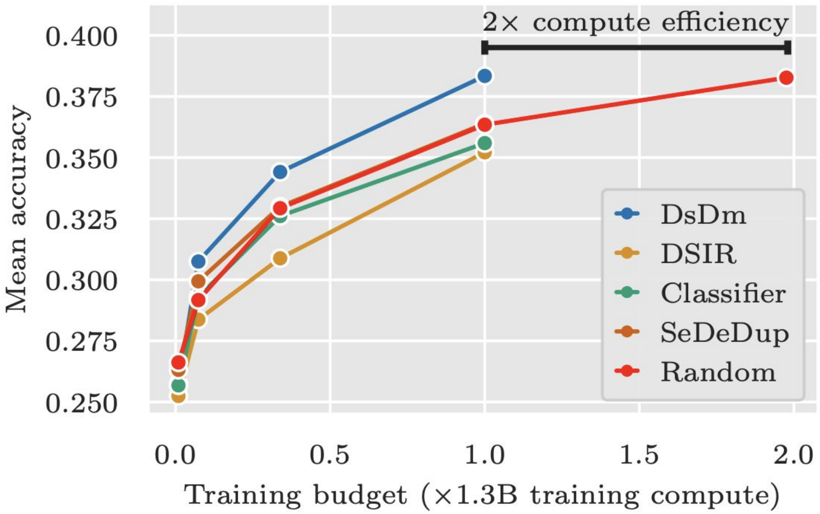
Unlimited data pool: Web

Limited FLOPs: Hardware



Fix a training budget

Maximize target performance



Engstrom, Logan, et al. DsDm: Model-aware dataset selection with datamodels. ICML 2024.

Gaps

Current data selection methods:

- Rule-based: C4, DSIR, SemDeDup
- Influence-based: TRAK, DsDm
- LLM-based: QuRating, FineWeb-Edu

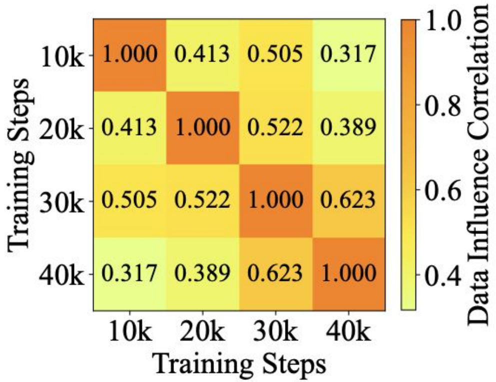
Static & not model-aware!

Raffel, Colin, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. JMLR: 1-67.
Xie, Sang Michael, et al. Data selection for language models via importance resampling. NeurIPS 2023.
Abbas, Amro Kamal Mohamed, et al. SemDeDup: Data-efficient learning at web-scale through semantic deduplication. ICLR 2023.
Park, Sung Min, et al. TRAK: Attributing model behavior at scale. ICML 2023.
Engstrom, Logan, et al. DsDm: Model-aware dataset selection with datamodels. ICML 2024.
Wettig, Alexander, et al. QuRating: Selecting high-quality data for training language models. ICML 2024.
Penedo, Guilherme, et al. The FineWeb datasets: Decanting the web for the finest text data at scale. arXiv 2024.

Motivations

Language models know what data to learn!

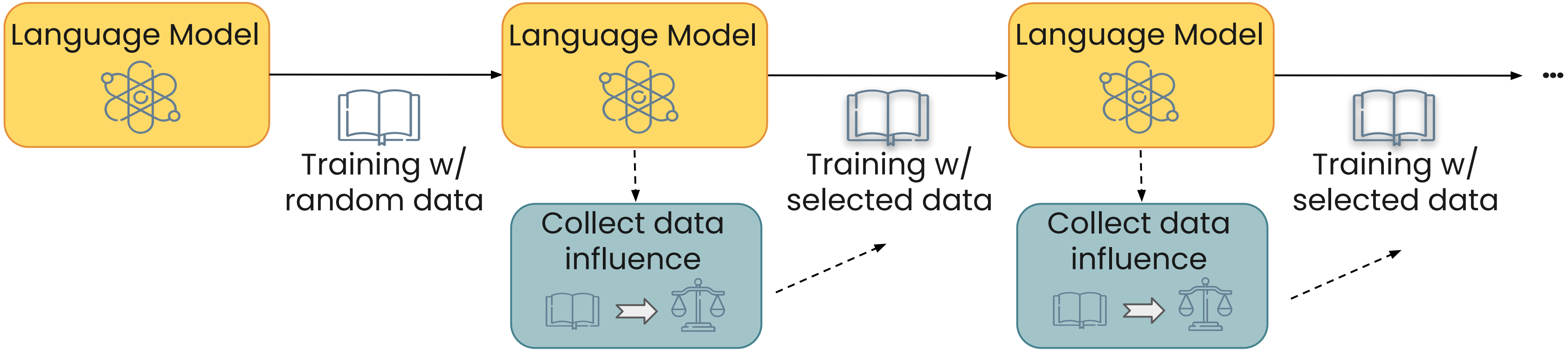
- Data influence can be collected with the pretraining model itself
- Data preferences of the model will evolve over time



(a) Preference correlation.

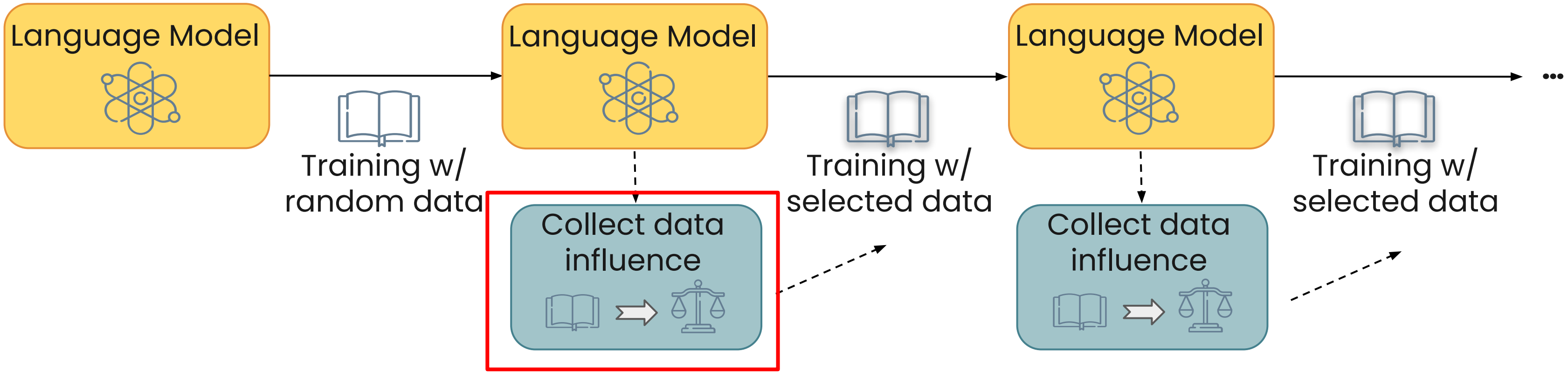
Hong, Xudong, et al. A surprisal oracle for active curriculum language modeling. arXiv 2023.

Model-Aware Data Selection Framework



- Collect the model's data influence along with the pretraining
- Use the collected influence to select the most useful data dynamically

Locally Probed Oracle Data Influence



Contrib #1: Locally probe the language model to collect precise oracle data influence via one-step training

$$\mathcal{I}_{\mathcal{M}^*}(x_i;\mathcal{D}_r) \approx n\nabla_{\mathcal{M}}\mathcal{L}(\mathcal{D}_r \mid \mathcal{M}^*)^\top(\mathcal{M}^*_{-\frac{1}{n},x_i} - \mathcal{M}^*)$$
$$\approx n(\mathcal{L}(\mathcal{D}_r \mid \mathcal{M}^*_{-\frac{1}{n},x_i}) - \mathcal{L}(\mathcal{D}_r \mid \mathcal{M}^*))$$
$$\propto \boxed{\mathcal{L}(\mathcal{D}_r \mid \mathcal{M}^*_{-\frac{1}{n},x_i})} - \boxed{\mathcal{L}(\mathcal{D}_r \mid \mathcal{M}^*)}.$$

Model's reference loss before training on x_i

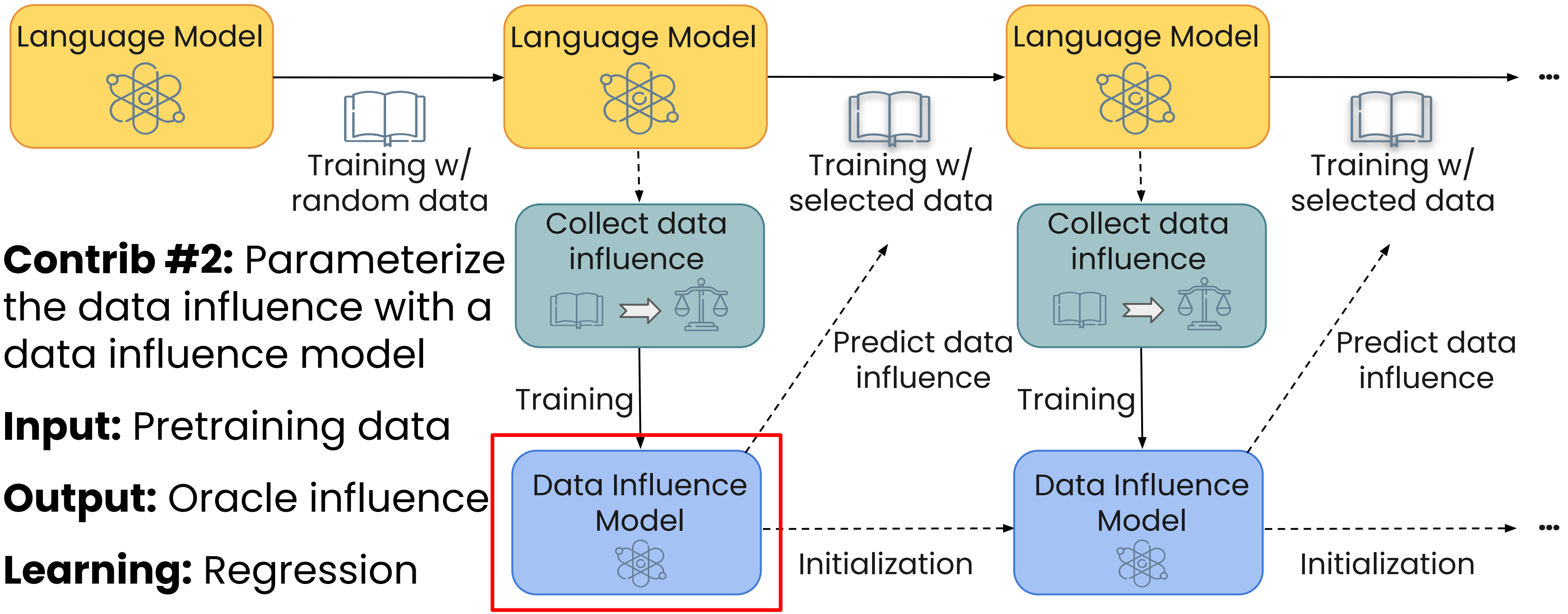
Model's reference loss after training on x_i

$\mathcal{M}^*$: Language Model

x_i : Pretraining Data

$\mathcal{D}_r$: Reference Data

Data Influence Model



Contrib #2: Parameterize the data influence with a data influence model

Input: Pretraining data

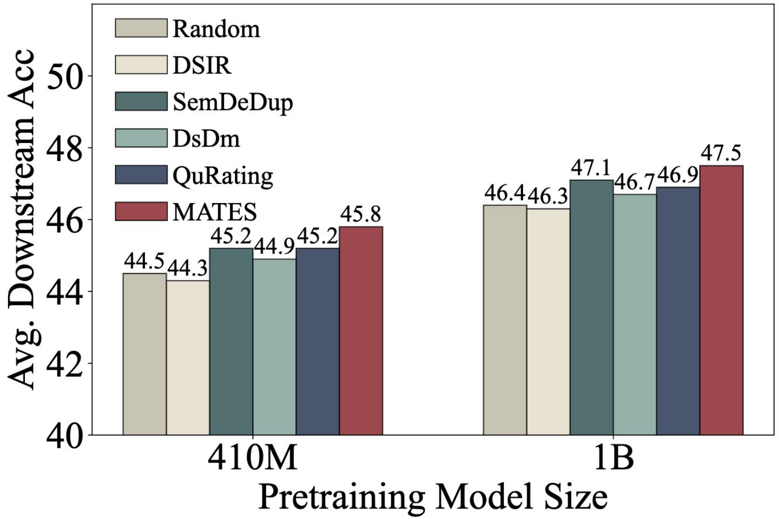
Output: Oracle influence

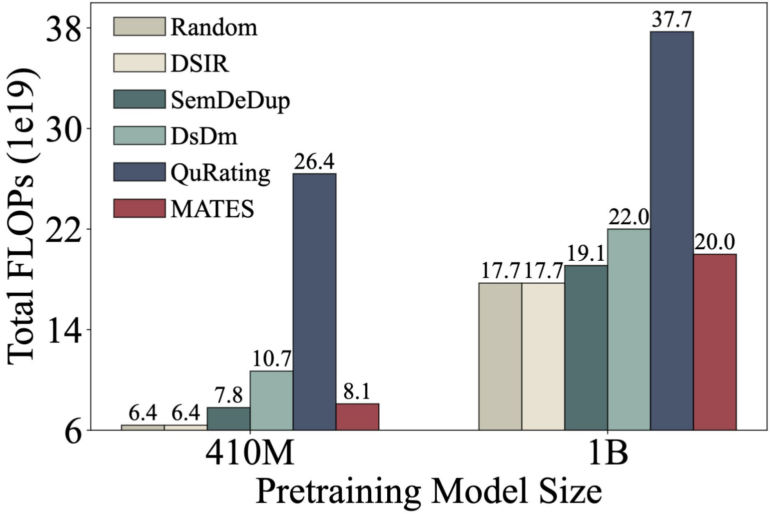
Learning: Regression

Experimental Setup

- Pretraining Model: 410M and 1B models
- Data Influence Model: Fine-tuned BERT-base (110M)
- Training Data: C4
- Reference Data: LAMBADA
- Evaluation: Avg. zero-shot accuracy across 9 downstream NLP tasks (not including LAMBADA)
- Baselines: Random, DSIR, SemDeDup, DsDm, QuRating

Main Results

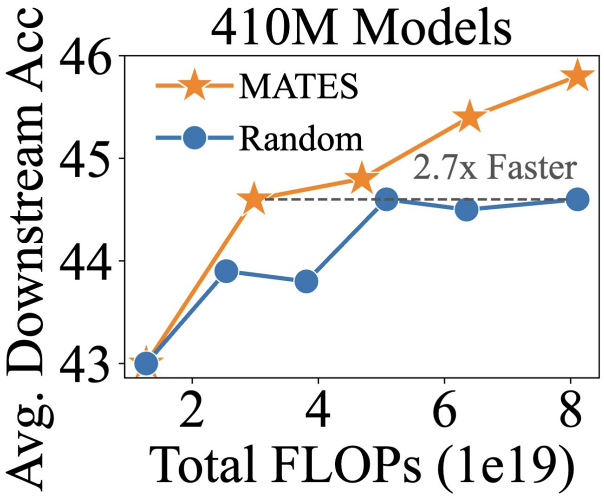


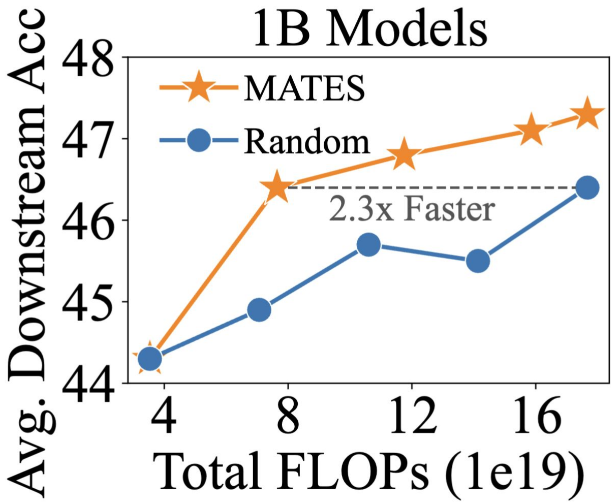


- MATES achieves higher downstream accuracy with relatively lower FLOPs
- MATES also ranks first in the DCLM 1B-1x setting (check their repo!)

Li, Jeffrey, et al. DataComp-LM: In search of the next generation of training sets for language models. NeurIPS 2024.

Scaling Curves





- MATES achieves the final random selection performance with less than half of the FLOPs

Effectiveness of Locally Probed Oracle Influence

Table 6: Performances of locally probed oracle data influence, MATES, and DsDm in 410M setting at 40k steps. We show zero-shot/two-shot results.

Methods	SciQ	ARC-E	ARC-C	LogiQA	OBQA
Oracle	$65.4_{(1.5)}/70.4_{(1.4)}$	$\mathbf{42.5}_{(1.0)}/43.6_{(1.0)}$	$\mathbf{25.2}_{(1.3)}/25.0_{(1.3)}$	$26.1_{(1.7)}/25.7_{(1.7)}$	$\mathbf{31.8}_{(2.1)}/\mathbf{30.4}_{(2.1)}$
MATES	$\mathbf{67.3}_{(1.5)}/\mathbf{76.7}_{(1.3)}$	$41.7_{(1.0)}/\mathbf{44.4}_{(1.0)}$	$24.7_{(1.3)}/24.0_{(1.2)}$	$\mathbf{26.9}_{(1.7)}/\mathbf{26.3}_{(1.7)}$	$28.8_{(2.0)}/28.0_{(2.0)}$
DsDm	$66.0_{(1.5)}/72.7_{(1.4)}$	$41.7_{(1.0)}/43.2_{(1.0)}$	$23.7_{(1.2)}/\mathbf{25.2}_{(1.3)}$	$24.4_{(1.7)}/23.3_{(1.7)}$	$29.2_{(2.0)}/29.4_{(2.0)}$

Methods	BoolQ	HellaSwag	PIQA	WinoGrande	Average
Oracle	$58.9_{(0.9)}/\mathbf{59.1}_{(0.9)}$	$\mathbf{41.1}_{(0.5)}/\mathbf{43.1}_{(0.5)}$	$\mathbf{68.2}_{(1.1)}/66.6_{(1.1)}$	$51.6_{(1.4)}/\mathbf{53.2}_{(1.4)}$	$\mathbf{45.6}_{(1.4)}/\mathbf{46.3}_{(1.3)}$
MATES	$59.6_{(0.9)}/57.0_{(0.9)}$	$40.1_{(0.5)}/39.6_{(0.5)}$	$67.6_{(1.1)}/\mathbf{67.7}_{(1.1)}$	$\mathbf{52.1}_{(1.4)}/51.3_{(1.4)}$	$45.4_{(1.3)}/46.1_{(1.3)}$
DsDm	$\mathbf{60.3}_{(0.9)}/58.1_{(0.9)}$	$40.4_{(0.5)}/40.2_{(0.5)}$	$67.2_{(1.1)}/66.5_{(1.1)}$	$50.4_{(1.4)}/52.2_{(1.4)}$	$44.8_{(1.3)}/45.6_{(1.3)}$

- **Oracle vs. DsDm:** Our locally probed oracle influence is more effective than DsDm (using TRAK to compute influence)
- **Oracle vs. MATES:** Our data influence model is able to approximate the oracle (almost) losslessly

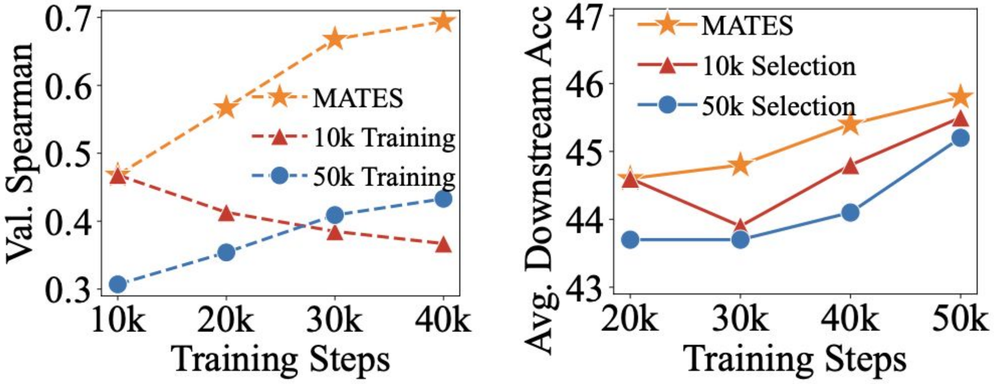
Robustness of Locally Probed Oracle Influence

Table 3: Performances of oracle selected data with different reference tasks in the 410M setting. We run the decay stage starting from the MATES model at 50k steps.

$\mathcal{D}_r$	SciQ	ARC-E	ARC-C	LogiQA	OBQA	BoolQ	HellaSwag	PIQA	WinoGrande	Average
LAMBADA	$66.0_{(1.5)}$	$42.2_{(1.0)}$	$24.8_{(1.3)}$	$27.2_{(1.7)}$	$30.8_{(2.1)}$	$\mathbf{59.1}_{(0.9)}$	$\mathbf{41.9}_{(0.5)}$	$\mathbf{68.5}_{(1.1)}$	$52.3_{(1.4)}$	$45.9_{(1.4)}$
ARC-E (MC)	$64.9_{(1.5)}$	$42.4_{(1.0)}$	$24.9_{(1.3)}$	$27.8_{(1.8)}$	$30.4_{(2.1)}$	$58.0_{(0.9)}$	$41.1_{(0.5)}$	$68.1_{(1.1)}$	$51.7_{(1.4)}$	$45.5_{(1.4)}$
ARC-E (LM)	$65.3_{(1.5)}$	$43.0_{(1.0)}$	$24.8_{(1.3)}$	$28.0_{(1.8)}$	$31.8_{(2.1)}$	$58.5_{(0.9)}$	$40.7_{(0.5)}$	$67.2_{(1.1)}$	$\mathbf{52.5}_{(1.4)}$	$45.8_{(1.4)}$
FLAN	$\mathbf{66.4}_{(1.5)}$	$\mathbf{45.1}_{(1.0)}$	$\mathbf{25.1}_{(1.3)}$	$\mathbf{28.7}_{(1.8)}$	$\mathbf{32.0}_{(2.1)}$	$56.2_{(0.9)}$	$40.5_{(0.5)}$	$67.9_{(1.1)}$	$52.3_{(1.4)}$	$\mathbf{46.0}_{(1.4)}$

- Our locally probed oracle influence is robust across different reference tasks
- Different reference tasks may strengthen different model abilities

Effectiveness of Model-Aware Data Selection



(a) Influence modeling. (b) Downstream accuracy.

Figure 5: Static (based on a 10k or a 50k random-pretrained model checkpoint) data selection versus model-aware data selection in influence modeling and downstream accuracy.

- Model-aware data selection is more effective than static one, either in influence modeling or downstream accuracy

Takeaways

- Data preference of the pretraining model is ever-changing
- Locally probed oracle data influence is effective to capture it
- A small data influence model can precisely learn the oracle and therefore efficiently select the effective data for the pretraining model

Paper

Code

Email: zichunyu@andrew.cmu.edu