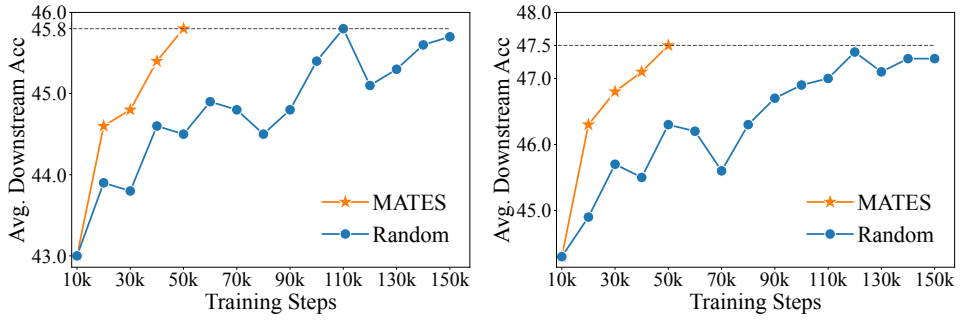
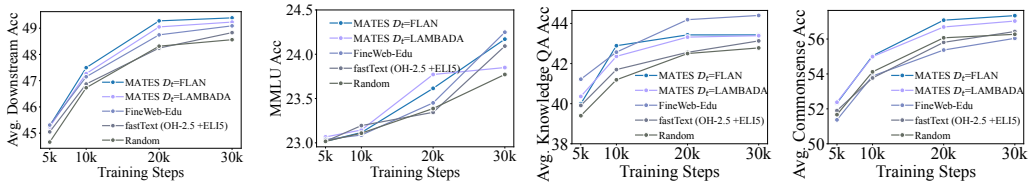
Table 1: Zero-/Two-shot evaluation of pretraining Pythia-1B with different data selection methods.

| **Methods** (#FLOPs ∗1e19) | **SciQ** | **ARC-E** | **ARC-C** | **LogiQA** | **OBQA** |
|---|---|---|---|---|---|
| Random (17.67) | $65.8_{(1.5)}/74.6_{(1.4)}$ | $43.7_{(1.0)}/45.1_{(1.0)}$ | $25.6_{(1.3)}/25.5_{(1.3)}$ | $27.5_{(1.8)}/24.6_{(1.7)}$ | $31.8_{(2.1)}/30.0_{(2.1)}$ |
| DSIR (17.67) | $65.8_{(1.5)}/75.0_{(1.4)}$ | $42.6_{(1.0)}/44.4_{(1.0)}$ | $24.7_{(1.3)}/25.8_{(1.3)}$ | $\mathbf{28.7}_{(1.8)}/\mathbf{25.8}_{(1.7)}$ | $29.2_{(2.0)}/\mathbf{30.6}_{(2.1)}$ |
| SemDeDup (19.13) | $66.8_{(1.5)}/75.9_{(1.4)}$ | $\mathbf{45.5}_{(1.0)}/45.1_{(1.0)}$ | $25.3_{(1.3)}/25.2_{(1.3)}$ | $27.6_{(1.8)}/24.1_{(1.7)}$ | $30.6_{(2.1)}/\mathbf{30.6}_{(2.1)}$ |
| DsDm (22.04) | $\mathbf{68.2}_{(1.5)}/\mathbf{76.6}_{(1.3)}$ | $45.0_{(1.0)}/45.5_{(1.0)}$ | $\mathbf{26.5}_{(1.3)}/26.5_{(1.3)}$ | $26.6_{(1.7)}/25.2_{(1.7)}$ | $29.4_{(2.0)}/30.0_{(2.1)}$ |
| QuRating (37.67) | $67.1_{(1.5)}/76.2_{(1.3)}$ | $\mathbf{45.5}_{(1.0)}/\mathbf{46.7}_{(1.0)}$ | $25.6_{(1.3)}/26.0_{(1.3)}$ | $26.9_{(1.7)}/24.4_{(1.7)}$ | $29.8_{(2.0)}/30.2_{(2.1)}$ |
| MATES (Ours) (19.97) | $67.3_{(1.5)}/75.7_{(1.4)}$ | $44.9_{(1.0)}/46.1_{(1.0)}$ | $25.9_{(1.3)}/\mathbf{26.8}_{(1.3)}$ | $\mathbf{28.7}_{(1.8)}/25.2_{(1.7)}$ | $\mathbf{32.2}_{(2.1)}/\mathbf{30.6}_{(2.1)}$ |
| **Methods** (#FLOPs ∗1e19) | **BoolQ** | **HellaSwag** | **PIQA** | **WinoGrande** | **Average** |
| Random (17.67) | $60.2_{(0.9)}/56.8_{(0.9)}$ | $43.8_{(0.5)}/42.9_{(0.5)}$ | $68.9_{(1.1)}/68.3_{(1.1)}$ | $50.7_{(1.4)}/52.1_{(1.4)}$ | $46.4_{(1.4)}/46.7_{(1.3)}$ |
| DSIR (17.67) | $59.7_{(0.9)}/55.6_{(0.9)}$ | $44.2_{(0.5)}/43.6_{(0.5)}$ | $68.3_{(1.1)}/67.4_{(1.1)}$ | $\mathbf{53.2}_{(1.4)}/52.8_{(1.4)}$ | $46.3_{(1.4)}/46.8_{(1.3)}$ |
| SemDeDup (19.13) | $60.2_{(0.9)}/57.3_{(0.9)}$ | $\mathbf{45.3}_{(0.5)}/44.4_{(0.5)}$ | $69.7_{(1.1)}/68.4_{(1.1)}$ | $52.5_{(1.4)}/\mathbf{53.5}_{(1.4)}$ | $47.1_{(1.4)}/47.2_{(1.3)}$ |
| DsDm (22.04) | $59.0_{(0.9)}/56.1_{(0.9)}$ | $44.8_{(0.5)}/44.6_{(0.5)}$ | $68.9_{(1.1)}/68.0_{(1.1)}$ | $51.9_{(1.4)}/52.7_{(1.4)}$ | $46.7_{(1.3)}/47.2_{(1.3)}$ |
| QuRating (37.67) | $60.3_{(0.9)}/58.0_{(0.9)}$ | $45.2_{(0.5)}/44.7_{(0.5)}$ | $\mathbf{70.2}_{(1.1)}/68.3_{(1.1)}$ | $51.6_{(1.4)}/51.8_{(1.4)}$ | $46.9_{(1.3)}/47.4_{(1.3)}$ |
| MATES (Ours) (19.97) | $\mathbf{60.9}_{(0.9)}/\mathbf{58.1}_{(0.9)}$ | $\mathbf{45.3}_{(0.5)}/\mathbf{44.8}_{(0.5)}$ | $69.5_{(1.1)}/\mathbf{68.7}_{(1.1)}$ | $52.4_{(1.4)}/51.6_{(1.4)}$ | $\mathbf{47.5}_{(1.4)}/\mathbf{47.5}_{(1.3)}$ |



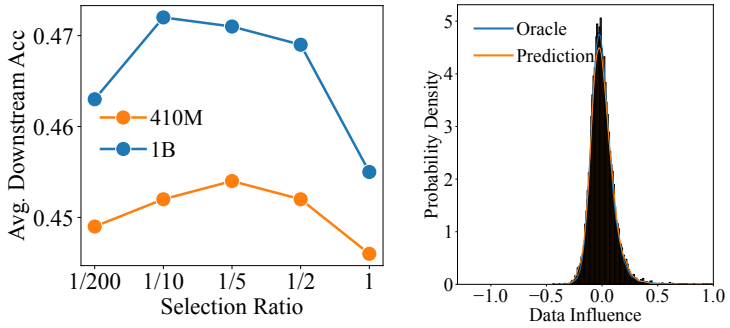

(a) Pythia-410M. (b) Pythia-1B.

Figure 1: Full data training performances of Pythia-410M and 1B.




(a) Average acc. (b) MMLU acc. (c) Knowledge QA acc. (d) Commonsense acc.

Figure 2: Zero-shot evaluation of pretraining Pythia-1B on FineWeb. We perform different data selection methods starting from 5k steps.




(a) Zero-shot performances of MATES with different data selection ratios at 40k steps. Main setup: 1/5; Random: 1.

(b) Distribution of oracle and predicted data influence. Experiments from Pythia-410M at 40k steps.

Figure 3: (a): Ablation of the selection ratio; (b): Data influence distribution.