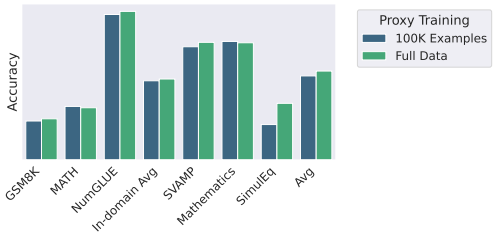


Figure 1: Per-dataset and average accuracy comparing proxy training on 100K examples and full data. S2L remains effective.

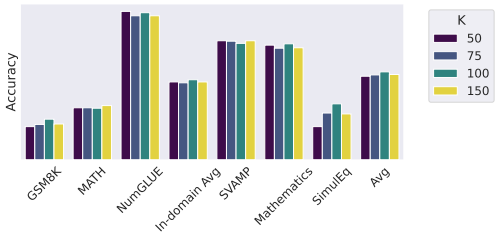
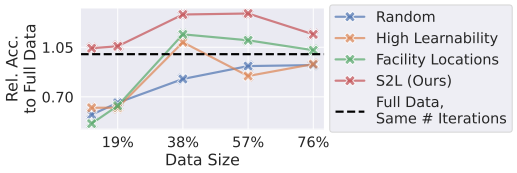
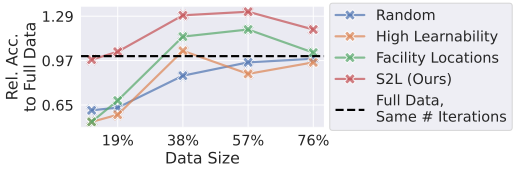


Figure 2: Per-dataset and average accuracy across different values of the clustering parameter K. S2L is relatively robust to the choice of K.

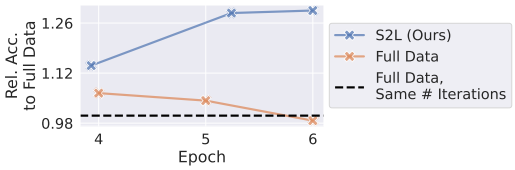


(a) In-domain Average Accuracy

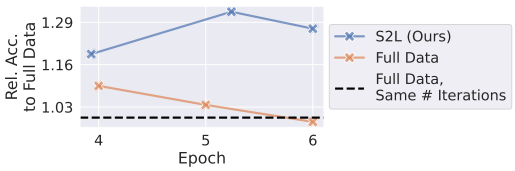


(b) Overall Average Accuracy

Figure 3: An extension of Figure 4 in the main submission, showing the relative accuracy to full data as the data size increases, with consistent total training iterations/steps for all results.



(a) In-domain Average Accuracy



(b) Overall Average Accuracy

Figure 4: Relative accuracy to full data across different epochs, comparing S2L-selected data and full data. S2L achieves superior performance with fewer data and fewer training iterations.

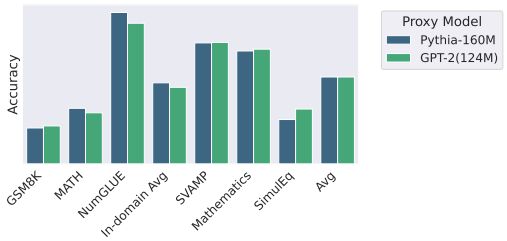


Figure 5: Per-dataset and average accuracy comparison between using different proxy models (Pythia-160M and GPT-2 (124M)) for data selection. Using both proxy models show comparable performance, demonstrating the effectiveness of different small models as reference models for S2L.