Table 1: Performance of instruction tuning with dataset selected by our method compared with the LESS. The dataset size is 4% of the candidate data repository and we train each model for one epoch on the selected set. The subscripts represent the standard deviations.

Model	LLAMA-2-7B			MISTRAL-7B		
Dataset	TydiQA	MMLU	BBH	TydiQA	MMLU	BBH
LESS	$54.4_{0.0}$	$46.5_{0.9}$	$40.4_{1.3}$	$60.5_{1.6}$	$\mathbf{60.8_{0.4}}$	$55.8_{1.5}$
Ours	$\mathbf{55.4_{0.5}}$	$\mathbf{47.9_{0.2}}$	$\mathbf{42.0_{1.1}}$	$\mathbf{63.6_{1.4}}$	$60.5_{0.8}$	$\mathbf{56.3_{2.1}}$

Table 2: F1 scores of the downstream tasks when the sample size varies. The size of the annotated data is set to 3K. Standard deviations are shown in the subscripts.

	——100K Sample Size ——				——300K Sample Size ——			
	ChemP.	IMDB	SCI.	AGNews	ChemP.	IMDB	SCI.	AGNews
Base	$77.1_{1.1}$	$88.7_{0.4}$	$75.8_{1.1}$	$87.7_{0.3}$	$77.1_{1.1}$	$88.7_{0.4}$	$75.8_{1.1}$	$87.7_{0.3}$
Rand	$77.8_{0.4}$	$88.9_{0.2}$	$78.7_{0.9}$	$88.5_{0.3}$	$78.2_{0.3}$	$89.2_{0.2}$	$78.8_{0.2}$	$88.5_{0.2}$
DSIR	$\mathbf{80.9_{0.9}}$	$89.0_{0.3}$	$79.9_{1.0}$	$89.0_{0.1}$	$\mathbf{82.1_{0.3}}$	$89.4_{0.3}$	$78.2_{0.4}$	$88.9_{0.2}$
Ours	$80.4_{0.8}$	$\mathbf{90.1_{0.1}}$	$\mathbf{80.5_{0.4}}$	$\mathbf{89.3_{0.1}}$	$81.6_{0.1}$	$\mathbf{90.2_{0.3}}$	$\mathbf{79.8_{0.2}}$	$\mathbf{89.2_{0.1}}$