T-SHIRT: Token-Selective Hierarchical **Data Selection** for Instruction Tuning LLMs

Yanjun Fu
(yanjunfu@umd.edu)

Faisal Hamman
(fhamman@umd.edu)

Sanghamitra Dutta
(sanghamd@umd.edu)

San Diego Poster

Wed 3 Dec 4:30 p.m. PST – 7:30 p.m. PST

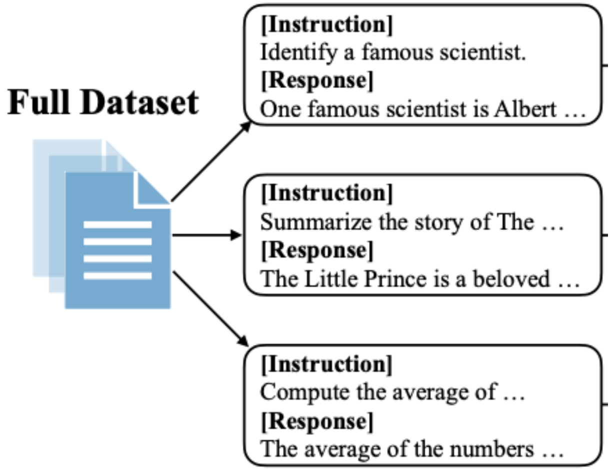
Exhibit Hall C,D,E

arxiv.org/abs/2506.01317

Motivation: Data Efficiency

- Instruction tuning is essential for LLMs for practical use

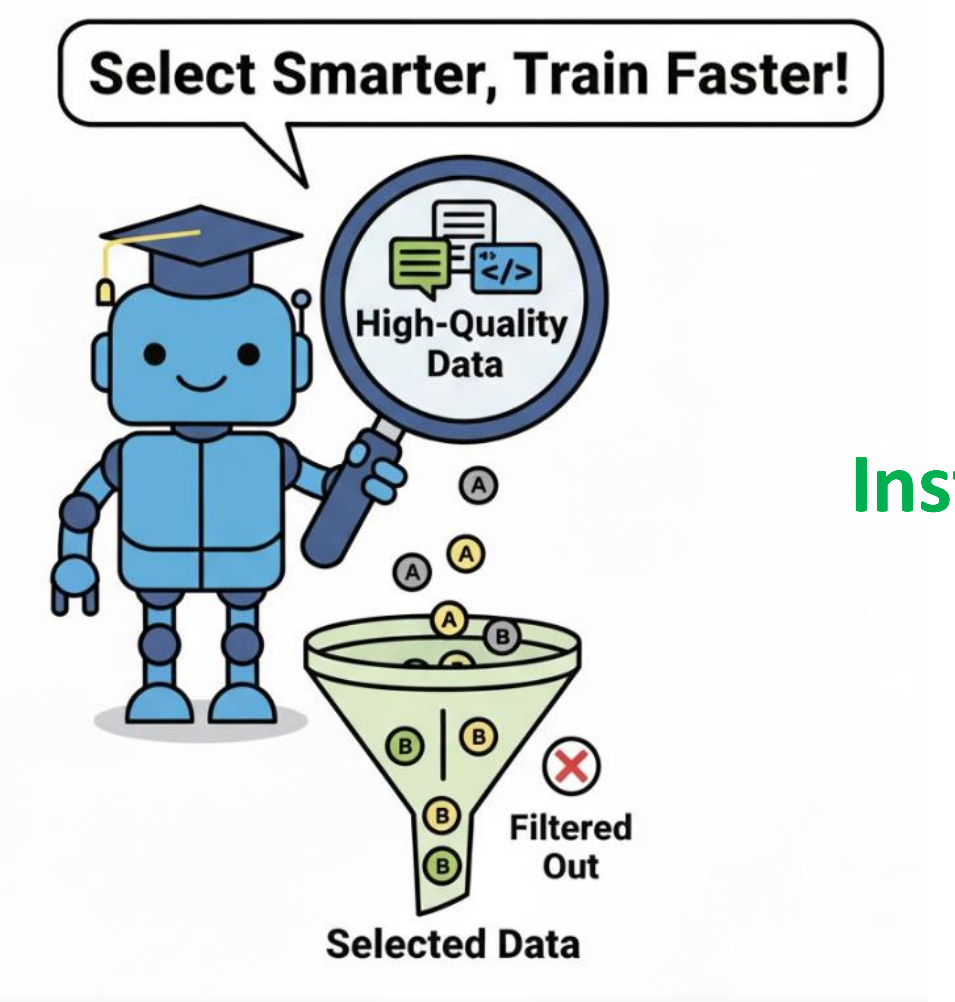
LLMs are typically fine-tuned on **massive** datasets of instruction-response pairs, which is **time-consuming**



- Shift in focus from **data quantity** to **data quality**

LIMA [NeurIPS’23] achieved strong performance after fine-tuned on just 1,000 manually selected samples.

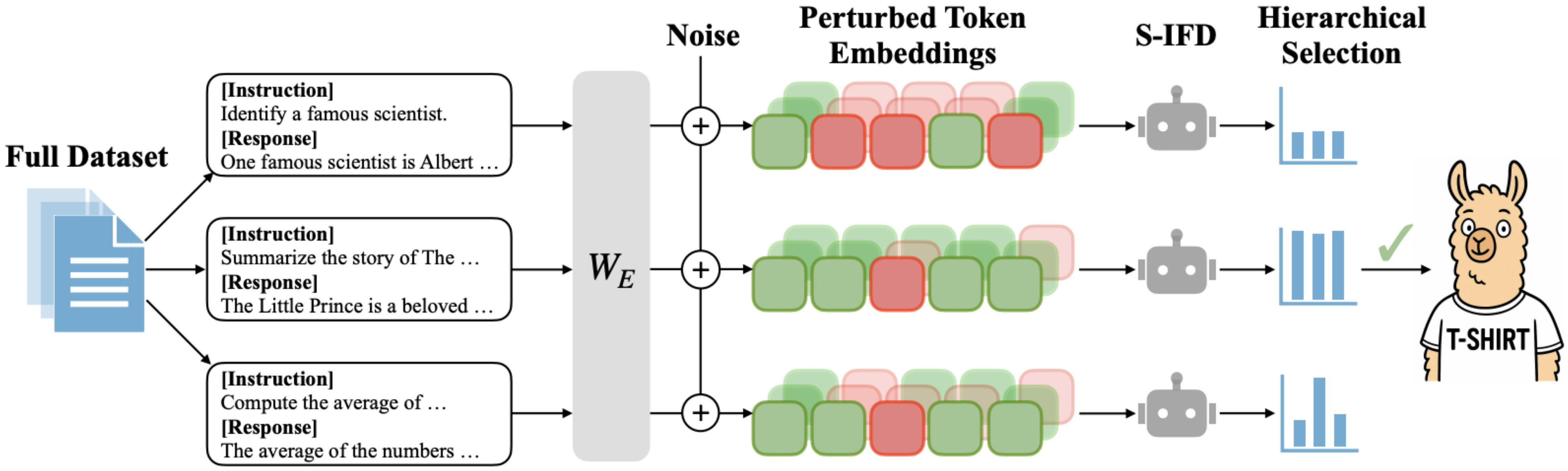
Motivation: Data Efficiency



Instruction tuning need not be data-intensive

Main Contribution:
A new fine-grained & robust data selection framework

T-SHIRT: **T**oken-**S**elective **HI**e**R**archical Data Selection for Instruction **T**uning



- New scoring method for data selection that includes only informative tokens
- Promotes robust and reliable samples whose neighbors also show high quality

Using only ~5% of the dataset, T-SHIRT outperforms existing baselines including those that use the full dataset!

Existing Scoring Functions for Data Selection

The standard Instruction-Following Difficulty (IFD) score for an instruction-response pair (x, y) is defined as:

$$\mathrm{IFD}(x,y) = \frac{\mathrm{PPL}_{\theta'}(y|x)}{\mathrm{PPL}_{\theta'}(y)} = \frac{\exp\left\{-\frac{1}{T}\sum_{t=1}^{T}\log P_{\theta'}\left(y_t|y_{<t},x\right)\right\}}{\exp\left\{-\frac{1}{T}\sum_{t=1}^{T}\log P_{\theta'}\left(y_t|y_{<t}\right)\right\}}$$

Captures ratio of perplexity of the response conditioned on the instruction to the unconditional perplexity of the response

Rewritten as:

$$\boxed{\mathrm{IFD}(x,y) = \exp\left(-\frac{1}{T}\sum_{t=1}^{T}\Delta_t\right)}$$

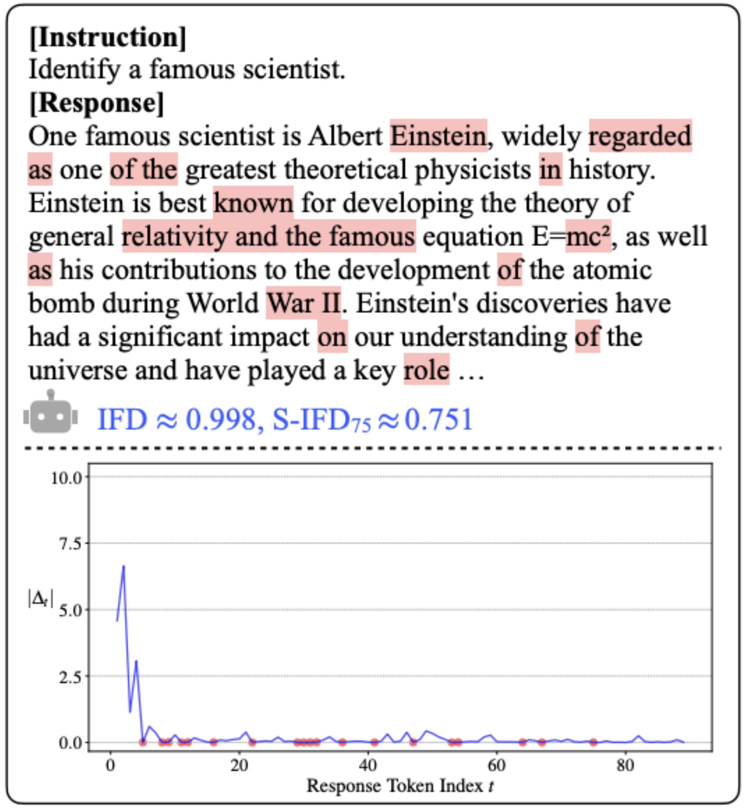
Our proposed token Informativeness

$$\Delta_t = \log P_{\theta'}(y_t|y_{<t},x) - \log P_{\theta'}(y_t|y_{<t})$$

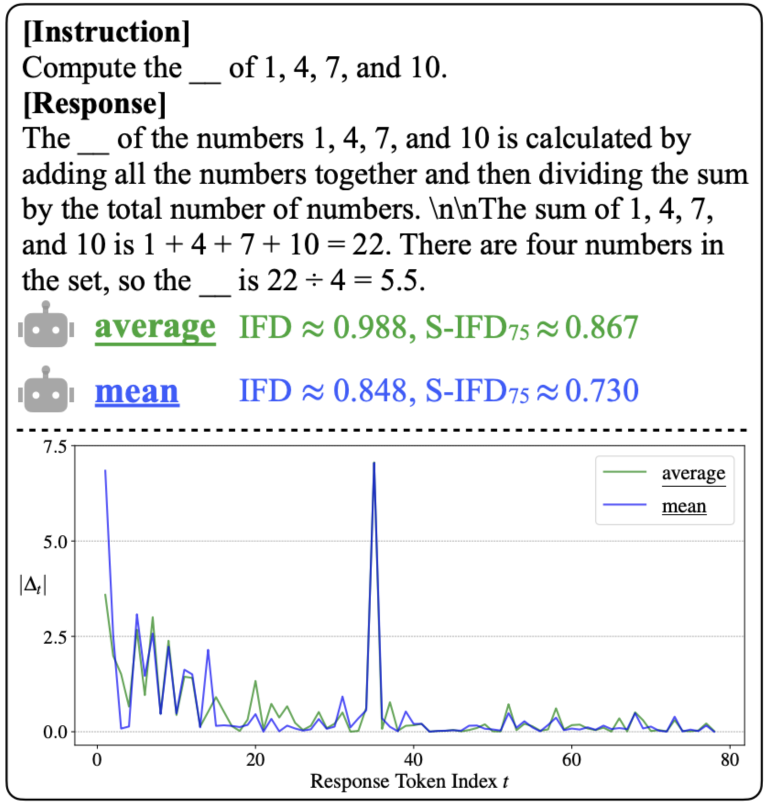
Δ_t measures the change in log-likelihood of token y_t when the instruction x is provided. A small $|\Delta_t|$ indicates that the instruction has little impact on generating y_t

We identify & resolve two key limitations of existing approaches

(1) Not all tokens are useful/informative in data quality evaluation



(2) Local neighborhood quality is important for reliable data selection



(1) Not all tokens are useful/informative in data quality evaluation

$$\mathrm{IFD}(x,y) = \exp\left(-\frac{1}{T}\sum_{t=1}^{T}\Delta_t\right)$$

Our proposed token Informativeness

Selective-IFD score:

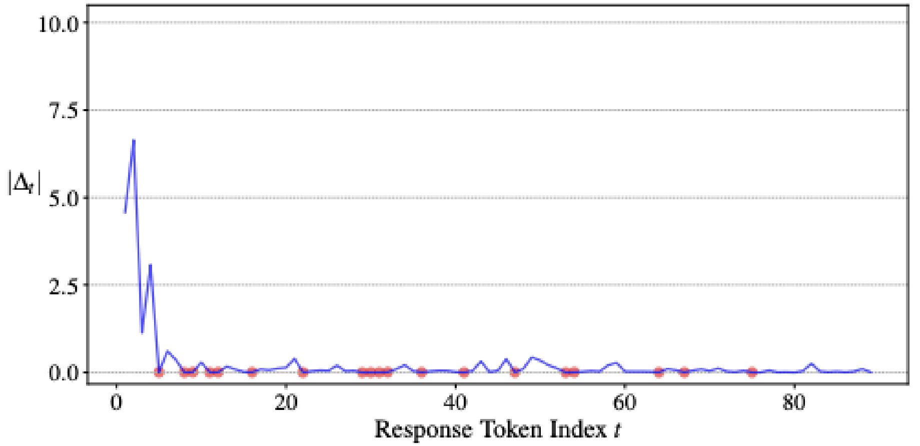
$$\text{S-IFD}_k\,(x,y) = \exp\left\{-\frac{1}{\sum_{t=1}^{T} w_t}\sum_{t=1}^{T} w_t\Delta_t\right\},$$

$$\text{where } w_t = \begin{cases} 1 & \text{if } |\Delta_t| \text{ ranks top } k\% \text{ in the dataset } \mathcal{D}, \\ 0 & \text{otherwise.} \end{cases}$$

Keeps only top k% tokens

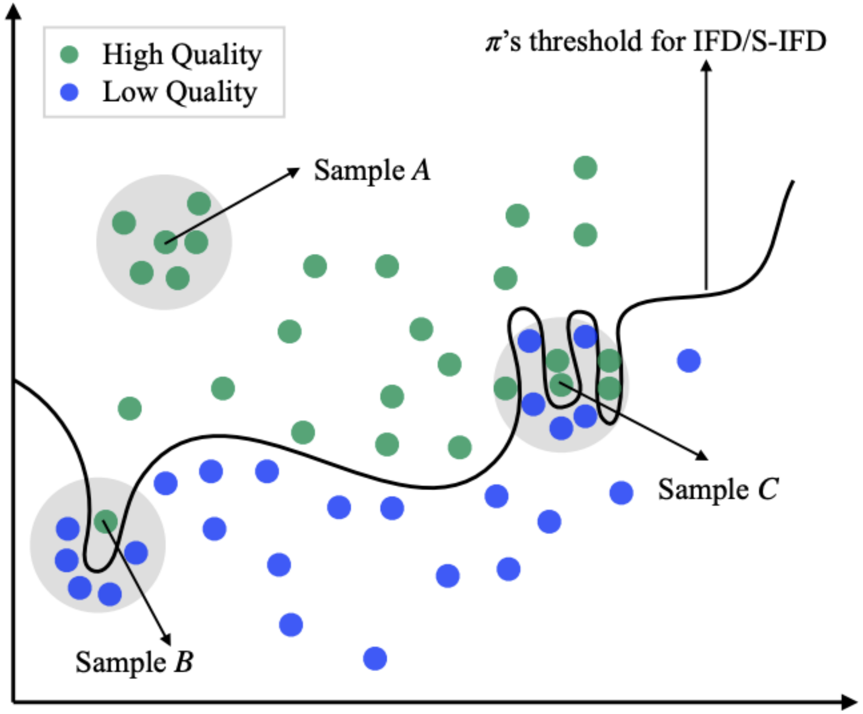
[Instruction]
Identify a famous scientist.
[Response]
One famous scientist is Albert Einstein, widely regarded as one of the greatest theoretical physicists in history. Einstein is best known for developing the theory of general relativity and the famous equation E=mc², as well as his contributions to the development of the atomic bomb during World War II. Einstein's discoveries have had a significant impact on our understanding of the universe and have played a key role …

IFD ≈ 0.998, S-IFD$_{75}$ ≈ 0.751



examples from the `Alpaca-GPT-4` dataset

(2) Local neighborhood quality is important for reliable data selection



$$\hat{\mu}(x, y) = \frac{1}{M} \sum_{i=1}^{M} \text{S-IFD}_k\big(x + \delta_x^{(i)}, y + \delta_y^{(i)}\big),$$

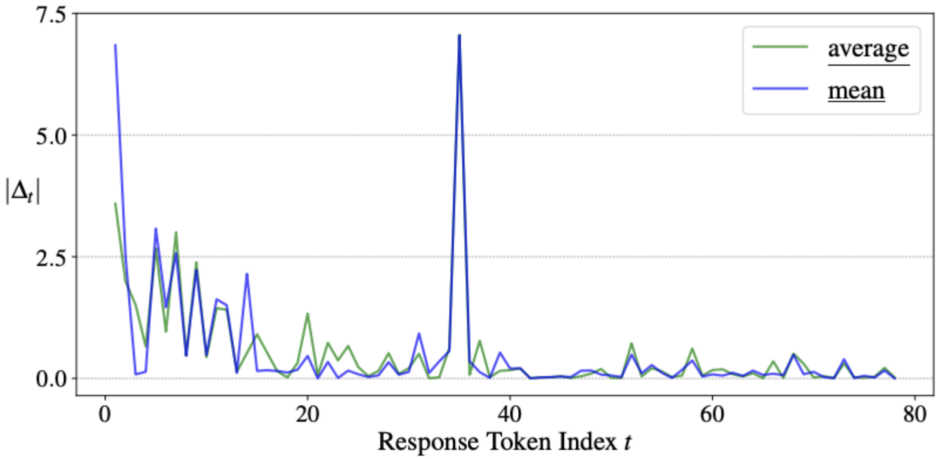
$$\hat{\sigma}^2(x, y) = \frac{1}{M} \sum_{i=1}^{M} \Big(\text{S-IFD}_k\big(x + \delta_x^{(i)}, y + \delta_y^{(i)}\big) - \hat{\mu}(x, y)\Big)^2,$$

where $\delta_x^{(i)} \sim \mathcal{U}^{L \times d}(-\epsilon, \epsilon)$, and $\delta_y^{(i)} \sim \mathcal{U}^{T \times d}(-\epsilon, \epsilon)$ with $\mathcal{U}$ denoting the uniform distribution.

[Instruction]
Compute the __ of 1, 4, 7, and 10.
[Response]
The __ of the numbers 1, 4, 7, and 10 is calculated by adding all the numbers together and then dividing the sum by the total number of numbers. \n\nThe sum of 1, 4, 7, and 10 is $1 + 4 + 7 + 10 = 22$. There are four numbers in the set, so the __ is $22 \div 4 = 5.5$.

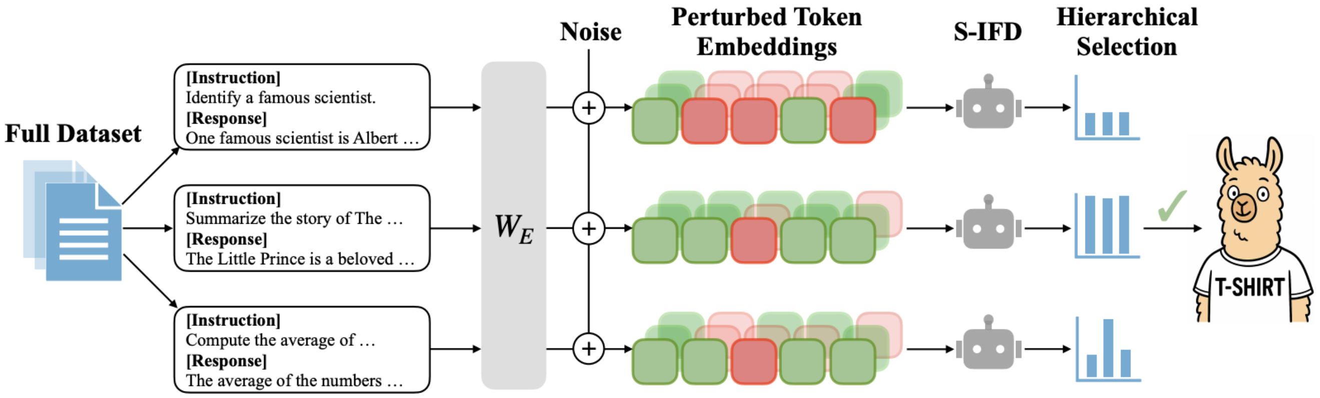
average IFD ≈ 0.988, S-IFD$_{75} \approx 0.867$

mean IFD ≈ 0.848, S-IFD$_{75} \approx 0.730$



examples from the `Alpaca-GPT-4` dataset

Our data selection algorithm T-SHIRT



Algorithm 1: Token-Selective Hierarchical Data Selection for Instruction Tuning (T-SHIRT)

Input: Dataset $\mathcal{D}$, selection budget b, token selection ratio $k\%$, oversampling factor γ, base noise scale α, and number of perturbations M

foreach $(x, y) \in \mathcal{D}$ **do**

  Compute $\epsilon \leftarrow \alpha / \sqrt{(L+T)d}$

  for $i \leftarrow 1$ **to** M **do**

    Sample noise $\delta_x^{(i)} \sim \mathcal{U}^{L \times d}(-\epsilon, \epsilon)$, and $\delta_y^{(i)} \sim \mathcal{U}^{T \times d}(-\epsilon, \epsilon)$

    Compute perturbed embeddings $x' \leftarrow x + \delta_x^{(i)}$, $y' \leftarrow y + \delta_y^{(i)}$

    Compute S-IFD$_k(x', y')$ via Equation (3)

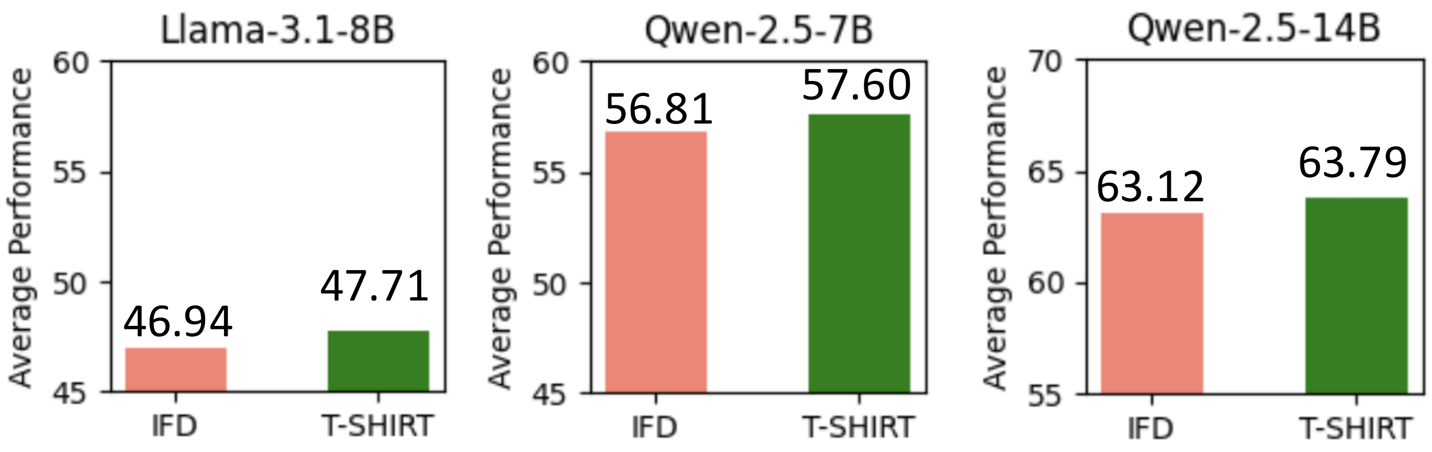
  Compute $\hat{\mu}(x, y)$ and $\hat{\sigma}^2(x, y)$ via Equation (4)

Select top γb samples from $\mathcal{D}$ with highest $\hat{\mu}(x, y)$ to construct $\hat{\mathcal{S}}$

From $\hat{\mathcal{S}}$, select final b samples with lowest $\hat{\sigma}^2(x, y)$ to construct $\mathcal{S}$

Output: Selected subset $\mathcal{S} \subset \mathcal{D}$ of size b

Experiments show strong performance



San Diego Poster

Wed 3 Dec 4:30 p.m. PST – 7:30 p.m. PST

Exhibit Hall C,D,E

	OpenLLM Leaderboard							LLM-as-a-Judge			μ_{ALL}
	ARC-C	HS	MMLU	TQA	BBH	GSM	μ_{OPEN}	AH[3]	AE-2	μ_{LLM}	
Qwen-2.5-14B											
LONGEST	68.17	84.01	79.23	**58.69**	61.00	**85.22**	72.72	34.40	35.01	34.71	63.22
IFD	**68.94**	**84.05**	79.07	57.87	61.78	83.02	72.46	33.10	37.15	35.13	63.12
T-SHIRT ($k = 50$)	68.60	83.90	**79.26**	58.60	**62.04**	84.15	**72.76**	**35.30**	**38.45**	**36.88**	**63.79**

Using only ~5% of the dataset, T-SHIRT outperforms existing baselines including those that use the full dataset!

	OpenLLM Leaderboard							LLM-as-a-Judge			μ_{ALL}
	ARC-C	HS	MMLU	TQA	BBH	GSM	μ_{OPEN}	AH	AE-2	μ_{LLM}	
Llama-3.1-8B											
FULL	55.97	77.89	59.77	53.32	43.86	32.75	53.93	5.20	[illegible]	[illegible]	[illegible]
RANDOM	57.94	81.29	61.37	53.39	45.62	30.71	55.05	[illegible]	[illegible]	[illegible]	[illegible]
LONGEST	58.45	83.07	61.83	54.86	46.40	51.25	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
DEITA	59.73	81.92	62.65	51.32	46.95	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
DS[2]	61.26	82.62	63.68	54.28	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
IFD	61.35	83.00	62.88	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	46.94
T-SHIRT ($k = 50$)	60.15	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	8.17	47.34
T-SHIRT ($k = 75$)	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	10.03	8.12	**47.71**
[illegible]											
FULL	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	77.10	66.62	11.50	8.81	10.16	52.50
RANDOM	[illegible]	[illegible]	[illegible]	[illegible]	53.34	33.13	61.08	14.70	16.45	15.58	49.70
LONGEST	[illegible]	[illegible]	[illegible]	61.48	53.52	84.61	70.27	14.30	**19.15**	16.73	56.88
DEITA	[illegible]	[illegible]	74.15	58.77	52.65	82.79	69.22	14.60	18.41	16.51	56.04
DS[2]	[illegible]	82.07	**74.35**	60.58	54.11	82.11	69.72	13.80	15.25	14.53	55.92
IFD	64.76	**82.66**	74.33	60.86	53.76	86.50	70.48	15.60	16.01	15.81	56.81
T-SHIRT ($k = 50$)	**66.21**	82.39	74.23	**61.58**	**54.21**	86.81	**70.91**	**16.40**	18.94	17.67	**57.60**
T-SHIRT ($k = 75$)	65.78	82.45	74.06	61.12	54.14	**87.19**	70.79	**16.40**	18.98	**17.69**	57.52

Several other experiments & ablations in full paper