Efficiently Learning at Test-Time: Active Fine-Tuning of LLMs

Jonas Hübotter, Sascha Bongni, Ido Hakimi, Andreas Krause

ETH zürich

Learning & Adaptive Systems

Background

- **Goal:** Learn a specific model, tailored to each prompt.
- This requires automatic data selection.

How can we select data that effectively reduces uncertainty about the response?

We find: Nearest neighbor retrieval selects redundant data ↓

Prompt: What is the age of Michael Jordan and how many kids does he have?

Nearest Neighbor:

1. The age of Michael Jordan is 61 years.
2. Michael Jordan was born on February 17, 1963.

SIFT (ours):

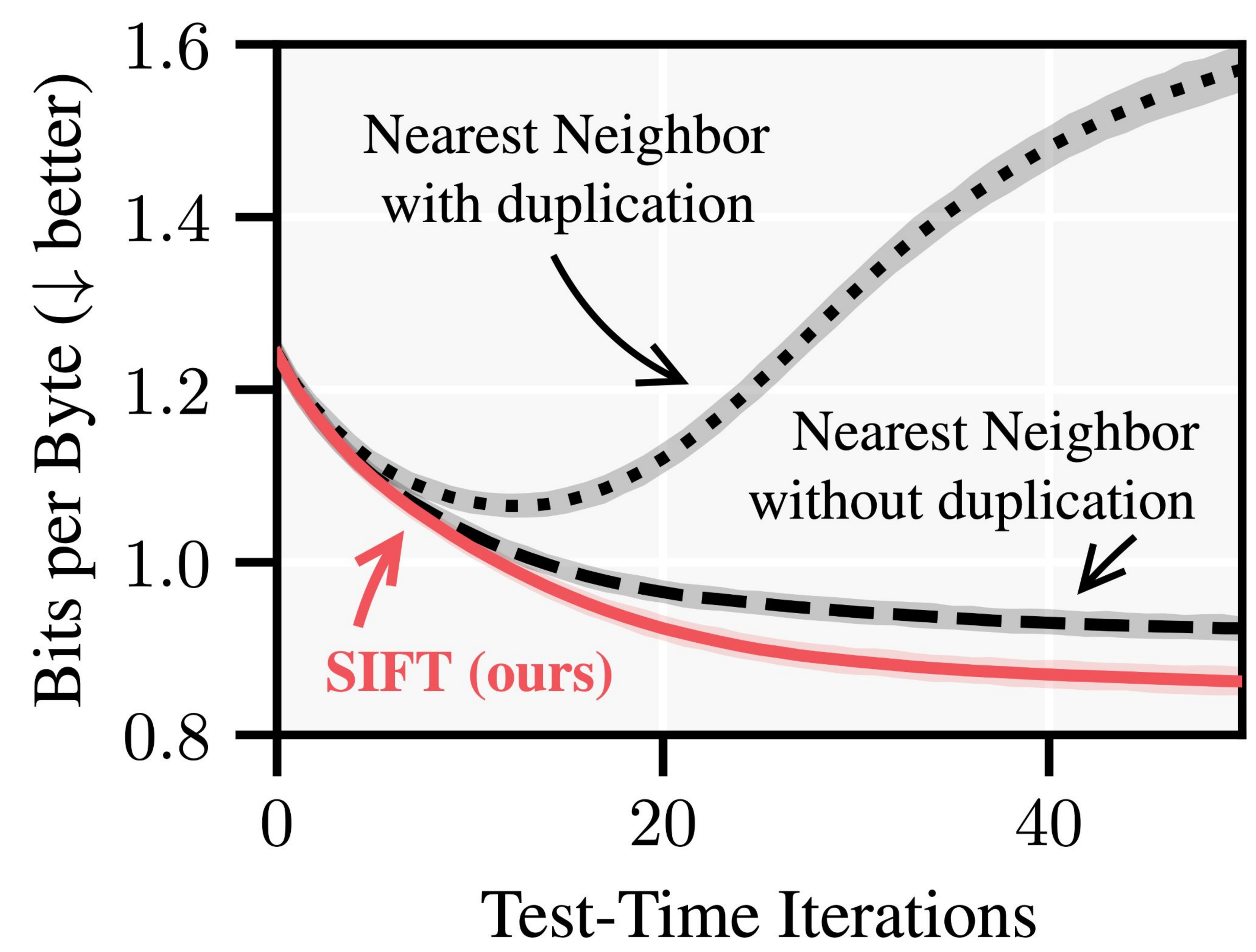
1. The age of Michael Jordan is 61 years.
2. Michael Jordan has five children.

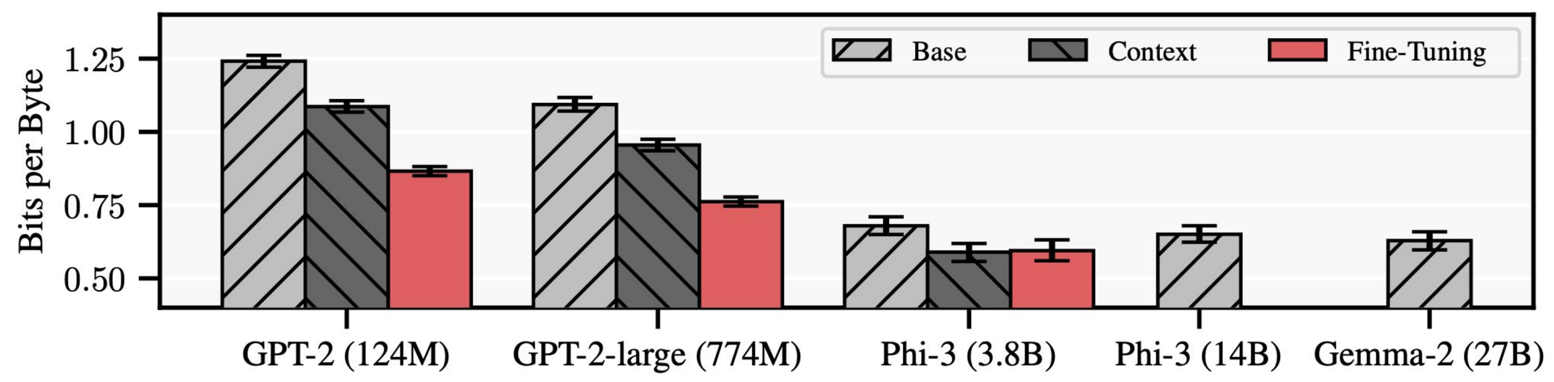
Contributions

- We propose **SIFT**, which selects data that maximally reduces the LLMs "uncertainty" about its response.
- SIFT tractably & effectively estimates the LLMs *relative* uncertainty.
- We show that test-time training can improve the performance of SOTA small language models.

LLMs improve by training at test-time.

Training on the most ***informative*** data yields the largest performance gains.

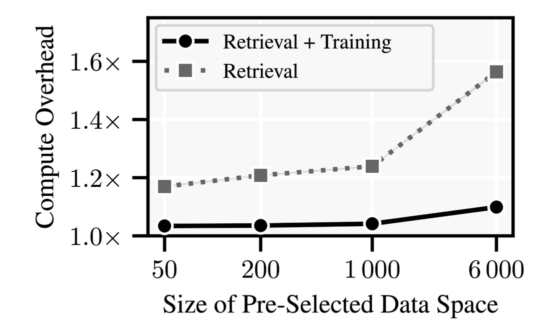


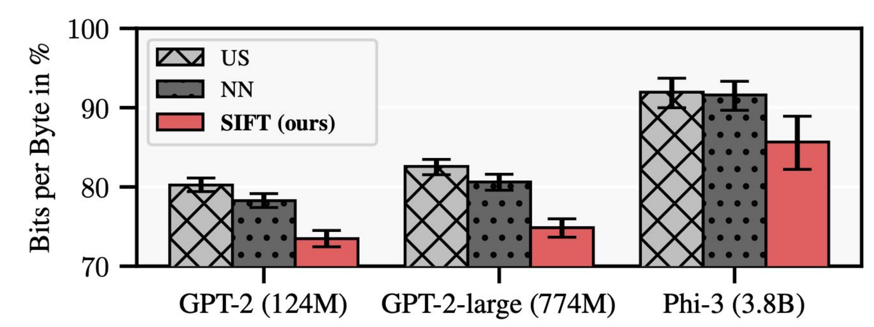


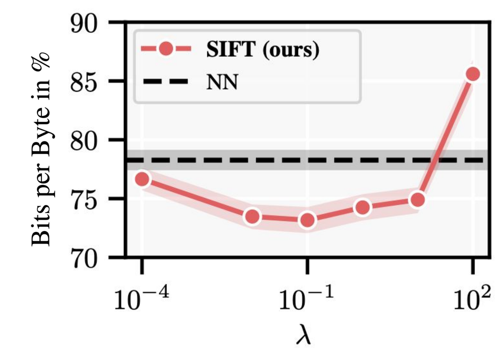
Details

- We evaluate on the broad Pile corpus.
- Test-time training with SIFT robustly outperforms base model and baselines.

	US	NN	NN-F	SIFT	Δ
NIH Grants	93.1 (1.1)	84.9 (2.1)	91.6 (16.7)	**53.8** (8.9)	↓31.1
US Patents	85.6 (1.5)	80.3 (1.9)	108.8 (6.6)	**62.9** (3.5)	↓17.4
GitHub	45.6 (2.2)	42.1 (2.0)	53.2 (4.0)	**30.0** (2.2)	↓12.1
Enron Emails	**68.6** (9.8)	**64.4** (10.1)	91.6 (20.6)	**53.1** (11.4)	↓11.3
Wikipedia	67.5 (1.9)	**66.3** (2.0)	121.2 (3.5)	**62.7** (2.1)	↓3.6
Common Crawl	92.6 (0.4)	90.4 (0.5)	148.8 (1.5)	**87.5** (0.7)	↓2.9
PubMed Abstr.	88.9 (0.3)	87.2 (0.4)	162.6 (1.3)	**84.4** (0.6)	↓2.8
ArXiv	85.4 (1.2)	**85.0** (1.6)	166.8 (6.4)	**82.5** (1.4)	↓2.5
PubMed Central	**81.7** (2.6)	**81.7** (2.6)	155.6 (5.1)	**79.5** (2.6)	↓2.2
Stack Exchange	78.6 (0.7)	78.2 (0.7)	141.9 (1.5)	**76.7** (0.7)	↓1.5
Hacker News	**80.4** (2.5)	**79.2** (2.8)	133.1 (6.3)	**78.4** (2.8)	↓0.8
FreeLaw	**63.9** (4.1)	**64.1** (4.0)	122.4 (7.1)	**64.0** (4.1)	↑0.1
DeepMind Math	**69.4** (2.1)	**69.6** (2.1)	121.8 (3.1)	**69.7** (2.1)	↑0.3
All	80.2 (0.5)	78.3 (0.5)	133.3 (1.2)	**73.5** (0.6)	↓4.8







1. Estimate uncertainty

Surrogate model: logit-linear model $s(f^\star(x))$ with $f^\star(x) = \boldsymbol{W}^\star\phi(x)$ [$\boldsymbol{W}^\star$ unknown, $\phi(\cdot)$ known]:

$$\underbrace{s^\star(x) = s(f^\star(x))}_{\text{"truth"}} \qquad \underbrace{s_n(x) = s(\boldsymbol{W}_n\phi(x))}_{\text{fine-tuned model on } n \text{ data points}}$$

Confidence sets: $\underbrace{d_{\mathrm{TV}}(s_n(x), s^\star(x))}_{\text{error}} \leq \underbrace{\beta_n(\delta)}_{\text{scaling}} \underbrace{\sigma_n(x)}_{\textbf{key obj.}}$

[with probability $1-\delta$]

$\rightsquigarrow$ $\sigma_n(x)$ measures **uncertainty** about response to x!

2. Minimize "posterior" uncertainty

$$x_{n+1} = \operatorname*{argmin}_x \sigma_{X_n \cup \{x\}}(x^\star) \quad \text{(prompt: } x^\star\text{)} \qquad \text{with } k(x,x') = \boldsymbol{\phi}(x)^\top\boldsymbol{\phi}(x')$$

$$= \operatorname*{argmax}_x \begin{bmatrix} k(x^\star,x_1) \\ \vdots \\ k(x^\star,x_n) \\ k(x^\star,x) \end{bmatrix}^\top \left(\begin{bmatrix} k(x_1,x_1) & \cdots & k(x_1,x_n) & k(x_1,x) \\ \vdots & \ddots & \vdots & \vdots \\ k(x_n,x_1) & \cdots & k(x_n,x_n) & k(x_n,x) \\ k(x,x_1) & \cdots & k(x,x_n) & k(x,x) \end{bmatrix} + \lambda I_{n+1} \right)^{-1} \begin{bmatrix} k(x^\star,x_1) \\ \vdots \\ k(x^\star,x_n) \\ k(x^\star,x) \end{bmatrix}$$

maximize relevance — minimize redundancy

Theory: $\sigma_n^2(x) - \sigma_\infty^2(x) \leq \frac{O(\lambda \log(n))}{\sqrt{n}}$