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

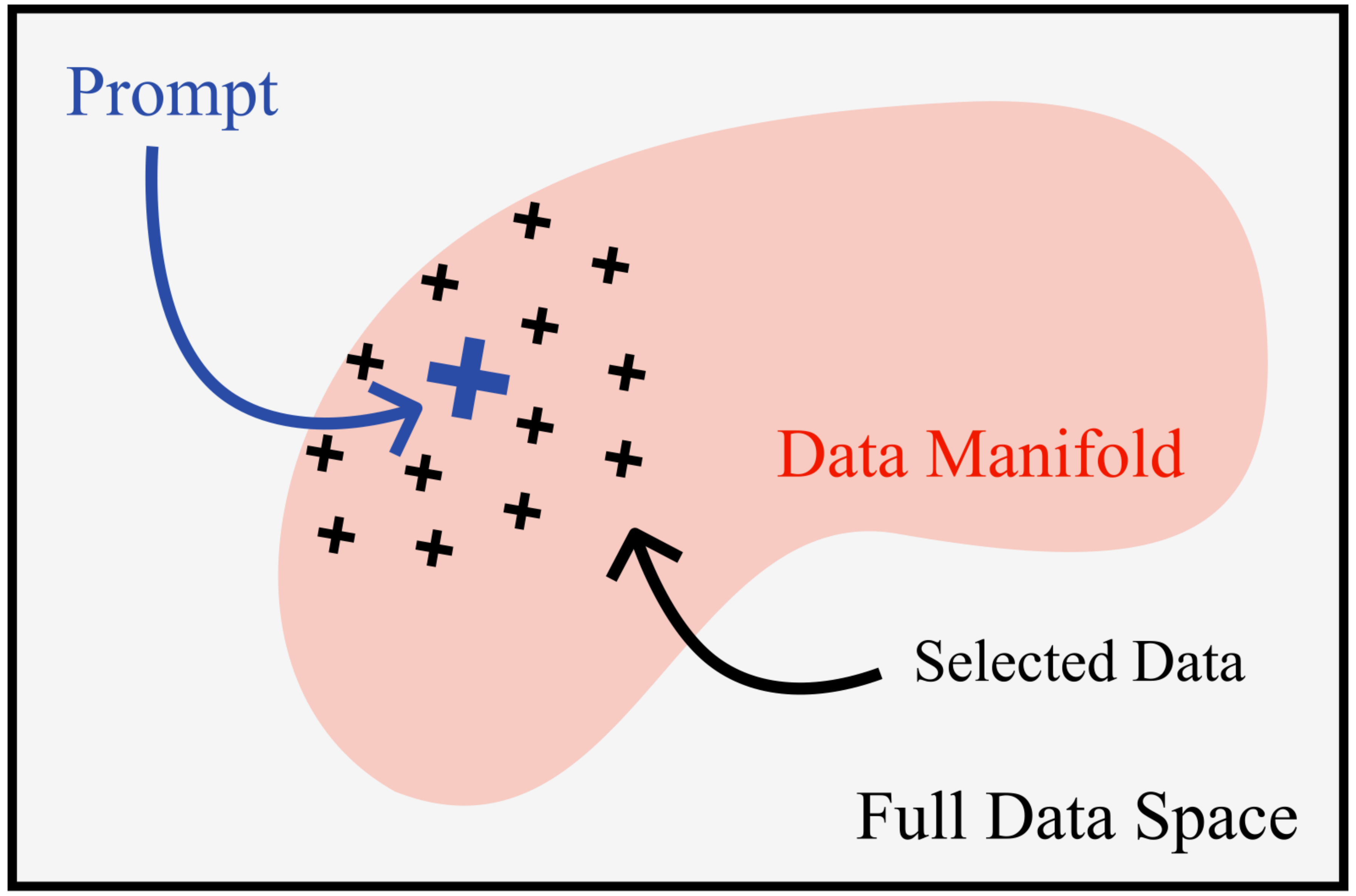
Jonas Hübotter, Sascha Bongni,
Ido Hakimi, Andreas Krause

ETH zürich




Test-time fine-tuning

(Bottou, Vapnik; ’92 & Hardt, Sun; ICLR ’24)



1. take pre-trained model f
2. given input x, find local data D_x from memory
3. fine-tune model f on local data D_x to get **local model** f_x
4. predict $f_x(x)$

Which local data D_x to use?

- previous work used Nearest Neighbor (NN) retrieval in some metric space
- **we show:** NN is suboptimal!

Which local data D_x to use?

- previous work used Nearest Neighbor (NN) retrieval in some metric space
- **we show:** NN is suboptimal!

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.

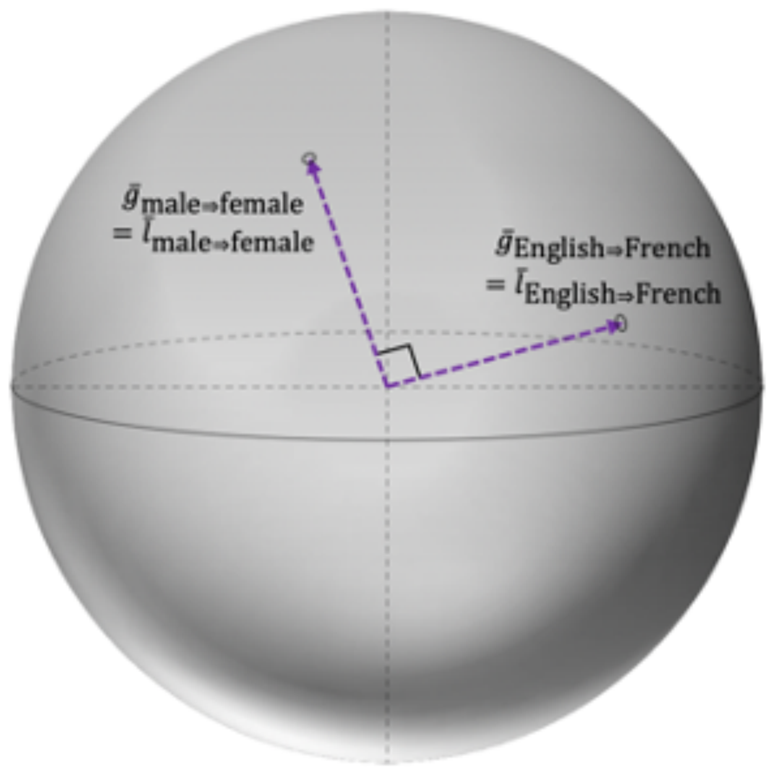
SIFT: **S**electing **I**nformative data for **F**ine-**T**uning

Principle:

Select data that maximally reduces “uncertainty” about how to respond to the prompt.

1. Estimate uncertainty
2. Minimize uncertainty

1) Estimating uncertainty



- Making this tractable…

  Surrogate model: approximate model f as logit-linear model in a known representation space

  → linear representation hypothesis (e.g., Park et al; ICML '24)

- **Error bound:** $\underbrace{d_{\mathrm{TV}}(f_n(x), f^\star(x))}_{\text{error}} \leq \underbrace{\beta(\delta)}_{\text{scaling}} \underbrace{\sigma_n(x)}_{\text{uncertainty}}$ (with prob. $1 - \delta$)

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

2) Minimizing uncertainty

- SIFT: minimize uncertainty about response to input $x^\star$

$$D_{x^\star} = X_n \cup \{x_{n+1}\} \quad \text{with } x_{n+1} = \operatorname*{argmin}_x \sigma_{X_n \cup \{x\}}(x^\star)$$

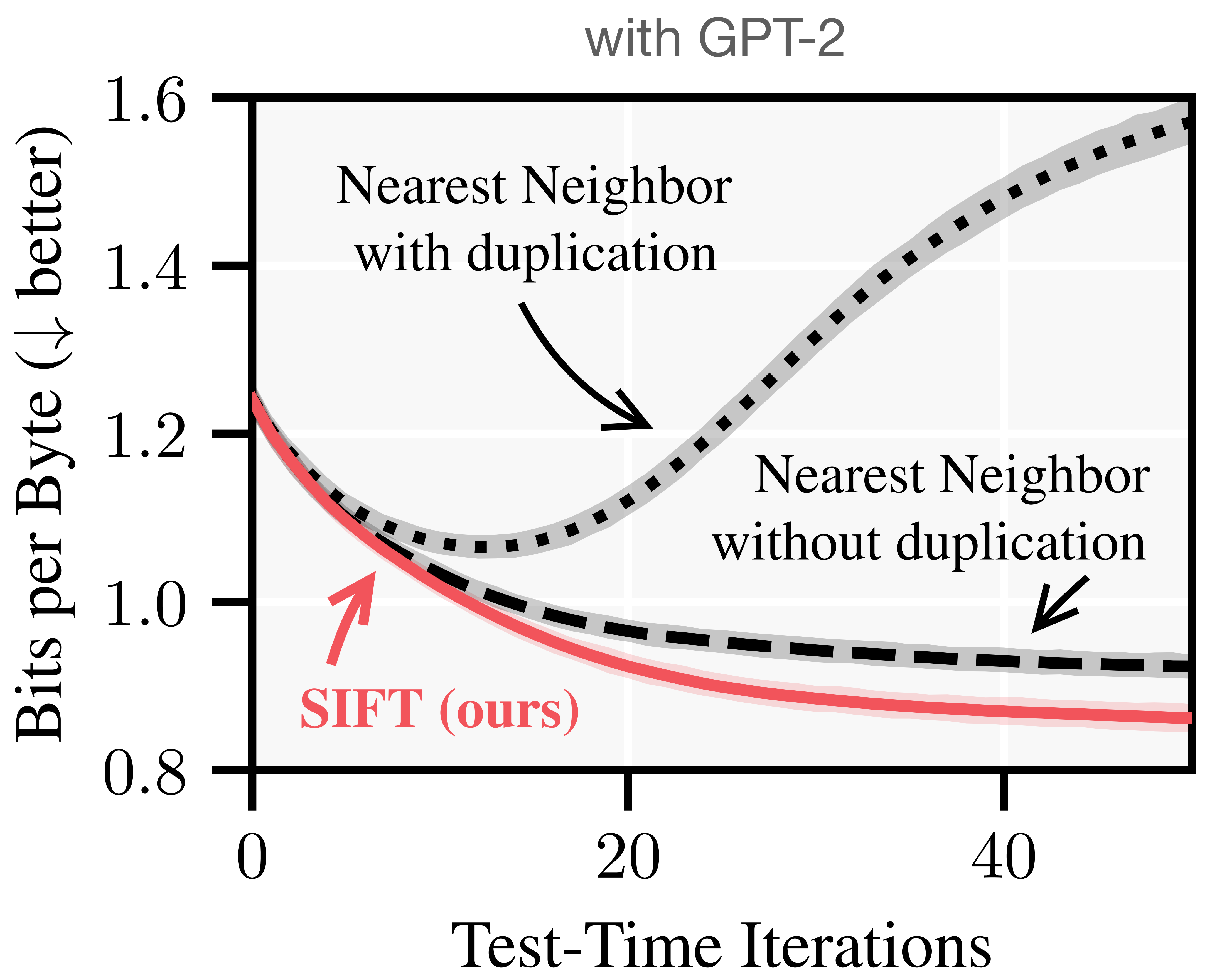
- convergence of uncertainty is guaranteed!

$$\sigma_n(x^\star) \to \underbrace{\boxed{\sigma_\infty(x^\star)}}_{\text{irreducible uncertainty}}$$

→ predictions can only be as good as the data and the learned abstractions!

Evaluation: language modeling on the Pile

Pile dataset
NIH Grants
US Patents
GitHub
Enron Emails
Common Crawl
ArXiv
Wikipedia
PubMed Abstr.
Hacker News
Stack Exchange
PubMed Central
DeepMind Math
FreeLaw
All

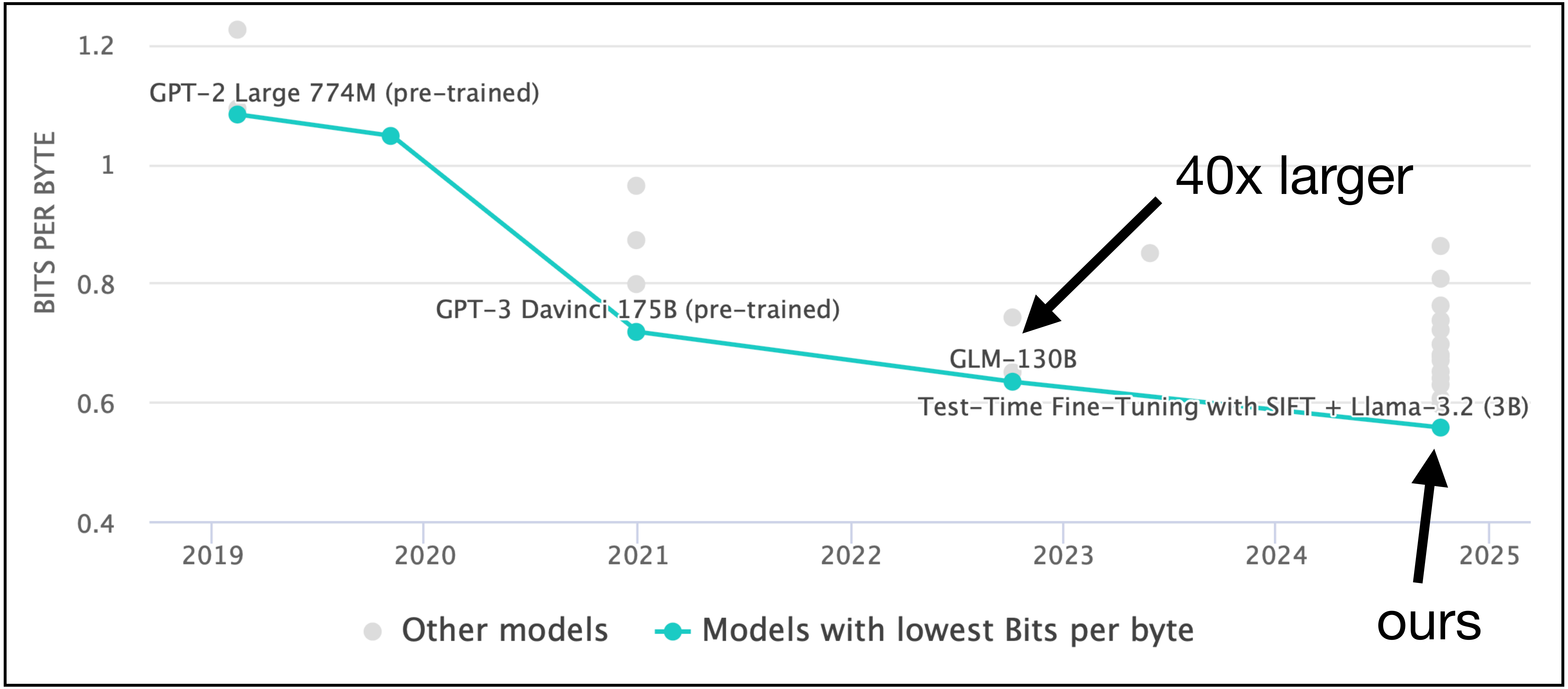


Observations

- larger relative gains with stronger base models
- larger relative gains with larger "memory"

	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

New SOTA on the Pile benchmark



https://paperswithcode.com/sota/language-modelling-on-the-pile

Conclusion

- SIFT selects more informative data than Nearest Neighbor retrieval
- Test-time fine-tuning is a promising paradigm to allocate compute to tasks we find interesting

Happy to discuss more

jonas.huebotter@inf.ethz.ch