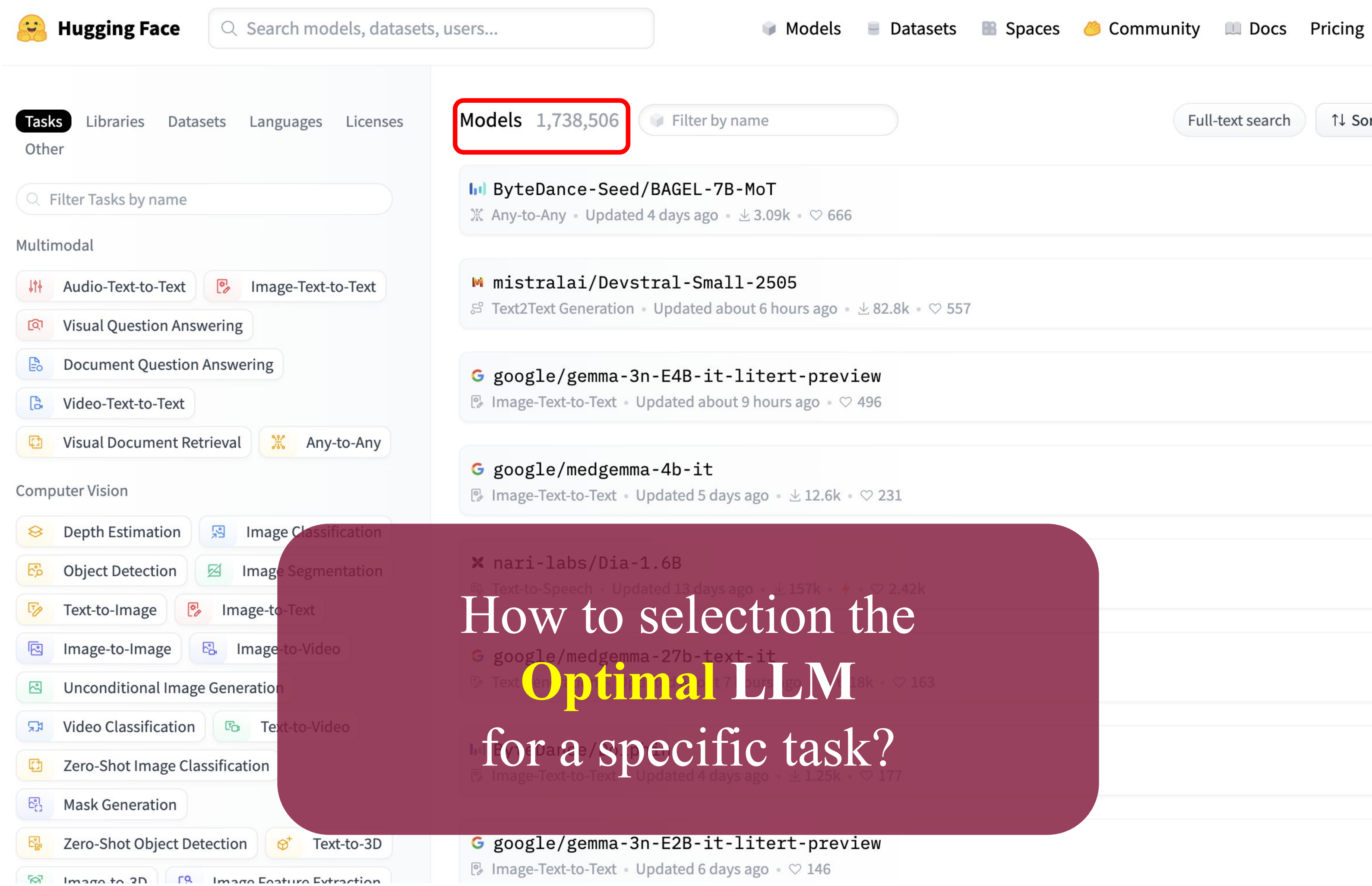
LensLLM: Unveiling Fine-Tuning Dynamics for LLM Selection

Presenter: Xinyue Zeng

Authors: Xinyue Zeng, Haohui Wang, Junhong Lin, Jun Wu, Tyler Cody, Dawei Zhou

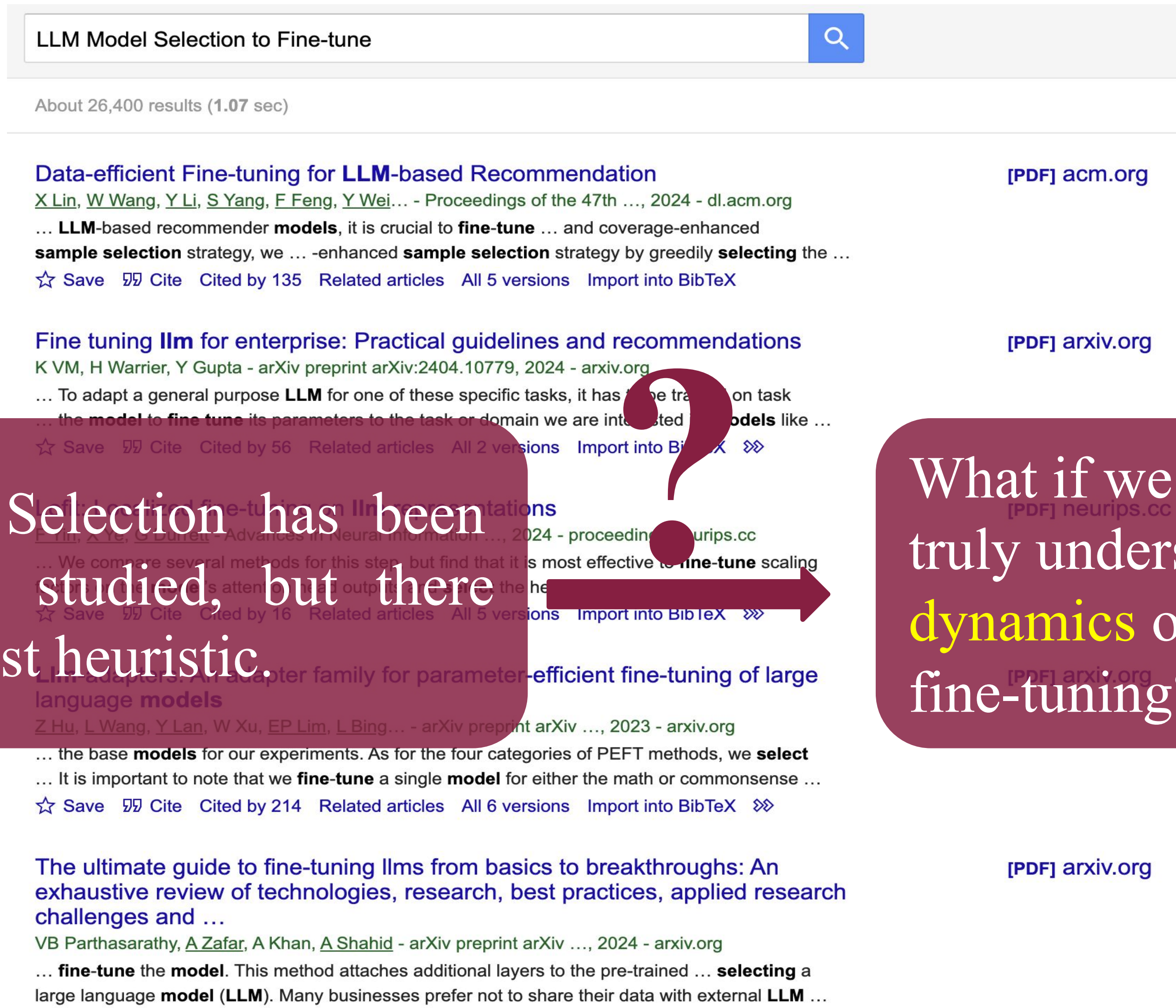
VIRGINIA TECH

Motivation



How to selection the **Optimal** LLM for a specific task?

Relative Work


LLM Model Selection to Fine-tune

About 26,400 results (**1.07** sec)

Data-efficient Fine-tuning for **LLM**-based Recommendation — [PDF] acm.org
X Lin, W Wang, Y Li, S Yang, F Feng, Y Wei… - Proceedings of the 47th …, 2024 - dl.acm.org
… **LLM**-based recommender **models**, it is crucial to **fine-tune** … and coverage-enhanced **sample selection** strategy, we … -enhanced **sample selection** strategy by greedily **selecting** the …
Save Cite Cited by 135 Related articles All 5 versions Import into BibTeX

Fine tuning **llm** for enterprise: Practical guidelines and recommendations — [PDF] arxiv.org
K VM, H Warrier, Y Gupta - arXiv preprint arXiv:2404.10779, 2024 - arxiv.org
… To adapt a general purpose **LLM** for one of these specific tasks, it has … on task
… the **model** to **fine tune** its parameters to the task or domain we are inte… **odels** like …
Save Cite Cited by 56 Related articles All 2 versions Import into BibTeX

… **llm** representations — [PDF] neurips.cc
… - Advances in Neural Information …, 2024 - proceedin… urips.cc
… We compare several methods for this step, but find that it is most effective … **fine-tune** scaling …
Save Cite Cited by 16 Related articles All 5 versions Import into BibTeX

Llm-adapters: An adapter family for parameter-efficient fine-tuning of large language **models** — [PDF] arxiv.org
Z Hu, L Wang, Y Lan, W Xu, EP Lim, L Bing… - arXiv preprint arXiv …, 2023 - arxiv.org
… the base **models** for our experiments. As for the four categories of PEFT methods, we **select** … It is important to note that we **fine-tune** a single **model** for either the math or commonsense …
Save Cite Cited by 214 Related articles All 6 versions Import into BibTeX

The ultimate guide to fine-tuning llms from basics to breakthroughs: An exhaustive review of technologies, research, best practices, applied research challenges and … — [PDF] arxiv.org
VB Parthasarathy, A Zafar, A Khan, A Shahid - arXiv preprint arXiv …, 2024 - arxiv.org
… **fine-tune** the **model**. This method attaches additional layers to the pre-trained … **selecting** a large language **model** (**LLM**). Many businesses prefer not to share their data with external **LLM** …

LLM Selection has been widely studied, but there are most heuristic.

?

What if we could truly understand the dynamics of LLM fine-tuning?

Conclusion

We need a theoretical framework to help us understand LLM behaviors in fine-tuning and assist to selection models.

Problem Formulation

LLM Selection in Resource-Constrained Scenarios: In the context of LLM selection, we aim to identify the optimal model from a set of candidate models for specific tasks under resource-constrained scenarios. Without loss of generalization, we denote S as training dataset from D, and $M = \{m_1, m_2, \ldots, m_k\}$ as a set of candidate models. For each model m_i, there is associated with a feature vector x_i representing its characteristics (e.g., model size, architecture, training data)

Given: (1) Limited training data S; (2) A set of candidate LLMs $M = \{m_1, m_2, \ldots, m_k\}$ with their corresponding feature vectors x.
Objective: The optimal model $\boldsymbol{m}^* \in M$ on S that has the best performance.

Proposed Framework

Theorem 2 (Hessian-based PAC-Bayes Generalization Bound). *For a fine-tuned transformer model $f_{\hat{w}}$ satisfying Assumptions 1-3, with probability at least 0.99 and any fixed $\epsilon > 0$:*

$$L(f_{\hat{w}}) \leq (\epsilon+1)\hat{L}(f_{\hat{w}}) + \frac{(1+\epsilon)\sqrt{C}\sum_{i=1}^{l}\sqrt{h_i}}{\sqrt{n}} + O(n^{-\frac{3}{4}}) \quad (4)$$

where $h_i \geq \max_{(x,y)\in D} \boldsymbol{v}_i^T \boldsymbol{H}_i^{+}[L(f_{\hat{w}})]\boldsymbol{v}_i$.

Proposition 1. *Let $\{h_i\}_{i=1}^{l}$ be a sequence of Hessian-related values that satisfy $h_i \geq \max_{(x,y)\in D} \boldsymbol{v}_i^T \boldsymbol{H}_i^{+}[L(f_{\hat{w}})]\boldsymbol{v}_i$:*

(1) By applying Cauchy-Schwarz theorem (Steele, 2004), we have the following inequality:

$$\sum_{i=1}^{l}\sqrt{h_i} \leq \sqrt{l\sum_{i=1}^{l} h_i} \quad (5)$$

(2) Based on the definition of h_i, we have the following upper bound:

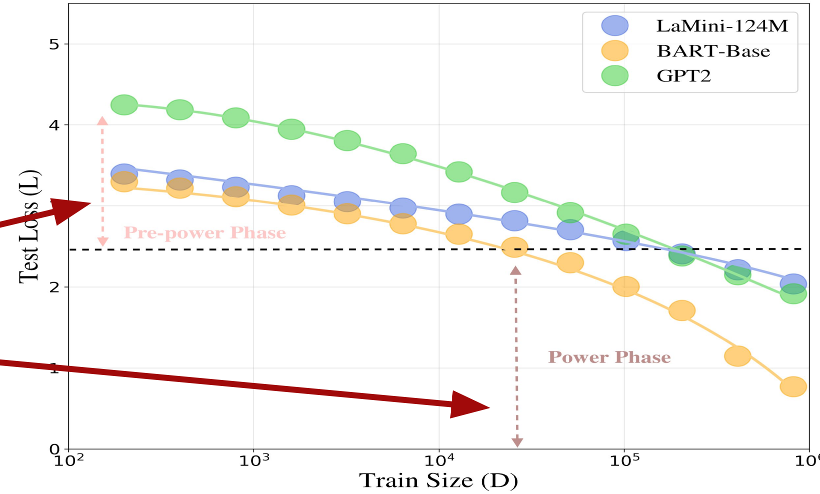
$$h_i \leq C_2 n^{-\beta_2} \quad (6)$$

where C_2 is independent of n and $\beta_2 \leq \beta_1$.

Corollary 1. *For any $\epsilon > 0$, with probability over 0.99, under Assumptions 1-3, considering the properties of the Hessian matrix, the Hessian-based PAC-Bayes generalization bound could be extended to:*

$$L(f_{\hat{w}}) \leq (1+\epsilon)\hat{L}(f_{\hat{w}}) + \boxed{C_3 n^{-\beta_3}} + \boxed{O(n^{-\frac{3}{4}})} \quad (7)$$

where $C_3 = \sqrt{C \cdot l \cdot C_2}$ and $\beta_3 = \frac{\beta_2+1}{2}$ are both model/task-dependent.



Our theoretical analysis successfully unveils the phases transition during Fine-tuning!

Proposed Framework

We propose **LENSLLM**, a Hessian-aware rectified scaling model:

$$L(D) = \frac{B}{F(\Theta, t) + D^{\beta}} + E$$

where (a) $F(\Theta, t)$ is the adapted NTK-based test loss function on transformer,

(b) D is the number of training data,

(c) β denotes the learning difficulty,

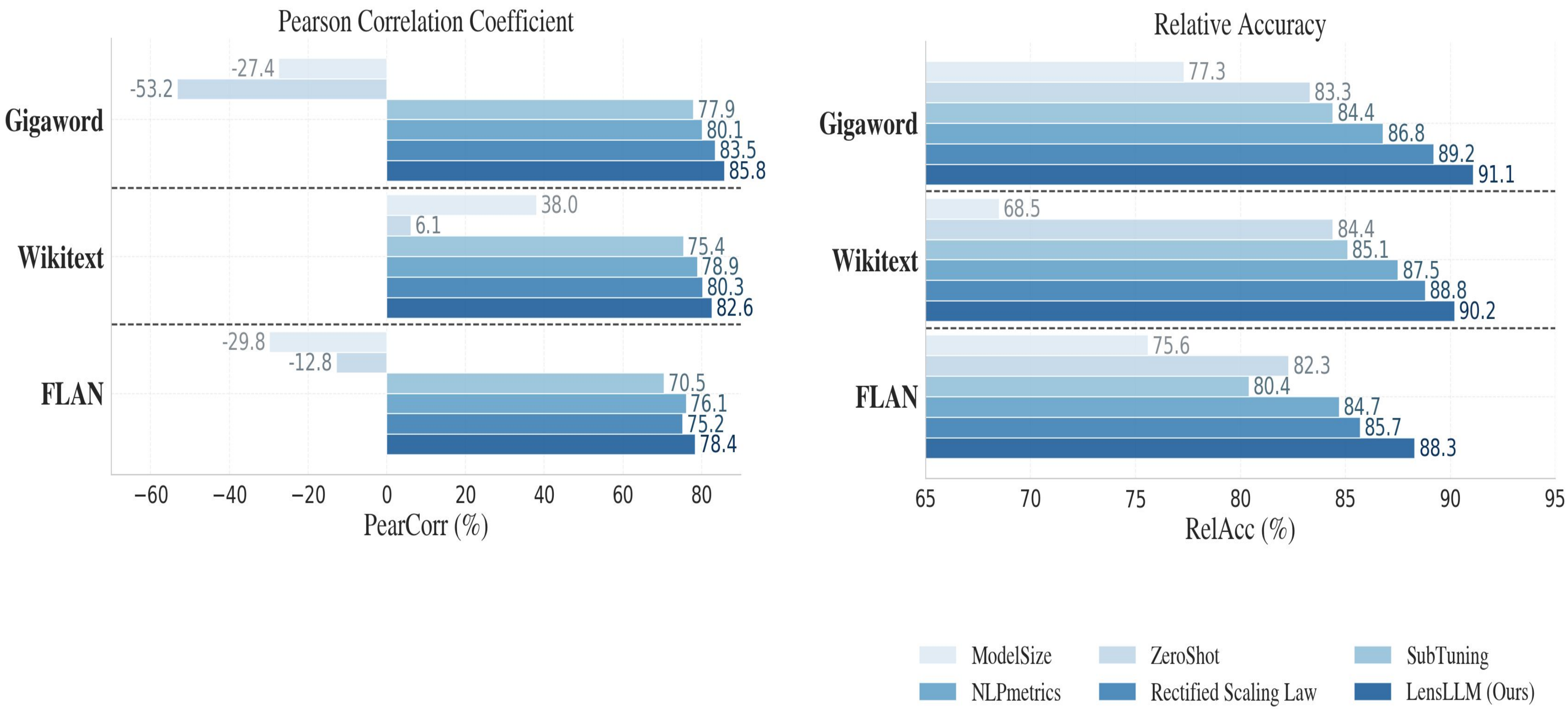
(d) B adjusts the initial test loss,

(e) E denotes the optimal loss of the model given an infinite amount of data.

VIRGINIA TECH

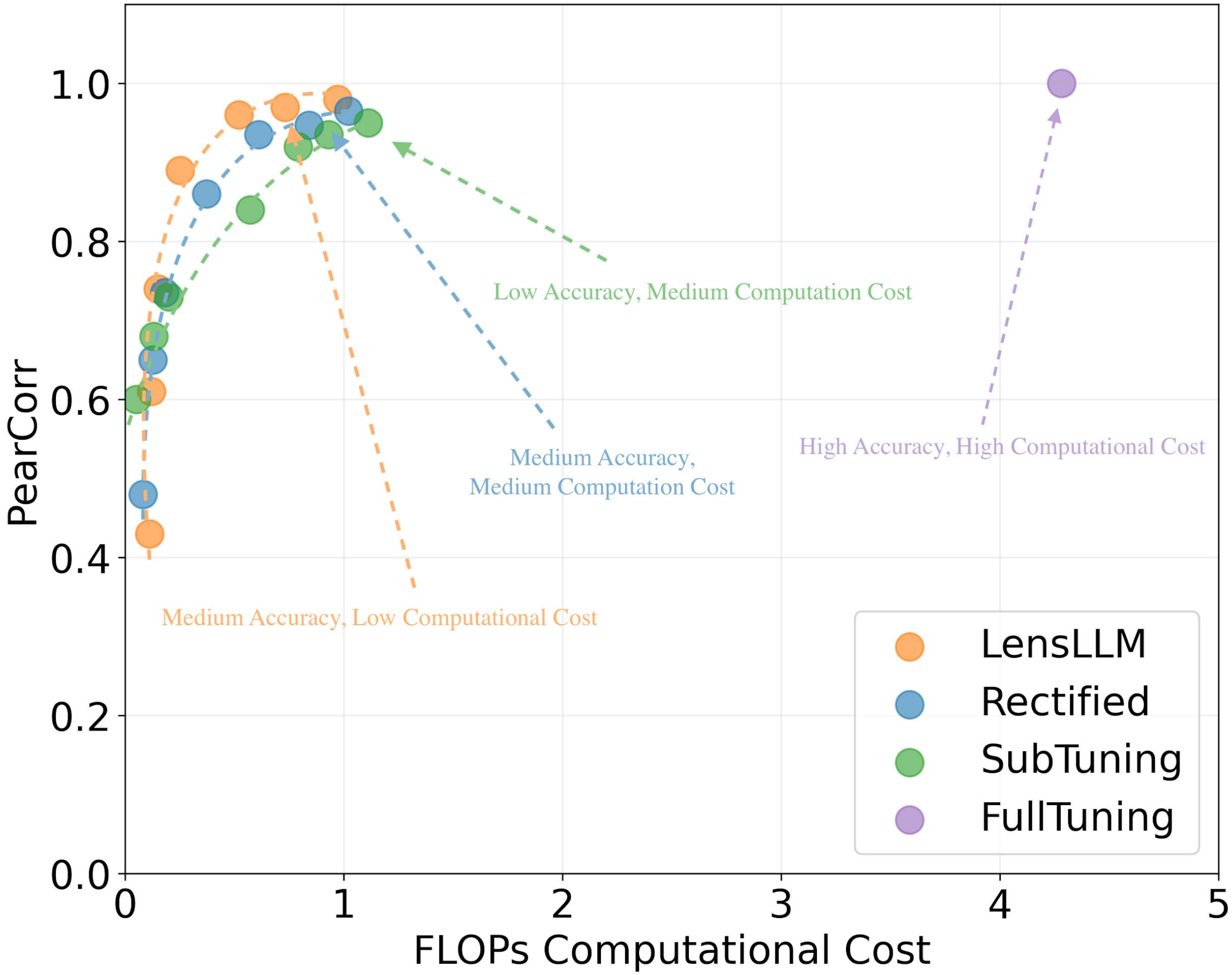
Experiments

More Accurate Model Selection



Experiments

More Efficient Model Selection



Summary

Key Contributions:

1. We propose a first-principled Hessian-based PAC-Bayes framework

Previous: Mostly focus on heuristic findings, and lack of theoretical foundation…

2. We introduce a more effective and efficient model LENSLLM

Better fit curving, Better model selection, Less computational cost…

3. Solid method and Comprehensive evaluation

7 diverse LLM family, Ablation study of Hyperparameters…

ICML
International Conference
On Machine Learning

Thank you!

Contact: Xinyue Zeng (xyzeng@vt.edu)

Paper

Code

Homepage

Our Group