Kaustubh Ponkshe

ponkshekaustubh11@gmail.com GitHub Scholar Webpage

Education

École Polytechnique Fédérale de Lausanne (EPFL), Switzerland *Sep '25-*

Ph.D. in Computer Science, Machine Learning and Optimization Lab (Advisor: Prof. Martin Jaggi)
– Contributor to Apertus: the biggest (70B) fully open and compliant training run & LLM to date
– My current research centers on rethinking training paradigms for foundation models across pretraining, midtraining, and post-training:
- Designing knowledge-free and retrieval-grounded pretraining to reduce parametric memorization and externalize memory in LLMs.
- Developing agentic pretraining methods that teach LLMs tool use and on-policy decision-making beyond passive next-token prediction.
- Understanding how midtraining and post-training data mixtures interact to shape reasoning, and generalization in language models.

Indian Institute of Technology Bombay (IITB), India *Jul '19 - Jul '24*

B.Tech in Electrical Engineering and M.Tech in Machine Learning + Artificial Intelligence; ***CPI:9.57/10***
Awarded Undergraduate Research Award 02 for one of the **best bachelor's thesis**

Selected Publications

** denotes equal contribution*

- **Project Apertus**; Apertus: Democratizing Open and Compliant LLMs for Global Language Environments [Paper] [HF]
  – *8B and 70B fully pretrained open-data open-weights models, multilingual in >1000 languages*
- **K. Ponkshe***, P. Prashant*, B. Salimi; TokenSwap: A Lightweight Method to Disrupt Memorized Sequences in LLMs; ***NeurIPS*** '25 - *Spotlight* 🏆 [Paper] [GitHub]
  – *Introduced a simple yet effective inference-time method to disrupt memorized patterns in LLMs without impacting overall utility.*
- **K. Ponkshe***, R. Singhal*, P. Vepakomma; FedEx-LoRA: Exact Aggregation for Federated and Efficient Fine-Tuning of Large Language Models; ***ACL Main*** '25 - *Oral* 🏆 [Paper] [GitHub]
  – *Achieved exact aggregation in distributed fine-tuning of LLMs without added costs, consistently improving over SOTA methods.*
- **K. Ponkshe***, S. Shah*, R. Singhal*, P. Vepakomma; Safety Subspaces are Not Distinct: A Fine-Tuning Case Study; ***ICLR** '26, Version accepted at INTERPLAY @ **CoLM** '25* [Paper] [GitHub]
  – *Demonstrated that safety alignment in LLMs is not confined to distinct subspaces, challenging basis of subspace-based defenses.*
- **K. Ponkshe***, R. Singhal*, R. Vartak*, P. Vepakomma; ABBA: Highly Expressive Hadamard Product Adaptation for LLMs; ***ICLR** '26, Version accepted at ES-FOMO @ **ICML** '25 -Spotlight* 🏆 [Paper] [GitHub]
  – *Enhanced update expressivity by parameterizing updates using a Hadamard product of low-rank adapters, achieving SOTA.*
- **K. Ponkshe***, R. Singhal*, R. Vartak, L. R. Varshney, P. Vepakomma; Fed-SB: A Silver Bullet for Comm. Eff. and Performance in (Private) Federated Fine-Tuning; ***TMLR** , J2C Certification (Top 10 % of accepted papers)* 🏆 [Paper] [GitHub]
  – *Established a new Pareto frontier for (private) distributed fine-tuning of LLMs, achieving SOTA performance, stronger privacy guarantees, and up to 230× lower communication costs.*
- **K. Ponkshe***, R. Singhal*, E. Gorbunov, A. Tumanov, S. Horvath, P. Vepakomma; Initialization using Update Apprx. is a Silver Bullet for Extremely Eff. Fine-Tuning; *SCOPE @ **ICLR** '25* [Paper] [GitHub]
  – *Achieved provably best approximation of full fine-tuning in low-rank spaces, outperforming LoRA with 90× fewer parameters.*

Selected Research & Work Experience

Massachusetts Institute of Technology & Mohamed bin Zayed University of AI | Researcher *Jun '24 - Jun '25*

Large Language Models, LLM Efficiency, AI Safety/Security, Post-Training, Large-Scale Distributed Learning

- Developed new techniques to make language model fine-tuning more efficient and effective, reducing parameter requirements by up to 90× without loss of performance on core arithmetic/commonsense reasoning and NLU benchmarks
- Designed and analyzed update schemes for distributed adaptation of LLMs, reducing communication overhead by up to 230× and enabling practical large-scale fine-tuning across a large number of clients, with a focus on privacy and reliability
- Investigated impact of supervised fine-tuning (SFT) on model safety/alignment, demonstrating limitations of current subspace-based defense methods and highlighting need for improved strategies to maintain safe behavior after adaptation

Entrepreneurs First | Founder in Residence *Jun '25 - Aug '25*

Startup incubator - Selected to join a curated community of top founders, collaborating to build globally impactful companies

- Exploring use of protein-language models to accelerate discovery of novel proteins for sustainable and nutritious food synthesis

Adobe Research | Research Collaborator *Aug '22 - Apr '23*

- Built a structure-aware corpus (100K arXiv docs) and pre-trained Longformer with global tokens; achieved **18%** gain over sparse-local baselines on SciREX IE by leveraging header/global attention patterns

Technical Skills & Awards

- **Languages / Libraries:** Python, PyTorch, Megatron-M, Nanotron, DeepSpeed, Transformers, PEFT, TRL, vLLM, Accelerate
- **Awards:** Regional Math Olympiad awardee (**State rank 3**) · All India rank **846** in JEE Advanced 2019 among 1.5M+ students · Graduated among **top 3** in the M.Tech (AI) cohort at IIT Bombay · Awarded the EDIC Fellowship at EPFL