AHMAD FARAZ KHAN

ahmadfk@vt.edu ♢ linkedin.com/in/ahmadfarazkhandurrani/ ♢ +1 540-4496457

EDUCATION

Ph.D. in Computer Science, Virginia Tech, Blacksburg, VA January 2021 - *present*
Research Focus: Machine Learning Systems

M.S. in Computer Science, Virginia Tech, Blacksburg, VA June 2024

B.S. in Computer Science, LUMS, Lahore, Pakistan May 2020

TECHNICAL PROFICIENCY

Programming Languages: Python, Javascript, C++.
Tools and Libraries: Pytorch, Tensorflow, Hugging Face, LangChain, Ollama, Pandas, SciPy, FLOWER, IBM Federated Learning, Spark MLlib, PySpark, Dask, Hadoop, DeepSpeed, MinIO, AWS Suite, Docker, OpenFaaS, SQL, Kubernetes

WORK EXPERIENCE

Graduate Research Assistant, DSSL, Virginia Tech Spring 2021 - *present*
Advisor: Dr. Ali Butt, Virginia Tech, Mentor: Dr. Ali Anwar, University of Minnesota

ML Algorithms and Optimization

- Devised a Direct Preference Optimization (DPO) approach for prompt optimization without separate reward modeling for Large Language Models (LLMs). **Enhanced score by 27%** compared to supervised fine-tuning.
- Created a DPO approach to mitigate sycophancy by fine-tuning LLMs on our curated dataset. **Reduced sycophancy by 64% in persona-based tests and 44% in preference-driven tests**.
- Designed an RLHF approach to fine-tune deep learning compression optimizations without sacrificing accuracy. Increased **resource utilization up to 81×**, **scalability by 78×**, and **accuracy up to 53%**.
- Developed clustering-based personalized learning solutions for distributed ML systems. Improved the **personalized accuracy by up to 45%**.
- Developed a reasoning-based Agentic AI-driven DevOps platform for adaptive online configuration of cloud systems, employing context-aware prompting for optimal resource efficiency and reduced human effort and cost.
- Built a benchmarking framework with 20000+ Python and Java programs for evaluating the capabilities of open and closed-source LLMs for semantic-preserving and semantic-altering mutated code understanding.

Impact: *Publications:* ***IPDPS'25, ACM EuroSys'24*** *and* ***IEEE BigData'24 (Best paper)****, submissions:* ***ACL'25*** *and* ***ICSE'26****.*

ML Infrastructure

- Created an adaptive aggregator server for collaborative learning with **one million+** nodes. Increased **scalability by 4×**, **latency by 8×**, and **cost reduction by 2×**.
- Developed a scheduler for collaborative learning that balances efficiency and accuracy tradeoff, improving **accuracy by 57%** and **reducing training time by 40%**.
- Designed an efficient, scalable, cost-effective cache with locality-aware execution for non-training workloads in distributed learning systems, decreased **average latency and cost by 71% and 98%** respectively.
- Improved secure AI systems by identifying and removing contributions from adversarial data sources, thereby enhancing accuracy through incentive-based systems. Raised the **accuracy by 7%**

Impact: *Publications:* ***MLSys'25, IEEE CLOUD'22, IEEE BigData'22 & 23, FL-AAAI'22.***

SELECT PUBLICATIONS

"FLStore: Efficient Federated Learning Storage for non-training workloads", **Ahmad Faraz Khan** et al. *8th Annual Conference on Machine Learning and Systems (MLSys 2025).*

"IP-FL: Incentive-driven Personalization in Federated Learning", **Ahmad Faraz Khan** et al. *39th IEEE International Parallel & Distributed Processing Symposium (IPDPS 2025).*

"FLOAT: Federated Learning Optimizations with Automated Tuning", **Ahmad Faraz Khan** et al. *19th ACM European Conference on Computer Systems (EuroSys 2024).*

"DynamicFL: Federated Learning with Dynamic Communication Resource Allocation", Qi Le1, Enmao Diao2, Xinran Wang, **Ahmad Faraz Khan** et al. ***Best Paper*** *in IEEE International Conference on Big Data (Best paper at BigData 2024).*

"Mitigating Sycophancy in Large Language Models via Direct Preference Optimization", Azal Ahmad Khan, Sayan Alam, Xinran Wang, **Ahmad Faraz Khan**, et al. *IEEE International Conference on Big Data (BigData 2024)*, pp. 1664–1671.

SERVICES

Reviewer for COLM 2025, USENIX ATC (2024), IEEE Transactions on SMC: Systems (2025), Springer Neural Processing Letters (2022 & 2023), IEEE TNSM (2024), and PeerJ Computer Science Journal (2024).

ADDITIONAL EXPERIENCES

Graduate Teaching Roles: Taught the Web/Cloud Development course (Summer 2024 & Fall 2023) and assisted with Advanced Operating Systems (Spring & Fall 2024), Python Programming (Spring 2020 & Fall 2021), and Computer Security (Spring 2022).