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, IBM Federated Learning, Spark MLlib, PySpark, Dask, Hadoop, DeepSpeed, MinIO, Selenium, AWS Suite, Docker, 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

- 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%**.
- 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**.
- Implemented a RAG-based AI-driven DevOps platform using LLM agents for adaptive online configuration of cloud systems, employing context-aware prompting for optimal resource efficiency and reduced human effort and cost.

***Impact:** Publications at **ACM EuroSys'24** and **IEEE BigData'25**, with current submissions at **OSDI'25** and **IPDPS'25**.*

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 at **IEEE CLOUD'22, IEEE BigData'22 & 23, FL-AAAI'22**, with current submissions at **FAST'25**.*

SELECT PUBLICATIONS

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

"Towards Cost-Effective and Resource-Aware Aggregation at Edge for Federated Learning", **Ahmad Faraz Khan** et al. *IEEE International Conference on Big Data (BigData 2023).*

"TIFF: Tokenized Incentive for Federated Learning", Jingoo Han, **Ahmad Faraz Khan** et al. *15th IEEE International Conference on Cloud Computing (CLOUD 2022).*

"Heterogeneity-Aware Adaptive Federated Learning Scheduling", Jingoo Han, **Ahmad Faraz Khan** et al. *IEEE International Conference on Big Data (BigData 2022).*

"Tokenized Incentive for Federated Learning", Jingoo Han, **Ahmad Faraz Khan** et al. *AAAI International Workshop on Trustable, Verifiable and Auditable Federated Learning (FL-AAAI 2022).*

SERVICES

External review committee for USENIX ATC (2024), reviewer for Springer Neural Processing Letters (2022 & 2023), IEEE Transactions on Network and Service Management Journal (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).