AHMAD FARAZ KHAN

ahmadfk@vt.edu ⋄ afkd98.github.io ⋄ linkedin.com/in/ahmadfarazkhandurrani/ ⋄ +1 540-449-6457

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

Research Intern, IBM Research, Almaden — May 2025 – *present*
Mentored by Dr. Taiga Nakamura and Dr. Swanand Ravindra Kadhe

- Designed a self-optimizing loop and domain-specific fine-grained synthetic data generation techniques guided by embedding-based similarity metrics.
- Enabled controlled distributional coverage, continual learning, and autonomous maintenance of foundational models, significantly improving robustness and performance in domain-specific tasks.

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 LLMs; **enhanced score by 27**% over supervised fine-tuning.
- Created a DPO pipeline to mitigate sycophancy, cutting it by **64%** in persona tests and **44%** in preference-driven tests.
- Implemented a RAG-based AI-driven DevOps platform using LLM agents for adaptive online cloud configuration, reducing human effort and cost.

Impact: *Publications at **IPDPS'25**, **ACM EuroSys'24** and **IEEE BigData'24 (Best paper)**, with current submission at **ACL'25**.*

ML Infrastructure:

- Created an adaptive aggregator server for collaborative learning with **one million+** nodes; improved **scalability by 4**×, **latency by 8**×, and **reduced cost by 2**×.
- Developed a scheduler balancing efficiency vs. accuracy; improved **accuracy by 57%** and **reduced training time by 40%**.
- Engineered a locality-aware cache for non-training workloads, decreasing **latency and cost by 71%** and **98%**, respectively.
- Improved secure AI systems by filtering adversarial contributions, raising **accuracy by 7%**.

Impact: *Publications at **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., *MLSys 2025.*

"IP-FL: Incentive-driven Personalization in Federated Learning", **Ahmad Faraz Khan** et al., *IPDPS 2025.*

"FLOAT: Federated Learning Optimizations with Automated Tuning", **Ahmad Faraz Khan** et al., *EuroSys 2024.*

"DynamicFL: Federated Learning with Dynamic Communication Resource Allocation", Qi Le, Enmao Diao, Xinran Wang, **Ahmad Faraz Khan** et al., *BigData 2024 (Best Paper).*

"Mitigating Sycophancy in LLMs via Direct Preference Optimization", Azal Ahmad Khan, Sayan Alam, Xinran Wang, **Ahmad Faraz Khan** et al., *BigData 2024.*

SERVICES

- Reviewer for COLM 2025; USENIX ATC 2024 (External Committee); Springer Neural Processing Letters 2022–23; IEEE TNSM 2024; PeerJ CS 2024.

ADDITIONAL EXPERIENCES

- **Graduate Teaching Roles:** Web/Cloud Development (Summer 2024 & Fall 2023); Advanced Operating Systems (Spring & Fall 2024); Python Programming (Spring 2020 & Fall 2021); Computer Security (Spring 2022).