AHMAD FARAZ KHAN

ahmadfk@vt.edu | afkd98.github.io | 73 Woodrow St, Daly City, CA 94014, **(Open to Relocation)**

EDUCATION

Ph.D. in Computer Science, Virginia Tech, Blacksburg, VA (**CGPA 3.8**) *December 2020 - May 2025*
Research Focus: Machine Learning Systems — Natural Language Processing

B.S. in Computer Science, LUMS, Lahore, Pakistan *2016-2020*
Advanced Courses: Distributed Systems, Deep Learning, Machine Learning, Cloud Computing, System Design

TECHNICAL PROFICIENCY

Programming Languages: Python, Javascript, C/C++, Java, Go.
Tools, Libraries: Pytorch, Tensorflow, PySpark, Spark MLlib, AWS Suite, Dask, Numba, Hadoop, Docker, Kubernetes, OpenFaaS, Selenium, MongoDB, ES6+, TypeScript, React, Node, Express, SQL, CUDA

WORK EXPERIENCE

Graduate Research Assistant, DSSL, Virginia Tech December 2020 - Present
Mentor: Dr. Ali Butt, PhD. Purdue University

Resource Constrained BigData Analytics

- Built a distributed system in *Pytorch* for resource-constrained privacy-aware analytics, focusing on enhancing resource utilization, scalability, and efficiency. Increased **resource utilization up to 81×**, **scalability by 78×**, and **accuracy up to 53%**.
- Designed a large-scale distributed analytics parameter server on *Hadoop Spark* to support up to **one million+** learning nodes with increased **scalability by 4×**, **latency by 8×**, and **cost reduction by 2×**.
- Designed a scheduler for distributed analytics systems in **Pytorch** to manage efficiency and accuracy tradeoff. Improved **accuracy by 57%** and **training time by 40%**.
- Designed an efficient, **infinitely scalable**, and cost-effective cache on *AWS Lambda, ElastiCache, SageMaker, and EC2* for non-training workloads in learning systems. Decreased **latency up to 99.9%** and **cost up to 99.6%**.

Impact: Publications: **BigData'22**, **IEEE Access'23**, **BigData'23**, and **EuroSys'24** — Submissions: **FAST'24**

Distributed Analytics Schedulers

- Improved distributed ML schedulers in *Pytorch* to identify and remove adversarial data sources to improve accuracy. Increased the **accuracy by 7%** by identifying and **mitigating 100% of malicious data sources**.

Impact: Publications in conferences including **IEEE CLOUD'22**, **FL-AAAI'22**.

Personalized foundational models — LLMs Fine-tuning

- Developed large-scale efficient clustering-based personalized learning solutions utilizing *Hugging Face and Pytorch* for distributed ML systems. Improved the **personalized accuracy by up to 45%**.
- Developed a Direct Preference Optimization (DPO) approach for prompt optimization without separate reward modeling for Large Language Models (LLMs). **Improved score by 27%** compared to supervised fine-tuning.
- Designed a RAG-based context-aware LLM framework, utilizing *Hugging Face and Pytorch*, to automate the adaptive online configuration of distributed cloud services. This enhances resource efficiency and minimizes human effort.
- Created a DPO approach utilizing *Hugging Face and Pytorch* 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**.

Impact: Submissions in conferences including **NeurIPS'24** and **ICML Workshop'24**.

PUBLICATIONS

– *FLOAT: Federated Learning Optimizations with Automated Tuning*. **Ahmad Faraz Khan**, Azal Ahmad Khan, Ahmed M. Abdelmoniem, Samuel Fountain, Ali R. Butt, and Ali Anwar. In Proceedings of the 19th ACM European Conference on Computer Systems **(EuroSys'24)**, Athens, Greece, 12 pages, April 2024. (AR: 16%).

– *Towards Cost-Effective and Resource-Aware Aggregation at Edge for Federated Learning*. **Ahmad Faraz Khan**, Yuze Li, Xinran Wang, Sabaat Haroon, Haider Ali, Yue Cheng, Ali R. Butt, and Ali Anwar. In Proceedings of the 2023 IEEE International Conference on Big Data **(BigData'23)**, Sorrento, Italy, 10 pages, December 2023. (AR: 17.49%).

– *A Survey on Attacks and Their Countermeasures in Deep Learning: Applications in Deep Neural Networks, Federated, Transfer, and Deep Reinforcement Learning*. Haider Ali, Dian Chen, Matthew Harrington, Nathaniel Salazar, Mohannad Al Ameedi, **Ahmad Faraz Khan**, Ali R. Butt, and Jin-Hee Cho. In **IEEE Access**: The Multidisciplinary Open Access Journal, vol. 11, pp. 120095-120130, October 2023. (AR: 30%).

– *TIFF: Tokenized Incentive for Federated Learning*. Jingoo Han, **Ahmad Faraz Khan**, Syed Zawad, Ali Anwar, Nathalie Baracaldo Angel, Yi Zhou, Feng Yan, and Ali R. Butt. In Proceedings of the IEEE International Conference on Cloud Computing **(CLOUD'22)**, Barcelona, Spain, 10 pages, July 2022. (AR: 22.4%).

– *Heterogeneity-Aware Adaptive Federated Learning Scheduling*. Jingoo Han, **Ahmad Faraz Khan**, Syed Zawad, Ali Anwar, Nathalie Baracaldo Angel, Yi Zhou, Feng Yan, and Ali R. Butt. In Proceedings of the IEEE International Conference on Big Data **(BigData'22)**, Osaka, Japan, 10 pages, December 2022. (AR: 19.2%).

– *Tokenized Incentive for Federated Learning*. Jingoo Han, **Ahmad Faraz Khan**, Syed Zawad, Ali Anwar, Nathalie Baracaldo Angel, Yi Zhou, Feng Yan, and Ali R. Butt. In Proceedings of the AAAI International Workshop on Trustable, Verifiable and Auditable Federated Learning **(FL-AAAI-22)** in conjunction with AAAI 2022, Vancouver, BC, Canada, 9 pages, March 2022.

SERVICES

- Served on the external review committee for USENIX ATC 2024.

- Reviewed for Neural Processing Letters 2022 & 2023.

ADDITIONAL EXPERIENCES

Teaching Roles, Virginia Tech: Instructed courses such as Web/Cloud Development (Summer'24 & Fall'23), Python Programming (Spring'20, Fall'21), and Principles of Computer Security (Spring'22).

Associate Data Engineer, i2c Inc. (May 2020 - December 2020): Spearheaded the development and upkeep of distributed sequential databases. Successfully accelerated query times for read-only tasks through database optimization techniques.