# Zeshan Fayyaz

647 551 5875 • zeshanf59@hotmail.com • zeshanfayyaz.com • zeshan-fayyaz • ZeshanFayyaz

## Education

**University of Waterloo** | Master of Mathematics, Computer Science — Sep 2022 - June 2025, Waterloo, ON
**Toronto Metropolitan University** | Bachelor of Engineering, Computer Engineering — Sep 2016 - April 2022, Toronto, ON

## Skills

**Programming Languages & Tools:** Python, SQL, C, Java, Bash, Git, LaTeX, Regex, REST APIs, FastAPI, NumPy, Pandas
**Machine Learning & AI:** Federated learning, RAG, MCP, LLMs, GRPO, Reinforcement Learning, PyTorch, TensorFlow, MLflow
**Data & DevOps Infrastructure:** PostgreSQL, FAISS, ETL pipelines, AWS EC2, Docker, GitHub Actions, Linux, Distributed Systems

## Work Experience

**AI (NLP) Engineer** — November 2025 – Present, Toronto, ON
*BoomerangFX*

- Built an NLP-powered RAG chatbot to provide domain-restricted, context-aware responses for healthcare and aesthetics clients, reducing hallucinations by 30%+ through rigorous prompt engineering and semantic grounding.
- Deployed production-ready FastAPI architecture for inference and response delivery while optimizing Azure cognitive search, embeddings, retrieval, and prompt-refinement pipelines to improve accuracy, conversation, latency, and reliability.

**Graduate Research Assistant** — Sep 2022 – June 2025, Waterloo, ON
*University of Waterloo*

- Designed a scalable distributed machine learning framework that used reinforcement learning to coordinate over 1,000 heterogeneous clients, reducing model convergence time by 24% and improving overall system stability under heterogeneous conditions.
- Improved privacy and computational efficiency by developing an adaptive encryption mechanism that optimized CKKS encryption levels during training, increasing convergence performance by 30% while preserving full security guarantees across all clients.

**ML Research Engineer** — Jan 2020 – Jan 2023, Toronto, ON
*Toronto Metropolitan University*

- Built a production grade image restoration model by implementing deep learning architectures in TensorFlow and BiGRU, improving PSNR by 22.9% and SSIM by 3.79% on large scale datasets and validating model robustness across multiple test suites.
- Built a reliable large scale training pipeline on AWS EC2 to handle over 100GB of image data, engineering dynamic batching, optimized data loaders, and parallel model tuning workflows that eliminated bottlenecks and reduced end to end runtime by 35%.

**Software Engineer Intern** — Jan 2022 – June 2022, Toronto, ON
*Royal Bank of Canada*

- Built Git-integrated features including credential caching, commit diff visualization, and unsaved change warnings that improved developer workflow reliability and reduced user-reported errors by 35%.
- Integrated CI/CD automation using Docker and GitHub Actions to streamline deployment workflows and cut setup time by 40%.

## Relevant Projects and Publications

*Author of 5 peer-reviewed papers with 650+ total citations on distributed machine learning, recommendation systems, and privacy-aware AI.*

**Knowledge Assistant: RAG + MCP Pipeline for Context-Grounded LLM Systems**

- Developed a retrieval augmented generation system using SentenceTransformers, FAISS, and Ollama, with an MCP powered FastAPI backend that supported structured tool use, efficient retrieval pipelines, and contextual reasoning over domain specific knowledge.

**GRPO Driven Optimization for Distributed LLM Workflows**

- Building a GRPO driven optimization pipeline that applies reinforcement learning to reduce straggler effects in distributed LLM workloads, with support for reward shaping, logging, evaluation, and PPO baseline comparisons.

**LLM Latency and Throughput Benchmarking Dashboard**

- Built a performance benchmarking dashboard that measures LLM latency, token throughput, and memory usage across local and API based models, using a FastAPI backend with SQL logging, parallel test execution, and real time visualization.

**Recommendation Systems: Algorithms, Challenges, Metrics, and Business Opportunities** | *Applied Sciences, 2020*

- Conducted a comprehensive technical analysis of modern recommendation algorithms, challenges, evaluation metrics, and real world deployment tradeoffs in a publication with 500 plus citations and over 30,000 downloads.

**HERL: Tiered Federated Learning with Adaptive Homomorphic Encryption using Reinforcement Learning** | *TPS 2025*

- Developed a tiered federated learning framework that used adaptive encryption, reinforcement learning, and clustering techniques to manage stragglers and heterogeneous clients, increasing model accuracy by 20% and reducing communication overhead by 30%.