Shahmeer Ali

647-829-8506 | shvhmeer@gmail.com | LinkedIn | GitHub | shahmeerali.com

Education

University of Waterloo — Waterloo, ON
B.A.Sc, Systems Design Engineering, ***3.90 GPA*** *- Concentration in Artificial intelligence* — *Sept 2022 - March 2027*

Technical Skills

Languages: Python, SQL, C++, Scala, Java, R
Technologies: PyTorch, TensorFlow, scikit-learn, Pandas, LangChain, Spark, Hadoop, PostgreSQL, FastAPI
Tools: AWS, Google Cloud, Codex, Claude Code, Cursor, Docker, Kubernetes, Linux, Git, CI/CD

Experience

Microsoft — Redmond, WA
Applied Scientist, Security Research — *May–Aug 2026*

- Developed an **LLM tagging agent** for impersonation emails using NLP intent models and a configurable taxonomy; **0.87 F1** against analyst labels
- Extended a synthetic-data curation agent to impersonation attacks using tagger outputs, adding **40K examples**
- Evaluated synthetic data with topic models/NLP classifiers: **92% real-cluster coverage**, 85% filter pass rate
- Built held-out real-email benchmarks; synthetic data raised impersonation recall from **69% to 81%** at fixed FPR

Remmie Health — Seattle, WA
Machine Learning Researcher — *Sept–Dec 2025*

- Quantized CNN diagnostic models to **INT8 for edge inferencing**, cutting **latency 60%** with < 1 pt accuracy loss
- Benchmarked YOLO, ResNet, MobileNet on **patient-separated holdouts**; selected at **0.91 ROC-AUC**

BMC Software — Santa Clara, CA
AI Engineer — *Jan–May 2025*

- Shipped **metadata-conditioned schema generation** with LangChain, Pydantic, and FastAPI, replacing **95% of manual schema work**
- Cut **hallucinated schema fields 35%** with embedding-based prompt normalization, evaluated on a labelled 1200-example test suite

BMC Software — Santa Clara, CA
Machine Learning Engineer — *May–Aug 2024*

- Built a **PyTorch/FAISS RAG system** serving 1K+ daily queries, cutting end-to-end response time 20%
- Cut DistilBERT **fine-tuning time 75%** on 2+ TB of text with **distributed multi-GPU training** on AWS
- Optimized embedding inference for production retrieval to < 40 **ms p99** via ONNX model compilation and batching
- Automated model deployment with **Docker and CI/CD**, cutting release time 40%

Royal Bank of Canada — Toronto, ON
Data Scientist, Risk Management — *Sept–Dec 2023*

- Optimized Spark/Hadoop jobs over **4+ TB of credit-risk data**, cutting processing time 30%
- Built an **outlier-detection model** flagging anomalous records in RBC's IFRS 9 regulatory reporting database
- Validated scikit-learn/TensorFlow credit-risk models on **temporal holdouts**, achieving **0.84 ROC-AUC** with $< 2\%$ calibration error

Royal Bank of Canada — Toronto, ON
Data Engineer, Risk Management — *Jan–Apr 2023*

- Implemented **50+ ETL pipelines** with **Apache Spark & Scala**, improving data transfer efficiency by 60%
- Automated 30 financial workflows with **Python and Jenkins**, reducing manual oversight by 90%

Projects

Undistracted- ML for Distracted Driving *Python, OpenCV, PyTorch, TensorFlow, Keras* — 2021-2023

- Built an ML model using CNN's for distracted driving, enhancing safety through **real-time image detection**
- Trained deep learning models on **100,000+ images** enabling high-precision driver behavior analysis