

Mohsin Qureshi

mohsinobaid@gmail.com | mohsinobaid.com | github.com/mohsinobaid | 613-869-8040

AI Engineer with 9+ years of experience across software engineering, machine learning, and data infrastructure. Currently building production multi-agent systems, LLM pipelines, and real-time AI features for enterprise clients, with full-stack delivery across Python, FastAPI, TypeScript, and cloud platforms. Proven track record delivering end-to-end AI solutions across financial services, telecommunications, and voice AI platforms.

Experience

Forward Deployed Engineer AI, Lazer Technologies – Toronto, ON Feb 2026 – Present

- **Avoca AI (Voice AI platform):** Built and shipped AI-powered features across a production voice AI platform, engineering multi-agent tool-calling systems, LLM prompt architectures, and real-time AI infrastructure handling thousands of inbound calls per day across three enterprise home services clients.
 - Built 6 AI agent tool-calling handlers integrated with enterprise CRMs (Salesforce, ServiceTitan) via VAPI and ElevenLabs — enabling autonomous record creation, contact management, case triage with AI-inferred priority, and live customer resolution mid-conversation. 80% of calls resolved without human handoff.
 - Owned testing, staged rollout, and Datadog monitoring for agent features in production.
 - Scoping pilots with client stakeholders, demos to client leadership, and integrating with the client's live systems.
- **AIQ Consultants (Fixed Income Analytics):** Designed and built a production multi-agent system of 5 specialized agents powered by Claude via the Anthropic Agent SDK, with silent handoffs, tool-calling orchestration, and shared compliance guardrails.
 - Built and deployed an MCP (Model Context Protocol) server exposing 22 domain-specific tools to external LLM clients, with session-scoped authentication and per-user rate limiting.
 - Built retrieval-augmented generation (RAG) over bond prospectuses and issuer filings using pgvector and OpenAI embeddings, grounding agent answers in source documents with citations.
 - Built end-to-end AI-powered analytics tools on a fixed income platform: liquidation speed analysis, VWAP analytics by CUSIP, and a Gemini-powered bond sector classification pipeline processing issuers at scale.
 - Shipped production AI features end-to-end: agent tool definitions, FastAPI/Python endpoints, Snowflake query layers, Neo4j graph queries, and interactive React/TypeScript UIs.
- **Stack:** Python, FastAPI, React, TypeScript, Next.js, Claude, Gemini, VAPI, ElevenLabs, MCP, Snowflake, Neo4j, PostgreSQL, Supabase, AWS ECS, AWS Lambda, EventBridge, S3, SSE, WebSockets, Pydantic, Inngest, FINRA TRACE, MarketAxess, Salesforce, ServiceTitan, Twilio, Terraform, Docker, Vercel, Datadog

Senior Software Engineer, Denologix – Toronto, ON Jan 2024 – Jan 2026

- **Shared Services Canada:** Developed cloud applications and ETL pipelines
 - Expanded an enterprise application and created reporting automation tools using Azure services. Improved overall reporting efficiency by 40%
 - Constructed ETL pipelines for new data ingestion using Data Factory and SQL Server
- **Bell Canada:** Developed the Kiosk Service Portal utilizing AWS services
 - Led the migration of the Kiosk Service Portal to a hybrid cloud environment, cutting infrastructure costs by 50% and retiring legacy components.
 - Managed AWS accounts, set up and configured EC2 instances, databases and Elastic Load Balancers
- **Stack:** EC2, PySpark, RDS, Redshift, S3, Lambda, Glue, JavaScript, Python, Databricks, SageMaker

Data Scientist, Scotiabank – Toronto, ON Jan 2020 – Jan 2023

- Contributed to creating a Machine Learning Supervised Human Trafficking model (Scikit-Learn, PyTorch) and data pipelines that identify customers who are trafficking, helping the Investigation Team with their work.
- Built developer tooling (Python, Dash, PostgreSQL, PySpark), enabling data scientists to deploy models faster,

reducing investigation time by 33%.

- Optimized PostgreSQL database performance and implemented caching strategies, reducing query latency by 60% for ML inference systems.
- Engineered production ML infrastructure and data pipelines supporting anti-money laundering models processing millions of daily transactions. (Python, PySpark, and SQL)
- Architected horizontally scalable ML serving platform using Kafka, Hive, and PySpark, increasing system capacity 10x to handle distributed workloads.
- Developed a Big Data Analysis tool to streamline and speed up data investigations of very large datasets (250GB)
- Decreased False-Positive testing in scenarios through threshold tuning and enriched data sources by identifying faults in Data Quality through comprehensive integrity analysis

Software Engineer II, Scotiabank – Toronto, ON

July 2016 – Dec 2019

- Developed Volcker regulation reports generation using Scala, Spark, and SparkSQL.
- Expanded and improved a full-stack risk reporting application by implementing new database manipulation/lookup features along with their front-end components, leading to an improved business user experience. (J2EE, Oracle, Spring, JavaScript).
- Improved the risk calculation process by integrating new source systems and business requirements, which enriched data and increased accuracy. Talend ETL and Java helper classes were used to transform, map, and load data from various source systems.

Software Developer Intern, BlackBerry – Toronto, ON

Sept 2013 – Aug 2014

Education

Toronto Metropolitan University– Master's in Computer Engineering (Machine Learning and Big Data) April 2022

Research Area: Deep Learning -The Role of Mutual Information in Deep Unsupervised Representation Learning.

(Skills: Deep Learning, Gen AI, CNN, Transformers, LLMs, Constructing Data Pipelines, TensorBoard visualizations, Hyperparameter Tuning, Data Parallelism, Utilizing SOTA codebases, Research writing, Presenting. DataSets: CIFAR10, NTU-R)

Carleton University – Bachelor's in Computer Engineering

April 2016

Personal Projects

- **Vision-language model** for autonomous-driving visual question answering on DriveLM-nuScenes under a strict 50M-parameter budget (PyTorch)
- **Data Science Analysis** on 350GB of Reddit Data by running the Latent Dirichlet allocation algorithm. (Spark, PySpark and SparkSQL) 15 meaningful latent dimensions were found from a set of 30 subreddits.
- **Machine Learning, Deep Learning:**
 - Computer Vision: Trained a Generative Adversarial Network on the Stanford cars dataset to construct new images.
 - Experimenting with various image augmentations in Contrastive Learning. (SimCLR model)
 - Exploring Mutual Information Regularization in Deep Learning models

Skills

Languages: Java, Python, JavaScript, SQL

Frameworks/Tools: React, Typescript, Next.js, Spring, Docker, Hadoop, Kubernetes, Spark, PySpark, Kafka, Django, FastAPI, Dash, TensorFlow, PyTorch, LLMs, HuggingFace, Unix, Git, Airflow, Jenkins, CI/CD

Databases: Oracle, Postgres, MySQL, NoSQL, Redshift, S3, Snowflake, Neo4j, Supabase