Jay Jahanzad

Passionate ML Engineer focused on bringing ideas to life with experience in AI/ML, full-stack, and mobile development.
jjahanzad@gmail.com • **https://soaapp.github.io** • **https://github.com/soaapp** • **https://linkedin.com/in/jayjahanzad**

Relevant Experience

Machine Learning Engineer - Team Lead, **CIBC** July 2023 - Present

- Led development of production-grade ML solutions including **RAG systems** with custom chunking strategies, advanced prompt engineering, and architecture design using open-source models (LLaMA 3.2, LLaMA 4, Mistral, Granite, Gemma); benchmarked and optimized embedding models (allMiniLM-v6) and retrieval strategies including MMR for enterprise applications
- **Deployed and scaled LLM infrastructure** on OpenShift using vLLM and RHOAI, building high-performance inference pipelines with vector store integration (Qdrant); implemented MLOps workflows for continuous model deployment and monitoring
- **Fine-tuned large language models** on proprietary financial data using LoRA techniques, optimizing models for fintech enterprise applications with domain-specific performance improvements
- Architected and implemented **data pipelines** leveraging Azure Document Intelligence and Docling with HuggingFace integration for document processing, extraction, and preprocessing at scale
- **Pioneered QA frameworks** for RAG system validation using DeepEval and RAGAS, establishing comprehensive testing pipelines for retrieval accuracy, generation quality, and end-to-end performance metrics
- Developed and tested **multi-agent systems** including supervisor agents, SQL database agents, and RAG agents using LangChain/LangGraph, implementing complex orchestration workflows for enterprise automation
- Translated legacy COBOL systems into modern Angular frontends and business logic using LLM-based **modernization tooling**; designed and executed transformation workflows spanning backend services and UI layers

Senior Application Developer, **CIBC** April 2019 - July 2023

- Developed enterprise-scale event-driven integration platforms connecting critical banking applications using Java, SpringBoot, and Quarkus with emphasis on secure token/certificate management and API design patterns
- Engineered Solace PubSub+ messaging solutions for real-time event streaming between distributed systems, ensuring reliable message delivery and system resilience across the bank's infrastructure
- Built and deployed full-stack API Marketplace platform with Java backend services and React/Redux frontend, enabling secure REST API discovery and consumption across the organization; implemented CI/CD pipelines using Jenkins and GitHub Actions

iOS Engineer, **Train Fitness Startup** Jan 2024 - July 2024

- Contributed as one of 5 iOS engineers to an AI-powered fitness app for iOS and watchOS, successfully launched to the App Store with revenue generation; delivered key features including Heart Rate Zones and Muscle Recovery tracking
- Developed using Swift, SwiftUI, and Combine with advanced MVVM architecture, implementing complex network services including third-party authentication, async operations, REST API integration, and local persistence with Core Data and UserDefaults
- Collaborated closely with design team to build and maintain the app's design system using Atomic Design Pattern, including comprehensive Dark Mode support, Typography system, and Asset Management
- Established robust testing practices including unit tests, domain logic tests, and view layer testing to ensure app reliability and performance

Technical Specialist, **Apple Inc** June 2017 - May 2019

- Worked at the Genius Bar to help diagnose and solve issues with iOS, macOS, and watchOS issues
- Hardware fixes on devices and tests done to ensure Apple device was functioning properly

Technical Expertise

Current AI/ML Tech Stack

Machine Learning & Generative AI:

RAG Systems, LangChain, LangGraph, PyTorch, Fine-tuning (LoRA), Prompt Engineering, Multi-agent Systems, Model Deployment (vLLM, RHOAI), MLOps, Vector Databases (Qdrant), Embedding Models (allMiniLM-v6), DeepEval, RAGAS
LLM Models: LLaMA (3.2, 4), Mistral, Granite, Gemma, GPT
Data & Infrastructure: Python, Azure Document Intelligence, Docling, HuggingFace, OpenShift, Databricks, MongoDB
Development Tools: Git, Docker, Jenkins, GitHub Actions

Professional Experience With

Backend & Cloud: Java, SpringBoot, Quarkus, Azure, AWS, OpenShift, REST
Frontend & Mobile: Swift, SwiftUI, Combine, React, JavaScript, React Native
Databases & Storage: PostgreSQL, Firebase/Firestore, Core Data
Design & Collaboration: Figma, Adobe Suite

Projects

Audr

Swift, SwiftUI, CreateML, and CoreML based iOS app that records doctor/dentist and patient interactions during in-person appointments, diarizes audio, trains baseline model as well as connects to varied clinic softwares and fills out patient charting and diagnosis

Resumai

Fullstack AI/ML based app that analyzes large numbers of resumes, outputs best candidates, and creates custom interview content

Orden

React, Prisma, Cypress stack for web application that is designed to be a marketplace for loaners + borrowers to find optimal loan rates

NEO-ML

NASA Open Source contribution using NEOs API data to fine-tune open-source HuggingFace model to measure NEO data + create predictions

Education

Toronto Metropolitan University

Bachelor's of Computer Science with Honours
Undergraduate Thesis in Computer Vision + ML