Akhil Shekkari

Portfolio | Certification | Email | GitHub | Linkedin | +1 (425) 426-8292

Master's student in **Applied Machine Learning** with a strong foundation in AI research and model development. While working as a Software Developer, my primary focus has always been on AI - fine-tuning transformers, optimizing LLMs, and deploying scalable ML systems. My research interests lie in **Reinforcement Learning** (RL), building true **personal AI assistants with memory** and adaptation, and advancing **Explainable AI** (XAI) for transparent decision-making.

Skills

Programming & Libraries: Python, PyTorch, Hugging Face Transformers, Vector databases, Langchain, Scikit-Learn, SQL
Optimization & Scaling: Deep Speed, Flash Attention, Vector Search, Distributed Training
Cloud and Specializations: AWS Sage Maker, Azure ML, Snowflake, Docker, GitHub, GenAI, RAG, NLP, Reinforcement Learning

---

Experience

Software Developer (Tezo) — July 2022 - July 2024

As I transitioned into a Software Developer role, my focus was more towards problem solving, fine-tuning, optimizing, and deploying domain-specific Generative AI models for internal enterprise applications. Developed a domain-specific AI assistant, enabling teams to query internal knowledge bases like project documents using LLM-powered search.

→ Integrated **LoRA for efficient fine-tuning**, reducing trainable **parameters by 90%**, enabling large transformers to be adapted to domain-specific data with minimal computational cost.
→ Adopted **Mixed precision training** with data parallelism techniques in early, later integrating Distributed Training reducing compute overhead and improving training speed by 2x while maintaining model accuracy.

→ Implemented a **CI/CD pipeline** for continuous model updates, enabling LLMs to incrementally learn from new domain knowledge, ensuring AI-generated insights remained accurate and up to date.

Junior Software Developer (Tezo) — July 2021 – July 2022

During my time as a Junior Software Developer, I worked on AI-powered data intelligence solutions, including the development of a Text2SQL model that enabled the sales team to query Snowflake data using natural language.

→ Built and fine-tuned a **Text2SQL model** with **85% query accuracy**. on internal Snowflake schemas, allowing sales team to retrieve insights without writing SQL, improving data accessibility by 60%.
→ **Optimized response time by 40%** by implementing efficient caching and query optimization techniques, ensuring minimal latency for real-time data retrieval and integrated the model into an internal chatbot interface, enabling sales teams to ask natural language.

Personal Projects and Interests

Research & AI Implementation

→ Implemented SOTA research papers from scratch in **PyTorch**, including Transformers, **Flash Attention.**

→ Exploring **Reinforcement Learning (RL) techniques**, focusing on training AI agents for autonomous reasoning, decision-making, and self-improving LLM interactions.

Developed and Deployed Multiple AI-driven applications leveraging LLMs.

AI-Powered Resume Analyzer — Project link

→ Developed an AI-driven Resume Analyzer using OpenAI embeddings & cosine similarity to assess resume-job description alignment, improving job application targeting.
→ Implemented **personalized resume feedback**, leveraging LLMs for content optimization, and deployed it on Streamlit Cloud.

LLM-Driven Customer Support System — Project link

→ Built an AI-powered customer support system using Large Language Models (LLMs) for ticket categorization, sentiment analysis, and response automation, reducing manual workload.
→ Integrated Retrieval-Augmented Generation (RAG) with **FAISS** & OpenAI embeddings to generate context-aware, accurate responses, improving resolution efficiency.

Multimodal RAG for Personalized Food Recommendations — Project link

→ Developed a **Multimodal Retrieval-Augmented Generation (RAG) system** using **AWS Bedrock, FAISS, and Claude-Sonnet** that combines **text & image data** to generate personalized food recommendations for a restaurant app.