Akhil Shekkari

Portfolio | *Email* | *GitHub* | *LinkedIn* | +1 (425) 426-8292

I am building AI systems that work, explain themselves, and are safe. I am deeply focused on making AI less of a **black-box** and more of a tool we can **trust**. My work revolves around **training deep learning models**, making them faster, smarter, and more **aligned** with human goals. I care about how **models reason**. My passion is to build powerful AI systems that are **transparent** and understandable.

Education

University of Maryland, College Park   August 2026 (Expected)
Master's in Applied Machine Learning   GPA: 3.9 / 4.0

Experience

Software Developer (ML Domain), Tezo   July 2022 – July 2024

- Developed a domain-specific **AI assistant** to help internal teams retrieve knowledge from project documents and support systems using LLM-powered search, reducing information lookup times and improving team efficiency.
- Fine-tuned large language models using **LoRA**, reducing trainable parameters by **90%** and enabling fast, low-cost adaptation of base models to internal company data without retraining the full model.
- Accelerated fine-tuning pipelines by adopting **mixed precision training** and **distributed training**, cutting model update times by **2x** and supporting faster iteration on new domain knowledge.
- Containerized ML training and inference pipelines using **Docker**, ensuring reproducibility and consistent behavior across development and production environments.
- Built and automated CI/CD pipelines integrating **GitHub Actions and Kubernetes**, streamlining model retraining, adapter deployment, and service rollouts with minimal downtime.

Junior Software Developer(Data Analytics Team), Tezo   July 2021 – July 2022

- Collaborated with the data analytics team, focusing on data modeling and visualization using **Snowflake** and **Power BI** to support business intelligence initiatives.

Independent Research and Publications

- **Technical Blogs** ⬈**:** Published articles as a ***Towards AI*** contributor, covering GPT internals, QLoRA fine-tuning, and topics related to memory footprint.
- **YouTube Channel** ⬈**:** Creating educational content focused on applied AI engineering practices.
- **Mechanistic Interpretability and AI Alignment (Ongoing):** Studying transformer circuits, sparse autoencoders, and scalable oversight techniques through the ARENA curriculum and independent projects, aiming to advance safe and transparent AI systems.

Personal Projects

- **AI-Powered Resume Analyzer:** Built an LLM system for resume-job alignment using OpenAI embeddings & cosine similarity.
- **Multimodal Food Recommendation:** Designed a text-image retrieval system using AWS Bedrock & Claude-Sonnet.
- **AI-Powered Code Reviewer (In Progress):** Developing a large-scale code review assistant based on research papers on pretraining and distillation for automating software code review activities.

Skills

- **Programming & Libraries:** Python, PyTorch, Hugging Face Transformers, Vector Databases, LangChain, Scikit-Learn, SQL
- **Optimization & Scaling:** DeepSpeed, Flash Attention, Vector Search, Distributed Training
- **Cloud & Specializations:** AWS SageMaker, Azure ML, Snowflake, Docker, GitHub, GenAI, RAG, NLP, Reinforcement Learning