Akhil Shekkari

Portfolio | *GitHub* | *LinkedIn* | *shekkari.akhil@gmail.com* | +1 (425) 426-8292

Professional Summary

AI engineer with over 3+ years of industry experience specializing in building production-grade AI systems that are fast, smart, and aligned with human goals. Proven track record of **training, optimizing, and deploying scalable AI systems**, while ensuring they adhere to ethical guidelines and are aligned with real-world applications.

Technical Skills

Programming & Tools: Python, PyTorch, SQL, ClearML, Docker, Kubernetes, GitHub actions, Llama parse, Llama index, Lang-chain, Lang-graph, Scikit-learn.
Cloud & Infrastructure: Snowflake, AWS SageMaker, Azure ML, CI/CD.
Core Competencies: Deep learning, Generative models, Machine learning, NLP, RAG, Agentic workflows, Model evaluation, Reinforcement learning.

Work Experience

AI Engineer | Atrium | Silver Spring, USA June 2025 – August 2025

- Collaborated with **Pfizer's** AI team to automate key components of **Statistical Analysis Plan** generation using a **RAG pipeline**, reducing statisticians' drafting time by **up to 60%**.
- Architected a custom **LLM-as-a-Judge** system combining prompting and structured rubric evaluation to detect hallucinations and guideline deviations in generated SAPs with over **80% precision**.
- Designed **Prompt Evaluation** Experiments by testing variations in temperature, system prompts, and input formatting resulting in a **20%** improvement in generation consistency for critical SAP tasks.

Software Developer(ML Domain) | Tezo | India July 2022 – July 2024

- Built a **RAG-powered chatbot** integrated with **SharePoint project repositories**, providing employees a unified interface to query across **10k+** internal documents including project charters, design specs, and status reports.
- Engineered a **semantic search pipeline with embeddings + vector DB**, reducing document lookup time by **60%** and improving cross-team visibility into ongoing projects.
- Extended capabilities with **LLM-based summarization** of lengthy project documentation, cutting manual review time for managers by **35%** and enabling quicker decision-making.
- Containerized and deployed the system with **Docker + CI/CD**, ensuring smooth secure on-prem access control.

Junior Software Developer(ML Domain) | Tezo | India July 2021 – July 2022

- Migrated policy and claims data to **Snowflake**, streamlining ETL pipelines and reducing average query times by **40%** compared to the legacy warehouse.
- Trained **ML models** to score policies and detect fraudulent claims, improving recall of fraud cases by **15%** and enabling earlier intervention to prevent financial losses.
- Built **SQL and BI dashboards** for stakeholders, cutting manual report preparation time by **25%**, increasing adoption to **50+ active users**, and improving decision-making speed across actuarial and operations teams.

Projects

Code Reviewer April 2025

- **Leveraged Distributed Training** to fine-tune(using LoRA) a **220M parameter Microsoft model** on a **150k** sample dataset, achieving realistic edit suggestions and serving the model via **containerized inference APIs**.
- Performed model distillation to reduce size to **80M parameters**, cutting inference cost by **60%** while maintaining high-quality review output in a staged deployment.
- Used **ClearML** for end-to-end experiment tracking, dataset and model versioning, and checkpoint management during training and deployment cycles.

Resume Analyzer June 2025

- Built an **AI-powered Resume Analyzer** using **OpenAI embeddings** and similarity scoring, improving candidate-job alignment accuracy by **20%** in benchmark evaluations.
- Delivered **actionable feedback with targeted improvement suggestions**, increasing resume shortlist scores by **15%** in trial runs.
- Analyzed **100+ resumes of peers and friends**, validating the system across diverse roles and collecting feedback to refine recommendations.
- Deployed on **Azure ML** for scalable and secure processing, ensuring reproducibility and enterprise-grade reliability.

Technical Blogs

- **Contributor at *Towards AI***, writing on GPT internals, fine-tuning, and memory-efficient model engineering.
- Implemented research papers such as Distillation and Microsoft Code Reviewer.

Education

University of Maryland, College Park Expected May 2026
Master of Science in Applied Machine Learning
Relevant coursework: Probability & Statistics; Advanced Machine Learning; Deep Learning; NLP; Optimization.