Akhil Shekkari

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

Summary

AI engineer with 3+ years of industry experience shipping RAG pipelines, LLM evaluation systems, and ML models to production. Deep hands-on expertise in **LLM post-training** (RLHF, GRPO), **agentic systems** (tool use, multi-step reasoning, MCP), and **inference optimization** (KV-cache, paged attention, Triton kernels).

Experience

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

- Automated **Statistical Analysis Plan (SAP)** generation from clinical trial documents for Pfizer's team, reducing drafting time by **up to 60%**.
- Evaluated knowledge graphs, naive chunking, and semantic chunking strategies; benchmarked **embedding models** (OpenAI, BiomedBERT, Cohere) and **hybrid retrieval** (dense + BM25 sparse), and selected the best-performing RAG configuration on internal test sets.
- Built a custom **LLM-as-a-Judge** evaluation system with structured rubrics scoring completeness, guideline adherence, and hallucination detection against ground truth SAPs, achieving **82% precision** and **78% recall**.
- Ran **50+ prompt experiments** (temperature, system prompts, few-shot examples, input formatting), improving generation consistency by **20%** and reducing manual revision cycles by **30%**.

Software Developer | Tezo | India July 2021 to July 2024

- Built a **RAG-powered chatbot** over SharePoint repositories, giving **200+ employees** a unified search interface across **10k+ internal documents** with sub-second retrieval.
- Engineered a **semantic search pipeline** (embeddings + vector DB) reducing document lookup time by **60%**, and added **LLM-based summarization** cutting manager review time by **35%**.
- Trained **ML models** for insurance policy scoring and fraud detection, improving fraud recall by **15%** and enabling earlier intervention across the claims pipeline.
- Containerized and deployed all services with **Docker + CI/CD** on secure on-prem infrastructure with **99.5% uptime**.

Projects

AI Agent Framework GitHub 2025

- Built a **multi-step reasoning agent** from scratch: agentic loop with function calling, **MCP integration** for external tool servers, Pydantic-validated structured outputs, and sliding-window memory with compaction.
- Implemented **5+ tools** (web search, code execution, multi-format file handling) with callback-driven observability and session persistence via FastAPI.
- **LoRA fine-tuned Qwen 3B** for chat and function calling using **QLoRA** (4-bit NormalFloat), then replaced the OpenAI backbone with the fine-tuned model as the agent's reasoning engine.
- Evaluated on the **GAIA benchmark** (GPT-4o vs Claude vs fine-tuned Qwen), achieving competitive scores on Level 1 to 3 multi-step tasks.

Building a Reasoning LLM GitHub 2025

- Implemented the full **post-training pipeline** to turn a base LLM (Qwen-2.5) into a reasoning model: inference-time scaling, self-refinement, and **GRPO reinforcement learning** from raw PyTorch (no TRL).
- Scaled training with **FSDP** and **gradient accumulation** across multiple GPUs; implemented **data, tensor, and pipeline parallelism** strategies for efficient rollout generation and policy updates.
- Built custom inference engine with **KV-cache** and sampling, plus a math evaluation harness with **SymPy symbolic equivalence** on GSM8K and MATH benchmarks. Implemented **chain-of-thought, self-consistency**, and self-refinement.

Mini Inference Engine GitHub 2025

- Built a modular LLM inference engine with **naive, streaming (online softmax), and Triton-fused attention** kernels, benchmarked across 4K to 16K token prefill and decode workloads.
- Designed **paged KV-cache** (on-demand allocation, zero fragmentation) achieving **O(T) memory** vs $O(T^2)$ for naive. Quantified Triton kernel-fusion speedups over Python page loops.

Technical Skills

Languages & Frameworks: Python, PyTorch, Triton, SQL, Pydantic, FastAPI, Gradio, LangChain, LlamaIndex

ML & Training: LoRA/QLoRA, FSDP, Hugging Face Transformers, Unsloth, W&B, ClearML, Scikit-learn

Infrastructure: Docker, Kubernetes, AWS SageMaker, Azure ML, Snowflake, GitHub Actions, CI/CD

Education

University of Maryland, College Park Expected May 2026

Master of Science in Applied Machine Learning

Coursework: Deep Learning, NLP, Advanced ML, Reinforcement Learning, Optimization, Probability & Statistics