Arun Kumar Chukkala

AI/ML & Generative-AI Engineer | LLM Researcher | MLOps Practitioner
arunkiran721@gmail.com | +1 (409) 549-3003 | Beaumont, TX, USA
github.com/arun3676 | Portfolio | Why Hire Me

Summary

Recent MSc-CS graduate with 3+ years of experience applying AI/ML and LLM technologies across fraud, finance, and search. Skilled in deploying LLMs (GPT-4, Claude, LLaMA), RAG systems with vector databases, and MLOps pipelines. Proven results improving accuracy by 25%+ and cutting latency by 40%+ through efficient model deployment and orchestration.

Core Competencies

- **LLM & Gen-AI:** GPT-4o, Claude 3, Gemini, LLaMA 3; fine-tuning with LoRA, QLoRA, Axolotl, TRL.
- **RAG & Vectors:** Pinecone, FAISS, Weaviate, ChromaDB; LangChain, LlamaIndex, hybrid retrieval.
- **LLMOps/MLOps:** LangSmith, Weights & Biases, MLflow, Ray Serve, BentoML, Docker, K8s.
- **GPU Inference:** vLLM, NVIDIA Triton, TensorRT-LLM, CUDA; latency/throughput optimization.
- **Dev Stack:** Python, PyTorch, TensorFlow, JAX, Git, SQL, REST, Linux.
- **Cloud:** AWS (Bedrock, SageMaker), Azure OpenAI, GCP Vertex AI, Terraform (IaC).

Professional Experience

AI/ML Engineer Jefferies Group — Remote, USA *Mar 2024 – Present*

- Deployed customer lifetime value (CLV) model on SageMaker using 15M+ records/month, boosting ROI by 25%.
- Integrated RAG system (LangChain + Pinecone) for instant document retrieval with GPT-4o summaries.
- Replaced HuggingFace inference with vLLM + Triton stack on A10G GPUs, cutting latency 60%.
- Automated evals + retraining using LangSmith + W&B; CI/CD via GitHub Actions.

Associate AI/ML Engineer Experian — Hyderabad, India *Jan 2021 – Dec 2022*

- Developed fraud detection using Isolation Forest + Autoencoders; reduced false positives by 40%.
- Used Kafka + Ray Serve for real-time streaming; inference SLA under 200ms.
- Designed explainability dashboards in Power BI and maintained Azure ML pipelines.

Projects

LLM-Powered Code Analyzer — GitHub *GPT-4 + Claude 3 + LangSmith for code review; increased issue detection by 40%.*

AI Learning Path Generator (RAG) — GitHub *Custom RAG system with ChromaDB + LlamaIndex; builds personalized curricula in seconds.*

Multimodal AI for Diagnosis — GitHub *ViT + GPT-4o on NIH X-rays; 92% top-1 accuracy.*

Education

M.S. in Computer Science Lamar University, TX *Dec 2024*
Research: Optimizing LLM Performance in Resource-Constrained Environments

B.Tech. in Computer Science Sri Indu Institute of Engineering & Technology, India

Certifications

- NVIDIA DLI: Accelerated Computing with CUDA C/C++ (2025)
- DeepLearning.AI: Generative AI for Everyone
- Stanford ML Specialization – Andrew Ng

Open Source

- Contributor to HuggingFace `trl` (LoRA memory optimization)
- PRs to LangChain (async streaming + callback support for RAG apps)