Arun Kumar Chukkala

AI/ML & Generative-AI Engineer | LLM Researcher | MLOps Practitioner
arunkiran72@gmail.com | +1 (409) 549-3003 | Beaumont, TX, USA
GitHub | Portfolio | LinkedIn | Why Hire Me

Summary

Recent MSc in Computer Science graduate with 3+ years of experience applying AI/ML and LLM technologies across fraud, finance, and search. Skilled in deploying LLMs (GPT-4o, Claude 3, Gemini, LLaMA 3), building RAG systems with vector databases, and orchestrating MLOps pipelines. Proven through projects and roles improving accuracy by 40%+ and reducing latency/memory by 45%+ via efficient model fine-tuning, deployment, and agentic orchestration. Eager to ship robust agents with tool use, memory persistence, and traceability.

Core Competencies

- **LLM & Gen-AI**: GPT-4o, Claude 3, Gemini, LLaMA 3, DeepSeek Coder; fine-tuning with LoRA, QLoRA, Axolotl, TRL; multi-model ensembles, consensus analysis, prompt strategies.
- **RAG & Vectors**: Pinecone, FAISS, Weaviate, ChromaDB, LangChain, LlamaIndex; hybrid retrieval, chunking for grounded agents.
- **LLMOps/MLOps**: LangSmith, Weights & Biases, MLflow, Ray Serve, BentoML, Docker, Kubernetes; CI/CD, automated evals/retraining.
- **GPU Inference & Optimization**: vLLM, NVIDIA Triton, TensorRT-LLM, CUDA basics; latency/throughput optimization, quantization/pruning, real-time quality scoring, security vulnerability detection.
- **Full-Stack Dev**: Python, Flask, Streamlit, HTML/CSS/JavaScript for interactive UIs; PyTorch, TensorFlow, JAX, Git, SQL, REST APIs, Linux.
- **Cloud & IaC**: AWS (Bedrock, SageMaker), Azure OpenAI, GCP Vertex AI, Render (health checks, env management); Terraform for infrastructure as code.

Professional Experience

AI/ML Engineer Mar 2024 - Present
Jefferies Group - Remote, USA
- 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 A100 GPUs, cutting latency 60%.
- Automated evals + retraining using LangSmith + W&B; CI/CD via GitHub Actions.

Associate AI/ML Engineer Jan 2021 - Dec 2022
Experian - Hyderabad, India
- 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.

Education

M.S. in Computer Science Graduated December 2024
Lamar University, Beaumont, TX
- Research Focus: Optimizing LLM Performance in Resource-Constrained Environments

B.Tech. in Computer Science Graduated May 2020
Sri Indu Institute of Engineering & Technology, India

Projects (All Built from ScratchGitHub Code Shows Agentic Builds)

- **AI Code Analyzer (Matrix Interface)** (GitHub | Live Demo)
  Developed a professional code analysis tool with a Matrix-themed UI using Streamlit; integrated state-of-the-art LLMs (OpenAI GPT-4, Anthropic Claude 3, DeepSeek Coder) for comprehensive reviews. Implemented multi-model comparison, consensus analysis, and real-time quality scoring; added file upload support, GitHub repository analysis, and security vulnerability detection. Deployed on Render with health checks, environment variable management, and production-ready configuration; features bug detection, security scanning, code quality assessment, and performance optimization suggestions, increasing issue detection by 40%.
- **AI Learning Path Generator (RAG Agent)** (GitHub)
  Developed RAG pipeline with ChromaDB and LlamaIndex for personalized curricula; implemented memory-like state persistence via embeddings, reducing usage 45%. Grounded answers from unstructured sourcesprototype for tool-using agents.
- **Multimodal AI for Medical Diagnosis** (GitHub | Live Demo)
  Built a full-stack AI application for chest X-ray analysis using Python, Flask, and Gemini AI. Implemented dynamic image analysis generating unique diagnoses for different X-rays; deployed on Render with a professional UI and complete user workflow. Achieved 92% accuracy via iterative fine-tuning and evals, optimized for sub-2s responses with quantization for on-device performance.

Certifications

- DeepLearning.AI: Generative AI for Everyone
- Coursera: Machine Learning Specialization by Andrew Ng (Stanford)

Open Source Contributions

- Contributor to HuggingFace TRL (LoRA memory optimizations for efficient agent fine-tuning).
- Pull Requests to LangChain (async streaming and callback support for RAG-based agents).