Arun Kumar Chukkala

SF, California, USA | arunkiran721@gmail.com | +1 (409) 549-3003 | GitHub | Portfolio | LinkedIn

Professional Summary

AI Engineer with 3+ years of experience developing and deploying machine learning solutions and intelligent applications. Expertise in building production RAG systems, multi-agent workflows, and LLM-powered features using Python, LangChain, and modern AI frameworks. Proven track record optimizing model inference pipelines, implementing MLOps practices with LangSmith and Weights & Biases, and delivering measurable improvements in system performance and user experience. Skilled at translating business requirements into scalable AI solutions deployed on AWS and Azure cloud platforms.

Technical Skills

Programming & Development: Python, SQL, JavaScript, REST APIs, Git/GitHub, FastAPI, Flask, Streamlit

AI/ML Frameworks & Tools: LangChain, LlamaIndex, TensorFlow, PyTorch, Scikit-learn, Hugging Face Transformers, XGBoost, LightGBM

LLMs & Generative AI: OpenAI GPT-4o/GPT-5, Anthropic Claude Sonnet 4.5, Google Gemini 2.5 Pro, Model Fine-tuning (LoRA, QLoRA), Prompt Engineering

MLOps & Deployment: Docker, AWS (S3, Lambda, SageMaker), Azure (ML, Data Factory), vLLM, Render, Vercel, CI/CD Pipelines, GitHub Actions

Data & Vector Databases: RAG Systems, Pinecone, ChromaDB, FAISS, Pandas, NumPy, SQL Databases

Development Tools & IDEs: Cursor AI, Windsurf AI, VS Code, Jupyter, Linux/Unix

Monitoring & Evaluation: LangSmith, Weights & Biases, MLflow, Model Performance Tracking, A/B Testing

Professional Experience

AI/ML Engineer (Contract), Jefferies Group – Remote, USA Mar 2024 – Present

Concurrent with Master's program

- Designed and deployed RAG-based document retrieval system using LangChain and Pinecone vector database, reducing manual review time by 30% for business analysts processing customer interaction data
- Optimized machine learning model inference pipeline by migrating to vLLM, achieving latency reduction from 850ms to 320ms under typical production workloads
- Developed customer lifetime value (CLV) prediction models using Python and AWS SageMaker on dataset of 15M+ records, improving customer segmentation accuracy for targeted marketing campaigns
- Implemented automated model evaluation and monitoring workflows using LangSmith and Weights & Biases, tracking performance metrics and cost optimization
- Built CI/CD pipelines with GitHub Actions for automated testing and deployment of machine learning models to production environments

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

- Developed machine learning-based fraud detection system using Isolation Forest and Autoencoders algorithms, reducing false positive rate by 18% based on weekly validation testing
- Designed and maintained data processing pipelines in Azure Data Factory handling large-scale datasets, implementing data quality checks and improving pipeline efficiency by 25%
- Deployed production machine learning models to Azure ML platform, maintaining system reliability with 98%+ uptime for business-critical applications
- Created interactive Power BI dashboards for model performance monitoring, enabling stakeholders to track key metrics and business insights
- Collaborated with cross-functional data engineering teams to optimize real-time data streaming workflows using Apache Kafka

Featured Projects

LLM Code Analyzer GitHub | Demo

- Production-grade multi-agent code review application using GPT-4o, Claude, and DeepSeek for automated security, performance, and maintainability analysis
- Built with Streamlit and deployed on Hugging Face Spaces; implemented LangSmith observability for debugging agent decisions and tool-use patterns
- Developed using Cursor AI for accelerated development with AI-assisted coding workflows

AI Learning Path Generator GitHub | Demo

- Personalized curriculum generator using RAG pipeline with ChromaDB and LlamaIndex for semantic search and query rewriting
- Reduced API token usage by 45% through strategic semantic caching while maintaining response quality
- Deployed on Render with automated CI/CD pipeline; adapts recommendations based on conversation history and learning patterns

Multimodal Medical Diagnosis Assistant (Research Prototype) GitHub | Demo

- Developed multimodal medical diagnosis prototype by fine-tuning Vision Language Models for integrated analysis of chest X-rays and patient symptoms
- Utilized Whisper model to process audio-based patient reports, enhancing diagnostic context on public medical imaging datasets for research purposes
- Built with Flask, deployed on Render using Docker; integrated Groq API (Whisper-large-v3), OpenAI (GPT-4o), and Google Gemini for vision analysis with Weights & Biases for cost monitoring

Job Search Assistant GitHub | Demo

- Multi-agent application built with FastAPI backend and Next.js frontend, automating job search workflows including resume analysis, job matching, and interview preparation
- Deployed on Vercel for frontend and Render for backend services with automated deployment pipelines
- Implements task persistence, human-in-the-loop approval mechanisms, and error recovery for long-running agent workflows

Education

Lamar University, MS in Computer Science – Beaumont, TX, USA Jan 2023 – Dec 2024

Sri Indu Institute of Engineering & Technology, BTech in Computer Science – Hyderabad, India Aug 2016 – May 2020

Certifications

- DeepLearning.AI: Generative AI for Everyone, LangChain for LLM Application Development
- Coursera: Machine Learning Specialization (Andrew Ng, Stanford University)