Arun Kumar Chukkala

SF, California, USA | arunkiran72@gmail.com | +1 (409) 549-3003 | GitHub | Portfolio | LinkedIn

Professional Summary

AI Engineer with 3+ years of experience building intelligent applications using large language models (LLMs) and retrieval systems. Skilled in taking AI features from prototype to production, with focus on practical solutions that improve business outcomes. Experienced with modern AI tools including GPT-4o/GPT-5, Claude Sonnet 4.5, and Gemini 2.5 Pro. Strong background in deploying scalable systems with measurable impact on performance and user experience.

Technical Skills

AI & Machine Learning: LangChain, LlamaIndex, OpenAI APIs, Anthropic Claude, Google Gemini, RAG Systems, Vector Databases (Pinecone, ChromaDB), Model Fine-tuning

Development & Deployment: Python, FastAPI, Flask, Streamlit, Docker, Git/GitHub, vLLM, AWS (S3, Lambda, SageMaker), Azure ML

Data & Analytics: Pandas, NumPy, SQL, XGBoost, LightGBM, MLflow, LangSmith, Weights & Biases

Tools: Jupyter, VS Code, Cursor AI, Linux, CI/CD Pipelines

Professional Experience

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

Concurrent with Master's program

- Built document retrieval system using RAG architecture (LangChain + Pinecone) that reduced manual review time by approximately 30% for business analysts
- Optimized model inference pipeline using vLLM, improving response latency from ∼850ms to ∼320ms under typical load conditions
- Implemented customer lifetime value (CLV) prediction models on AWS SageMaker using 15M+ customer records, contributing to improved customer segmentation
- Developed automated evaluation workflows with LangSmith and Weights & Biases to monitor model quality and track performance metrics
- Set up CI/CD pipelines using GitHub Actions for streamlined model deployment and testing

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

- Developed anomaly detection system for fraud prevention using Isolation Forest and Autoencoders, reducing false positive rate by approximately 18% based on weekly evaluation sets
- Built and maintained data processing pipelines in Azure Data Factory with data quality checks, improving pipeline efficiency by ∼25%
- Deployed machine learning models to Azure ML for production use, maintaining high system uptime (98%+)
- Created Power BI dashboards for model performance monitoring and business insights
- Collaborated with data engineering team to optimize real-time data streaming workflows

Featured Projects

LLM Code Analyzer GitHub | Demo

- Multi-agent code review application using GPT-4o, Claude, and DeepSeek for automated analysis
- Identifies security vulnerabilities, performance issues, and maintainability concerns
- Implemented LangSmith observability for debugging and monitoring agent decisions

AI Learning Path Generator GitHub | Demo

- Personalized curriculum generator using RAG pipeline (ChromaDB + LlamaIndex)
- Reduced API token usage by 45% through semantic caching while maintaining response quality

- Adapts recommendations based on user conversation history and learning patterns

Medical Imaging Assistant (Research Prototype) GitHub | Demo

- Research prototype for medical imaging analysis using Gemini 2.5 Pro with RAG-enhanced knowledge base
- Evaluated on public medical imaging datasets; not intended for clinical use
- Retrieves relevant medical literature from PubMed to support image interpretation

Job Search Assistant GitHub | Demo

- Multi-agent application (FastAPI + Next.js) for automating job search workflows
- Includes resume analysis, job matching, application tracking, and interview preparation features
- Implements task persistence and human-in-the-loop approval mechanisms

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)