Ram Chandra Ramaraju

(571)474-6977 | San Jose,CA | rc.ramaraju17@gmail.com | ram-chandra17

Software & AI Engineer with 5+ years of experience delivering production-grade AI systems and scalable web applications. Skilled in deploying LLMs (GPT-4, LLaMA 2) with LangChain, FastAPI, and React, optimizing ML pipelines, and building secure, low-latency data workflows on cloud platforms. Proven track record in reducing latency, automating processes, and generating actionable AI-driven insights across healthcare and finance.

Technical Experience

Molina Healthcare **April 2025 — Present**
Software AI Engineer *USA*

- Developed AI-driven RAG models to summarize patient health records, reducing manual chart review time by 45% and improving care response speed.
- Integrated GPT-4 and LLaMA 2 with internal EHR dashboards using LangChain, FastAPI, and React, enabling clinicians to query patient histories conversationally.
- Built a secure document search system using Pinecone and FAISS for 20M+ records, achieving semantic retrieval under 200ms latency.
- Automated FHIR API and HL7 data ingestion via Apache Airflow and AWS Glue, increasing pipeline reliability by 30% and reducing data lag in dashboards.
- Implemented model observability with Prometheus and Grafana, tracking inference latency, token usage, and drift metrics for compliance reporting.
- Designed model retraining workflows using MLflow, SageMaker, and TensorRT, shortening iteration cycles by 35% while ensuring HIPAA-compliant PHI handling.

First Tennessee Bank **Jan 2024 — April 2025**
Software Developer *USA*

- Optimized a fraud detection module by batching transaction scoring requests, reducing average response time from 4.8s to 1.6s.
- Converted Python scripts to Java Spring Boot microservices with Kafka streaming and PostgreSQL backend, improving reliability under load.
- Reduced ML inference time by 60% by profiling bottlenecks and moving models to a persistent in-memory cache.
- Implemented Kubeflow pipelines for ML lifecycle orchestration, reducing model deployment lead time by 50%.
- Built real-time Grafana dashboards for suspicious transactions, enhancing compliance team visibility and response times.

Virtusa **June 2019 — July 2022**
Software Developer *Remote*

- Worked on a React.js dashboard module to track certification progress; fixed state sync issues and optimized API calls, improving page load by 25%.
- Reworked Spring Boot APIs for certification validation and user enrollment, reducing failed certificate issuance from 15% to near zero.
- Wrote automated BDD test cases in C# (SpecFlow), catching data inconsistencies before release and minimizing hotfix cycles.
- Tested integrations with third-party LMS providers using Postman and monitored deployments via Kubernetes CI/CD, reducing post-deploy failures by 40%.
- Tuned SQL queries and indexing for certification reports, cutting generation time from 18s to under 6s and improving audit responsiveness.

Education

Master of Science in Computer Science, *George Mason University* GPA 3.7

Bachelor of Science in Computer Science, *Jawaharlal Nehru Technological University* GPA 3.8

Skills

- **Programming & Scripting:** Python, TypeScript, JavaScript, SQL, PySpark, Pandas, Java, C, C++
- **Web & Frontend Development:** React, Next.js, GraphQL, Apollo, WebSockets, HTML5, CSS3
- **Backend & Cloud:** FastAPI, Flask, Node.js, Spring Boot, REST, Docker, Kubernetes, AWS (S3, IAM, SageMaker, Glue), Vertex AI, CI/CD (Jenkins, GitHub Actions)
- **Machine Learning & AI:** PyTorch, TensorFlow, scikit-learn, Hugging Face Transformers, CLIP, OpenAI Vision, XGBoost, Random Forest, Logistic Regression, LangChain, GPT-4, LLaMA 2, RAG, Semantic Search, Summarization, ONNX, TensorRT, AutoML
- **Data & Analytics:** Airflow, Databricks, Delta Lake, Kafka, Spark Streaming, Kinesis, PostgreSQL, MySQL, MongoDB, Redshift, Snowflake, BigQuery, Pinecone, FAISS, Weaviate, Tableau, Power BI, Looker
- **Monitoring & Governance:** SHAP, LIME, Bias Detection, Explainable AI, MLflow, Prometheus, Grafana, Apache Atlas, Real-time AI pipelines, Multimodal AI applications
- **Collaboration & Version Control:** Git, GitHub, Bitbucket, JIRA, Confluence