Narasimha Rao Kadimi
kmlnrao2008@gmail.com
linkedin.com/in/kmlnrao
+1 214 286 5599

PROFESSIONAL SUMMARY

Innovative Application Architect with 18+ years of experience delivering scalable enterprise software and cloud-native solutions. Expert in building scalable AI systems powered by LLMs, agentic frameworks, secure data engineering pipelines, and cloud-native deployments. Expert in designing robust data flows between autonomous agents and multi-modal data sources. Skilled in fine-tuning large language models, implementing RAG pipelines(textual, tabular, and multimodal), and building production-grade AI solutions integrated with human-in-the-loop feedback systems, semantic search, and facial analysis. Specialized in secure AI deployments across Azure, AWS, and GCP ecosystems.

- Specialized in training and fine-tuning LLMs (OpenAI, Hugging Face, Ollama) using structured and unstructured datasets.
- Designed Retrieval-Augmented Generation (RAG) pipelines—textual, tabular, and multimodal—leveraging vector databases such as FAISS, Pinecone, and ChromaDB.
- Developed predictive models using Logistic Regression, SVM, Decision Trees, Random Forests, Gradient Boosting, Deep Learning, and Reinforcement Learning.
- Deployed CNN-based image classifiers for facial and visual data analysis using Azure ML and TensorFlow.
- Strong background in NLP techniques including word embeddings, summarization, clustering, classification, and named entity recognition.
- Engineered scalable ELT/ETL pipelines across Azure, AWS, and GCP using Spark, Databricks, and Data Factory.
- Architected and deployed agentic AI frameworks using LangChain with human-in-the-loop workflows.
- Implemented secure and compliant AI workflows with OAuth2, JWT, RBAC, and audit logging via FastAPI, Docker, and Kubernetes.
- Leveraged big data ecosystems (Hadoop, Kafka, Delta Lake, Azure Data Lake, Data bricks) for advanced analytics and ML training workflows.
- Automated processes and pipeline orchestration using Python, Airflow, MLflow, and Prefect.
- Expertise in secure AI system design, scalable cloud deployments with Terraform, Docker, and Kubernetes.
- Experience mentoring junior AI engineers and leading AI projects aligned to strategic business goals.
- Designed composite agentic AI solutions for multi-agent collaboration, integrating forecasting, recommendation systems, and scenario-based reasoning.
- Proven expertise in Loading and deploying end-to-end machine learning models using Python, TensorFlow, PyTorch, and cloud-native MLOps tools (Azure ML, Vertex AI, SageMaker).

TECHNICAL SKILLS

- **Programming:** Python, SQL, Java, C#, Dart
- **AI/ML Frameworks:** PyTorch, TensorFlow, Hugging Face, Crew AI, LangChain
- **LLMs & NLP:** OpenAI, Ollama, Groq, Generative AI, RAG, PEFT, LoRA
- **Databases & Storage:** PostgreSQL, MongoDB, Redis,

- **Big Data:** Apache Spark, Azure Databricks, Hadoop, Kafka
- **Graph Databases:** Neo4j, Gremlin
- **Cloud Platforms**: Azure (Blob, Data Lake, ML, OpenAI, AI Search), GCP, AWS
- **Deployment Tools:** Docker, Kubernetes, Terraform, FastAPI, Flask, Azure DevOps
- **Security:** OAuth2, JWT, Role-based Access, Rate Limiting, API Security
- **Visualization:** Power BI, Streamlit, Tableau
- **Version Control:** Git, GitHub, GitLab

PROFESSIONAL EXPERIENCE

Technical Architect & AI Data Engineer | Advanced Integrated Systems | Dec 2021 - Present

Product: HIMS & SLIMS (Cloud-native Health & Lab Systems)

Responsibilities:

- Designed and developed agentic AI systems for multi-agent orchestration using Crew AI and LangChain.
- Built asynchronous pipelines supporting tool usage, agent memory, feedback loops, and human-in-the-loop supervision.
- Developed fine-tuned LLMs using custom data (structured & unstructured) for summarization, reasoning, and automation tasks.
- Integrated Groq and Ollama to optimize latency and model inference for real-time agent responses.
- Developed scalable ELT pipelines with Spark and Azure Synapse to enable high-volume model training workflows.
- Deployed AI systems using Azure Kubernetes Service, Docker, and Terraform for scalability and automation.
- Created vector embeddings with FAISS and Pinecone to power RAG pipelines (textual and tabular document processing).
- Integrated facial recognition modules using Azure CV, Azure Video Indexer

Key Agentic Projects

- **MedScribe Agent:** Real-time transcription and summarization pipeline using Whisper for speech-to-text and fine-tuned LLMs for context-aware summarization. Outputs are structured into JSON schemas compatible with Retrieval-Augmented Generation (RAG). Enables integration into EHR or downstream analytics engines.

- **Smart Intake Agent** Performs OCR on scanned ID documents using Azure Computer Vision and Tesseract, followed by metadata extraction and schema alignment using rule-based and embedding-based matching. Outputs structured data directly into document or relational databases for workflow automation.

- **Insights Search Agent:** Implements hybrid search using ChromaDB for dense vector retrieval and LLMs for semantic re-ranking and synthesis. Supports contextual QA over unstructured document sets with RAG-based query pipelines. Built with LangChain and supports document chunking with metadata tagging.

- **Billing Reasoning Agent**: Designed and implemented a multi-step LLM prompt-chaining pipeline using contextual embeddings to automate ICD/CPT code mapping. Integrated reinforcement learning for claim validation optimization and aligned outputs with HL7/FHIR standards for seamless billing system interoperability.

- **Ops Optimization Agent:** Analyzes historical utilization data using Prophet/LSTM-based forecasting and agentic workflows for dynamic scheduling. Crew AI coordinates multi-agent decisions for resource allocation, ensuring coverage and efficiency across operational nodes.

Responsibilities:

- Designed and deployed ELT pipelines to support agentic logic.
- Built APIs with FastAPI integrated with secure authentication and access controls.
- Developed agentic systems with retry logic, failure handling, audit logs, and metrics.
- Collaborated with data scientists to preprocess datasets for LLM training and inference.
- Created dashboards visualizing model predictions and agent actions using Streamlit and Power BI.

Lead Power BI Developer | Puma | Jul 2018 - Nov 2021

Responsibilities:

- Built enterprise dashboards in Power BI for healthcare ops, diagnostics, and finance analytics.
- Connected to SQL Server, Excel, Azure SQL; used DAX, Power Query for transformation.
- Scheduled data refresh using Power BI Gateway; implemented RLS and optimized models.

Lead Developer & Mobile Architect | VF Global | Jun 2014 - Jul 2018

Responsibilities:

- Developed modular vendor/customer portals using ASP.NET, Angular, Flutter, and SQL Server.
- Built mobile apps for stakeholder engagement; CI/CD via Azure DevOps; secured APIs with

Lead Developer | Molina Healthcare | Mar 2007 - May 2014

Responsibilities:

- Built digital platform for customer and vendor onboarding using ASP.NET and SQL Server.
- Implemented authentication, role-based access, responsive web portals.
- Designed and optimized stored procedures and SQL queries to support high-performance data retrieval and reporting.
- Collaborated with cross-functional teams including QA, Product, and Business Analysts to deliver secure, scalable modules aligned with HIPAA compliance.
- Integrated third-party APIs for eligibility verification and insurance claim processing workflows.
- Implemented reusable service components using .NET for secure communication across internal healthcare systems.