

UMA AJAY KUMAR REDDY P S

Senior AI/ML & Software Engineer | LLM/RAG Systems | Python, FastAPI, Cloud

+91 9148041589 • umaveena.alvas@gmail.com • [LinkedIn](#) • [GitHub](#) • India

PROFESSIONAL SUMMARY

Senior AI/ML and software engineer with 5+ years across technology delivery, including hands-on work on LLM/RAG applications, NLP pipelines, backend APIs, full-stack products, and cloud deployment workflows. Experience spans fintech, legal technology, healthcare, analytics, and SaaS, with responsibilities covering architecture, technical roadmaps, secure engineering practices, CI/CD, and team leadership. Core technologies include Python, FastAPI, LangChain/LangGraph, vector databases, React/Node.js, AWS/Azure, Docker, and Kubernetes.

TECHNICAL SKILLS

Generative AI & ML: LLMs, RAG, NLP, Transformers, fine-tuning, embeddings, hybrid search, prompt engineering, model evaluation, PyTorch, TensorFlow, Hugging Face

AI Frameworks & Retrieval: LangChain, LangGraph, LlamaIndex, OpenAI API, FAISS, Pinecone, tool/function calling

Backend & Architecture: Python, FastAPI, Django, Node.js, NestJS, Express.js, Java, REST APIs, microservices, data structures & algorithms

Cloud, DevOps & Data: AWS, Azure, Docker, Kubernetes, Jenkins, CI/CD, monitoring, PostgreSQL, MongoDB, MySQL, Supabase, Pandas, NumPy, ETL

Frontend: React, Next.js, Angular, Vue.js, JavaScript, HTML, CSS, Bootstrap

CORE AI/ML & ENGINEERING EXPERIENCE

Chief Technology Officer

Payism Global

Jan 2025 - Present

Hyderabad, India

- Define the technology roadmap, architectural direction, engineering standards, and security practices for a global payments platform.
- Lead a cross-functional engineering team of **15+ developers** and strengthen delivery through CI/CD, peer review, and secure coding practices.
- Architect AI-driven transaction routing and real-time fraud-detection capabilities associated with a **15% improvement in transaction success rates**.
- Implement Kubernetes, AWS, Docker, and CI/CD workflows associated with a **40% increase in release velocity** and a **six-month reduction in product time-to-market**.

Senior AI Lead Developer

Legal Soft

Aug 2024 - Present

California, United States

- Architect hybrid RAG systems combining vector and keyword retrieval for legal-document processing, litigation research, and workflow automation.
- Lead LLaMA and Mistral fine-tuning workflows on proprietary legal data and develop secure FastAPI services for enterprise integrations.
- Design high-concurrency API architectures for low-latency responses and integration with corporate CRM and ERP platforms.

Senior Software Engineer

MedVirtual

Sep 2024 - Apr 2026

United States

- Developed healthcare applications using LLMs, RAG, NLP, and transformer models for clinical text classification, entity extraction, and analytics.
- Built and fine-tuned models with PyTorch and TensorFlow, and deployed ML services through Docker, Kubernetes, and automated CI/CD pipelines.
- Delivered full-stack AI features using React/Next.js and backend services for data visualization and predictive analytics.

Full-Stack Developer

FixTman

Nov 2023 - Dec 2024

United States

- Built full-stack applications, backend APIs, transaction workflows, database integrations, and customer-facing product features using the MERN stack.
- Supported debugging, implementation fixes, and production release delivery across frontend and backend modules.

LEADERSHIP, PRODUCT & FOUNDER EXPERIENCE

- Technology Officer - Culinary Analytics

Sep 2024 - Present

United States

Translate business requirements into technical roadmaps and microservices architectures for a data-driven analytics platform. Establish automated testing, CI/CD, linting, and secure coding practices, and lead predictive ML delivery for demand and supply-chain analytics.
- Founder & Chief Executive Officer - Syncrocore

Dec 2024 - Present

India

Lead an AI consulting and development firm delivering LLM, RAG, automation, and multi-tenant SaaS systems. Architect agentic workflows and enterprise document-retrieval solutions using LangChain, LangGraph, LlamaIndex, Pinecone/FAISS, FastAPI, and cloud deployment patterns.
- Founder & Chief Executive Officer - SanjeevTech

Sep 2021 - Present

India

Deliver client and startup products from requirements through deployment using React, Next.js, Angular, Node.js, Express, Python, Django, relational/NoSQL databases, Azure, VPS, and cloud hosting.

EARLIER EXPERIENCE

- Lead Full-Stack Developer, FLAPLIVE INNOVATIONS (May-Dec 2023)

- Led AI-enabled MERN product delivery, reusable APIs, business logic, and release coordination.
- Full Stack Trainer, PixxelKey (May-Dec 2023)

- Trained learners in MERN, Next.js, Python, data structures, algorithms, debugging, and project implementation.
- Lead Full Stack Developer, Way2Agribusiness (Jan-May 2023)

- Developed React/MERN and e-commerce features and contributed to Django, machine-learning, and AI-driven web functionality.
- Full-Stack Developer, Refactor Academy (Oct 2022-May 2023); Full Stack Developer, Leadics (Apr-Jul 2022); Programming Tutor, CodeKaroYaaro (Sep 2021-Jun 2022).
- Earlier Design Experience:

Graphic Designer, Blitz Jobs (Oct 2020-Jul 2021) and SuccessR HR Tech Pvt Ltd (Nov 2019-Sep 2020).

SELECTED AI/ML PROJECT WORK

- Enterprise RAG-Based Document Intelligence System

Designed semantic search, grounded question answering, document summarization, metadata filtering, and retrieval workflows for large enterprise knowledge bases.

LangChain | LlamaIndex | FAISS/Pinecone | FastAPI | Embeddings | Docker
- AI-Powered Customer Support Agent with Tool Calling

Built an LLM-driven support workflow with external API actions, function calling, conversation-state handling, escalation logic, and observability.

OpenAI API | LangGraph | FastAPI | Tool Calling | REST Integrations
- Healthcare RAG & Clinical NLP System

Implemented document ingestion, chunking, vector retrieval, clinical question answering, text classification, and entity extraction for healthcare information workflows.

Python | Transformers | PyTorch/TensorFlow | Pinecone | LangChain | NLP
- Real-Time Fraud Detection & Anomaly Monitoring

Created streaming feature pipelines, fraud scoring, alerting, audit logging, model monitoring, and retraining triggers for production-oriented financial workflows.

Python | FastAPI | Feature Engineering | AWS | Monitoring | CI/CD

EDUCATION

- B.Tech, Computer Science & Engineering

- Ramaiah University of Applied Sciences, Bengaluru | Aug 2019 - May 2023

GPA: 9.35 / 10.0, Graduated with Honors | Focus: AI/ML, Data Structures, Cloud Computing

PUBLICATION

- Cloud-Native Real-Time Health Monitoring with Bio-Robotic Sensors and MERN-Based Edge Analytics

- International Journal of Science and Research, Vol. 15, Issue 3, pp. 666-669, Mar 2026. DOI: 10.21275/SR26310004056

CERTIFICATIONS & LANGUAGES

- Certifications:

IBM Full Stack Web Development (Dec 2021); Triplebyte Certified; Git & GitHub Essentials; Front-End Development with React; Introduction to Cloud Computing; Node & Express Essentials.
- Languages:

English, Hindi, Telugu, Kannada.