Priya Raja

Full-Stack AI Engineer — LLMs • RAG • Agents • Cloud-Native Deployment

Dubai, UAE • +971 565809793 • workinfo.priya@gmail.com
LinkedIn • GitHub • Medium

PROFESSIONAL SUMMARY AI/ML Engineer building production SaaS platforms with LLMs and RAG systems. Recently deployed medical AI SaaS on AWS App Runner featuring automated clinical documentation, intelligent summarization. Reduced clinical documentation time by 70% through LLM-powered automation. 5+ years full-stack experience with proven ability to ship end-to-end AI products from concept to production with authentication, payments, and scale. Seeking AI/ML Engineering roles in Dubai.

CORE TECHNICAL SKILLS

- **AI/ML Production:** LangChain, RAG Systems, FAISS, ChromaDB, Semantic Search, Embeddings (Ollama), Prompt Engineering, Agent Architectures, Function Calling, LlamaIndex, Vector Databases
- **LLMs & APIs:** OpenAI (GPT-4/GPT-4o), Claude (Anthropic), Gemini, Groq, Model Selection, Cost Optimization, Token Management
- **SaaS & Cloud:** AWS (App Runner, Lambda, CloudFront, Bedrock), Docker, Kubernetes, Terraform, Clerk Auth, Stripe Billing, Multi-tenancy, RBAC, Redis, PostgreSQL
- **ML/Data Science:** PyTorch, scikit-learn, Pandas, NumPy, Feature Engineering, Model Evaluation, A/B Testing
- **Backend:** Python, FastAPI, Flask, Node.js, NestJS, REST APIs, GraphQL, WebSockets, Microservices, Event-Driven Architecture
- **Frontend:** React, Next.js, TypeScript, Tailwind CSS, Redux Toolkit, Chrome Extension Development, PWAs
- **DevOps & Testing:** CI/CD, GitHub Actions, Terraform, pytest, Jest, Docker Compose, Monitoring (Datadog, Sentry), Load Testing

AI ENGINEERING PROJECTS

Priya's Digital Twin - AI-Powered Portfolio Assistant *Live: CloudFront — Nov 2025*

- Built autonomous AI agent that answers questions about my skills, projects, and AI/ML expertise
- Deployed full-stack AI application showcasing RAG systems and agentic AI implementation
- Architected scalable cloud infrastructure with AWS Lambda, CloudFront CDN, and Terraform IaC
- Implemented CI/CD pipeline using GitHub Actions for automated testing and deployment
- Demonstrates real-time knowledge base integration with conversational AI capabilities
- **Tech Stack:** React, Next.js, TypeScript, AWS Lambda, CloudFront, Bedrock, Terraform, GitHub Actions, AI/LLM Integration

MedScript AI - Clinical Documentation SaaS *Live: AWS App Runner — Nov 2025*

- Built production SaaS platform automating medical documentation for healthcare providers
- Implemented LLM-powered prescription summarization reducing documentation time by 70%
- Built subscription billing system with usage-based tier management and Stripe integration
- Processing with 99.9% uptime on AWS App Runner
- **Tech Stack:** Python, FastAPI, LangChain, OpenAI API, Clerk Auth, Stripe, PostgreSQL, React

Smart Property Search AI - Intelligent Real Estate Assistant *GitHub — Oct 2025*

- Deployed AI rental assistant processing 100+ Dubai listings in real-time with GPT-4o-mini
- Achieved $0.001 per search through optimized token usage and intelligent caching strategies
- Reduced property search time from hours to seconds with 85% preference matching accuracy
- Implemented web scraping pipeline with BeautifulSoup for live Bayut/Dubizzle data ingestion
- **Tech Stack:** Python, BeautifulSoup, OpenAI API, FastAPI

Enterprise RAG System with Semantic Search *Oct 2024*

- Architected production RAG pipeline processing 10,000+ documents with sub-second retrieval
- Implemented FAISS vector database achieving 70% faster retrieval than keyword search
- Built hybrid search combining semantic and BM25 ranking for optimal relevance
- Reduced token costs 40% through intelligent context windowing and caching strategies
- **Tech Stack:** LangChain, FAISS, Ollama embeddings, ChromaDB, FastAPI, Docker

PROFESSIONAL EXPERIENCE

Excellence Driving School *Dubai, UAE* **Full Stack Engineer** *Jan 2023 – Sep 2024*

- Increased organic traffic by 45% through SEO-optimized Next.js implementation with Strapi CMS
- Built Python-based ML pipeline for predictive analytics on instructor performance metrics
- Implemented semantic search for 1000+ training documents using vector embeddings
- Engineered REST APIs with FastAPI, optimizing dashboard data delivery by 40%

e-Zest Solutions *Pune, India (Remote)* **Frontend Engineer** *Feb 2021 – Dec 2022*

- Delivered D3.js dashboards for UNICEF data across 30+ countries, improving usability 35%
- Created data preprocessing pipelines handling multi-country datasets for ML-ready formats
- Built React visualization layer for AI model outputs with real-time updates
- Achieved full WCAG accessibility compliance and localized WHO portal into 5 languages
- Integrated .NET + SQL backends improving API response times 25%

Addteq Software *Pune, India* **Junior Developer** *Dec 2012 – Feb 2014*

- Automated build workflows with Groovy scripting, boosting developer efficiency 20%
- Migrated legacy systems to dynamic templates, reducing manual update time 60%

EDUCATION & CERTIFICATIONS

- **Applied Data Science with Python Specialization** - Coursera 2024
- **AWS Cloud Practitioner** - Amazon Web Services In Progress
- **Bachelor of Information Technology** - Pune University, India 2012

TECHNICAL WRITING & COMMUNITY

- Published technical articles on AI/ML implementation on Medium

KEY ACHIEVEMENTS

- Best Performer Award 2024 - Excellence Driving School for technical innovation
- Client Appreciation - e-Zest Solutions for UNICEF & WHO project delivery
- Hackathon Participant - Built production-grade Portfolio app for HCL