Yuvaraj Kannan

Software Engineer – AI
9894970783 | ai.yuvaraj.career@gmail.com | linkedin.com/in/yuvaraj-so-kannan | github.com/yuvaraj3855

Summary

AI Engineer with 4+ years of experience building and scaling LLM-driven production systems. Specialized in advanced RAG architectures, LLM inference optimization (vLLM, SGLang), and real-time multimodal AI platforms deployed in healthcare environments. Strong expertise in scalable, robust, and structured hybrid retrieval architectures and secure, HIPAA- and FHIR-aligned production AI system design.

Skills

AI & LLM Systems: Generative AI, RAG Pipelines, LangChain, LangGraph, LlamaIndex, Agentic Workflows, Prompt Engineering, Hybrid Search, MMR Reranking, Applied NLP
LLM Infrastructure: vLLM, SGLang, Self-Hosted LLM Serving, Model Benchmarking, Inference Optimization
Speech & Multimodal AI: Whisper (STT), LiveKit (Video + STT/TTS), Streaming AI Workflows
Backend & System Design: NestJS, FastAPI, Flask, Django, Event-Driven Architecture, API Design, Role-Based Access Control, Performance Optimization
Real-Time Systems: WebSockets, Server-Sent Events (SSE), Redis Pub/Sub, Kafka, RabbitMQ
Databases & Vector Stores: MySQL, MongoDB, Redis, Qdrant
Cloud & DevOps: Docker, Docker Compose, CI/CD (GitHub Actions, GitLab CI), AWS, GCP
Programming: Python, TypeScript, JavaScript, C++
Frontend: React.js, Angular.js
Healthcare Compliance: HIPAA-Aware Architectures, FHIR/EHR Data Models

Engineering Strengths: Analytical Thinking, Structured Problem-Solving, System Ownership, Adaptability, Cross-Functional Collaboration

Experience

Software Engineer – AI Feb 2025 – Present
Meril (Nuvo AI) *India*

- Leading the development of **Healthjini**, a hospital-deployed AI healthcare platform, owning end-to-end AI system architecture and production deployment.
- Designed and implemented scalable **NestJS backend architecture** delivering secure, high-performance APIs for clinical workflows.
- Built and deployed **Python-based AI services** on in-house infrastructure, orchestrating LLM inference using **vLLM and SGLang**.
- Benchmarked and evaluated open-source LLMs including **MedGemma, Mistral, and Gemma**, selecting models based on latency, accuracy, and healthcare-domain relevance.
- Designed advanced **RAG pipelines** using **Qdrant**, implementing hybrid search, metadata filtering, and MMR-based reranking.
- Achieved **93–97% evaluation accuracy in document summarization and retrieval** across diverse real-world clinical datasets by optimizing chunking, hybrid search, and prompt strategies.
- Implemented **real-time streaming chat architecture** using Redis Pub/Sub for AI orchestration and **Server-Sent Events (SSE)** for low-latency frontend streaming.
- Developed **real-time multimodal AI workflows** using LiveKit for video conferencing and live STT/TTS integration.
- Integrated **Whisper Large v3** for WebSocket-based streaming speech-to-text pipelines.
- Implemented **FHIR/EHR-aligned data models** and HIPAA-aware architecture for secure handling of healthcare data.

Senior Full Stack Developer Sep 2024 – Dec 2024
Adshi5.Com Private Limited *Chennai, India*

- Led and mentored a team of developers, providing technical direction and ensuring high-quality, on-time delivery.
- Owned backend development using **Python**, designing APIs, data pipelines, and integrating AI models into production workflows.

- Performed advanced **prompt engineering** to improve the accuracy and reliability of AI-generated **SOAP Notes** for healthcare use cases.
- Deployed and operated scalable, serverless services on **Google Cloud Platform (Cloud Run, Pub/Sub)**, ensuring performance, reliability, and **HIPAA-compliant** data handling.

Full Stack Developer Dec 2021 – Jun 2024

Adshi5.Com Private Limited *Chennai, India*

- Led the integration of **Generative AI** solutions using **ChatGPT APIs**, **Google Vertex AI**, and **LlamaIndex**, with **Redis**-backed conversational memory, improving system intelligence by **30%**.
- Designed and optimized high-performance REST APIs using **MERN stack**, **Python**, and **FastAPI**, reducing system response times by **25%**.
- Built and optimized dynamic, responsive user interfaces using **React.js**, increasing user engagement by **40%**.
- Improved database performance and reliability through **MySQL** query optimization, increasing data handling efficiency by **20%**.

Executive Mar 2019 - Dec 2021

Udaan India Pvt Ltd *Chennai, India*

- Managed a team of 16 members, achieving key performance metrics across multiple states.

Honors & Awards

Above & Beyond Award Nov 2025

Meril (NUVO AI – Annual Connect) *India*

- Recognized for ownership and high-impact contributions in delivering production-grade AI healthcare systems beyond defined responsibilities.

Projects

preocr – Document Pre-OCR Decision Library | *Python, Document AI* 2025 – Present

- Created and published **preocr**, a Python library on PyPI to determine whether documents require OCR processing.
- Implements fast heuristics to skip OCR for machine-readable files, reducing compute cost and preprocessing latency.
- Designed a modular API for seamless integration into document ingestion and AI pipelines.

AI-Powered SOAP Notes Generator | *React.js, Node.js, Python, GCP* Sep 2024 – Dec 2024

- Built an AI-powered system to convert doctor–patient conversations into structured, accurate **SOAP Notes**, improving clinical documentation efficiency and reducing manual effort.

AI-Powered Chatbot Platform | *Python, FastAPI, React.js, LlamaIndex, MySQL* May 2022 – Dec 2024

- Developed a production-grade AI chatbot using **Vertex AI and OpenAI**, improving response quality through prompt engineering, model tuning, and systematic evaluation.
- Implemented **RAG pipelines with LlamaIndex** and built high-performance **FastAPI** services optimized for low latency and efficient retrieval.

Education

University of Madras Chennai, India

B.Sc. Computer Science *2013–2016*