Yuvaraj Kannan

AI Systems Engineer
9894970783 | ai.yuvaraj.career@gmail.com | linkedin.com/in/yuvaraj-kannan-ai | GitHub | Portfolio

Summary

AI Systems 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, Prompt Engineering, Hybrid Search, MMR Reranking

LLM Infrastructure: vLLM, SGLang, Self-Hosted LLM Serving, Model Benchmarking, Inference Optimization

Speech & Multimodal AI: Whisper (STT), LiveKit, Sarvam, Streaming AI Workflows

Backend Engineering: FastAPI, Flask, NestJS, Node.js, GraphQL, REST APIs, Event-Driven Architecture

Real-Time Systems: WebSockets, SSE, Redis Pub/Sub, Kafka, RabbitMQ

Databases & Vector Stores: MySQL, MongoDB, Redis, Qdrant

Cloud & DevOps: Docker, Docker Compose, AWS, GCP, CI/CD (GitHub Actions, GitLab CI)

Languages: Python, TypeScript, JavaScript, C++

Healthcare Systems: HIPAA-aware architectures, FHIR/EHR data models

Experience

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

- Leading the architecture and 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%** retrieval accuracy across clinical datasets by designing advanced RAG pipelines using Qdrant with hybrid search and MMR reranking.
- 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

HealthJini – AI Healthcare Platform | *NestJS, Python, LLMs, Whisper STT, RAG* | healthjini.ai 2025 – Present

- Developed a hospital-deployed AI healthcare platform connecting patients, doctors, and hospitals with intelligent clinical workflows and medical data management.
- Built scalable backend services using **NestJS** and **Python**, supporting appointment systems, patient records, and AI-driven clinical interactions.
- Designed **RAG-based knowledge retrieval pipelines** using **Qdrant** to enable contextual AI assistance for healthcare queries.
- Integrated **Whisper Large v3** for real-time speech-to-text processing in clinical AI workflows.
- Implemented **FHIR-aligned healthcare data models** and HIPAA-aware architecture for secure medical data processing.

CareScribe – AI Medical Scribe | *GCP Cloud Run, Google Gemini, LLMs* | carescribe.health 2021 – 2024

- Developed an AI-powered medical scribe platform that converts doctor–patient conversations into structured **SOAP notes** and clinical documentation.
- Integrated **Google Gemini** models to generate structured clinical summaries and improve medical documentation workflows.
- Deployed scalable AI inference services on **GCP Cloud Run**, enabling containerized and serverless execution of clinical AI pipelines.
- Designed LLM pipelines for contextual medical note generation using clinical knowledge retrieval.

Simmer – Healthcare AI Search Platform | *Python, LLMs, RAG, Hybrid Search* | trysimmer.com 2021 – 2024

- Developed an AI-powered enterprise search and conversational assistant platform for healthcare and dental manufacturers.
- Implemented semantic search and RAG-based retrieval to enable intelligent discovery of healthcare product documentation.
- Designed scalable backend services for document ingestion, indexing, and AI-powered knowledge retrieval.

PreOCR – Document Pre-OCR Decision Library | *Python, Document AI* | PyPI | GitHub 2025 – Present

- Created and published **preocr**, a Python library that determines whether documents require OCR processing.
- Implements heuristic detection to bypass OCR for machine-readable files, reducing compute cost and preprocessing latency in document AI pipelines.
- Designed a modular API enabling integration into document ingestion and AI pipelines.

Education

University of Madras Chennai, India

B.Sc. Computer Science *2013–2016*