

SURYAANSH RATHINAM

Senior AI Engineer | LLMs · Agentic AI · Full-Stack Engineering

suryaansh2002@gmail.com | +65 8278 2863 | Singapore | [Portfolio](#) | [LinkedIn](#) | [GitHub](#)

PROFESSIONAL SUMMARY

Senior AI Engineer with production experience across LLMs, agentic AI, RAG, MCP integrations, model fine-tuning, and full-stack engineering. Currently building applied AI systems within ST Engineering's R&D division. Combines an NUS Masters in AI with hands-on delivery of production-grade intelligent systems at high-growth startups, backed by three peer-reviewed publications.

SKILLS

AI / ML: LLMs, AI Agents, MCP, RAG, Vector Databases, Prompt Engineering, Model Fine-Tuning, Whisper, LangChain, OpenAI API, Transformers, Computer Vision (CNN)

Languages: Python, JavaScript, TypeScript, Java, C++, C, PHP

Backend & APIs: FastAPI, Django, Node.js, Express, NestJS, REST, GraphQL

Data & Cloud: AWS (Lambda, S3, EC2, RDS), GCP, Docker, MongoDB, PostgreSQL, MySQL, Redis, Pinecone, ChromaDB, Firebase

Frontend: React, Next.js, Vue.js, TypeScript, Tailwind CSS, Redux

Testing: Selenium, Appium, Pytest, Jest, Cypress, TestRail

WORK EXPERIENCE

Senior AI Engineer | ST Engineering, R&D Division

May 2026 – Present | Singapore

- Working on securing AI guardrails to ensure safe, reliable behavior of production AI systems.
- Building generative UI capabilities, enabling AI systems to dynamically generate user-facing interfaces.

Founding Engineer (Backend & AI Lead) | ThoutAI

Aug 2025 – Apr 2026 | Singapore

- Led the backend and AI team building a Meeting Intelligence Platform that turns conversations into action, at the intersection of AI agents, real-time systems, and product engineering.
- Architected backend services for real-time meeting capture, transcription, and intelligence.
- Built AI agents for action extraction, summarization, and downstream workflow automation.
- Developed the frontend interface surfacing agent outputs to end users.

AI Agents · Python · FastAPI · Node.js · TypeScript · PostgreSQL · Redis · LangChain · OpenAI · WebSockets · AWS · Docker

AI Developer | Datalens AI

May 2025 – Jul 2025 | Remote

- Engineered a diarization and transcription pipeline to process large volumes of customer service call recordings using Whisper.
- Improved Whisper model accuracy through fine-tuning on a labeled dataset of call audio and transcripts.
- Built an end-to-end natural language-to-SQL interface: parsed schema, generated SQL from prompts, executed queries, and returned formatted results.
- Deployed tools as containerized microservices with interactive front-ends for internal teams.

Whisper · Python · PyTorch · HuggingFace Transformers · Streamlit · FastAPI · PostgreSQL · SQLAlchemy · OpenAI API

Full Stack Developer | Asian Institute of Digital Finance (AIDF)

Oct 2024 – Jun 2025 | Singapore

- Architected and launched 'Caesars', a credit risk assessment report platform, automating report generation and cutting manual processing time.
- Built a Credit Research Tool leveraging OpenAI's GPT-4 for on-demand financial insights, reducing analyst workload.
- Designed and implemented secure, scalable REST APIs with Django and interactive dashboards in Next.js using Recharts and Ant Design.
- Integrated CI/CD pipelines using GitLab, Docker, and AWS Lambda for zero-downtime deployments.

Next.js · Django · OpenAI GPT-4 · GraphQL · PostgreSQL · Docker · AWS (Lambda, S3, RDS) · GitLab CI/CD

Full Stack Developer | Moneyflo

Mar 2023 – Jul 2024 | Remote

- Implemented an AI assistant using OpenAI's GPT-4o model, integrated into a client dashboard, increasing time spent using the product by 25%.

- Built REST API endpoints with NestJS and FastAPI; used AWS S3, Lambda, and EC2 to store and process large volumes of client data.

Next.js · FastAPI · NestJS · TypeORM · PostgreSQL · AWS · OpenAI

Research Intern | [IIT Kharagpur, Medical Imaging & Theragnostic Lab](#) Jun 2023 – Aug 2023 | Kharagpur, India

- Fine-tuned a pre-trained U-Net model on 300 ultrasound images for kidney region segmentation and classification, achieving 99.2% accuracy.

QA Automation Engineering Intern | [Ridecell](#) May 2022 – Aug 2022 | Remote

- Tested and debugged API endpoints and mobile applications using Python, Selenium, and Appium; managed test cases in TestRail.

SELECTED PROJECTS

Gmail RAG System | [Personal Project](#) Apr 2025

- Full-stack RAG application enabling natural-language queries over Gmail content via secure OAuth login, sentence-transformer embeddings, Pinecone vector storage, and an Ollama-powered LLM backend.

Next.js · FastAPI · Python · Pinecone · Ollama · OAuth

Document & Video Q&A RAG System | [Personal Project](#) Feb 2025

- Web app for uploading PDFs and YouTube videos, chunking content, and enabling chat-based Q&A using MongoDB for storage, Pinecone for retrieval, and the OpenAI API for responses.

React · MongoDB · Pinecone · OpenAI

PUBLICATIONS

- Exploring IoT-Blockchain Integration in Agriculture: An Experimental Study. IEEE Access, 2023.
- Survey of AI Models and Techniques in the Analysis and Prediction of Neuro-Degenerative Diseases. Presented at AICECS 2023; published in IOP Journal of Physics, Vol. 2751.
- Analysis and Comparison of Different Frontend Frameworks. Presented at ATIS 2022; published in Springer CCIS, Vol. 1804.

EDUCATION

Masters of Computing, Artificial Intelligence Specialisation | [National University of Singapore](#) Aug 2024 – Dec 2025

- GPA: 4.63 / 5.0
- Teaching Assistant for CS5224 (Cloud Computing) and CS3219 (Software Engineering Principles and Patterns).

B.Tech, Computer Science Engineering | [Manipal Institute of Technology](#) Sep 2020 – Jul 2024

- GPA: 9.47 / 10.0