Soumith Ganji

Jersey City, NJ | 551-325-9714 | soumithganji@gmail.com | www.linkedin.com/in/soumithg

Education

Stevens Institute of Technology — Hoboken, NJ
Master of Science in Computer Science, CGPA: 3.9/4 — *September 2024 – December 2025*

Manipal Institute of Technology — Manipal, India
Bachelor of Technology in Information Technology — *July 2019 – May 2024*

Experience

AI Engineer — May 2025 – Current
SQOR.ai — *New York, NY*

- Designed and deployed a multi-agent orchestration with LangChain, LangGraph, OpenAI Agents SDK, FastAPI, AWS, and Docker to automate trend detection, risk alerts, and KPI monitoring across business functions.
- Worked on MCP servers and pipelines for seamless LLM tool-calling and integration with 500+ KPIs from diverse SaaS platforms, eliminating manual ETL.

AI Engineer — May 2023 – July 2024
Zobaze — *Hyderabad, India*

- Designed and deployed multi-agent AI systems and RAG chatbots with LangChain, Pinecone vector DB, and AWS Bedrock, enhancing POS customer support and operational efficiency.
- Fine-tuned open-source LLMs(Mistral/Llama 2) with LoRA for domain-specific tasks like receipt parsing and inventory classification, boosting accuracy by 18%.
- Built scalable GenAI pipelines for multilingual query handling, reducing support latency by 22%.
- Collaborated with product and ops teams to integrate GenAI features into retail workflows, increasing SMB adoption and satisfaction.

Founder — November 2021 – May 2023
Dotfood — *Manipal, India*

- Founded Dotfood, a food delivery platform serving 10,000+ university students
- Developed and launched the Android application using Kotlin with Firebase as the backend
- Collaborated with restaurant partners to optimize order management, reducing fulfillment time by over 30%
- Scaled to 5000+ monthly orders, demonstrating a strong demand among students

Projects

CareerCraft – AI Powered Resume Analyzer & Career Coach — June 2025 – July 2025

- Built CareerCraft, an AI-driven platform for resume and career path analysis using FastAPI, LangGraph, and OpenAI APIs, orchestrating multi-agent workflows with persistent state management to evaluate resumes on clarity, skills relevance, and market competitiveness.
- Designed a type-safe backend with PostgreSQL and Prisma ORM, integrated Supabase Storage, and deployed an OCR-enabled document pipeline for parsing PDF, DOCX, and TXT resumes at scale.
- Developed an AI coaching system using LangGraph's multi-step reasoning to simulate recruiter screening, generate personalized improvement feedback, and recommend role alignment based on FAANG and Big 4 hiring heuristics.

CareGuide – HIPAA-Aware Healthcare RAG Chatbot — Jan 2025 – Feb 2025

- Built a chatbot using LangChain & LangGraph with multi-agent RAG to answer queries from hospital PDFs and FHIR data.
- Designed ingestion pipeline with semantic chunking, PHI redaction, vector dbs, and hybrid retrieval, improving precision by 23%.
- Deployed a FastAPI backend with guarded responses, citations, and Redis caching, reducing latency to 1.4s.
- Evaluated with RAGAS & SME review, achieving 0.91 groundedness and cutting hallucinations by 37%.

Technical Skills

Languages & Frameworks: Python, Java, Kotlin, C/C++, SQL, HTML/CSS, JavaScript, FastAPI
AI/ML & GenAI: LLMs (OpenAI GPT, Grok, Mistral, Llama, BERT), Hugging Face Transformers, LangChain, CrewAI, MCP, Pinecone, Redis
Data, Cloud & Devops: Git, CI/CD, Docker, Kubernetes, AWS(EC2, S3, Bedrock), Google Cloud Platform(including Vertex AI), PostgreSQL, MongoDB, Supabase, Firebase