SANSKAR GARODIA

213-675-1503 | sanskargarodia05@gmail.com | Linkedin | Portfolio

SUMMARY

AI Engineer with 3+ years of experience building enterprise-grade, production-ready AI systems spanning autonomous multi-agent orchestration, retrieval-augmented generation (RAG), and cloud cost intelligence. Proven expertise in designing stateful conversational architectures with human-in-the-loop controls, persistent checkpointing, and long-running agent workflows using LangGraph and LangChain. Skilled in architecting MCP client–server integrations for seamless tool interoperability across AWS, Jira, and Snowflake. Passionate about shipping scalable AI infrastructure on AWS and Docker that drives measurable business impact.

EDUCATION

University of Southern California — Los Angeles, CA
Master of Science in Computer Science (GPA: 3.70 / 4.0) — Aug 2023 – May 2025
Coursework: Artificial Intelligence, Database Systems, Machine Learning for Data Science, Web Technologies

MIT World Peace University — Pune, India
Bachelor of Technology in Computer Science (GPA: 3.80 / 4.0) — Jul 2018 – Jun 2022
Coursework: Machine Learning, Distributed Computing, Computer Networks, Data Structure

WORK EXPERIENCE

AI Engineer | Finopsly | Cincinnati, OH, USA — Jun 2025 – Apr 2026

- Engineered an autonomous Service Parking module using **LangGraph's stateful, cyclic agent graphs** with **persistent checkpointers** (PostgreSQL-backed) and **human-in-the-loop** approval gates to schedule **multi-cloud resources**, reducing idle compute costs by **up to 30%**
- Transitioned the parking system from a reactive architecture into a **proactive, guardrailed Autopilot** framework supporting **long-running conversations** that survive restarts via durable state persistence
- Architected an **AWS Recommendation Agent** via **Amazon Bedrock** and **MCP**; built a custom **MCP client** connecting to **AWS Cost Explorer API** and **Jira MCP servers** for **closed-loop reasoning** and automated **tool-calling**, decreasing manual analysis time by **85%**
- **Built Ask-FI,** an **enterprise multi-agent platform** orchestrating sub-agents via **LangGraph's supervisor–worker pattern** with **conditional branching and state handoffs**; integrated **Snowflake Cortex** and a **custom RAG pipeline** with **semantic chunking** to shrink LLM context by **75%**
- Deployed Ask-FI on **Docker containers** with **AWS Elastic Beanstalk** auto-scaling for enterprise-grade availability and fault tolerance

AI Software Engineer Intern | Zavvis Technologies | Miami, FL, USA — Sept 2024 – Jan 2025

- Integrated machine learning models into web application to deliver scalable financial insights, designed to support data-driven decision-making for **up to 10,000 users post-launch**
- Engineered a scalable web application leveraging React, **Next.js** and **PostgreSQL**, transforming early-stage designs into functional interfaces, decreasing development time by **50%**
- Collaborated with **founding team** to build an investor-ready pitch deck, aiding in **securing $500k** in funding

AI Engineer Intern | Finopsly | Cincinnati, OH, USA — May 2024 – Aug 2024

- Analysed cloud utilization datasets across **AWS and Azure** using **cost intelligence tools** to identify optimization opportunities, reducing costs by **$1.5M annually**
- Designed and optimized scalable **RAG pipelines and LLM systems** using **LangChain's document loaders, text splitters, and retrieval chains**, improving retrieval performance by **25%** and enhancing expertise in production ML architecture

Software Engineer | Infinite Computer Solutions | Pune, India — Aug 2022 – Jun 2023

- Developed **RESTful Microservices** using **Java (JDK 17) and Spring Boot**, implementing robust **API Security (Authentication & Authorization)** for a global vehicle loan management platform serving **170+ enterprise clients**
- Applied MVC and Repository design patterns to build a scalable, modular software architecture; conducted **functional and system integration testing** across layered system components
- Managed **CI/CD pipelines** via **GitHub Actions** with automated code quality gates, achieving a **25% gain** in team productivity and **10% faster** release cycles
- Reduced API response time by **35%** via an AJAX-based notification system, boosting user engagement by **20%** globally

Full Stack Developer Intern | Volstory | Pune, India — Mar 2021 – Jul 2021

- Launched Vrinda app with Spring and Maven backend, integrating **NLP-based content recommendation logic** to personalize user feeds, **boosting engagement by 50%** and reducing crashes by **25%**
- Led cross-platform Flutter development, consuming **third-party REST APIs** and implementing **structured data parsing pipelines** to enrich the experience for **1,000+ Android and iOS users**
- Transformed product **sketches into a polished, production-ready prototype** with Flutter, applying **rapid prototyping and iterative feedback loops** to cut development time by **30%**

PROJECTS

Trojan Square AI Assistant | *Python, RAG, LangChain, LangGraph, Groq, OpenAI*

- Built a scalable **Retrieval-Augmented Generation (RAG) pipeline** using **LangChain's retrieval QA chains** with **recursive character text splitting** and **semantic embedding generation**, increasing answer relevancy by **35%** compared to keyword-based search
- Engineered end-to-end pipeline, web scraping 50+ USC sites into **1M+ semantic embeddings** stored in **PostgreSQL (pgvector)**, cutting query latency **50%** for 1,000+ USC queries
- Implemented **stateful conversation management** using **LangGraph checkpointers** to maintain multi-turn context and deliver coherent, context-aware responses across **long-running user sessions**

Facesham | *Python, Flask, Flutter, API, CNN, AI, Git*

- Created a scalable solution for **Deepfake detection**, enhancing video authenticity verification by **40% using Google's Xception Classification Model**, built on CNN architecture
- Integrated deep fake detection algorithm having an **accuracy of 85%** with Python application using **Flask API**

LEADERSHIP AND ACHIEVEMENTS

- Vice President, GRIDS: Led 10+ members to develop Trojan Square AI assistant, organized tech workshops for 100+ attendees
- Led a team of 5+ members and won **Special Mention Award** in Techtatva'20 held by MIT among 100,000+ participants

SKILLS AND CERTIFICATIONS

Languages: Python, TypeScript, JavaScript, C++, HTML, CSS, Java

AI/ML & Frameworks: **LangGraph, LangChain, RAG, MCP (Model Context Protocol),** Amazon Bedrock, Snowflake Cortex, OpenAI, Groq

Web Frameworks: Next.js, React, Node.js, Angular, Flask, .NET, Spring

Databases, Tools & Platforms: **Docker, Kubernetes, AWS, Azure, PostgreSQL,** MySQL, MongoDB, NoSQL, Git, Visual Studio

Certifications: **AWS Certified Solutions Architect**, Advanced Object Oriented Programming, Java For Android