MOHITH MANOHAR

NYC | 6466623508 | mohith.m.venkat@gmail.com |github.com/mohith2017| linkedin.com/in/mohithmanoharcu
https://mohithmanohar.vercel.app/

EDUCATION

Columbia University, *Master of Science in Computer Science* — New York, NY
Ramaiah Institute of Technology, *B.E. in Information Science & Engg* — Bangalore, India

RELEVANT WORK EXPERIENCE

Flagship Pioneering — New York, NY
AI Platform & Applied AI Engineer / Forward Deployed AI Engineeer — Feb 2025 - May 2026

- Architected & Built Shared Enterprise Large Language Models (LLMs) deployment platform for 10+ Artificial Intelligence applications, 5+ teams, and 100+ users using front-end Next.js/React, TypeScript, Python/FastAPI, AWS CDK, CI/CD, authN/authZ, observability & reusable deployment patterns for production agentic systems
- Built a shared agent infrastructure layer with service templates, tool-calling, model routing,prompt/context management, retrieval & evaluation hooks, logging/monitoring, and AWS IaC modules, cutting new app setup from 2–3 weeks to under 2 days and accelerating delivery by 60%+
- Developed a Biotech Scientific analysis agent system with FastAPI services, Postgres-backed agentic workflows, retrieval pipelines, vector/RAG retrieval pipelines, execution tracking & AWS CDK deployment pipelines, supporting 100+ scientific workflows per month over portfolio-company datasets
- Owned shipping Full-Stack development, Product Engineering & analytics engineering across 10+ production AI apps spanning RAG-based portfolio analytics, LLM evaluations/A/B testing, AI-assisted competitive intelligence, MCP-style agent workflows, agent registry patterns, chat assistants & Chrome-based research curation
- Standardized platform security, developer tooling, and customer-facing needs with multi-tenant app templates, version-controlled design docs/runbooks, safety/model-risk notes & evaluation playbooks while benchmarking 15 + agent, observability, vector DB, backend/frontend & AWS modules, saving 300+ engineering hours/year, improving reliability and cutting onboarding by 50%

Simply Fixable Inc. — New York, NY
Founding Software Engineer - Gen AI/LLM/Machine learning & Full stack — May 2023 - Jan 2024

- Built Open AI LLM - GPT-3.5 based LangChain powered AI Chatbot with Next.js using >15 Custom Tools & Agents for automated appointment scheduling & Lead generation of customers increasing customer traffic
- Enhanced the accuracy of the LLM by fine tuning it using Python with HuggingFace Transformers & integrated with 3+ Retrieval Augmented Generation chains including NLP improving user output by 40%
- Programmed Vector embeddings based RAG chains using Pinecone database and AWS S3 improving accuracy 30%
- Built and setup the AWS Architecture, developed the Database initial schema and developed relevant APIs documenting using Swagger & updating using a CI/CD pipeline to AWS RDS(SQL), S3 improving efficiency 20%
- Developed and deployed the AI chatbot using AWS EC2 automating it using a PM2 CI/CD pipeline and Nginx web server. Built the internal Javascript/Typescript REST APIs, integrated with the Pinecone vector database

CoreStack.io — New York, NY
Software Engineer – AI/Machine Learning and Backend — Jul 2022 - Dec 2022

- Developed a custom Python based Spacy ML model for a RASA powered AI Chatbot integrated with external & internal APIs to give set of suggestions on input improving user efficiency with 90% accurate keywords
- Collaborated with a team of ML and Software engineers to build the Python based NLP and ML model powered Custom Conversational AI improving user experience and customer retention rate by 20%
- Developed RASA powered AI chatbot using React.js with HTML5 ,CSS3 & TypeScript integrated with custom external API functions & internal RESTful APIs to generate cloud costs, forecasts filtered based on input
- Integrated custom Spacy Machine learning Model and ETL pipeline into the AI Chatbot to give a set of relevant keyword suggestions seamlessly improving user traffic by 30% using Linux serverless architecture

WIT Legal LLC New York, NY
Software Engineer - AI/Machine Learning Jun 2021 - Feb 2022

• Built Python NLP search and clustering pipelines over 10K+ records using OpenAI-based keyword generation and retrieval workflows, improving search relevance and legal workflow efficiency by 50%.

Columbia University New York, NY
Software engineer - Machine Learning Research Jan 2021 - Jul 2021

• Developed Python NLP models and a JavaScript Chrome extension to classify encrypted network traffic and detect phishing across 10K+ pcap files and 100+ domains, improving detection efficiency by 20%.

Cognizant Technology Solutions Bangalore, India
Software Engineer - Python & Application Security Jan 2020 - Jan 2021

• Automated Governance, Risk & Compliance workflows with Python scripts on Dockerized applications, securing 10,000+ IAM roles across enterprise systems.
• Deployed Okta Java SDK, CyberArk YAML, and SailPoint IdentityIQ JavaScript for privileged IAM across cloud infrastructure, improving security controls by 80%.

PROJECTS AND RESEARCH

Almanax - LLM platform for Smart contract auditing Jul 2024 - Aug 2024

• Built and developed features for the Smart contract auditing platform using React and Vue.js to upload, download and receive PDF reports in real-time achieving the output website
• Programmed Go based API routes with resulting in the backend being integrated with the LLM Based summarization

GitHub Org History Website Mar 2024

• Built a GitHub organization history based website to real-time plot the star history of the repositories using Vue.js
• Integrated the pepy.tech and Nixtla API to fetch and plot the real-time statistics with respect to the Python package
• Displayed Github based org projection forecasts based on the current state of the Organization repository based stars

Student Loan Eligibility Rating using Machine learning/Deep learning Mar 2023 - May 2023

• Built an end-to-end student-loan eligibility platform on 100K+ financial records and 100GB+ historical data, combining predictive scoring with REST API and SQL-backed backend workflows, React Native + Redux frontend interfaces, Firebase/Square payment integrations, and AWS-scalable infrastructure for approval decision workflows.

TECHNICAL SKILLS

• **Languages**: Python, TypeScript, JavaScript, SQL, Java, Go, Swift, C/C++, HTML/CSS
• **Tools & Frameworks:** FastAPI, Flask, Node.js, React, Next.js, Vue.js, React Native, Redux, LangChain, LangGraph, LlamaIndex, OpenAI API, Anthropic Claude API, MLOps, OpenAI Agents SDK, Model Context Protocol (MCP), MCP Servers, Langfuse, LangSmith, OpenAI Evals, Pinecone, pgvector, HuggingFace Transformers, spaCy, RASA, TensorFlow, PyTorch, Keras, scikit-learn, Apache Spark, Chrome Extensions (Manifest V3), Swagger/OpenAPI, Claude Code, Codex, Cursor, Figma
• **Platforms & Infrastructure:** AWS, AWS CDK, EC2, S3, RDS, DynamoDB, CloudWatch, IAM, Docker, Kubernetes, GitHub Actions, GitLab CI/CD, PostgreSQL, MongoDB Atlas, Firebase, Google Cloud Platform (GCP), Azure, Linux, Nginx, PM2, Okta, Entra ID, CyberArk, SailPoint IdentityIQ, RSA Archer, Networking