KOLLURU KRISHNA MOULI

+91 9177732023 ◇ kollurukrishnamouli@gmail.com

LinkedIn GitHub

INTRODUCTION

Software Engineer with Industry experience in Full-Stack Gen-AI, Agentic AI and web development

PROFESSIONAL EXPERIENCE

- **Deloitte USI** *Nov 2025 - Present, Hyderabad, Telangana, India*
  - **AI Engineer 2** *Nov 2025 - Present*
    - Currently building and designing AI Agents, evals and agentic workflows for US Federal Governament's general services administration GSA (manage, support, and supply federal agencies).
      - Designed and deployed event driven agentic orchestration workflows where autonomous agents interacted exclusively through asynchronous messaging via event buses (Google Cloud Pub/Sub), promoting loose coupling, scalability, and fault tolerance.
      - Architected and implemented Vertex AI Search based ingestion and retrieval pipelines to improve agent contextual grounding using structured and unstructured market data from SAM.gov, FAR, and other enterprise data sources. Designed and implemented custom chunking strategies to optimize indexing quality and retrieval accuracy.
    - Part of the rapid prototyping team for AI, GenAI, and Agentic solutions. Drove innovation across core delivery teams by identifying gaps and whitespace opportunities, and designing accelerators to address them through accelerated development.
      - **SNAP (USDA):** Developed a multimodal (voice and chat) LLM powered assistant that converted natural language queries into structured eligibility and benefits lookups for USDA SNAP systems. Engineered low latency voice interactions using WebSockets to enable real time bidirectional communication.
- **Sonata Software Limited** *Jan 2023 - Nov 2025, Bengaluru, Karnataka, India*
  - **Software Design Engineer** *Mar 2025 - Nov 2025*
    - Designed and developed **Product Intelligence**, an **Enterprise Search** solution for a Talent Management product company. The system leverages **LangGraph** for building search agents and **Azure AI Search** for knowledge indexing, with seamless LLM integration through Azure AI Foundry.
    - Worked on a **PaaS Platform (MAP-Enai)** for a Pharma client, built with AWS services which helps the company seamlessly integrate their custom AI solutions with **AWS AI Services**, such as Bedrock, Textract, AWS Kendra and others.
    - Designed and led the development of **Agent Bridge**, an innovative agentic framework integrating cutting edge AI orchestration technologies including **LangGraph**, **LangChain** and **LangFuse** to streamline and enhance intelligent agent workflows.
    - Worked with **Advanced Retrieval-Augmented Generation (RAG Pielines)**, **Prompt Engineering**, and **Model Fine-Tuning** to enhance the accuracy and responsiveness of AI-driven solutions.
    - Designed Custom AI Solutions for for clients, including:
      - **Finance Client**: Designed and developed an AgenticAI-based financial transaction fraud detection solution, integrating **Neo4j knowledge graphs** for entity relationships and NLP techniques for entity extraction, **Agent Memory** for contextual reasoning, and **evaluation pipelines (Evals)** for precise relevance and anomaly measurement.
      - **Pharma Client**: Designed and Developed a comprehensive solution to migrate their existing search platform to a new, scalable model using AWS services. Prepared functional and technical overview, effort estimates, and commercial plans.
      - **Research Assistant Bot**: Designed and developed a memory module for maintaining bot memory and created a framework to measure and improve response accuracy, optimizing researchers' efficiency.
      - **Implementation Blueprint**:Designed and Developed a solution for creating Technical Guides for a Talent Management Product company to implement internal products and services using Agentic AI and Azure AI search.
      - **Fabric Search Agent**: Designed a Fabric-based search solution leveraging Azure AI Foundry agents as the orchestration layer, Fabric Data Agents as macro agents for data operations, and Microsoft Purview as the data governance layer.
  - **Full Stack GenAI/Web Developer** *Jan 2023 - Mar 2025*
    - Part of the core development team of **Harmoni.AI**, Sonata Software's responsible-first GenAI product, contributing from its inception and driving key technical decisions.

* Built an internal **HR Policy Chat Bot** using Python Azure Functions, Azure Bot Framework, and React.js, which provided employees with rapid access to internal policy information.
* Developed a novel content extraction algorithm using **Azure Document Intelligence**, converting documents into vectors for effective Q&A. Created a re-ranking algorithm, improving response accuracy by **80%**.
* Conceptualized and developed an assistant using Copilot and Autogen, culminating in an in-house solution called **AgentIQ** for dynamic, multi-skill AI solutions.
* In addition to leveraging frameworks like Autogen, developed an in-house agentic framework, **Agent Nexus**, which incorporates further enhancements and optimizations to better meet our specific needs.
* Developed an **ITSM Chat Bot** for automating IT support queries, including ticket creation for unresolved issues and assisting support agents in quickly extracting answers from past tickets. Introduced a self-service feature for agents to generate and upload Technical Solution Guides (TSGs) using AI.

SKILLS

- **Technical:**
  - **Programming:**
    * **Languages:** Python, C#, JavaScript
    * **Web server Frameworks:** Django, Flask, Fast-API, .NET web api
    * **Front-End:** ReactJS, MaterialUI, TailWindCSS
  - **Cloud & Data:**
    * **Hyperscalers:** GCP, AWS, Azure, OCI
    * **DevOps:** Kubernetes, Github Actions, Terraform
    * **Databases:** PostgreSQL, MongoDB, CosmosDB, Azure Tables
  - **AI & Generative AI:**
    * **Platforms:** Google Vertex AI, Azure AI Foundry, AWS Bedrock
    * **LLM Application Frameworks:** LangChain, LlamaIndex, OpenAI SDK
    * **Agent Orchestration Frameworks:** LangGraph, Google ADK V2, Microsoft Agent Framework, OpenAI Agents SDK, AWS Strands SDK
    * **LLM Engineering & Ops:** LangFuse, Opik, BrainTrust, Phoenix Arize
    * **AI Evaluation and Guardrails:** LangFuse Evals, Nemo Guardrails, RAGAS, G-Eval, MEGA, OpenAI-Evals,
    * **AI Protocols:** MCP, A2A
    * **Core Concepts:** Prompt Engineering, RAG, Fine-Tuning(LoRA, Q-LoRA, SFT, distillation), Vector Search, Context Engineering
  - **Developer Tools:** VS Code, Cursor, AntiGravity, Claude Code, GIT, PowerShell, Bash
- **Business:** Proposal Creation, Market Research, Pitch Deck Creation
- **Soft Skills:** Communication, Team Leadership, Problem Solving

EDUCATION

- **Bachelor of Technology** *Computer Science and Engineering*
  - **Swami Vivekananda Institute of Technology, affiliated to (JNTU) Hyderabad** *2018 - 2022*

PROFESSIONAL CERTIFICATIONS

- **Anthropic Certified**
  - Claude Certified Architect - Professional [ref] — Agentic AI/Claude/GenAI in Business
  - Claude Certified Architect - Foundations [ref] — Agentic AI/Claude
  - Claude Certified Developer - Foundations [ref] — GenAI/RAG/MCP/Claude
  - Claude Certified Associate - Foundations [ref] — Agentic AI
- **Microsoft Certified**
  - AB-100 Agentic AI Business Solutions Architect [ref] — Agentic AI/Copilot/Azure
  - AZ-305 Azure Solutions Architect Expert [ref] — Cloud
  - AZ-104 Azure Administrator Associate[ref] — Cloud
  - AI-900 Azure AI Fundamentals [ref] — AI/ML
  - AI-102 Azure AI Engineer Associate [ref] — AI/Azure AI Services
  - DP-100 Azure Data Scientist Associate [ref] — AI/ML/Data-Science
  - AZ-204 Azure Developer Associate [ref] — Azure Cloud Infra/Development
  - AZ-400 Designing and Implementing Microsoft DevOps Solutions [ref] — Azure DevOps
- **AWS Certified**
  - AWS Certified Developer - Associate [ref] — AWS Cloud Development

- **Google Certified**
  - Google Data Analytics Professional Certificate [1] [2] [3] [4] [5] [6] [7] — AI/ML/Data-Science
- **Oracle Certified**
  - Oracle Database Design [ref] — Database Management
  - Oracle Database Programming with SQL [ref] — Database Management
  - Oracle AI Cloud Database Services 2025 Certified Professional [ref] — Database
  - Oracle Cloud Infrastructure 2025 Certified Al Foundations Associate [ref] — Generative AI
  - Oracle Cloud Infrastructure 2025 Certified Generative Al Professional [ref] — Generative AI
  - Oracle Cloud Infrastructure 2025 Certified Al Vector Search Professional [ref] — Vector Search/Gen-AI
  - Oracle Cloud Infrastructure 2025 Certified Data Science Professional [ref] — AI/ML

PROJECTS

GCP-Based HritiGT.ai [ref] *September 2024*
Developed a versatile, GCP-based AI solution for enterprises which aids the organization across domains such as HR, Sales and others. Used, GPT-4o, Text-Embedding-3-large, GCP Cloud bucket, functions (serverless), openAI Function calling, and PineCone Vector store.

AI Based Resume Shortlister [ref] *November 2024*
An AI-powered resume shortlisting application that enables HR professionals to chat with a corpus of resumes, generate interview question suggestions, and evaluate candidate profiles.

GNN-Based Software Development Effort Estimation [ref] *January 2022*
Designed a framework using Graph Neural Networks (GNN) to enhance software cost estimation accuracy in a dynamic industry.
- Demonstrated improved estimation accuracy compared to traditional models, aiding in resource allocation and cost control.

Patron-Procuration-Target-Prognosis: A Deep Learning Solution [ref] *June 2020*
Built a deep learning framework for predicting user intentions on e-commerce platforms to bridge online and physical store experiences.

ACHIEVEMENTS

Winner of Sonata Spark Hackathon [more] *November 2024*
Winner of Sonata Tech Fest [more] *June 2024*

LECTURES

Programming for Industry - SVIT, Telangana [ref] *October 2022*
Delivered a comprehensive lecture on programming practices relevant to the industry, focusing on practical applications and real-world problem-solving.

Programming with Python - SVIT, Telangana [ref] *December 2022*
Presented an in-depth lecture covering Python programming fundamentals, advanced features, and its use in data analysis and automation.

MongoDB Essentials - Rajalakshmi Engg College, Chennai [ref] *September 2024*
Conducted a lecture on MongoDB, exploring its architecture, querying, indexing, and its application in modern web development.

PUBLICATIONS

"Supply Chain Management: An Innovative Application of Edge Computing" [ref] *ICIHPE 2020, Tamil Nadu*
"Graph Neural Networks for AI-Driven Insights" [ref] *AI Conference (Govt. Telangana), 2022*
Presented findings at an AI conference, discussing the use of Graph Neural Networks (GNN) for advanced data analysis and prediction models.