Mounika Grandhi

grandhimounika29@gmail.com | +91 9493328585 | Hyderabad, India | github.com/mouni298 | linkedin.com/in/mounika-grandhi

PROFESSIONAL SUMMARY

Senior GenAI Developer with 3+ years of experience specialising in Generative AI, multi-agent systems, and RAG-based architectures. Proven record of architecting and delivering production-grade AI solutions using LangChain, LangGraph, DeepAgents, and MCP - reducing manual effort by 70% and improving AI response accuracy by 30%. Experienced in building end-to-end Agentic AI workflows, LLM orchestration across OpenAI, Anthropic, Gemini, and deploying scalable AI products on AWS and GCP.

SKILLS

Programming Languages: Python, Java
AI / ML Frameworks: LangChain, LangGraph, LangSmith, Google ADK, Deep Agents, LlamaIndex, FastMCP
Generative AI / LLMs: RAG, Agentic AI, Prompt Engineering, NLP, Machine Learning, Deep Learning, MCP, A2A
LLM Providers: OpenAI, Anthropic (Claude), Google Gemini, Grok
Databases & Vector Stores: MySQL, MongoDB, Milvus, ChromaDB, Vector Databases, Semantic Search
Cloud Platforms: AWS, Google Cloud Platform (GCP)
Backend & APIs: FastAPI, RabbitMQ, REST APIs
Tools & Version Control: Git, Docker, LLMOps, CI/CD, Monitoring, HuggingFace
Soft Skills: Problem Solving, Cross-functional Collaboration, Agile Workflows

PROFESSIONAL EXPERIENCE

AI Engineer | Inncircles *Jul 2025 – Present*

- Built a production-grade multi-agent AI framework using Deep Agents and LangChain, orchestrating multi-step specialized LLM agents (Compare, Summary, KPI) with hierarchical sub-agent delegation and human-in-the-loop interrupts.
- Designed and implemented a RAG pipeline and Knowledge Base Search system using Milvus vector database, incorporating document chunking strategies, semantic search, token-aware batching, and guardrails for output safety and response grounding.
- Implemented multi-provider LLM orchestration across OpenAI, Anthropic, Gemini, and Grok with role-based model routing, structured output enforcement using Pydantic, token usage tracking, and intelligent fallback strategies for latency and cost optimization - integrated LangSmith for end-to-end monitoring and observability across providers.
- Integrated 13+ external tool connectors using MCP (Model Context Protocol) and Composio, enabling agent access to Gmail, Google Sheets, Google Calendar, Tavily Web Search, and image generation APIs.
- Engineered a B2C WhatsApp AI assistant for construction project queries using hybrid RAG and LangGraph workflows, enabling automated call scheduling with sales teams.
- Built a conversational analytics system for construction stakeholders using LangGraph agent workflows, enabling natural-language exploration of leads data with real-time dashboard insights.

Digital Specialist Engineer | Infosys *Aug 2022 – Jun 2025*

- Spearheaded the Infosys Migration Platform, an AI-driven solution transforming legacy systems using Generative AI, building end-to-end data pipelines enabling seamless migration to modern frameworks and cutting manual effort by 70%.
- Pioneered an end-to-end code summarization pipeline leveraging LangChain summarization chains and language parsers, significantly improving content extraction and retrieval accuracy.

- Implemented RAG-based QA retrieval chains across 5+ legacy project codebases, enhancing efficiency and scalability in reverse engineering and knowledge extraction.
- Designed LLM evaluation pipelines using BLEU, CodeBLEU, and custom benchmarks to assess model output quality, accuracy, and hallucination rates ensuring robust post-migration validation.
- Crafted AI-powered Business Requirement Documents for legacy applications using LangChain, RAG, VectorDB, and embeddings, substantially enhancing document processing and knowledge retrieval.
- Refined prompt engineering strategies delivering a 30% improvement in AI-generated response accuracy and processing speed.

PROJECTS

Customer Support Assistant using LangGraph and Agentic AI

- Designed an AI-powered customer support assistant leveraging LangGraph and Agentic AI for dynamic, autonomous workflow execution, reducing manual support intervention through intelligent multi-agent routing.
- Deployed LLM-powered services on Google Cloud Platform using Vertex AI for model hosting and Cloud Functions for event-driven agent workflow execution.
- Integrated multi-agent retrieval mechanisms - vector database search, document retrieval for FAQs, database-powered product lookup via LangChain tools, and real-time web search enabling comprehensive, accurate responses across query types.

RAG Chatbot for Database with LlamaIndex

- Built a natural language interface for querying MySQL databases using LlamaIndex NLSQLTableQueryEngine and vector database semantic search, enabling non-technical users to extract insights without writing SQL.
- Implemented real-time SQL query generation via Groq LLM 70B model with HuggingFace embeddings for enhanced semantic query understanding and retrieval accuracy.

Conversational Chatbot with Long-Term Memory using LangChain

- Built a conversational AI system using LangChain and Google Gemini 2.0 LLM with long-term memory capabilities, enabling persistent context retention across sessions using a Vector Store Retrieval system.
- Integrated ChromaDB for efficient storage and retrieval of conversation history with Vertex AI embeddings on GCP for semantic search and context understanding.

EDUCATION

Bachelor of Technology (B.Tech) — Information Technology *2018 – 2022*
SRKR Engineering College, Bhimavaram, Andhra Pradesh **CGPA: 8.67**

Class XII — MPC | Narayana Junior College, Andhra Pradesh *2018* | **97.2%**

Class X — SSC | Narayana English Medium School, Andhra Pradesh *2016* | **CGPA: 10.0**

CERTIFICATIONS

- AWS Certified Cloud Practitioner - Amazon Web Services
- Google Cloud Certified Associate Cloud Engineer - Google
- Salesforce Certified AI Associate - Salesforce
- Applied Generative AI Professional - Infosys
- Certified Python Programmer - Infosys
- Certified Mongo Developer - Infosys

AWARDS & RECOGNITION

- EnGenius Best Engineer Award - Infosys (Sep 2024 & Mar 2025)
- Rise Insta Award - Infosys (Mar 2024)