Devaraj Kudumula

devaa.kudumula@gmail.com | 6281003863 | linkedin.com/in/devaraj-kudumula

Professional Summary

Results-driven AI/ML Engineer with about 3 years of experience designing, developing, and deploying enterprise-grade AI solutions. Expertise in LLM orchestration (LangChain, LangGraph), advanced RAG techniques, and full ML lifecycle development, including fine-tuning and deployment. Demonstrated success in delivering high-impact projects, achieving quantifiable improvements in user engagement, operational efficiency, and system performance. Skilled in Python, TensorFlow, PyTorch, and cloud platforms like GCP, with a strong background in NLP and scalable system architecture. Strong communicator and team player with a strong track record in driving innovation, improving execution reliability, reducing operational costs, and streamlining business decision making with AI-first solutions.

Skills

Languages: Python, SQL, Bash

Technologies & Tools: LangChain, LangGraph, LiteLLM, FastApi, Hugging Face, Git, Jenkins, MongoDB, Docker, Kubernetes, GitHub Actions

Other: Machine Learning, Natural Language Processing, Deep Learning, Computer Vision

Experience

AI/ML Engineer, Intellect Design Arena – Chennai, Tamil Nadu June 2023 – Present

- Architected and deployed scalable, enterprise-grade LLM applications using **Google Gemini** and **OpenAI**, supporting over 10,000 daily active users and driving adoption across multiple business units. Managed end-to-end model lifecycle, from pre-production optimization to final deployment.
- Spearheaded the development of **Enterprise GPT** - a powerful enterprise grade conversational assistant with integrated intelligent **document processing and Q&A**, **real-time web search** capabilities, and user-configurable agent marketplace, enabling domain experts to add custom task-specific agents/tools into the agent.
- The innovative **LangGraph** orchestration system was implemented that allows users to create conditional workflows by adding multiple agents and tools together and is executed without LLM overhead, reducing operational costs, increased efficiency and reduced execution time.
- Engineered and developed advanced RAG capabilities, **query planner**, metadata-driven **self-query** filtering, intelligent conversation flow management(**Agentic RAG**), **Follow-up Questions** ensuring complete coverage in multi-entity queries and delivering context-aware responses by intelligently distinguishing between general and knowledge-base-driven questions, resulting in higher accuracy and relevance.
- Built a scalable **Agentic AI(ReACT)** framework with adaptive reasoning and **self-healing** mechanisms, enabling seamless custom multi-tool and multi-agent integration by users. Enhanced system reliability with automated error recovery by self healing and intelligent workflow optimization.
- Conducted applied NLP research into three successful production implementations, including the deployment of **memory-augmented Q&A** systems and specialized domain-specific chat agents.
- Built the **OCR Service** to reconstruct document layouts from **Azure Doc Intelligence** output and the Chunk Service to generate Markdown-based table structures, extract image content, and aggregate all elements with precise positional metadata and chunks data into page level,block level and word level, enabling accurate document recreation and downstream AI ingestion.

Data Science Intern, Caterpillar – Chennai, Tamil Nadu Jan 2023 – Jun 2023

- Constructed a Transformer-based NLP pipeline for large-scale customer feedback analysis, processing over 100,000 comments monthly with 92% sentiment classification accuracy. Delivered predictive insights that enabled proactive, data-driven decisions.
- Engineered a fine-tuned Transformer model for domain-specific comment summarization, which decreased manual review time by 90% and accelerated insight generation.

- Designed and built a real-time data visualization dashboard, providing interactive analytics to monitor service diagnostics and detect anomalies, thus improving operational oversight across distributed systems.
- Migrated and modernized legacy SAS analytics pipelines to Python, which streamlined data processing and led to significant improvements in scalability, execution speed, and infrastructure cost-efficiency.

Projects

Enterprise Conversational AI Assistant

- Developed a scalable conversational AI platform leveraging advanced LLM orchestration via LangChain, integrating real-time web search and secure document processing.
- Architected a user-configurable agent marketplace, enabling subject matter experts to create and deploy custom, task-specific agents and accelerating internal solution development.
- Engineered a self-healing ReACT agentic framework with adaptive reasoning, significantly increasing the platform's reliability and operational uptime.
- Deployed the solution to serve over 10,000 daily active users, providing measurable improvements in internal communication and knowledge retrieval efficiency.

Advanced RAG Enhancements

- Engineered an adaptive RAG system leveraging multiple retrieval strategies, including query planning, agentic routing, multi-query generation, and metadata-driven self-query filtering. This significantly improved relevance and completeness of responses for complex, multi-entity user queries
- Developed and implemented a Query Planner that dynamically breaks down intricate, multi-entity questions into focused sub-queries, orchestrating parallel retrieval and re-ranking for comprehensive information synthesis.
- Pioneered an Agentic RAG system with intelligent routing logic that distinguishes between general knowledge and retrieval-dependent questions. This optimized latency and inference cost by bypassing unnecessary retrieval steps.
- Integrated a Multi-Query Retriever to address ambiguous user inputs by generating multiple rephrased queries. This broadens the retrieval net, increasing the probability of fetching the most relevant documents for improved accuracy.
- Created a Self-Query Retrieval capability enabling dynamic metadata filtering in MongoDB based on user intent. This dramatically improves retrieval precision by focusing the search on highly relevant document subsets.

Education

Amrita Vishwa Vidyapeetham, Coimbatore Jul 2019 - Jun 2023
B.Tech. in Computer Science and Engineering Specialization in Artificial Intelligence

Awards and Certificates

- **Spot Light Award(2):** Recognized with two Spotlight Awards, a quarterly distinction for delivering high-impact enterprise AI solutions, specifically for contributions to Enterprise GPT and RAG pipeline enhancements.
- Google Data Analytics Certification, Coursera.