E Hemanth Nagesh

AI Systems Engineer | Agentic AI | RAG Architect

✉ hemanthnagesh082@gmail.com ☎ 9980144503 📍 Bangalore, India in linkedin.com

🔗 Agentic AI Demo Github

PROFESSIONAL SUMMARY

AI Systems Engineer with 2+ years of experience designing, building, and deploying Retrieval-Augmented Generation (RAG) pipelines and multi-agent AI systems in enterprise environments. Strong expertise in LLM orchestration using LangChain and LangGraph, vector-based retrieval with pgvector, and context-aware prompt engineering. Experienced in developing production-grade AI microservices using FastAPI, integrating STM/LTM memory, and automating deployments via CI/CD pipelines. Proven ability to collaborate across data, platform, and product teams to deliver scalable, reliable, and context-aware AI systems with measurable business impact.

EDUCATION

PES College of Engineering, Mandya, 10/2021 – 07/2023
Master of Computer Applications

Ramaiah Institute of Business Studies, Bangalore, 05/2018 – 09/2021
Bachelor of Computer Applications

PROFESSIONAL EXPERIENCE

Tata Consultancy Services, *AI Engineer* 12/2023 – Present | Bangalore

- Designed and engineered enterprise-grade AI system components supporting retrieval, reasoning, and generation workflows across distributed environments.
- Built performance-aware AI services and internal tooling to analyze execution behavior, latency, and resource utilization in AI-driven workloads.
- Developed structured logging, metrics collection, and observability pipelines to support monitoring, debugging, and reliability of production systems.
- Implemented RESTful AI microservices using FastAPI, enabling scalable integration of LLM-powered workflows into enterprise platforms.
- Contributed to CI/CD automation using GitLab, ensuring reliable builds, automated testing, and deployment validation for AI services.
- Collaborated cross-functionally with research, platform, infrastructure, and product teams to align AI system design with enterprise scalability, security, and performance requirements.
- Applied strong understanding of compute, memory, and concurrency models to optimize system behavior and eliminate performance bottlenecks.

BOTSIO Chatbot LLP, *Al Engineer Intern* 03/2023 – 05/2023 | Bangalore

- Designed and developed a Generative AI chatbot using LangChain, OpenAI APIs, vector databases, and RAG for real-time document intelligence.
- Implemented semantic search pipelines using embeddings, significantly improving retrieval accuracy and contextual relevance.

- Optimized prompt engineering and retrieval strategies to enhance LLM response quality and user experience.
- Integrated AI workflows into backend services, enabling real-time query handling and document-based reasoning.

TECHNICAL SKILLS

AI & LLM Systems — Retrieval-Augmented Generation (RAG), Multi-Agent Systems, LangChain, LangGraph, Prompt Engineering, Context Engineering, LLM Fine-Tuning (PEFT/LoRA – familiarity), Agentic AI, Semantic Search, **Cloud platform's** — Azure (AI Foundry, Azure Web Apps), AWS (EC2, S3, Lambda), **Backend** — Python, FastAPI, Flask, REST APIs, Microservices Architecture, **DevOps & MLOps** — Docker, GitLab CI/CD, Kubernetes (basic), Helm (exposure), Model Evaluation & Testing, **Vector database** — pgvector, Pinecone (familiarity), Milvus (familiarity), Weaviate (familiarity), **Observability & Reliability** — Structured Logging, Metrics Collection, Monitoring (Prometheus/Grafana – exposure)

PROJECTS

Multi-Agentic AI System – Travel & Retail Automation 11/2023 – 12/2023

- Architected a multi-agent AI system using LangChain and LangGraph, enabling autonomous task execution across travel and retail domains.
- Designed Agent-to-Agent (A2A) coordination workflows for collaborative planning, reasoning, and execution.
- Implemented Short-Term and Long-Term Memory (STM/LTM) using PostgreSQL with vector similarity search.
- Integrated Mem0-based user profiling to extract key facts from conversations and enable persistent personalization.
- Built a Deep Research Agent for real-time web data retrieval and browser automation to support hotel and product booking workflows.

DocuBot System - Agentic RAG for Document Intelligence 05/2023 – 07/2023

- Built an AI-powered document intelligence system using LLMs and Agentic RAG, enabling intelligent Q&A across PDF, DOCX, and TXT files.
- Implemented multi-vector-database retrieval, dynamically routing queries to the most relevant vector store based on user intent.
- Enhanced system accuracy through prompt engineering, contextual templates, and retrieval optimization.
- Delivered a scalable solution for enterprise document search and knowledge extraction.

PUBLICATIONS AND CERTIFICATIONS

01/05/2023

Published: A technical paper on integrating LLMs and vector embeddings with chatbots to enhance document retrieval across formats.
Certification: certified AI Engineer by Microsoft Azure