TALARI PAVAN KUMAR

Hyderabad, Telangana, India | +91 949-12-15-197
pavankumartalari01@gmail.com | https://www.linkedin.com/in/pavankumar-talari/| https://github.com/talaripavan

SUMMARY

Motivated Software Engineer with over 1 year of experience at an AI startup, specializing in building end-to-end **RAG (Retrieval-Augmented Generation)** systems and **LLM-powered agents**. Skilled in **Python**, Django, and **LlamaIndex**, with experience in developing **Prompt Studios**, **Agent Evaluation** frameworks, and Multi-Agent Workflows. Proficient in **implementing vector databases** (Milvus), prompt engineering techniques, and integrating **MCP servers** for database interaction. Passionate about experimenting with emerging AI technologies, **turning ideas into working prototypes**, and delivering scalable, intelligent solutions that enhance automation and decision-making.

EDUCATION

Sree Nidhi Institute of Science and Technology **2019 - 2023**
Computer Science Engineering GPA (6.9/10.0)

PROFESSIONAL EXPERIENCES

AI Engineer [Scorpius Networks] **Feb 2024 - Nov 2025**

- Led the end-to-end development of Retrieval-Augmented Generation (RAG) Agents for enterprise data — from concept to deployment.
- Designed and implemented the RAG pipeline: document ingestion, chunking, embedding generation, and storage using Milvus Vector Database.
- Built a chat interface enabling users to query documents and retrieve context-aware responses powered by LLMs.
- Developed a Prompt Studio using Jinja2 Templates, Promptify, and Prompt Source for creating, testing, and comparing prompt templates.
- Integrated multiple prompt engineering techniques, including Chain of Thought, One-Shot Prompting, and ReAct for improved response reasoning.
- Created an Agent Response Evaluation system using LlamaIndex and DeepEval to measure accuracy and reduce hallucinations through context-based scoring metrics.
- Implemented Intent Query Transformation, optimizing tool selection and improving agent performance for multi-tool workflows.
- Built MCP Servers for databases like InfluxDB (Flux queries) and connected them with agents for natural language interactions.
- Experimented with Google MCP ToolBox for PostgreSQL to enable conversational database querying.
- Added Source Node tracing to display the exact document context used by agents for generating responses, improving transparency and reliability.
- Engineered a Multi-Agent Orchestrator for workflow automation, where agents coordinate tasks dynamically.
- Developed a Workflow Engine to automate actions using triggers and events through simple natural language queries.
- Deployed the complete Django-based RAG application on AWS EC2 using Nginx and Gunicorn for production readiness.
- Explored Prometheus metrics to monitor system performance and identify key operational insights.
- Developed a Pods Analyzer Tool that connects to the Kubernetes API to track and analyze active pods in real time.

SKILLS

- **Languages:** Python, HTML, CSS.
- **Frameworks & Tools:** LlamaIndex, Django, Streamlit, DeepEval, Milvus, Jinja2, Promptify, AWS EC2, Model Context Protocol (MCP) , Google MCP Tool Box , Prometheus , Grafana, Kubernetes
- **Databases:** PostgreSQL, InfluxDB , Milvus Vector Store .
- **Version Control:** GitLab.
- **Core Skills:** Prompt Engineering, Agent Development (RAG Systems), People Management, Team Collaboration, Experimentation & Proof of Concept Development