Venkata Lokesh Panthangi

☎ +919573580571 ✉ lokeshpantangi@gmail.com LinkedIn GitHub 🌐 lokeshpanthangi

Education

MisogiAI By Masai - Bengaluru **Jun 2025 – Present**
AI Engineering

Vasireddy Venkatadri Institute of Technology - Guntur **Nov 2021 – Apr 2025**
B.Tech in Computer Science Engineering - Artificial Intelligence and Machine Learning

Technical Skills

Monitoring and Evaluation Tools: Grafana, LangSmith, Loki, Jaeger, Prometheus, RAGAS
Gen AI: AI Agents, n8n, LangChain, LangGraph, CrewAI
RAG: Traditional, Multimodal, Knowledge Graphs, Agentic RAG
LLM: Transformers, Fine-Tuning (Unsloth)
Machine Learning: Supervised, Unsupervised, Deep Learning
Databases: MySQL, SQL, Redis, Familiarity with GraphDB
Development: Python, React, FastAPI, Pydantic, Streamlit, Docker, AWS

Projects

JumppApp | *Python, FastAPI, Next.js (TypeScript), MongoDB, Gemini-2.0, LangChain* **Aug 2025**

- **Architected and Created** scalable AI workflows using Python, LangChain, and **Gemini-2.0**, **Integrated multilingual translation across 16+ languages**, reducing average **response latency to less than 2.5s** for global users.
- **Developed and deployed** an **AI-powered YouTube video-specific chatbot** with interactive Q&A features that **improved user engagement by 40%** in pilot testing with **500+ active sessions**.
- **Integrated frontend and backend systems**, resolving complex **Git merge conflicts** and **optimizing performance** across the **FastAPI–Next.js pipeline**.

Advanced RAG with Evaluation Pipeline | *Pinecone, AWS, LangSmith, Prometheus, Grafana, Jaeger***Nov 2025**

- Designed and deployed a **dual-EC2 architecture** — one for RAG inference and another for monitoring — eliminating resource contention and system conflicts.
- Integrated **OpenAI GPT-4 with Pinecone vector store** to build a retrieval-optimized RAG pipeline achieving **Precision: 80%** and **Recall: 90%** (k=10) on domain benchmarks.
- Implemented a complete **observability suite using Prometheus, Loki, Grafana, and Jaeger**, collecting real-time logs, traces, and metrics with **less than 150ms response latency**.
- Developed an **automated evaluation layer with LangSmith** for tracking prompt quality, token efficiency, and retrieval performance, enabling data-driven model iteration.

Experience

Full Stack Machine Learning Intern - BrainoVision **Jul 2024 – Dec 2024**
Hyderabad

- Designed and implemented a part of the Django application for real-time monitoring and management of 50+ pressure pumps across multiple locations and optimizing operational efficiency through self-initiated alerts and data analytics.
- Tracked and analyzed performance data from 50+ pumps, optimizing operational efficiency and reducing latency of the dashboard by 18% by modifying data retrieval algorithms.
- Generated automated notifications and reporting with Django Channels, cutting response times to critical issues by 30% and improving maintenance scheduling via an event-driven architecture.

Achievements

- Solved over 200+ LeetCode problems, showcasing strong algorithmic and coding proficiency.
- Wrote an Unofficial Expansion for the game RDR2, adding innovative gameplay elements.
- Conducted a Hackathon in CSM branch, fostering innovation and collaboration among participants.