Perumalla Surya Pranav

Hyderabad, India | suryapranavperumalla@gmail.com | +91 7780688133 | linkedin
github.com/surya-p7 | portfolio

Objective

AI Practitioner & ML Engineer with expertise in Generative AI, NLP, and MLOps. Skilled in bridging AI research, full-stack development, and DevOps to deliver scalable, production-ready solutions.

Experience

Intern, L&T Technology Services — Hyderabad May 2025 – Aug 2025

- Conducted research on IBM Granite 3.2 LLM and finalized its adoption for deterministic and reliable prompt responses, reducing hallucination cases by ∼25%.
- Designed and optimized prompt templates for summarization, abbreviation expansion, and QA tasks, improving contextual accuracy and response consistency.
- Integrated FAISS vector retrieval with LangChain orchestration for document understanding workflows.
- Contributed to dashboard development using Node.js and React, enabling visualization of extracted insights and LLM outputs for business teams.

Projects

AI Image Caption Generator GitHub | Live Demo

- Built a full-stack GenAI application to generate smart captions for uploaded images using Google's Gemini Vision Pro model.
- Developed FastAPI backend with secure `.env` key management and CI/CD integration.
- Deployed on Render with GitHub Actions auto-builds, achieving 95% uptime during testing.
- Implemented unit testing and coverage reports using `pytest` and `pytest-cov` for production readiness.
- **Tech Stack:** FastAPI, Gemini Vision Pro API, Python, React, GitHub Actions, Render.

Offline RFP Summarizer using GenAI GitHub

- Built an offline RAG pipeline with IBM Granite 3.2 via Ollama, using PDFMiner & PaddleOCR—**reduced RFP review time by 70%**.
- Optimized retrieval with FAISS & prompt tuning, adding abbreviation expansion & tagging—**boosted context precision by 25%**.
- Delivered a **production-ready, offline enterprise solution** with Python, FAISS, LangChain, Ollama, and YAML.
- **Tech Stack:** Python, FAISS, LangChain, Ollama, YAML.

Skills

Programming & Data: Python, C, Pandas, NumPy, SQL
Machine Learning & AI: TensorFlow, PyTorch (basic), MLflow, CV, NLP
LLMs & RAG: LangChain, Ollama, FAISS, Prompt Engineering
MLOps & Deployment: Docker, FastAPI, Git, GitHub Actions (CI/CD), Linux, pytest
Cloud Platforms: AWS, Render
Databases: MongoDB, MySQL, SQLite3
Visualization: Power BI, Tableau
Other: OOP, Design Patterns, Multi-threading

Education

B.Tech in CSE(DS) Vignan Institute of Technology and Science, Hyderabad 2022 – Present

Hackathons

VHack2K25 — Finalist (Top 10) Mar 18 – Mar 19, 2025