URVIL PANCHAL

Ahmedabad | 9016760721| urvilpanchal13@gmail.com | LinkedIn | Github | Portfolio

Work Experience

Junior AI Developer @ Eternal Web PVT. LTD. Apr 2025 - Present

- Hosted the open-source LLM (**Llama-3.2-11b**) locally using **Ollama.**
- Created a MCP server and connected it with client appllications like Claude Desktop and Cursor IDE.
- Developed the Backend services for mobile application using **FastAPI** and hosted it on **AWS**.

Data Science Trainee @ Nexuslink Services PVT. LTD. Jun 2024 – Jan 2025

- Developed an end-to-end object detection system using the yolo-v8, with deployment on **RunPod** serverless.
- Built a **RAG** system leveraging large language models (LLMs) for enhanced information retrieval. Created and deployed scalable APIs using **FastAPI** and containerized applications with **Docker**.
- Managed cloud infrastructure and deployments on **AWS**, optimizing for performance and cost, while handling version control and code management through **github** to streamline team collaboration.

Skills

Programming Languages: Python
Core Competencies: Deep Learning, Computer Vision, Natural Language Processing, Machine Learning
ML Frameworks: PyTorch, TensorFlow, Langchain, LangGraph, Scikit-Learn
Tools & Platforms: Git, Github, Docker, Amazon Web Services (AWS)
Additional Skills: LLM Fine-tuning, Ollama, Transformers, RAG Techniques, MCP Protocol, A2A Protocol

Projects

Agentic Chatbot

- Built an agentic chatbot integrating standard database APIs and a RAG pipeline with vector DB as **MCP tools.**
- Created two separate agents for fetching information from both standard database and vector database.
- Leveraged **Google's A2A** protocol to orchestrate these two agents into a single application.
- Created a streamlit UI for chat interface.

MCP Powered Chatbot | Github

- Developed a MCP Server for a local SQL database and weather api.
- Hosted it on **AWS EC2 Instance** and connected the it with clients applications like Claude Desktop and Cursor IDE using the **Proxy Server**.
- Created a separate custom client using **mcp-use**, facilitating robust client-server communication and seamless data exchange.

Fine-Tuned DeepSeek-R1 on Bhagwad_Geeta | Github

- Fine-tuned **DeepSeek-R1-Distil-Qwen-1.5B** on the Bhagwad Geeta using **LoRA**, optimizing for efficient adaptation with low-rank updates using parameter-efficient fine-tuning (**PEFT**) technique.
- Leveraged PEFT techniques to reduce memory usage while maintaining model performance.
- Developed a Streamlit-based chat UI, enabling interactive conversations.

LLM-Powered Document Q&A System

- Developed a **Retrieval-Augmented Generation (RAG)** application over private data, where documents were pre-ingested into a vector database using a **semantic chunking** strategy for improved retrieval quality.
- Used **GPT-OSS** with **Groq API** and **Qdrant Vector Database** for efficient storage and retrieval, enhanced by a **ColBERT reranker**.
- Used **Langchain** as the framework to seamlessly integrate the LLM, document retrieval, and user interaction for an interactive Q&A experience.

NSFW Detection System

- Trained a **YOLOv8** object detection model to detect explicit content by identifying adult imagery. Data was manually annotated to ensure high accuracy.
- Created a **Docker** application for the detection system, enabling seamless integration and scalability.
- Hosted the system on **RunPod**, providing efficient and scalable cloud-based detection for NSFW content.

Education

B.E in Information Technology • Gujarat Technological University 2020-2024