

URVIL PANCHAL

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

Work Experience

AI Developer @ Eternal Web PVT. LTD.

Apr 2025 - Present

- Created a **RAG Application** on AWS Eco System.
- Created a **MCP server** and connected it with client applications like Claude Desktop and Cursor IDE.
- Developed the Backend services for mobile application using **FastAPI** and hosted it on **AWS**.
- Migrated the legacy data from MySQL to Postgres database.

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.
 - Manually annotated the the image data to ensure high accuracy.
 - 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, Backend Development

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

Project Management System | Eternal Web

- A backend services for Project Management System (Mobile Application) enabling Admins and Project Managers to assign, track, and manage tasks for employees, with real-time notifications, file management, live location tracking, and integrated chat.
- Developed a scalable backend using **Python**, **FastAPI**, and **PostgreSQL**, and **role-based access control** for Admins, Project Managers, and Employees.
- Implemented a full task and subtask management module with automatic priority logic, multi-user task assignment, file attachments using **AWS S3**, and real-time notifications using **Firestore Cloud Messaging** for all task activities.
- Integrated **Google Maps API** for employee location tracking and built a **real-time chat system** to enable seamless communication between Project Managers and Employees.

Apex Tenders | Eternal Web

- A system designed to automate the extraction, cleaning, and processing of government tender data from PDFs and HTML files, and store it in a PostgreSQL database for analysis and reporting.
- Created parsing scripts using **pdfplumber** for PDFs and **BeautifulSoup** for HTML files to extract structured tender information.
- Implemented an **AWS S3 + Lambda** workflow where files uploaded to S3 are read by the Lambda function, parsed, and the extracted data is stored in the database via **FastAPI** APIs.
- Created FastAPI endpoints that interact with PostgreSQL using **psycopg2** to store, manage, and update all parsed tender data.

LLM-Powered Document Q&A System | Eternal Web

- 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.

Agentic Chatbot | Eternal Web

- 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.

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.

NSFW Detection System | Nexuslink Services

- 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.

Certifications

AWS Certified AI Practitioner | Verify Nov 2025- Nov 2028

Education

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