RIYA SHELKE

+91 84850 86457 riyasshelke824@gmail.com Kothrud, Pune, India
LinkedIn GitHub LeetCode

EDUCATION

BTech, Computer Science Engineering 2022 – 2026
Dr. Vishwanath Karad MIT-WPU, Pune CGPA: 7.43/10

EXPERIENCE

Tech Mahindra Jan 2026 – May 2026
AI/ML Engineer Intern — *i.Sustain – AI-Enabled ESG Intelligence Platform*

- Contributed to the development of an **AI-powered Hybrid RAG chatbot** for i.Sustain, retrieving real-time ESG data from multiple **PostgreSQL tables** for context-aware query handling.
- Implemented **prompt engineering and rule-based intent detection** to identify user intent and route queries between **text responses and structured chart data**, with responses generated using a **Groq-hosted LLaMA model**.
- Developed and integrated Python-based **FastAPI REST APIs** for the chatbot workflow, connecting ESG data retrieval, LLM processing, and response handling with the platform's backend.

SKILLS

Languages	Python, C++, SQL
AI & Generative AI	Machine Learning, Deep Learning, NLP, RAG, LLMs, MCP, Prompt Engineering
Libraries & Frameworks	Pandas, NumPy, Scikit-learn, Matplotlib, Hugging Face, FastAPI, Streamlit
Backend & Databases	REST APIs, PostgreSQL, MySQL, Docker
Core Concepts	Data Structures & Algorithms, Operating Systems, DBMS, OOP

PROJECTS

Smart Meeting Summariser

- Built an AI system to convert meeting audio and video into transcripts using **OpenAI Whisper ASR**.
- Used **PEGASUS** and **BART** from Hugging Face for concise meeting summaries.
- Applied **BERT-based NER and rule-based logic** to extract action items, deadlines, and assignees via **Streamlit**.

Corn Disease Detection – Multimodal Deep Learning System

- Designed a **CNN–LSTM multimodal deep learning model** for early detection of corn plant diseases from leaf images.
- Deployed the system using **Flask**, allowing users to upload leaf images and receive real-time disease predictions.
- Trained the model on **3,000+ labeled agricultural images** and achieved **92% classification accuracy**.

RESEARCH WORK

A Lightweight Framework for Prompt Redundancy Removal and Semantic Compression in Open-Source Language Models *Springer – ETMIS 2025 Best Paper Award*

- Proposed a lightweight NLP-based semantic compression framework to reduce redundant LLM prompt tokens, achieving approximately **30% prompt reduction** with less than **2% accuracy loss** and **15–20% lower inference latency**.

Enhancing Model Context Protocol (MCP) Performance via Built-In Intelligent Caching

- Proposed an in-built intelligent caching layer for MCP servers, achieving an **88.89% reduction in execution time** through optimized tool-call reuse.

CERTIFICATIONS [Certificates]

- IBM Data Science Professional Certificate – Coursera
- Applied Generative AI – Infosys Springboard
- Machine Learning & Deep Learning – NPTEL