Soham Walimbe

Electronics and Communication Engineer with AIML

sohamwalimbe@gmail.com | 7083541780 | Pune | Portfolio | LinkedIn | Github

Summary

GenAI builder with 9 months of R&D experience at Tech Mahindra's Makers Lab prototyping Agentic RAG pipelines, LLM orchestration workflows, and self-evolving knowledge base PoCs for 3+ client applications. Shipped production-grade systems reducing manual effort by 80% and improving sprint delivery by 20%. IEEE-published researcher with hands-on expertise in RAG, embeddings, vector databases, prompt engineering, and rapid PoC development under senior mentorship.

Education

MIT World Peace University — B Tech • CGPA- 7.09
Electronics and Communications: AI/ML — Pune • 2022 - 2026

Vishwakarma College of Arts Commerce and Science — HSC • 68.67%
Pune • 2020 - 2022

Sinhgad Spring Dale Public School — CBSE • 84%
Pune

Experience

Tech Mahindra Ltd. — Hinjewadi, Pune
AI ML Project Intern — April 2025 - December 2025

- Rapidly prototyped GenAI PoCs and demo flows for 3+ client applications under senior mentorship in an R&D innovation lab
- Built Agentic RAG frameworks with dynamic context assembly, retrieval strategies, and LLM generation for context-aware client responses
- Designed and deployed modular AI task orchestration pipelines chaining LLM calls, retrieval, and generation with structured outputs and retry logic
- Led computer vision projects automating routine client tasks, reducing manual effort by 80%
- Implemented REST APIs using Django, MongoDB, and SQL for scalable backend services powering AI features Optimized development workflows by integrating AI-assisted code reviewers, increasing sprint delivery efficiency by 20%
- Authored technical documentation for AI-powered workflows, including prompt engineering strategies and orchestration logic for knowledge transfer and scalability
- Streamlined R&D-to-production transitions using Git and CI/CD workflows with end-to-end testing and debugging

Wordlab Multilingual Translation Enterprises — Pune
Data Annotator Intern — October 2022 - July 2023

- Automated data extraction pipelines using Python, incorporating data validation and preprocessing techniques to process 15+ daily sources, reducing manual workload by 90%
- Improved LLM performance using RLHF, prompt engineering, and evaluation metrics for NLP tasks
- Ensured high-quality annotated datasets for model training and evaluation

Projects

IoTide- Underwater Image Recognition and Marine Life Conservation Project
Underwater marine life detection using YOLOv8 with real-time fish recognition and ML analytics dashboard. ESP32-CAM integration for live underwater streaming with multi-sensor water quality monitoring. Runner-up at HACK MIT WPU 2025.
https://ieeexplore.ieee.org/document/11278148

AI Ops platform
Real-time anomaly detection (Isolation Forest) with LLM-based incident explanation via Ollama. Scalable FastAPI backend for telemetry ingestion, Apache Kafka for decoupled streaming, and React dashboard for live monitoring. Demonstrates end-to-end ML pipeline from data ingestion to human-readable output.

Python, scikit-learn, Ollama, FastAPI, Kafka, React

PDF Detail extraction Python Library
Python library for structured extraction from PDFs: text, tables, and image-based graphs. Hybrid rule based and AI-based approach using PyMuPDF, Camelot, and PDFParser. Converts image-based graphs to digital plots with AI post-processing for structured JSON output. *Python, PyMuPDF, Camelot, PDFParser, AI Post-Processing*

LLM-Assisted COREP Reporting Prototype
Self-evolving knowledge base PoC: ingests regulatory text into ChromaDB using LLM embeddings, retrieves relevant rules via vector similarity search, and generates structured JSON output aligned to COREP schema through a local LLM (Ollama). Maps directly to production RAG-over- documents architectures.

Python, ChromaDB, Ollama, LLM Embeddings, JSON Schema, Vector Search

Skills

GenAI & LLM
Agentic RAG, Prompt Engineering, LLM APIs (Gemini & Ollama), LangChain, LLM Orchestration, Vector Databases

Machine Learning
DNNs, YOLO, PyTorch, scikit-learn, Image Processing, Pattern Recognition

Software & Backend
Python, OOP, FastAPI, Django, REST APIs, Webhooks, MongoDB, SQL

Data & Systems
Pandas, NumPy, Data Pipelines, Data and API Integration, Kafka

Tools and Workflow
Git, CI/CD, Jupyter Notebooks, n8n Automation, Cursor IDE, Antigravity

Cloud
AWS Academy (Cloud Foundations); Azure-adjacent

Certifications

AWS Academy Graduate - Cloud Foundations 10 April 2026
AWS Academy

Introduction to LLM
Swayam NPTEL

National Academic Immersion Program in Full Stack Development 24 June 2024 - 3 July 2024
IIT Jodhpur

MathWorks Matlab Onramp Course and Signal Processing Onramp Course 28 November 2023
Mathworks

Awards

HACK MIT WPU 2025 Runner Ups for Ideathon and Workathon Project Iotide 28 March 2025
MIT WPU

Publications

IoTide- Underwater Image Recognition and Marine Life Conservation – Project
IEEE
https://ieeexplore.ieee.org/document/11278148

Languages

English **Hindi** **Marathi**