Vedant Rajpurohit

+91 9979508828 / vedantrajpurohit3907@gmail.com | LinkedIn | Github| Huggingface | Portfolio

Experience

WebOsmotic Private Limited — Surat, India

AI / ML Engineer — *Jan 2025 – Present*

- Built 6 production AI systems from scratch spanning Deep Learning, Agentic AI, RAG pipelines, and Voice & Speech owning the full AI architecture from zero across all projects.
- Mentored and managed a team of 3 AI interns, conducting regular code reviews, guiding system design decisions, and overseeing delivery across active projects.
- Fine-tuned and evaluated LLMs for FAQ generation using SFT and DPO training strategies, Iterating on model behaviour, alignment, and response quality for a client-facing production use case.
- Designed and developed Python FastAPI backends for all AI projects, handling real-time data flows, WebSocket communication, REST APIs, and integration layers between AI models and client-facing products.

Rocket.new — Surat, India

AI / ML Engineer Intern — *Dec 2023 – Jul 2024*

- Delivered 5 end-to-end AI products across 3 × 48-hour hackathons - co-led a 3-person team to ship an AI Ticketing System with auto-generated tickets from Mail, Discord, and YouTube, and a Recruitment Platform with GitHub and resume scoring, automated email communication, and round/position management.
- Built a Conversational AI Assistant (LlamaIndex + Pinecone + OpenAI) with auto-drafting of responses from a vector-indexed document store; integrated ticket management across Mail, Discord, and YouTube.
- Independently shipped an LLM Evaluation & Testing Platform: concurrent multi-model benchmarking from a single prompt with streaming output, GPT-4 judging, per-model parameter control, and prompt versioning.
- Developed a VS Code extension that auto-generates acceptance criteria and execution plans from PDFs, images, web links, and GitHub codebases using LangChain + LlamaIndex loaders.

P.P. Savani University — Surat, India

Teaching Assistant — *Jan 2023 – May 2023*

- Instructed 40+ first and second-year B.Tech AIML students in Python, SQL, and core ML concepts; designed practical lab sessions integrating emerging AI techniques to bridge theory and implementation.

IBM SkillBuild Edunet — Surat, India

Computer Vision Intern — *Sep 2022 – Nov 2022*

- Developed a real-time hand sign language recognition system covering ASL alphabets, Hindi Varnmala, and numbers with integrated voice modulation using MobileNet + MediaPipe.
- Built and published a custom annotated dataset on Kaggle; achieved high accuracy across 3 language sets with an interactive dashboard for real-time feedback.

Projects

Guidy AI | *Python, FastAPI, Milvus, MediaPipe, RAG, Langchain* — Jan 2025 – Present

- Engineered a 5-pipeline agentic AI orchestrator, LLM intent router classifies every message and dispatches to specialised agents for Website RAG Q&A, UI accessibility commands, URL redirect, form auto-fill, and on-page scroll, embeddable as a chat widget on any website.
- Built voice navigation and MediaPipe hand gesture control as primary inputs for users with visual or motor disabilities speak or gesture to trigger 50+ WCAG commands (dark mode, contrast, magnification), navigate pages, or auto-fill forms by describing values aloud, without keyboard or mouse.
- Designed a stateful SSE layer that parses the LLM's JSON token stream and emits field-by-field events (intent, commands, form values, URL) the moment each value completes frontend executes UI toggles and redirects before text finishes generating.

Recruitment Smart | *Python, FastAPI, C# .NET, OpenCV, STT, Mediapipe,* Mar 2025 – Dec 2025

- Helped build a Microsoft Teams bot using Graph Communications API to capture real-time participant audio/video streams for live interview analysis with per-speaker attribution.
- Built the behavioral analysis pipeline scoring candidates across 5 dimensions (Suspiciousness, Positivity, Confidence, Stress, Focus) - video detects facial emotions, head pose, custom eye gaze tracking, and smile duration; audio extracts pitch, jitter, voice steadiness, and voice-break patterns; transcript measures sentence-level sentiment and filler word frequency.
- Built a FastAPI WebSocket backend for concurrent live interviews, enabling real-time per-speaker transcription and live VISTA classification that instantly tags interviewer questions into Vision, Interest, Salary, Technology, or Availability—for example, "When can you join?" is marked as Availability and synced live to all recruiters in the meeting.

AssureAI | *Python, FastAPI, LLM Agents, RAG, AWS Bedrock,* Feb 2025 – Present

- Built a Python backend platform for UK Local Authorities automating Children's Social Care Stage 2 statutory complaint investigation workflows, covering case setup, document classification, multi-document ingestion, and end-to-end report generation.
- Engineered a large-document processing pipeline for 2,000+ page statutory case bundles, converting raw PDFs into structured markdown to reduce token overhead while preserving evidence integrity across complaint points.
- Designed an agentic case-memory and report-generation system with structured case state, evidence retrieval, legislation lookups, and automated QA for coverage and compliance.

Technical Skills

Languages: Python, SQL

Backend: FastAPI, Node.js, WebSocket, REST APIs

AI/LLM Frameworks: LangChain, LlamaIndex, CrewAI, Hugging Face, Transformers, Unsloth, Claude code

Fine-Tuning: SFT, DPO, LoRA

ML/DL: TensorFlow, MediaPipe, OpenCV, RoBERTa, MobileNet, Deep Learning, Computer Vision

Cloud & DevOps: AWS (Bedrock, S3, EC2, Lambda, CloudWatch), Docker, Git, CI/CD Pipelines

Databases: PostgreSQL, Redis, Milvus (Vector DB), Pinecone

Education

P.P. Savani University Surat, India

Bachelor of Technology in Computer Science Specialization AI & ML (9.25/10) *Oct 2020 – Apr 2024*

Opensource Contribution

Llama 3.1 8B for Text-to-SQL

- Fine-tuned Llama 3.1 8B for Text-to-SQL generation using LoRA + SFT; ranked top 3 in Text-to-SQL category on Hugging Face with ~4k downloads, adopted by 3 inference providers (TensorBlock, mradermacher, Featherless AI) and quantized by the community for llama.cpp, Ollama, and LM Studio.

Llama 3.2 1B for persona classification and Llama 3.2 0.2B Prompt Generator model

- Fine-tuned Llama 3.2 1B and 0.2B using LoRA + SFT; published with LoRA adapters for community use.

Sign Language & Linguistics Datasets

- Published 4 open-source datasets covering Numbers Sign Language, Hindi Sign Language, Gujarati Sign Language, and Gujarati POS Tags totalling 38K+ samples across linguistics and computer vision domains.

Certifications

Deep Learning Fundamentals (Cognitive) Sep 2022

Machine Learning Specialization (DeepLearning.AI) Jun 2023

AWS Cloud Foundations (AWS) Jul 2023

IBM Data Science Professional Certificate (IBM) Feb 2024

Google Data Analytics Professional Certificate (Google) Feb 2024

Achievement

Published research paper: "Human–Computer Interaction with Detection of Hand Gesture to Improve Artistic Creativity" Jul 2024

Azure AZ-900 Fundamentals (Microsoft) Jul 2022