ADITYA JETHANI

+91 9328223890 | Surat, Gujarat, India
adityajethani11@gmail.com | LinkedIn | Portfolio | GitHub

SUMMARY

Cloud AI/ML engineer with 1.5+ years shipping production AI and data solutions for customers across legal-tech, healthcare, and e-commerce. Combine deep-learning and data-pipeline engineering with customer-facing delivery, turning ambiguous business problems into measurable, deployed outcomes.

EDUCATION

Pandit Deendayal Energy University — Gandhinagar, India
B.Tech in Computer Engineering, **CGPA: 9.44/10** — *August 2021 – May 2025*
Coursework: Data Structures & Algorithms, OS, DBMS, Machine Learning, NLP, Probability & Statistics, Big Data Analytics

TECHNICAL SKILLS

- **Languages:** Python, SQL, C++, Go, JavaScript, Bash
- **ML & AI:** PyTorch, TensorFlow, scikit-learn, LLM Fine-tuning (QLoRA), Agentic RAG, LangChain, OpenCV
- **Cloud & Data:** GCP (Vertex AI, BigQuery, Dataflow), AWS (EC2, S3, EMR), Docker, ETL/ELT Pipelines, FastAPI, PostgreSQL, MongoDB, Qdrant
- **Solutions & Delivery:** Distributed/GPU Training, Data Visualization & Dashboards, MLOps, Technical Presentations

EXPERIENCE

AI/ML Solutions Engineer — *December 2025 – Present*
Logicwind Technologies — Surat, India

- Engineered an agentic **NL-to-SQL voice interface** converting natural-language and voice queries into governed SQL with auto-generated chart dashboards in **under 2s**, giving non-technical stakeholders self-serve analytics
- Architected an **Agentic RAG** platform over **360K** multilingual documents (45% scanned, Surya OCR), reaching **4.3x** throughput via async batching and Qwen3 embeddings; benchmarked 16+ chunking strategies and 6+ embedding models
- Delivered production AI on **OpenAI Agents SDK, Google ADK, and Mastra AI** with LangFuse observability and Docker, cutting client agent-delivery time **70%**

AI/ML Engineer — *June 2025 – December 2025*
CybraneX Technologies — Delhi, India

- Engineered a high-throughput **RAG inference pipeline** for QLoRA fine-tuned LLMs, achieving **9x speedup** (latency **110s** → **12.2s**, 139.6 tok/s) via GPU-batched inference, FP16 quantization, KV-cache, and FlashAttention-2
- Built a clinical **tongue-analysis REST API** and interactive UI on an OpenCV pipeline, enabling scalable deployment of AI diagnostics into customer clinical workflows

Software Engineering Intern — *January 2025 – June 2025*
Yanolja Co. Ltd. — Surat, India

- Architected an **image-to-image translation** system for multilingual e-commerce that lifted accuracy **50%** over the Cycle-GAN baseline while preserving native typography

PROJECTS

Grade Flow | *FastAPI, LangChain, LangGraph, pgvector, PostgreSQL*

- Automated **90%** of an exam-grading workflow with plagiarism detection and auto paper generation, cutting manual evaluation time **75%**

Prompt Detective | *TypeScript, OpenCV, CLIP, pgvector, Sentence-Transformers*

- Built a multimodal pipeline inferring generation prompts from images and videos, reaching **85% Top-K accuracy** at **under 5s** per 1-minute video

LLM-TTS Suite | *PyTorch, HuggingFace, FAISS, QLoRA, Coqui TTS*

- Fine-tuned LLMs with FAISS dense retrieval and GPU-optimized TTS over 15+ textbooks; **91% Recall@K**, **under 2s** latency, **under 1hr** training on A100

ACHIEVEMENTS & LEADERSHIP

- **Google Solution Challenge 2025:** Top 105 of **4,000+** international teams | Finalist, JPMorgan Code for Good 2024; Smart India Hackathon 2023
- **Leadership & Communication:** Chair & Advisor, ACM-PDEU Student Chapter (2023–2025) — led technical workshops & talks; **15+** merged open-source PRs (Hacktoberfest, GSSoC)
- **Competitive Programming:** LeetCode Knight (Rating **2071**, Top 1.75%), AIR 61 in Weekly Contest 462 | **Certifications:** NVIDIA Deep Learning Institute, IEEE CIS & ACM Summer Schools