ADITYA JETHANI

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

PROFESSIONAL SUMMARY

Researcher interested in evaluation, faithfulness, and failure analysis of large language models in retrieval-augmented and multilingual settings. 2 years of production NLP and LLM systems experience, oriented consistently toward diagnosing why systems fail rather than demonstrating that they work. GATE CS 2026, 95.61 percentile. B.Tech CGPA 9.44/10, PDEU.

EDUCATION

Pandit Deendayal Energy University — Gandhinagar, India
B.Tech in Computer Engineering, **CGPA: 9.44/10** — *August 2021 - May 2025*

Relevant Coursework: Data Structures & Algorithms, Design & Analysis of Algorithms, Object-Oriented Programming, Operating Systems, Database Management Systems, Artificial Intelligence, Machine Learning, Natural Language Processing, Statistics, Big Data Analytics

Shree Swaminarayan Academy — Surat, Gujarat, India
Higher Secondary Certificate (HSC), **Score: 97.3%** — *April 2019 - July 2021*
Relevant Coursework: Applied Mathematics, Python Programming, SQL

TECHNICAL SKILLS

- **Programming Languages:** Python, C++, Go, JavaScript, SQL, MATLAB, Bash
- **ML/AI Frameworks:** TensorFlow, PyTorch, scikit-learn, LangChain, LangGraph, Transformers, Keras
- **Web Frameworks:** React.js, Node.js, Express.js, Flask, Django, FastAPI, Next.js
- **Cloud & DevOps:** AWS (EC2, S3, EMR), GCP (BigQuery, Vertex AI, Compute Engine, DataFlow), Docker, Git, Linux
- **Databases:** MySQL, MongoDB, ChromaDB, FAISS, pgvector
- **Developer Tools:** OpenCV, Selenium, Postman, Swagger, VS Code, Uvicorn, Alembic, FFmpeg, Pydantic
- **Specializations:** Computer Vision, Natural Language Processing, RAG Systems, LLM Fine-tuning, Time Series Analysis

PROFESSIONAL EXPERIENCE

Machine Learning Engineer — *December 2025 – Present*
Logicwind Technologies — Surat, India

- Built a production multi-agent AI system using the OpenAI Agent SDK with LangFuse for observability, traceability, and prompt versioning across concurrent agent workflows, **reducing manual prompt optimisation time by 60%** and achieving 90% user satisfaction on chatbot responses.
- Designed and evaluated **8+ chunking strategies and 5+ embedding models** for NCOSE-aligned document retrieval pipeline, systematically analysing retrieval consistency degradation across query types and input distributions
- Implemented blocklisted organisation detection at **92% accuracy using fuzzy matching and regex**, and built automated vector database cleaning scripts achieving 99.95% accuracy **across 145K chunks indexed from 8K website URLs**, with AWS S3 fallback for pipeline resilience

Machine Learning Engineer — *June 2025 - December 2025*
CybraneX Technologies — Delhi, India

- Engineered a high-throughput RAG inference pipeline for fine-tuned LLMs (QLoRA), **achieving 9 times parallel speedup** by replacing serialised per-thread generation with GPU-batched inference, **reducing latency from 110s to 12.2s**

- Optimized runtime efficiency through **FP16 quantization, KV-cache, flash attention**, and context truncation, sustaining **139.6 tokens/sec generation speed and 0.82 queries/sec throughput** in production-scale workloads
- Implemented model safety and evaluation tooling **surfacing 27% more quality violations** through statistical drift detection and structured observability

Software Engineering Intern *January 2025 - June 2025*
Yanolja Co. Ltd. Surat, Gujarat, India

- Architected **image-to-image translation system for e-commerce** that elevated accuracy by **50% over baseline**, translating product images while preserving native language typography and visual aesthetics
- Revamped Yanolja's internal **LLM-based code analysis workflow** by mining commit histories and PR diffs, mitigating **27% of security vulnerabilities** through automated RAG-based pattern detection
- Engineered scalable **computer vision pipelines using Go, Python, and FFmpeg**, seamlessly handling large image datasets and reducing average processing time by 1.3x

AI Engineering Intern *May 2024 - October 2024*
CybraneX Technologies Delhi, India

- Engineered **AI-based telemarketing system processing 50,000+ daily calls**, harnessing custom LLM models to achieve **89% accuracy in legitimate lead identification** with sub-200ms latency
- Developed **real-time analytics dashboard** integrating purchase history and geolocation data, boosting client's **inventory turnover rate by 22%** through actionable insights
- Accelerated healthcare document processing workflows by consolidating pipeline for PDF extraction, reducing processing time from 5s to 0.5s per document through batch optimization

PROJECTS

Neural ODE for Heart Rate | *Python, PyTorch, torchdiffeq, NumPy, Pandas, Matplotlib, Jupyter, MIMIC-III*

- Built a continuous-time ODE pipeline for irregularly sampled ICU heart-rate trajectories (MIMIC-III), integrating masked-point imputation and short-horizon forecasting. Achieved **MSE = 0.44263** on held-out masked points (**RMSE $\approx$ 7.01 bpm**), demonstrating accurate heart-rate reconstruction/prediction in a clinically realistic missing-data setting.

LLM-TTS Suite | *Python, PyTorch, Transformers, FAISS, RAG, CUDA, A100, TTS Inference*

- Engineered an integrated LLM fine-tuning with FAISS-based retrieval and GPU-optimized TTS pipeline over **15+ curriculum-level textbooks**. Achieved **more than 90% Recall@K and Hit@K**, **<2 s** TTS inference latency, and **less than 1 hour** fine-tuning time on an **NVIDIA A100**, demonstrating efficient, high-accuracy retrieval-conditioned speech generation.

TalentScout AI | *Python, NLP, Transformers, LangChain, LangGraph, Groq Cloud*

- Delivered production-ready **candidate validation system with 85% accuracy** through NLP pipeline with real-time processing. Reduced hiring **evaluation time by 60%** with automated question generation

Prompt Detective | *Python, TypeScript, Computer Vision, Multimodal AI, Semantic Retrieval, PostgreSQL/PLpgSQL*

- Built an advanced multimodal reverse-engineering pipeline to infer original generation prompts from images/videos, combining CV-based signal extraction. Achieved **>85%** prompt extraction accuracy (Top-K and semantic similarity) and **<5 s** processing time per **1-minute** video, enabling fast and reliable prompt reconstruction at scale.

Grade Flow | *Python, FastAPI, pgvector, FAISS, LangChain, LangGraph, SQL*

- Automated **90% of grading workflow**, reducing test evaluation time by **75%** by engineering robust platform for test paper submission, plagiarism detection algorithms, and automated question paper generation

Clever Query Bot | *Python, LangChain, FAISS, Hugging Face, Keras-tuner, FastAPI*

- Built a lightweight **CSV analytics assistant** converting natural language queries into SQL with **92% accuracy on 300+ datasets**, cutting manual reporting time by 80%

RESEARCH & PUBLICATIONS

Advanced Forecasting of Solar Power Generation Using Bi-LSTM with Fourier Features

- Built Bi-LSTM with Fourier feature models for solar power forecasting, achieving **R^2 = 0.86, RMSE = 128.09 KW** on May–June 2020 datasets. Incorporated geo-temporal and temperature features to enhance predictive accuracy for daily AC power generation. *(Under Review)*

A Comprehensive Approach for Efficient Labelling of Hinglish Dataset and Hate Speech Classification

- Curated a code-mixed Hinglish dataset with a consensus-based labeling pipeline using RoBERTa, mBERT, and VADER. Applied active learning with RLHF feedback to resolve annotator disagreement, achieving **86% test accuracy** for low-resource hate-speech classification. *(Manuscript in Preparation)*

ACHIEVEMENTS

- **Competitive Programming:** LeetCode Knight (Max Rating: 2071, Top 1.75%), 190+ problems solved, AIR 61 in Weekly Contest 462 among 35K+ participants
- **CodeChef:** Secured a rank of 936 in Starters 113D (Top 5% among 19K participants)
- **Kaggle Contributor:** Ranked 278 in Season 4, Episode 8
- **Google Solution Challenge 2025:** Top 105 among 4000+ international teams
- **Major Hackathons:** Finalist in JPMorgan Code for Good 2024, Smart India Hackathon 2023, ByteVerse 2025 (IIT Patna)
- **Open Source Contributions:** Merged 15+ production-quality Pull Requests in Hacktoberfest 2024, GSSOC 2024 and 2023
- **Leadership:** Advisor and Chair at ACM-PDEU Student Chapter, *July 2023 - June 2025*

CERTIFICATIONS

- **NVIDIA Deep Learning Institute:** Fundamentals of Deep Learning Certificate
- **IEEE CIS Summer School:** Attended and Presented comprehensive project work at IIT Indore
- **ACM Summer School:** Presented projects at summer schools at IIT Gandhinagar, NIT Goa, IIT Patna
- **Machine Learning with Python:** Fundamental to Advanced concepts of machine learning by freecodecamp
- **Generative AI:** Introduction to Prompt Engineering for Generative AI
- **Git and Github**: Fundamentals of version control and project management
- **Hackathon Certifications:** Samsung Solve for Tomorrow, DotSlash 6.0, DotSlash 7.0, Flipkart GRID 5.0

DOMAIN-SPECIFIC PROJECTS

Healthcare AI & Computer Vision

- **Med Genie** | *Python, FastAPI, TensorFlow, OpenCV, React*: Built comprehensive healthcare solution with AI-powered symptom analysis and medical recommendation system, achieving 87% diagnostic accuracy
- **HerbiSense** | *Python, CNN, OpenCV, Flask*: Developed real-time medicinal plant identification system with 94% accuracy on 10K+ plant images, integrated chatbot for guidance reducing consultation time by 60%
- **BlinkWink** | *Python, OpenCV, MediaPipe, TensorFlow*: Created real-time drowsiness detection system using eye-aspect ratio analysis, achieving 96% accuracy with sub-100ms latency for workplace safety monitoring

Educational Technology & Automation

- **Meeting Assistant** | *Python, Transformers, Whisper, LangChain*: Developed AI-powered conference summarization system with speaker diarization and key-point extraction, reducing meeting review time by 75%
- **TimeFy** | *Python, Genetic Algorithms, Django, PostgreSQL*: Engineered automated timetable generation system for universities, reducing manual scheduling effort by 85% while handling 200+ course-faculty constraints
- **Algorithm Visualizer** | *JavaScript, D3.js, React, Node.js*: Built interactive visualization platform for 15+ sorting and graph algorithms