Harsh Ahir

AI Engineer | LLM Applications, RAG, Evals, Speech
Valsad, India | harsh@datamace.in
harsh.datamace.in | linkedin.com/in/harsh-ahir-336449229 | github.com/Ahir7

AI engineer building LLM applications in Python, LangGraph, and FastAPI: agentic workflows, RAG, LLM evaluation, and guardrails. Fine-tuned F5-TTS for Gujarati speech (4.2% CER, 4.21/5 MOS).

Experience

Software Engineer | Contrl, Inc. Apr 2026 – Present
Remote, India

- Prototyped an agentic LangGraph workflow turning contracts into policy rules, with guardrails (compiler checks, bounded LLM repair, human-in-the-loop approval) and 20-document evals (precision, recall, latency).
- Shipped a production vision-model pipeline that classifies client documents and extracts their data, with human review and resumable ERP loading.
- Own the production FastAPI ERP backend: PostgreSQL schema, tenant isolation, auth, and audit logging.
- Set up GitHub Actions CI on self-hosted runners: database-backed tests, SAST, SBOM, and image publishing.

AI/ML Engineer Intern | Linde Engineering India Pvt. Ltd. Jan 2026 – Mar 2026
Vadodara, Gujarat

- Built a LangGraph workflow that maps English/German questions on industrial time-series data to Elasticsearch DSL or KQL: it retrieves schemas and examples (RAG), then generates, validates, and executes the query.
- Added template routing for high-confidence English queries to bypass LLM inference, with rule-based fallbacks.

Software Developer | Odoo Oct 2025 – Dec 2025
Gandhinagar, Gujarat

- Fixed Python backend bugs and developed custom modules with the Odoo core team.

Selected Projects

Gujarati Text-to-Speech | *PyTorch, F5-TTS, Flask*

- Fine-tuned F5-TTS on 3,300 Gujarati audio-text pairs (about 10 hours, IIT Madras corpus) with audio/text preprocessing and an 80/10/10 split; integrated the model into a Flask speech-generation app.
- Measured 4.2% CER on 330 held-out utterances and a 4.21/5 mean opinion score (MOS) from 15 native listeners.

Manuscript Reformatting Pipeline | *FastAPI, Ollama, GROBID | github.com/Ahir7/giga-papers/tree/new*

- Built the backend of a team project that turns journal guidelines into validated rules with a local LLM, then applies 10 deterministic document transformations; extracted metadata and checked reference DOIs through Crossref.

SynthMed: Medical Report Analysis | *FastAPI, MedGemma*

- Built the Android app and backend for a team prototype using locally served MedGemma; added OCR retries and fallback processing, plus unit-of-measure, drug-interaction, and guideline-threshold checks on model requests.

Meeting Monitor AI | *WhisperX, GLiNER, Gemini, FastAPI*

- Wrote the live audio backend of a team project: WhisperX transcription and speaker diarization, GLiNER entity extraction, and Gemini-generated sales cues, with filters for silence and likely transcription hallucinations.

Image-Based 3D Object Measurement | *PyTorch, COLMAP, FastAPI*

- Co-developed an academic prototype combining multi-view reconstruction, monocular depth, and scale recovery to estimate object dimensions; evaluated results against manual measurements of real objects.

Technical Skills

Languages: Python, SQL, Go
GenAI: LangGraph, RAG, vector search, prompt engineering, LLM evaluation, vLLM, Ollama, Gemini, Azure AI
ML: PyTorch, TTS fine-tuning, ASR (WhisperX), speaker diarization, NLP, OCR, computer vision, NumPy, Pandas
Backend: FastAPI, REST APIs, PostgreSQL, Elasticsearch, Docker, GitHub Actions, Azure, GCP, Cloudflare

Education

Charotar University of Science and Technology (CHARUSAT) May 2026
B.Tech. in Artificial Intelligence and Machine Learning, CGPA 8.44/10 Anand, Gujarat

Achievements

Winner, Intuitive.AI x DA-IICT Hackathon 2026; 3rd Runner-Up, Odoo x CHARUSAT Hackathon 2025; Finalist, Gujarat Robofest 3.0 (Jetson Nano rover navigation).