Harshal Bhambhani

bhambhaniharshal@gmail.com +91 8561972727 Bangalore, Karnataka
https://www.linkedin.com/in/harshal-bhambhani-hhbb24 github.com/Harshaaalll
My Portfolio

Professional Summary

AI engineer building production AI systems across voice, language and documents. Has shipped a live Gemini Live voice agent on Pipecat, fine-tuned transformer models to 98.6% recall, and quantised a 3B-parameter model from 6GB to 2GB for CPU-only hardware. Builds multi-agent systems with human approval gates and measures them honestly, from Word Error Rate on low-resource audio to about 1,300 checks on a live agent.

Skills

AI & Agents: Gemini Live, Pipecat, LangGraph, LangChain, Tool Calling, RAG, Fine-Tuning, Quantisation
ML & NLP: PyTorch, Scikit-learn, XGBoost, Transformers, Whisper, OCR, OpenCV, A/B Testing, Evals
Languages & Tools: Python, SQL, Git, FastAPI, PostgreSQL, Pandas, Playwright, Odoo, dbt, Power BI
More: Flutter, Firebase, Dataiku DSS, Streamlit; coursework in Docker and AWS (S3)

Work Experience

AI Engineer - OTO Capital Aug 2026 – Present
- Built a production Gemini Live voice agent on Pipecat, serving 5 client deployments across 10 languages.
- Moved negotiation rules into a deterministic Python gate that approves or refuses the model's tool calls.
- Evaluated every model change against about 1,300 checks across five suites and a 500-call simulator.
- Measured agent variants through A/B control arms, isolating behaviour shifts before a production release.
- Compressed per-turn prompt cost 79% to 3,490 tokens by moving situational context into tool responses.

Data & Analytics Consultant - Goel Exports July 2026 – Aug 2026
- Mapped an exporter's manual order-to-cash process into a documented, 4-phase Odoo 19 workflow model.
- Configured field-level data rules across 6 process stages, agreeing each with the operations owners.

Data Science Intern - Standard Chartered Global Business Services Pvt. Ltd. Jul 2025 – Dec 2025
- Deployed a production NLP pipeline on Dataiku DSS for 100+ page contracts, cutting audit time by 90%.
- Fine-tuned DistilBERT under a weighted loss, reaching 98.6% recall on rare clauses in noisy contracts.
- Calibrated confidence thresholds on model output, filtering 95%+ of false positives before human review.
- Adapted SBERT embeddings for semantic clause retrieval, ranking candidate matches by cosine similarity.

Machine Learning Engineer Intern - DG Liger Consulting Jun 2024 – Jul 2024
- Constructed a LangChain RAG assistant over 30+ client PDFs, answering business questions in plain language.
- Indexed MiniLM embeddings into a FAISS store, returning grounded answers in under a second per query.
- Quantised a 3B-parameter LLM to 4-bit GGUF, shrinking it from 6GB to 2GB for CPU-only client hardware.

Technical Projects

Sanwaad — Multi-Agent Complaint Resolution Sep 2026
- Built a LangGraph case pipeline with hybrid RAG, where code validates every fix and a person approves payouts.
- Passed 15 of 15 offline trajectory scenarios with 0 safety violations and reached 1.00 recall@5 in retrieval.

OpsAgent — Autonomous Browser Operator 2026
- Built a plan-act-verify agent over a real browser with Playwright and Gemini, gated by approval policy in code.
- Verified results with a no-LLM check of the ERP, passing 11 of 12 evaluation scenarios at about $0.03 per run.

Medical Bill OCR — Computer Vision & Vision-Language Extraction Nov 2025
- Processed scanned medical bills via PaddleOCR with CLAHE and 4x DPI upscaling, at 90%+ extraction accuracy.
- Engineered schema-constrained Qwen 7B prompts that turned noisy OCR text into consistent JSON records.

AutoTransFlow — Document Layout Detection & Multilingual Translation Aug 2025 – Nov 2025
- Detected page layout blocks with a doc-layout-yolo object detector at a recall-first 0.25 confidence cutoff.
- Designed an on-premise translation pipeline across 200 languages under a zero-external-API data policy.

Multilingual ASR System — Speech Recognition & Error Correction Aug 2024 – Dec 2024
- Transcribed low-resource Urdu speech with zero-shot Whisper, applying language-biasing tokens in decoding.
- Corrected low-confidence ASR tokens via IndicBERT masked language modelling, cutting Word Error Rate 14%.

Education

B.E. Electrical & Electronics, Minor in Computer Science **BITS Pilani, Hyderabad (2022–2026)**
CGPA: 6.8 – *CBSE:* Class 12 - 92.2% – Class 10 - 91.2%
Activities: State representative in basketball and cricket; volleyball gold medalist; captain of 20+ person logistics teams