

Lekkala Ganesh

AI/ML Engineer | LLM Applications, RAG & Data Science

Bangalore, India | +91-8688479455 | lekkalaganesh14@gmail.com | [LinkedIn](#) | [GitHub](#) | [Portfolio](#)

SUMMARY

AI/ML engineer building LLM applications and data-science models in Python: RAG pipelines (chunking, embeddings, vector search), conversational AI — chatbots and voicebots with speech-to-text and text-to-speech — LangGraph agents with guardrails, and the REST APIs behind them, shipped with statistical evaluation and structured logging. 500+ automated tests and CI across public repos. B.Tech CSE.

TECHNICAL SKILLS

Languages: Python, TypeScript, JavaScript, SQL, HTML5, CSS3

Machine Learning & Statistics: scikit-learn, Pandas, NumPy, SciPy, Regression, Classification, Feature Engineering, Cross-Validation, Hyperparameter Tuning, Model Evaluation, EDA, TF-IDF, Cohen's Kappa, Wilcoxon Signed-Rank, Class Imbalance, Precision/Recall

Generative AI & Agents: GPT-4o, Gemini, Ollama (local inference), Hugging Face Transformers, PyTorch, LangChain, LangGraph, MCP, Tool Calling, Guardrails, Prompt Engineering, Multi-Provider Fallback, SSE Streaming

RAG & Conversational AI: Embeddings (sentence-transformers), Chunking, Vector Search (LanceDB, pgvector, ChromaDB), Semantic Search, Chatbots, Voicebots, Speech-to-Text (Whisper, Web Speech API), Text-to-Speech, Intent Classification

Backend & Data: FastAPI, Flask, REST APIs, OpenAPI/Swagger, JSON Schema Validation (Pydantic, Zod), WebSockets, MERN (MongoDB, Express.js, React, Node.js), PostgreSQL (Row-Level Security, Triggers, Indexes), SQLite, Supabase, Schema Design, Joins & Aggregations

Engineering & Ops: pytest, Vitest, Playwright, GitHub Actions (CI), Docker, Git, Linux, Structured Logging, Correlation IDs, Token-Cost Tracking, ONNX Runtime, Vercel

EXPERIENCE

Freelance AI/ML & Full-Stack Engineer

Self-Employed

Remote

Jan 2025 – Present

- **Conversational AI:** Built a browser voicebot — speech-to-text → LLM → streaming text-to-speech, barge-in, VAD, 10-turn context ([VoxAI](#)) — and an **ElevenLabs** agent placing outbound calls over **Twilio**; multi-provider LLM fallback (GPT-4o → Gemini → OpenRouter).
- **AI Training Data & Annotation:** Author scoring rubrics and hand-label model outputs, blind-scored against anchored criteria before automated judging; adversarial probes for sycophancy, verbosity, position bias.
- **Backend & Delivery:** Designed **PostgreSQL** schemas (CHECK constraints, composite indexes, triggers) with **Row-Level Security** and **JWT**; shipped **REST APIs** (**OpenAPI/Swagger**, **HMAC**, request-scoped **correlation-ID logging**) and React/Next.js clients to **Vercel**, verified with **Playwright**.

Full Stack Web Development Intern

Slash Mark IT Solutions

Remote, Part-time

Dec 2024 – Apr 2025

- Delivered **6 full-stack projects** on the **MERN stack** (MongoDB, Express.js, React, Node) — including a URL Shortener **REST API** and a Food Delivery app with **authentication**, geolocation and payment flow.

Machine Learning Intern

Bharat Intern

Virtual

Oct 2023 – Nov 2023

- Built **regression** and **classification** models in **Python** with **scikit-learn** and **Pandas** — **data cleaning**, **EDA**, **feature engineering**; compared estimators on held-out metrics via **cross-validation** and **hyperparameter tuning**.

PROJECTS

Cost-Efficient RAG Service | Python, FastAPI, LanceDB, sentence-transformers, litellm

[GitHub](#)

- **Chunking** → **embeddings** (all-MiniLM-L6-v2) → **LanceDB vector search** → generation, with a second **pgvector** backend; **0.71 Recall@5**, **34.6 ms p95**, documented **token-cost/TCO** comparison.

Agentic Hallucination Guardrail | Python, LangGraph, LangChain, Ollama, SSE

[GitHub](#)

- **LangGraph** 6-node **agent** state machine detecting, validating and correcting LLM hallucinations inline — temperature-decay retries, circuit breaker, **Pydantic**-validated output; every run **traced end-to-end** (latency, retries, confidence) over **SSE**.

LLM-as-Judge Evaluation Pipeline | Python, scikit-learn, SciPy, Rubric Design

[GitHub](#)

- Measures 5 judge biases (position, verbosity, self-enhancement, sycophancy, score clustering) using **Cohen's kappa** and **Wilcoxon signed-rank** tests against **105 hand-labeled** human judgments; **200 tests**.

ATS Resume Optimizer | Python, scikit-learn, Streamlit, GitHub Actions

[GitHub](#)

- Resume-to-job **semantic matching** via **scikit-learn TF-IDF** and **cosine similarity** with keyword-gap extraction; **276** **pytest** tests and **GitHub Actions CI** (3.11/3.12, ruff) over **~20K lines of Python**.

EDUCATION

Parul University

Bachelor of Technology in Computer Science & Engineering; GPA: 7.53/10

Vadodara, Gujarat

Jul 2022 – Jun 2025