

MOUNIKA BHARATHA

AI Engineer | LLM Applications | Agents & RAG | Voice AI

+91-9542200881 | mounikabharatha2081@gmail.com | linkedin.com/in/mounikabharatha | github.com/Mounikabharatha

SUMMARY

AI Engineer with a year of production experience building LLM-powered applications, RAG systems, agents, and real-time voice agents. Builds end to end, from Python and TypeScript services to deployment, and measures quality with evaluation sets, automated tests, and CI.

TECHNICAL SKILLS

LLMs & GenAI: Anthropic Claude API, Gemini API, Llama 3.1 (Ollama, Workers AI), Groq, prompt engineering, grounding checks
RAG: pgvector, embeddings, chunking, hybrid search (RRF), cross-encoder reranking, RAG evaluation (recall@k, MRR, faithfulness)
Agents: LangGraph, multi-step agents, self-critique loops, human-in-the-loop, checkpointing, Tavily
Voice AI: Streaming ASR/TTS, NVIDIA Riva (Nemotron, Magpie), Deepgram, VAD, endpointing, barge-in, Retell, Voiceflow
Backend & Languages: Python, TypeScript, JavaScript, SQL, FastAPI, asyncio, WebSockets, gRPC, REST APIs
Frontend & Data: React, Next.js, Tailwind CSS, PostgreSQL (Supabase, Neon, Row-Level Security), MySQL, SQLite, Pandas
DevOps & Automation: Git, GitHub Actions CI/CD, Cloudflare Workers, Vercel, Render, Netlify, n8n, Zapier

EXPERIENCE

Vertical Systems

Nov 2025 – Present

AI Engineer

- Built a B2B lead-generation platform end to end (React, TypeScript, Supabase/PostgreSQL) used daily by the sales team.
- Developed LLM features with the Anthropic API: natural-language prospect search, AI-generated company research briefs, and personalized outreach drafts.
- Built automated data-ingestion pipelines from public sources (scheduled scrapers, HTML/PDF parsing, GitHub Actions cron) with deduplication and entity resolution, feeding a configurable lead-scoring engine.
- Developed AI voice agents for live phone conversations using Deepgram, Voiceflow, and Retell.
- Secured the platform with Row-Level Security, role-based access control, JWT-verified endpoints, and restricted CORS; automated approvals, notifications, and lead routing with n8n and Zapier.

PROJECTS

GrantScope – RAG Search over Federal Grant Documents | *Python, pgvector, Llama 3.1, Cloudflare Workers* [Live](#) | [Code](#)

- Built a RAG engine over 200 federal grant PDFs (15K chunks) with section-aware chunking, metadata filters, and hybrid pgvector + full-text search merged by reciprocal rank fusion; section-aware chunks beat fixed-size on recall@5 (82% vs. 63%).
- Raised recall@5 from 82% to 90% with a cross-encoder reranker, with hybrid search putting the right grant in the top 5 for every question; CI rebuilds the index on each push and fails the build if recall drops.
- Answers use only word-for-word verified quotes with PDF page citations: 97% faithful and 13 of 14 unanswerable questions refused on a 93-question eval set. Deployed as a TypeScript Cloudflare Worker on Neon Postgres at \$0/month.

Voice Interview Coach – Real-Time Voice Agent | *Python, FastAPI, NVIDIA Riva, Groq* [Live](#) | [Code](#)

- Built a full-duplex voice agent for behavioral interview practice: asks 3 questions from a bank of 14, follows up when an answer is thin, and scores feedback on 5 competencies.
- Streamed browser audio into NVIDIA Nemotron streaming ASR, Qwen3 on Groq (reasoning off so replies can be spoken at once; model picked from a preference list at startup), and NVIDIA Magpie TTS; cut voice-to-voice latency from ~3s to ~1.5s with sentence-by-sentence synthesis and model prewarming.
- Implemented adaptive end-of-turn detection (1.3–2.6s) that tolerates mid-answer pauses, and ~300ms barge-in that cancels the agent's reply when the user interrupts.
- Added a grounding check that verifies model-cited quotes against the transcript; it rejected 42 of 70 quotes (60%) across two scripted test sessions. Covered by 66 automated tests.

LangGraph Research Agent – Agentic RAG | *LangGraph, Gemini, SQLite* [Live](#) | [Code](#)

- Built a multi-step agent that splits questions into sub-queries, searches the web with Tavily, and writes cited answers, with a bounded self-critique loop, checkpointing, and human approval.
- Added semantic memory: past reports are embedded (gemini-embedding-001) into a SQLite vector store, and the top 3 matches above a 0.80 cosine cutoff (unrelated questions scored ~0.70) feed the planner. 100 tests in CI.

Chillville Bakery & Boba – Freelance Client Website | *Next.js, TypeScript, Tailwind CSS, Vercel* chillvilletreats.com

- Built and deployed a multi-page bakery site with a serverless pickup-ordering flow, Resend email confirmations, and JSON-LD SEO.

INTERNSHIPS

EPAM Systems – Front-End Development Intern, Hyderabad Feb 2025 – May 2025

- Built responsive e-learning components from Figma designs and integrated REST APIs with Python backend services.

Infoz IT Solutions – Data Science Intern, Hyderabad Oct 2024 – Jan 2025

- Built data-cleaning pipelines and predictive ML models in Python (Pandas, NumPy), with SQL analysis and Tableau dashboards.

Avantel Limited – Front-End Development Intern, Hyderabad: responsive, cross-browser e-learning interfaces Jul 2022 – Dec 2022

Inquilab Inventions – UX/UI Intern, Hyderabad: prototyped a WhatsApp chatbot for remote learning May 2021 – Aug 2021

EDUCATION & CERTIFICATIONS

Icfai Foundation for Higher Education

Hyderabad

B.Tech in Computer Science and Engineering

2019 – 2023

Certifications: Data Science, NASSCOM certified (A+) | Front-End Development (EPAM) | Tableau