Parth Tiwari

AI Engineer | GenAI Application Developer

Bengaluru, India | +91-7000181882 | parthti2003@gmail.com

parth-tiwari-1.vercel.app | github.com/parthtiwari-dev | linkedin.com/in/parth-tiwar1

Summary

AI engineer who ships production GenAI systems end to end and measures what they do. Built a music generation and stem-separation platform solo in 24 days and took its generation time from 500 seconds to 56. Works across diffusion, retrieval, agents and the GPU infrastructure underneath, and builds the evaluation harness that decides what is good enough to ship.

Skills

GenAI and LLMs: LangGraph, LangChain, RAG (hybrid vector + BM25), multi-agent orchestration, RAGAS evaluation, prompt engineering, Groq, OpenAI API

Diffusion and Audio: FLUX.1-dev and FLUX.2-dev, PuLID, LoRA fine-tuning, ACE-Step, BS-RoFormer source separation, Diffusers, PyTorch, GPU inference optimization

ML Engineering: XGBoost, leakage-safe feature engineering, ROC-AUC, SHAP, backtesting, evaluation harness design, class imbalance handling

Systems and Infra: FastAPI, Modal, Docker, PostgreSQL, Neon, Drizzle, ChromaDB, Qdrant, DuckDB, Next.js, React, Cloudflare R2, GitHub Actions, Python, SQL

Experience

AI and ML Development Intern

Mar 2026 - Present

Stick and Dot | Early-stage AI startup, creative tools for writers and filmmakers | Remote

- Built and shipped BeatMind (beatmind.tech) solo in 24 days: it generates a track, separates it into stems, and lets a producer rebuild it section by section with every substitution held in key and in tempo. Four independently deployed GPU and CPU services on Modal behind a durable job state machine with lease tokens, per-stage timeouts and fencing, so a stale worker cannot overwrite a newer attempt.
- Cut generation from 500 seconds to 56 for a two-minute track and the full pipeline from roughly 1000 seconds to 250, after root-causing a missing CUDA library that left ONNX Runtime silently on CPU while the GPU idled. Re-verified on two GPU classes separately.
- Migrated Vivid, the studio storyboard tool, to FLUX.2-dev and built the evaluation harness it lacked: per-character ArcFace identity, prompt adherence, realism, latency and cost across a 12-scene benchmark. A distilled schedule cut GPU time by 5.9x and cost per run from $2.81 to $0.48. It did not ship: the same measurements showed identity degrading on text-heavy scenes, so the base model stayed in production.
- Shipped the company web platform in Next.js and React in three weeks: dashboards for writers, readers and business clients, a commission marketplace, an article system with dynamic routing, and session-aware authentication.

Projects

QueryPilot - Self-Correcting Text-to-SQL Agent

GitHub

- Text-to-SQL tools hallucinate table names, break silently on schema changes, and have no recovery when a query fails. QueryPilot addresses all three: schema-aware RAG injects only the relevant tables, a four-layer critic validates SQL before execution, and a three-stage loop repairs failures by regex first and by LLM second.
- 95.7% execution success across an 82-query benchmark, of which 90.0% succeeded first attempt and 5.7 percentage points came from the correction loop, with a 0.0% hallucination rate. A separate 15-query library schema scored 100% with no re-tuning, which is evidence it generalizes rather than fitting one database.

Stack: Python, LangGraph, FastAPI, Groq, OpenAI, SentenceTransformers, ChromaDB, PostgreSQL, Docker, GitHub Actions

Tathya - Autonomous Public-Record Pipeline

Live | API | GitHub

- A non-partisan record of India's Union Government. Configured public sources are watched on a schedule, snapshotted, clustered, and persisted as sourced case files showing what government, media and citizens each said about a topic. No topic is chosen by hand and the system issues no verdict of its own.
- FastAPI service typed against the frontend's own TypeScript definitions, with a Next.js reader wired to the real API and no mock data anywhere in the deployed path. Supabase Postgres with versioned migrations, scheduled ingestion, and a run-health alert that fires when a run's source count falls below 20% of its recent median.

Stack: Python, FastAPI, Next.js, TypeScript, Supabase, PostgreSQL, Vercel, Render, pytest

Also: Evidence-Bound Medical RAG grounded only in 20 FDA and NICE PDFs with a hard refusal policy; 0.80 RAGAS faithfulness across a 20-query benchmark at $0.168 total cost. Real-Time UPI Fraud Engine 480+ point-in-time features over 590K transactions; backtesting exposed a 0.895 to 0.60 ROC-AUC training-serving drift; 92% precision inside a 0.5% alert budget.

Education

Institute of Engineering and Science, IPS Academy, Indore *2021 - 2025*

B.Tech in Computer Science, AI and ML

Great Learning, Bangalore *Jul 2025 - Feb 2026*

Post Graduate Program in Data Science, specialization in GenAI