

# Vansh Goenka

## Applied AI Engineer

+91 9663377550 | vanshgoenka007@gmail.com | Kolkata, India | Available immediately  
linkedin.com/in/vanshgoenka | github.com/unusual9guy

### SUMMARY

Applied AI engineer specialising in production GenAI systems: retrieval, evaluation and LLM infrastructure. **Python** first; shipped this year's production backend in TypeScript. Sole backend engineer on a GenAI mobile app: retrieval, streaming chat, payments, ingestion, observability and LLM cost controls. Improved retrieval **recall 0.53→0.98** and cut latency from 5–8 s to **0.6 s**, verified by a 71-question eval harness and gated on SLOs in CI. Build with **Claude Code** and **agent harnesses**; wrote the pricing model and PRDs.

### EXPERIENCE

- **TinyCheque (Saroya AI) | AI Engineer** *February 2026 – September 2026*
  - **Cut retrieval latency from 5–8 s to 0.6 s and lifted recall 0.53→0.98 (MRR 0.89).** Graph-database queries were slow enough to time out streams, so I moved retrieval to Postgres pgvector with full-text search, RRF fusion and mem0 memory, and wrote a 71-question golden set and eval harness to prove quality held.
  - **Built and ran the entire backend of a GenAI product, alone.** Saroya lets creators build an AI avatar that users chat with. Wrote the TypeScript/Node/Express API from zero (auth, Razorpay payments, creator onboarding, chat) and operated it from May 2026 through the release to testers on Play Store and TestFlight.
  - **Reached a 78% cached-input ratio and cut LLM spend after real routing costs overran the model.** Instrumented every call for cost, added prompt and reply caching with a weekly CI check that alerts below 70%, and pinned OpenRouter to 5 providers with fallback. Streamed tokens to mobile over SSE with pacing and keepalive.
  - **Built the ingestion pipeline behind every avatar.** Pulled creator content from X, Reddit, YouTube and LinkedIn, and from PDF, DOCX, XLSX and PPTX uploads with OCR. Added KYC identity checks and DPDP-aligned hard delete across every data store.
  - **Wrote the pricing model.** Originated the creator pricing and credit model the CPO adopted, plus the PRDs, ADRs and cost analyses the team planned against.
  - **Shipped solo at team pace by building my own tooling.** Set up a Claude Code harness with 14 subagents, 22 slash commands, MCP tools and hooks that enforce spec-first and test-first work, backed by a 169-test suite.
- **Chitra Goenka Crafts & Creations | Freelance AI Consultant** *January 2025 – February 2026*
  - Replaced manual product photoshoots with a **Gemini** + image-generation pipeline producing photorealistic assets.
  - Built a B2B lead-finder over the **Google Maps API** with a custom location database (sub-second queries); shipped both behind a **Streamlit** dashboard that non-technical staff use daily.

### PROJECTS

- **Market Research Agent | Python, LangChain, OpenAI, Tavily, Exa**
  - **Multi-agent** tool that researches companies and industries, proposes AI use-cases and surfaces datasets; GPT-4o-mini synthesis with Tavily and Exa search in a Streamlit app.
- **Pizzeria Review Agent | Python, LangChain, Ollama, ChromaDB**
  - Local **RAG pipeline** over unstructured reviews with ChromaDB and Llama 3.2 on Ollama, so inference runs without cloud latency or per-call cost.

### SKILLS

- **AI & LLM Systems:** RAG, hybrid retrieval, reranking (RRF), chunking, LLM evaluation (recall@k, MRR, golden sets), prompt engineering, prompt caching, guardrails, agents, tool calling, MCP, long-term memory, OCR
- **LLM Stack:** OpenAI, Gemini, OpenRouter, LangChain, mem0, Ollama, Claude Code (custom subagents, skills, hooks)
- **Languages & Backend:** Python, TypeScript, JavaScript, SQL, Cypher; FastAPI, Node.js, Express, REST, SSE/WebSocket, Razorpay
- **Data:** PostgreSQL, pgvector, Supabase, FalkorDB, Redis, ChromaDB, Cloudflare R2, Apify, Kreuzberg, Tesseract
- **Infrastructure:** Docker, GCP Cloud Run, Railway, GitHub Actions, OpenTelemetry, SigNoz, AWS (EC2, S3), Linux
- **Practices:** TDD, SDD, PRDs, ADRs, RCAs, LLM cost modelling, DPDP-aligned data deletion

### EDUCATION

- **BSc (Hons) Artificial Intelligence | University of Manchester, UK** *September 2021 – July 2024*
  - Dissertation: Simulation and Optimization of the Beta-Algorithm for Autonomous Swarm Robotics