

ANURAG SINGH

AI / GenAI Engineer | RAG · Agentic AI · LLMs · MCP

Gurugram, India | +91 93195 31330 | anuragsingh2445@gmail.com | [Portfolio](#)
linkedin.com/in/anurag2050 | github.com/anuragroque | leetcode.com/u/anuragsingh2445

PROFESSIONAL SUMMARY

AI / GenAI Engineer with 2+ years building production LLM systems - retrieval-augmented generation (RAG), agentic AI with the Model Context Protocol (MCP), and local/air-gapped LLM deployments (Ollama). Sole developer of **5 enterprise platforms serving 100+ users and 200K+ monthly transactions**, fully on-premise with zero data leaving client infrastructure. Anti-hallucination by design: hybrid retrieval, cross-encoder reranking, grounded refusal, prompt-injection filtering, and evaluation. 300+ DSA problems solved.

TECHNICAL SKILLS

Languages: Python, SQL, JavaScript/TypeScript, C/C++

Generative AI & LLMs: RAG, Agentic AI, LangChain, LangGraph, MCP, Prompt Engineering, Guardrails / Prompt-Injection Filtering, Evaluation, LoRA / Fine-Tuning, Ollama, Hugging Face, NLP, spaCy (NER)

Retrieval & Search: Hybrid & Semantic Search, BM25, Vector DBs (pgvector, FAISS, Pinecone, ChromaDB), Cross-Encoder Reranking, RRF, MMR, Grounded Refusal, Embeddings

ML Frameworks: PyTorch, TensorFlow, scikit-learn, Isolation Forest, MLOps

Backend & APIs: FastAPI, Flask, REST APIs, SQLAlchemy, Pydantic, Server-Sent Events, Uvicorn, Nginx

Databases: PostgreSQL (+pgvector), MongoDB, MySQL, Redis, Vector DBs

Cloud & DevOps: AWS (S3, EC2, IAM), GCP, Docker, Docker Compose, Linux, Git, CI/CD, Render, Vercel

PROFESSIONAL EXPERIENCE

AI & Data Engineer (Data Analyst) - TRPW Strategic Partners, Gurugram, India

Mar 2024 - Present

- Sole developer of **5 production enterprise platforms** - Limestone (reconciliation), Betel TMS (task management & audit scheduling), Betel VMS (vendor billing), Excellia AI (spreadsheet intelligence), SOP Intelligence Portal (RAG chatbot) - serving **100+ users and 200K+ monthly transactions**.
- Deployed local large language models (**LLMs**), RAG pipelines, ML anomaly detection, and OCR across all platforms - **100% on-premise**, zero data leaving client infrastructure.
- Designed the shared security layer for all platforms: multi-tenant RBAC, OTP authentication, session management, and multi-layer prompt-injection filtering.
- Paytm KYC Audit:** built an OCR pipeline (Google Cloud Vision + Tesseract) with SequenceMatcher/ML name-matching and bulk deduplication across large Excel datasets, reducing manual audit effort by **90%**.
- Automated reconciliation of 10,000+ monthly transactions, eliminating **32+ hours of weekly manual finance work** and cutting processing time by **80%**.
- OYO / PRISM Fraud Audit:** automated an Excel audit pipeline (Python, Pandas, NumPy, SQL, Regex) with anomaly checks, cutting analyst cycle time by **60%**.

KEY PROJECTS & PRODUCTS

Enterprise Agentic RAG Chatbot - Multi-Layered Hybrid Retrieval Engine

Python, FastAPI, PostgreSQL/pgvector, Ollama (llama3.2), Cross-Encoder, SSE

- Architected a privacy-first, multilingual AI document assistant across **4+ enterprise environments (100+ users)** with a six-stage hybrid retrieval funnel - semantic + BM25 + RRF + cross-encoder reranking + MMR + grounded refusal - streaming answers over an OpenAI-compatible API with real-time updates and no retraining. [Link](#)

Excellia AI - Local-LLM Spreadsheet Intelligence (19-tool MCP Server)

Python, FastAPI, MCP, Ollama, scikit-learn (Isolation Forest), pandas

- Cut analyst review time by **70% on 100K+ row datasets**. Air-gapped spreadsheet intelligence - profile, validate, anomaly-check, reconcile, fraud-score, KYC-match - exposed through a 19-tool **MCP server**, Excel add-in, web chatbot, and offline AI agent; scales to 500K-row files via a background job queue. [Link](#)

FilloAI (Product) - AI Form-Filling Chrome Extension - live on Chrome Web Store

JavaScript, Chrome MV3, Ollama, pdf.js

- Shipped a privacy-first extension that auto-fills any web form with a locally running LLM and zero cloud calls; a 9-layer field-resolution hierarchy handles React/Vue/Angular inputs, custom dropdowns, Shadow DOM, and iframes on LinkedIn, Indeed, Greenhouse, Workday, and Lever. [Link](#)

Additional: KYC OCR pipeline (Paytm, 90% less audit effort); Notepad PRO-AI (offline AI editor, Ollama/Gemma); self-hosted Raspberry Pi edge server (Docker, Nextcloud).

EDUCATION & ACHIEVEMENTS

Master of Computer Applications (MCA) - Galgotias College of Engineering & Technology

2021 - 2023 | CGPA 7.5

Bachelor of Computer Applications (BCA) - Singhania University, Rajasthan

2018 - 2021 | CGPA 8.3