Ansh Kumar Singh

AI Engineer

Pune, India (open to relocation & remote) | anshkumarsingh107@gmail.com
linkedin.com/in/anshks | github.com/Ansh-66 | ansh-66.github.io

Experience

AI Engineer — AI & Service Transformations Jan 2025 – Present
Tata Consultancy Services (designation: Assistant Systems Engineer) *Pune, India*

- **Designed and built four AI systems end to end for enterprise clients** on the Rapid Build Team, taking each from client use case to a working solution *(Python, FastAPI, LangChain, Gemini, DuckDB, React).*
- **Loan & policy onboarding**: Gemini extracts submitted PDFs into Pydantic schemas, rules validate every field, cases route junior → senior for approval, and decisions are emailed via the Gmail API.
- **Grounded GenAI analytics platform**: end-to-end ingestion with rule-based validation and quarantine into layered DuckDB storage with lineage; Gemini turns questions into a query spec checked against a semantic layer, and every figure in AI narratives is verified against SQL-computed facts. Added an explainable price-anomaly service (robust z-scores, Isolation Forest) routing flagged survey entries to human review.
- **Licence application validator**: licence, passport and address proof cross-checked against rule sets in a background job, producing a pass/fail decision and downloadable report.
- **Conversational intake agent**: guided question flow with prompt-level guardrails that generates a Bill of Materials and session summary PDF, replacing manual resource-booking sheets.
- **Presented AI enablement to executive leadership**: 4 sessions for 35–40 board members and senior executives at two enterprise clients, including a live multi-agent demo I built.

Projects

Restaurant Reservation Agent | *Python, MCP, Groq, Sarvam* | github.com/Ansh-66/goodfoods-agent

- **Booking agent for a 75-outlet chain, no agent framework**: own tool loop, 8 MCP tools, 11-language voice.
- **Gated writes in code, not the prompt**: no booking or change until the guest confirms; full outlets reroute.
- **Built a 16-journey eval**: 11/16 → 13/16 = 2 agent fixes + 2 eval fixes − 1 false pass caught − 1 untraced.

Graph RAG for Multi-Hop QA | *Python, Pydantic, Groq, NumPy* | github.com/Ansh-66/Insurance-Graph-RAG

- **Built an end-to-end Graph RAG system over 110 scraped news articles** — relation extraction with deal status and ownership intervals, entity resolution (222 names into 136 entities), cited answers.
- **Benchmarked against vector RAG on an 11-question set**: graph RAG answered 10/11 (vector 9/11, model alone 3/11) and 4/4 cross-document multi-hop questions vs 3/4, with every citation checked against its source.
- **11-case extraction test suite (9/11 exact)**; multi-run pass rates after output varied at temperature 0; **round-robin assembly** stopped one deal filling 14 of 30 context slots.

Retrieval-Augmented Generation Pipeline | *Python, NumPy, Gemini* | github.com/Ansh-66/rag-from-scratch

- **Diagnosed an inert LLM reranking stage, raising hit@1 from 13/17 to 17/17** — results identical to the no-rerank baseline exposed a silent fallback to cosine order; fixed the prompt contract (17 questions, 6 chunks).
- **Hand-written with AI code generation off**: chunking, embeddings, cosine retrieval, reranking, hit@k harness.

Skills

Languages: Python, JavaScript, SQL, HTML, CSS
AI / ML: RAG and Graph RAG (embeddings, reranking, relation extraction, entity resolution, graph traversal), LLM evaluation, LLM APIs (Gemini, OpenAI, Groq, Sarvam), MCP, tool calling, prompt engineering, agentic workflows
Frameworks & Libraries: FastAPI, SQLAlchemy, LangChain, Pydantic, NumPy, scikit-learn, sentence-transformers, React
Tools: Git & GitHub, SQLite, DuckDB, Ollama

Certifications *(all Sep 2026)*

Databricks: Generative AI Engineer Associate, ML Associate | **Anthropic**: Claude Certified Developer — Foundations

Education

The National Institute of Engineering Mysore, India
Bachelor of Engineering in Electronics & Communication *2020 – 2024*