ARSHDEEP SINGH

GenAI Developer | Generative AI · Agentic AI · Backend (Java, Python)
Noida, India | arshdeep.si.3010@gmail.com | github.com/arshdeep3010 | linkedin.com/in/arshdeep-singh30

Summary

GenAI Developer at TCS, building agentic AI for insurance claim workflows for a US healthcare client. I build AI systems where code, not the model, has the final say: RAG, LangGraph agents, durable workflows on PostgreSQL and evaluations with measured results, on Java (Spring Boot) and Python (FastAPI) backends.

Experience

Tata Consultancy Services (TCS) Feb 2026 – Present
GenAI Developer, US healthcare client *Noida, India*

- Build agentic AI for insurance claim workflows: agents that analyse claim information, retrieve the relevant policy knowledge and recommend a decision with its reasons.
- Build agent workflows with LLMs, RAG, LangChain and LangGraph, and integrate them into backend services with structured, validated outputs.
- Investigate production issues through Java logs, SQL data and backend behaviour; work with Java, Spring Boot, REST APIs and SQL alongside development and business teams.

Projects

AgentOps | *durable workflow engine for AI agents* Python, FastAPI, PostgreSQL, Docker, LLMs

- Built an event-sourced workflow engine on PostgreSQL alone: `SKIP LOCKED` job queue, worker leases with fencing tokens, idempotent tool calls, human approvals that hold no worker, replay from any step, live updates over SSE.
- Chaos test: **500 runs, 435 random worker kills, 0 runs lost, 0 duplicate side effects**; median takeover 1.9 s.
- Built a 12-step SDLC pipeline on top: requirements → design → tests first (must fail, then hash-locked) → build → 9 quality gates in a network-less Docker sandbox → AI review → PR with a traceability matrix. A live run took a coding task to **merged code in 5.5 min** on a free NVIDIA model.

FitCheck | *evidence-grounded job-fit agent* Python, LangGraph, FastAPI, pgvector, NVIDIA Nemotron

- LangGraph pipeline that parses a job description, retrieves matching code from my repositories (RAG over pgvector) and scores each requirement; a "met" must cite a retrieved chunk, linked as a GitHub permalink, or it is marked unverified.
- On 25 hand-labelled requirements: **100% accuracy and 0% false claims** (keyword baseline: 54% false claims).
- Found and fixed a pgvector filtered-search bug (0–2 of 3 rows returned); syncs only changed repos; 73 tests.

Health Claim Decision Agent | *AI-assisted claim decisions* Java 21, Spring Boot 3.5, PostgreSQL, pgvector, React

- 7 deterministic rules run first (REJECT > REVIEW > ALLOW); an LLM is consulted only when they allow, and low confidence or any AI or retrieval failure goes to human review.
- RAG over policy PDFs (PDFBox, NVIDIA NIM embeddings, pgvector HNSW); every decision stores its evidence, with an append-only audit trail.
- JWT with 3 roles, rate limits, idempotency keys, Resilience4j, live SSE stage stream to a React UI; **100 backend + 22 frontend tests** (Testcontainers, LocalStack).

VibeAudit | *security scanner for AI-written apps* Python, FastAPI, Docker, bandit, semgrep

- Secret scanner, bandit and semgrep first, then an LLM audit with verification, and a Docker sandbox run.
- Cut a snippet scan from about 12 model calls to **2**, and the sample scan from 75–90 s to **39.8 s**.

Technical Skills

Languages	Python, Java, SQL, JavaScript, TypeScript
GenAI	LLMs, RAG, Embeddings, pgvector, Structured Outputs, Prompt Engineering, LLM Evaluation
Agentic AI	LangGraph, LangChain, Tool Calling, Human-in-the-Loop, Durable Agent Workflows
Backend & Data	Spring Boot, Spring Security (JWT), JPA, FastAPI, REST, SSE, PostgreSQL, pgvector, MySQL
Cloud/DevOps	Docker, GitHub Actions, AWS (S3, CloudFront, Lambda, ECR), Terraform, Linux
Testing	pytest, JUnit 5, Mockito, Testcontainers, Vitest
Core CS	Data Structures & Algorithms, OOP, DBMS, System Design

Education & Certifications

Chandigarh University 2021 – 2025
B.E. in Computer Science, CGPA 7.07/10 *Mohali, India*
Claude Certified Architect (Anthropic) · Claude Certified Developer Foundations (Anthropic) · The Complete Agentic AI Engineering Course (Udemy) · Generative and Agentic AI in Production (TCS)

Publication

Automated Network Vulnerability Assessment with Nmap: A Comprehensive Approach, IEEE ICACCTech 2024, first author.