

Rakesh Singh

Bengaluru, India | [+91-626-433-2150](tel:+91-626-433-2150) | rakeshnitb23@gmail.com
[linkedin.com/in/rakesh-singh-58926a116](https://www.linkedin.com/in/rakesh-singh-58926a116) | github.com/rakeshnitb23

SUMMARY

AI Agent Engineer with 4 years of production backend experience (Java/Spring Boot, 2M daily transactions), now shipping agentic RAG: LangGraph, MCP, hybrid retrieval (pgvector + BM25 + RRF), eval-gated releases and a multi-provider LLM gateway (OpenAI, Anthropic Claude, AWS Bedrock). Author of [toolism](#), an open-source token-cost optimizer for LLM tool definitions.

OPEN SOURCE

[toolism](#) — LLM tool-schema token optimizer | Python, MCP, PyPI

github.com/rakeshnitb23/toolism

- Built an open-source Python library + CLI that measures and rewrites tool / MCP schemas to cut the input tokens re-sent on every agent turn; cut tool-definition tokens by **38%** across **12** public MCP servers (o200k_base + Anthropic count_tokens).
- Designed slimming as decorator-registered rule passes (redundant Pydantic titles, anyOf-null collapse, single-use \$defs), each proven schema-equivalent by bidirectional property-based tests (Hypothesis).
- Shipped a GitHub Action that diffs tool-token cost against the main-branch baseline and fails PRs over budget; published to PyPI via trusted publishing; **45** stars, **3** external contributors.

EXPERIENCE

AI Backend Engineer, Genpact

Feb 2026 – Present

GroundedDocs — private document Q&A that answers only with faithful citations from the user's files, or abstains.

- Designed idempotent PDF ingest (PyMuPDF) with content hashing and stable chunk IDs, so changed documents re-index without duplicate vectors or stale citations; treated chunk boundaries as product decisions.
- Built tenant-scoped hybrid retrieval in PostgreSQL (pgvector + BM25 fused with RRF, bge-reranker-base cross-encoder rerank), filtered by `user_id` and workspace; **92%** recall@5 on a frozen 40-question set.
- Shipped Pydantic-validated answers with citations and calibrated abstention as the default path: **94%** citation faithfulness, **89%** / **86%** abstention precision / recall (18 answerable / 11 partial / 11 unanswerable).
- Threat-modeled with DFD + STRIDE (fail-closed auth, rate limits, secret isolation, retrieved text as untrusted); gated releases on a 30+ case adversarial suite (prompt injection/extraction, cross-user access, hostile PDFs, conflicting sources) at **97%** pass rate.
- Added a bounded LangGraph agent path only for questions hybrid RAG failed: loop limits, one read-only MCP tool, at most one ephemeral subagent; kept only where it beat deterministic RAG on quality, latency, cost and failures.
- Built eval gates for retrieval quality, citation faithfulness, abstention, security pass rate, structured-output success, latency and cost per successful task (custom harness + Ragas / DeepEval, traced in Langfuse), with version-tagged prompts and corpus.
- Built an OpenAI-compatible LLM gateway over OpenAI, Anthropic Claude, AWS Bedrock: streaming + structured outputs, tenant API keys, token budgets, retries, circuit breakers, privacy-aware caching; **\$0.004** cost per successful task under failure injection.

Backend Java Developer, Genpact (Client: Shutterfly, USA)

Oct 2022 – Jan 2026

- Architected Java 11 / Spring Boot microservices processing 2 million daily transactions, reducing API response times by 65%.
- Implemented Resilience4j fault tolerance across 15+ microservices, reaching 99.9% uptime; optimized PostgreSQL, MongoDB and Redis, increasing retrieval speed by 70% and cutting costs by 30%.
- Built React.js + TypeScript internal tools and a photo pipeline for 5 million daily uploads (sub-200 ms); smart compression cut storage cost by 40%, and paginated APIs over 100k+ records improved search by 60%.
- Automated release management with a TypeScript service (85% less manual coordination); mentored 3 developers, raising code coverage from 65% to 92% through TDD (JUnit 5, Mockito).

TECHNICAL SKILLS

Agentic AI	LangGraph, LangChain, MCP (servers and clients), tool calling, subagents, Pydantic structured outputs, OpenAI, Anthropic Claude, AWS Bedrock (Knowledge Bases, Guardrails), AgentCore
RAG	Hybrid search (pgvector + BM25 + RRF), cross-encoder reranking, citation grounding, agent memory (Redis, Postgres)
Evals & Observability	Eval gates, golden datasets, faithfulness, LLM-as-judge, Langfuse, Ragas, DeepEval, OpenTelemetry
LLMOps & Cost	LLM gateway, token budgets, caching, retries, circuit breakers, cost/successful task, model routing, toolism
AI Security	DFD + STRIDE, prompt-injection defense, adversarial test suites, fail-closed auth, least-privilege tools
Backend	Python (FastAPI), Java 8/11/17 (Spring Boot, Resilience4j), TypeScript, JavaScript, SQL, React.js
Data	PostgreSQL, pgvector, MongoDB, Redis, SQL optimization, DynamoDB
Cloud & DevOps	Docker, Kubernetes, AWS (EC2, RDS, S3, Bedrock, Lambda), Terraform, Git, CI/CD (Jenkins, GitHub Actions)
Testing & Tools	JUnit 5, Mockito, TDD, pytest, Hypothesis, Claude Code, Cursor

EDUCATION

National Institute of Technology (NIT) Bhopal

Jan 2018 – Jan 2022

Bachelor of Technology (B.Tech)

ADDITIONAL

- Leadership:** Led ISTE chapter initiatives at NIT Bhopal, organizing technical events for 200+ students; won 2 internal hackathons.
- Community:** Volunteered 100+ hours at Arushi NGO, Bhopal, working with specially abled children.