

Nikunj Khitha

📞 +91 9540234616

✉ njkhitha2003@gmail.com

🌐 nikunj.codenex.dev

📄 Nikunj Khitha

🔗 Nikunj2003

Professional Summary

Full-Stack GenAI Engineer focused on Java/Spring Boot backends, secure agent infrastructure, retrieval systems, LLM proxying, and automation platforms. Built ArmorCode's central Knowledge Graph RAG brain over 1,000,000+ entities with tenant-scoped MCP access for platform and internal agents. Delivered 6 production MCP servers, OpenCode automation, multi-provider LLM infrastructure, and resilient APIs, improving retrieval accuracy by 40% and eliminating 200+ manual hours per month.

Technical Skills

Languages: Java, TypeScript, JavaScript, Python, Go, SQL, Bash

Backend and Full Stack: Spring Boot, Spring AI, Spring Security, Spring Data JPA, Hibernate, Node.js, FastAPI, Gin, Next.js, React, SQLAlchemy, REST APIs, Microservices, SSE, RBAC, JWT, OAuth2, Stripe

GenAI and Agents: RAG, GraphRAG, Knowledge Graph RAG, LightRAG, Graphiti, Temporal Graph Memory, LangGraph, LangChain4j, CrewAI, DeepEval, MCP, OpenAI, Claude, Gemini AI, LLaMA, LLM Proxying, Prompt Engineering

Data and Storage: PostgreSQL, pgvector, Neo4j, MongoDB, Elasticsearch, Redis, MSSQL, Supabase, MinIO, NFS

Reliability and Platform: Resilience4j, Docker, Kubernetes, Fabric8, Kubernetes Ingress, Kafka, LocalStack, GitHub Actions, Jenkins, AWS, Azure, Vercel, LiteLLM, Grafana, Nginx, n8n, OpenCode Server, Kiro, Claude Code, Codex

Professional Experience

ArmorCode

Dec 2025 – Present

Associate Engineer, Full-Stack GenAI

Gurugram, India

- Architected ArmorCode's **central Knowledge Graph RAG brain**, ingesting 1,000,000+ entities from Zendesk, QMetry, Jira, Chorus, and codebase sources into Neo4j and pgvector, improving retrieval accuracy by 40%.
- Built the **tenant-scoped MCP access layer** around the graph brain, enabling platform and internal agents to retrieve customer-specific context and execute authorized tools without cross-tenant leakage.
- Deployed and maintained ArmorCode's internal AI runtime stack with Knowledge Graph RAG, LiteLLM, OpenCode Server, n8n, and 6 production MCP servers, standardizing retrieval, model access, agent execution, and workflow orchestration.
- Shipped the ArmorCode platform agent memory layer with Graphiti temporal knowledge graphs, combining session-scoped context with tenant and person-level long-term Knowledge Graph recall for multi-step reasoning workflows.
- Contributed to development of ArmorCode **AI Exposure Management (AIEM)**, building frontend and backend for AI visibility, ownership, governance, shadow AI discovery, and auditable risk tracking.
- Built the **OpenCode gateway** in front of OpenCode Server using Go and Gin with observability, load balancing, task queueing, main-server routing, and scheduled and PR-triggered workflow orchestration.
- Delivered 30+ automations and 25+ AI workflows across 20+ systems with fallback paths, approval gates, and Slack/Jira handoffs, eliminating 200+ manual hours monthly and earning ArmorCode's AI Ninja Award in the second month after FTE conversion.

ArmorCode

Jan 2025 – Nov 2025

Software Development Engineer Intern

Gurugram, India

- Owned backend integrations for 5+ security tools, including Black Duck, Snyk, and Checkmarx, on an AppSec platform with 130+ connectors for consolidated security finding management.
- Created AI-assisted code generation utilities with template engines and AST parsing to automate new integration scaffolding, reducing per-integration boilerplate setup time by 30%.
- Built a shared backend service-orchestration library with parallel container startup for MySQL, Elasticsearch, Redis, Kafka, MongoDB, and LocalStack across backend services.
- Built an OpenAI-compatible proxy in Go for Gemini CLI, Codex, and Claude Code auth with multi-account load balancing, provider fallbacks, and Redis-backed LLM caching, cutting LLM indexing costs by 70% and saving \$15,000+ annually.
- Shipped a reusable HTTP client using Resilience4j for failure isolation, configurable backoff, timeout handling, and SSRF protection adopted across multiple services.

Xansr Media (AIKO)

Jun 2024 – Dec 2024

SDE Intern, Backend and AI

Remote

- Built Node.js and FastAPI microservices, improving API performance by 40% and reducing deployment time by 42% with Docker and GitHub Actions.
- Engineered Fantasy GPT with RAG, LangGraph, backend APIs, agents, and DeepEval quality checks to resolve 98% of complex sports queries.
- Built Python-based ETL pipelines to collect sports data from multiple sources and ingest it into MS SQL for Fantasy GPT SQL RAG workflows.
- Built AIKO personalization for voice-based sports media with Azure Speech SDK, live commentary in 20+ languages, and on-the-fly highlight reels.

Projects

CodeNex: AI Builder: *Java, Spring Boot, Spring AI, React, TypeScript, SSE, Kubernetes, MinIO, Stripe*

[Live](#) | [GitHub](#)

- Built a distributed AI code generation SaaS that turns natural-language prompts into full React applications with Spring Boot, Spring AI, SSE streaming, and MinIO/NFS-backed code persistence.
- Implemented Kubernetes/Fabric8/Ingress preview pods with sub-2 second cold starts, 10,000+ concurrent streaming sessions, sub-200 ms token latency, token quotas, RBAC, autoscaling, and Stripe subscriptions.

Serenify: *React, TypeScript, Supabase, pgvector, Gemini AI, Vercel*

[Live](#) | [GitHub](#)

- Built an open-source AI wellness product with empathetic Gemini-powered chat, mood tracking, journaling, guided sessions, crisis-help flows, privacy-aware analytics, and pgvector-backed personalization.
- Designed a state-rich consumer UX around return usage, privacy-minded AI interactions, Supabase-backed product workflows, and a polished experience users can return to consistently.

Education

The NorthCap University

Aug 2021 – Jun 2025

B.Tech in Computer Science Engineering, Full Stack Development

Gurugram, India