

DEVANSHU CHICHOLIKAR

SOFTWARE ENGINEER | AI ENGINEER | FORWARD DEPLOYED ENGINEER | BOSTON, MA (OPEN TO RELOCATION)

(857) 339-8803 | chicholikar.d@northeastern.edu | [linkedin.com/in/devanshuchicholikar](https://www.linkedin.com/in/devanshuchicholikar) | github.com/DevanshuNEU | devanshuchicholikar.com

PROJECTS

- OpenCodeIntel** | *Creator & Maintainer* | *Python, FastAPI, MCP, Pinecone, Redis, React, TS* | opencodeintel.com **Nov 2025 - Present**
- Shipped a **production MCP server** (12 tools: semantic code search, dependency graph, impact analysis) enabling **AI agents** (**Claude, Cursor**) to retrieve codebase context via a hybrid AST + embedding **RAG** pipeline (**p50 641ms cold, 242ms** cached, in production)
 - Benchmarked retrieval to **94% average Hit@1** across **14 OSS codebases** (4 languages, **665-query** research eval), improving accuracy **+8.4 points** with **reranking** (Voyage rerank-2) isolated via a **98-run ablation**
 - Hardened the public **REST API** with dual-mode auth (JWT + SHA-256 keys), **multi-tenant isolation** (per-repo Pinecone metadata filters, Postgres row-level security as a backstop), and Redis tiered rate limiting; **700+ commits, 450+ tests** solo
- Overhear** | *Creator* | *TypeScript, Next.js, Retell, Claude API, Postgres, Drizzle, Vitest* **Sep 2026**
- Engineered an **AI QA analyst for voice agents** in 7 days: every **Retell** call is graded on 7 rubric dimensions, 3 (**50% of the score**) decided in code from the action log and 4 by an **LLM-as-a-judge**
 - Validated the judge on a **39-call golden dataset**: **23 of 23** planted failures caught, **macro-F1 0.89, 0.99 self-consistency** over 3 runs, with adversarial cases and published limits
- CallBudget** | *Creator* | *Python, FastMCP, FastAPI, scikit-learn, Optuna, Pipecat, Deepgram, DuckDB* **2026**
- Deployed an **agentic system** end-to-end against a real pharmacy-stock workflow: a **FastMCP server** (predict / plan / eval) over an sklearn ranker, cutting expected calls-to-find from **4.3 to 2.3 (47%)** on a 19-pharmacy simulation
 - Designed an **abstention layer** on the stock extractor (**self-consistency voting**) that drove false in-stock answers from **10% to 0%** under a deliberately 25%-unreliable extractor; at 5 votes it held 0% while abstaining on 51% of calls instead of 69%
 - Hardened the tuning loop with an **Optuna**-constrained auto-research loop and a held-out validation guard that rejected its own overfit gains; one engine serves **MCP, CLI, and a FastAPI web UI**
- SecureScale** | *AWS, Terraform, Packer, Docker, Lambda, RDS, KMS, CloudWatch* **2025**
- Provisioned **multi-AZ AWS infrastructure** (VPC, ALB, ASG, RDS, S3 with KMS) as modular **Terraform** with **IAM least-privilege** defense-in-depth, cutting provisioning from **2 hours to 10 minutes** and **cloud spend 30%** at **99.9% uptime** with CloudWatch

SKILLS

AI-Assisted Development: Claude Code (subagents, hooks, custom skills), MCP (Model Context Protocol) server development, **agentic workflows**, spec-first development with AI agents, LLM evals; Cursor

AI & Agents / Eval: MCP servers (FastMCP), agent tool-calling, **RAG** (hybrid AST + embedding, BM25 + reranking), **LLM evaluation** (pre-registration, calibrated abstention, self-consistency, guardrails, LLM-as-judge), voice agents (Pipecat, Deepgram), tree-sitter, context engineering

Languages: Python, TypeScript, Java, JavaScript, Go, SQL, C++

Backend & APIs: FastAPI, Spring Boot, Node.js, REST, gRPC, WebSocket, microservices; React, Next.js

Cloud & DevOps: AWS, GCP, Docker, Kubernetes, Terraform, GitHub Actions, CI/CD, Railway, Vercel

Databases: PostgreSQL, Redis, Pinecone, DuckDB, Supabase, MongoDB

Foundations: data structures & algorithms, distributed systems, OOP & design patterns, system design

EXPERIENCE

- Northeastern University** **Sep 2025 - May 2026**
Graduate Teaching Assistant, CSYE 6225 Network Structures and Cloud Computing *Boston, MA*
- Taught **AWS, Terraform, and distributed systems** to **100+ graduate students** across two semesters; authored a **Docker + GitHub Actions CI/CD lab** adopted as official course content across 3 sections (**180+ students**)
 - Led **system-design and code reviews** for **15 cloud-native API project teams**, coaching on scalability, fault tolerance, and API design
- Jaksh Enterprise** **Aug 2022 - Jul 2024**
Software Engineer | *B2B industrial-equipment platform* *Ahmedabad, India*
- Developed a **Java / Spring Boot** product-catalog and quotation engine for **590+ customizable products** serving a **10K+ customer base**, over PostgreSQL and a rules-based pricing module
 - Cut quote-page **p95 latency 65%** (800ms to 280ms) via PostgreSQL indexing, Redis caching, and async processing
 - Owned delivery end-to-end as one of **two engineers**: gathered client requirements, scoped functional specs, and shipped via Dockerized **CI/CD**, cutting release cycles from **2 weeks to 3 days**
- Pitney Bowes** **Jan 2022 - Jul 2022**
Software Development Engineer Intern *Pune, India*
- Built **REST APIs** for PitneyShipPro with idempotent endpoints, eliminating **40%** of manual data entry; raised automated **test coverage to 85%** with Jest and Cypress

EDUCATION

- Northeastern University** | **GPA: 3.85** **Sep 2024 - May 2026**
M.S. in Software Engineering Systems *Boston, MA*
- Coursework:** Generative AI, MLOps, Distributed Systems, Cloud Computing, Database Management Design, Algorithms
- Dhirubhai Ambani Institute of ICT** **Aug 2018 - May 2022**
B.Tech in Information and Communication Technology *Gujarat, India*