

ISHAN S.

AI / GenAI Engineer | RAG | Agentic Systems | Production LLM Systems

Boston, MA | 617-652-6644 | srivastava.ish@northeastern.edu | [Portfolio](#) | [LinkedIn](#) | [GitHub](#) | [Medium](#)

EDUCATION

Northeastern University

Boston, MA

Master's in Applied AI, GPA: 4.0/4.0

Expected May 2027

Coursework: Artificial Intelligence, Machine Learning, Deep Learning, Neural Networks, NLP, Generative AI, Chatbot Application.

Bachelor of Computer Application, MGKVP

July 2021

EXPERIENCE

Tata Consultancy Services

Noida, India

Developer

Jul 2021 to Aug 2025

- Deployed LLM-powered chatbots across 3 enterprise clients using GPT-3.5/GPT-4 with intent classification, prompt engineering, and retrieval-grounded responses; achieved 60% query deflection and +9 point CSAT improvement.
- Built end-to-end NLP pipelines for ticket classification, multi-turn FAQ handling with entity extraction, and document summarization; processed 10K+ tickets/month, cutting resolution time 40% at 99.5% uptime.
- Led LLM orchestration workflows on internal tooling with conditional routing and automated decision points; cut manual effort 55% and lifted SLA compliance from 88% to 96%.
- Designed reusable AI workflow patterns and prompt templates adopted across multiple delivery teams, accelerating deployment cycles 40%.

PROJECTS

Findmejob (CareerForge): AI Career Platform with Cost-Aware Multi-Model Routing

Apr 2026

Live: [findmejob-nu.vercel.app](#) | Repo private (available on request)

- Shipped AI career platform on Next.js (App Router) + Vercel Fluid Compute + Supabase Auth (Google OAuth + email) with RLS, rubric-grounded profile assessment, JSearch-backed job feed, and on-click tailored resume generation.
- Designed cost-aware routing via Vercel AI Gateway (Sonnet 4.6 for moat tasks, GPT-4.1-mini for support) with prompt-caching targeting 70%+ hits; projected per-user cost ~\$0.48/mo against ~\$10 all-Sonnet baseline.
- Built edit-via-JSON resume engineering pattern: LLM emits structured edit ops, a deterministic transformer applies them to a stable LaTeX template (Tectonic compiled in Vercel Sandbox), eliminating LaTeX-syntax-from-LLM bugs.

UpMyRank: RAG and Socratic Tutoring Engine for K-12

Jan 2026

Live: [upmyrank-poc.vercel.app](#) | Repo private (available on request)

- Shipped Socratic tutoring engine using OpenAI native function calling to plan multi-turn tutor decisions: classify learner intent, choose retrieval scope, pick next pedagogical move, escalate model tier when needed.
- Designed two-layer policy guardrails: regex solution-seeker detector blocks direct-answer requests, CONTEXT_LOCK prompt scaffold pins agent persona; routes between GPT-4o-mini, GPT-4.1-mini, and GPT-4o by task complexity.
- Built RAG layer on FastAPI + Render with pgvector retrieval over Hugging Face NCERT corpus, Docker, Redis caching, GitHub Actions CI/CD; covers three subject domains in production.

OmegaTK Code Assistant: Production RAG Chatbot (Built for OpenEye, a Cadence company, Boston, MA)

Feb 2026

Repo: github.com/Ishan2036924/omega-tk-chatbot

- Built production RAG assistant generating Python for the OpenEye Omega Toolkit (cheminformatics) on FastAPI + React + Supabase pgvector, grounded in 200+ pages of API docs. Cut time-to-working-script from 20-40 minutes to under 30 seconds at ~\$0.01 per query.
- Designed three-layer pipeline: GPT-4o-mini intent classifier (4 routes), FAISS retrieval over 574 doc chunks, AST and API allowlist validation. Wired up to 5 corrective retry passes on validation failure.
- Authored 70-pair gold-standard eval set; achieved RAGAS 0.85 pipeline score and 1.00 groundedness with hard refusal on low-confidence retrievals.

SKILLS

LLM System Design: RAG pipelines (chunking, retrieval, reranking), agentic flows with function calling, multi-agent orchestration, multi-stage guardrails, prompt caching, cost-aware model routing, token and context-window management

LLM Frameworks and Protocols: OpenAI API, Claude API, LangChain, LangGraph, LangSmith, Vercel AI Gateway, Anthropic Model Context Protocol (MCP), Google A2A protocol

Retrieval and Embeddings: FAISS, pgvector, sentence-transformers, Hugging Face, OpenAI embeddings

Evaluation and Reliability: RAGAS, gold-standard dataset design, groundedness scoring, latency optimization, monitoring, hard-refusal guardrails

Full-Stack Development: FastAPI, Next.js (App Router), React, server-side rendering, server actions

Cloud and Agent Hosting: AWS SageMaker, Google ADK, Vercel (Fluid Compute, AI Gateway, Sandbox), Render, DigitalOcean, Supabase, Google Cloud Console

Infra and DevOps: Docker, GitHub Actions CI/CD, Redis, Postgres

ML / DL and Tools: PyTorch, scikit-learn, spaCy, NLTK, Pandas, NumPy, Python, Git, VS Code, Claude Code, conda

CERTIFICATIONS

- Advanced Data Science and Machine Learning, IIIT-Bangalore (May 2025).