

# Venkat Neelraj Nitta

Boston, MA    [venkatneelraj.vercel.app](https://venkatneelraj.vercel.app)

Email: [vneelraj.nitta@gmail.com](mailto:vneelraj.nitta@gmail.com)

[linkedin.com/in/v-neelraj-nitta](https://linkedin.com/in/v-neelraj-nitta)

[github.com/Subhash-269](https://github.com/Subhash-269)

---

## SUMMARY

Software, data and ML engineering under one AI umbrella: 2+ years across e-commerce, finance and insurance, taking data from storage to decisions. I want to keep building on that as an AI Engineer.

---

## WORK EXPERIENCE

- **Staples** Framingham, MA, US  
*AI Engineer Intern* *Jan 2026 – Jul 2026*
  - Designed the **evaluation** for a multi-turn LangGraph agent with function-calling tools and its product tagger: labeled sets plus an LLM judge; tagging accuracy averaged about 0.8, closer to 0.7 on buyer persona.
  - Maintained and extended the agent's **RAG** layer for agentic e-commerce: three Databricks Vector Search indexes and a Neo4j product graph, implementing its product-to-product relationships.
  - Built a fine-tuned multi-label DeBERTa tagger in PyTorch with an LLM fallback; moving batch inference over about 300K SKUs onto Ray on one GPU node cut runs from **about 4 hours to 90 minutes**.
  - Owned the daily pipeline feeding the agent: Databricks SQL merging more than 15 Snowflake tables into one Delta table, with validation, failure alerting, on-call support, a runbook and Azure DevOps CI/CD.
- **Cognida.ai** Hyderabad, India  
*Software/AI Developer* *Jul 2022 – Aug 2024*
  - Built embedding-based **RAG** for financial document search, cutting query resolution from 10 days to minutes; fine-tuned and quantized an LLM for an insurance-company chatbot to reduce serving compute cost.
  - Shipped unsupervised anomaly detection over millions of insurance claims (PySpark, topic modeling, K-means), flagging an additional 8% of new cases; ran distributed training on Ray and monitored GPU usage.
- **Lead Teaching Assistant**, Northeastern University, Boston, MA, US *Sep 2026 – Present*
  - AAI 5004 (Applications of AI for Professionals): review course material for technical accuracy; grade submissions.

---

## SKILLS

**GenAI & Agents:** LangGraph, LangChain, RAG, LLM-as-Judge Evaluation, LLM Fine-Tuning, Hugging Face

**ML & Data:** PyTorch, Scikit-learn, Pandas, NumPy, SQL, PySpark, Databricks, Snowflake, Delta Lake, Ray, Neo4j

**MLOps & Cloud:** Docker, Kubernetes, MLflow, Azure DevOps, CI/CD, Databricks Model Serving, AWS, Azure

**Programming:** Python, TypeScript, FastAPI, Django, REST APIs, Unit & Integration Testing, Git, Linux

**AI-Assisted Development:** Claude Code, GitHub Copilot, Google Antigravity, Cursor, Codex

---

## PROJECTS

- **InsurAsst (FAISS, Kustomize, Gemini API) [GitHub]:** Claims assistant citing policy PDFs, with LLM coverage routing, a 12-scenario evaluation harness, per-turn traces and deployment manifests with health probes.
- **rl-code-sandbox (pytest, JUnit XML) [GitHub]:** Sandboxed code grader with reward-hacking checks: 3 of 5 cheating agents earned full reward with checks off, and all 5 scored 0 with checks on across 5 tasks; 53-test suite passing.

---

## EDUCATION

- **Northeastern University** Boston, MA, US  
*Master of Professional Studies in Applied AI (STEM); GPA: 3.92* *Sep 2024 – Dec 2026 (Expected)*
  - Relevant Coursework: Fundamentals of Artificial Intelligence, AI System Technologies, Recommender System
- **Mahindra Ecole Centrale** Hyderabad, India  
*Bachelor of Technology in Electrical and Electronics Engineering* *Aug 2018 – Jun 2022*
  - Relevant Coursework: Machine Learning, Natural Language Processing, Data Structures, Statistics

---

## AWARDS & CERTIFICATIONS

- **Hackathon:** Won the Workhuman-sponsored hackathon with Lighthouse (4-person team) [GitHub]
- **Certifications:** Generative AI with LLMs (DeepLearning.AI); AI Agents on Databricks (Databricks Academy)