

Anmol Vijay Bhatia

anmolvb1@gmail.com | 312-826-2020 | linkedin.com/in/anmol-vb | github.com/abhati27 | New York Metro Area

EDUCATION

New Jersey Institute of Technology

Newark, NJ

Master of Science in Artificial Intelligence

September 2026 – December 2027

- **Relevant Coursework:** Artificial Intelligence, Machine Learning, Reinforcement Learning

University of Illinois at Chicago

Chicago, IL

B.S. in Computer Science, Minor in Mathematics

Aug 2019 – May 2023

TECHNICAL SKILLS

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

ML/DL: PyTorch, HuggingFace Transformers, LoRA/QLoRA Fine-tuning, Quantization, XGBoost, NumPy, pandas

LLMs & Agents: LangChain, LangGraph, MCP, RAG, Langfuse, Vector Search (ChromaDB, FAISS, Elasticsearch)

Systems & Cloud: AWS, Docker, Spring Boot, Node.js, React, PostgreSQL, MongoDB, Jenkins, Git

PROFESSIONAL EXPERIENCE

Incedo Inc

Boston, MA

Software Engineer (AI/ML Enablement)

October 2025 – July 2026

- Built a LangGraph-powered multi-agent Network Configuration Summarizer in Python that extracts structured insights from raw telecom data, reducing per-site change analysis from 30+ minutes to under 2 minutes.
- Architected a multi-agent NL-to-SQL system that decomposes user intent, maps schema context, and validates generated queries, enabling 10,000+ non-technical users to self-serve enterprise data.
- Shipped an MCP server exposing safe SQL execution, schema introspection, data validation, retrieval, and audit logging tools, standardizing agent integration across 4 engineering teams.
- Partnered with the Chief Solutions Officer to define AI product strategy across Incedo's BrainSpark and Lighthouse platforms; led discovery workshops that directly shaped platform architecture.

Associate Software Engineer (Full Stack)

July 2023 – October 2025

- Diagnosed redundant API chains causing 40-second response times; restructured microservice communication to eliminate blocking dependencies, reducing latency to 2 seconds across core client-facing endpoints.
- Redesigned role-based access control (RBAC) across a 150K+ user enterprise platform using Java/Spring Boot, reducing unauthorized access incidents by 85%.
- Introduced tiered request queuing stratified by request size to prevent large requests from starving time-sensitive operations, improving overall throughput by 78% and reducing average processing delays from 5 minutes to under 30 seconds.

PROJECTS

Legal Clause Auditor LLM | Python, PyTorch, QLoRA, LLaMA 3 8B, GGUF, FastAPI

HuggingFace

- Fine-tuned LLaMA 3 8B using QLoRA on California court rulings, achieving 89% accuracy in legal clause detection. Quantized the model to GGUF and deployed a self-hosted FastAPI service, reducing inference costs.

Document Intelligence System | Python, PyTorch, Transformers, LayoutLMv3, Donut, LangChain

GitHub

- Built a multi-modal document understanding pipeline using LayoutLMv3 and Donut for layout analysis and key information extraction from legal and financial forms. Achieved 90% extraction accuracy, outperforming traditional OCR + NLP methods by 30%.

Deep Agentic Research Engine | Node.js, LangChain, ChromaDB, Multi-LLM Routing, SSE

GitHub

- Developed an autonomous research agent capable of multi-step reasoning, dynamic query generation, web crawling, source verification, and structured report synthesis using orchestrated calls across OpenAI, Anthropic, and local LLMs.

AWARDS & CERTIFICATIONS

Individual Excellence Award, Incedo R&R Program, North America

2024, 2025

AWS Certified Cloud Practitioner, Amazon Web Services

May 2025