

Bhanu Harsha Yanamadala

Boston, MA (Open to relocation) | (401)-368-9812

bhanuharsha.ya@gmail.com | [linkedin.com/in/bhanu-harsha-y](https://www.linkedin.com/in/bhanu-harsha-y) | github.com/BhanuHarshaY | [Portfolio](#)

EDUCATION

Northeastern University | Boston, MA

Jan 2025 – May 2027

Master of Science in Computer Science

GPA: 3.85/4.0

Coursework: Machine Learning, Generative AI, MLOps, DBMS, Algorithms, Web Development

GITAM University | Visakhapatnam, India

Aug 2020 – May 2024

Bachelor of Technology in Computer Science and Engineering

GPA: 3.80/4.0

Coursework: OOP, Data Structures, Deep Learning, Operating Systems, Computer Networks, Cloud Computing

PROFESSIONAL EXPERIENCE

AI Software Developer Intern

May 2026 – Present

Aosenuma

Boston, USA

- Architected a unified **Infra/LLM observability platform** leveraging OTel, LiteLLM, Langfuse, Prometheus, and Grafana, establishing end-to-end tracing, monitoring, and evaluation of production AI applications.
- Engineered role-based authorization and data-access controls across **Next.js**, **FastAPI**, and **PostgreSQL/Supabase**, enabling secure cross-region visibility and confidential record protection in a multi-tenant case management platform.
- Developed a framework for **automated LLM evaluation** using **RAGAS**, golden datasets, and CI-integrated regression tests to catch model quality regressions before deployment, enabling consistent quality measurement.

Research Assistant – Generative AI

Jan 2026 – Present

Insight X Research Lab, Northeastern University

Boston, USA

- Built a fine-tuning-free prompt optimization system with **DSPy** and **GPT-4o-mini** to generate counter-narratives against online hate speech; ranked **3rd at LT-EDI@ACL 2026 global leaderboard**, with **paper accepted**.
- Designed a **16-metric LLM evaluation framework** across **6 categories** integrating Detoxify, RoBERTa-MNLI, SBERT, BERTScore, and GPT-2 for **automated counter-speech quality assessment**.

Machine Learning & Data Science Intern

May 2023 – Jul 2023

Fluentgrid Limited

Visakhapatnam, India

- Engineered an **NLP chatbot** using Python, **TensorFlow/Keras**, and NLTK, training a sequential neural network on 35+ intent categories with 150+ patterns, enabling partner detection across 15+ organizations and structured lead capture.
- Automated data preparation workflows** using **Tabula-py** and **pdfplumber** to parse multi-section regulatory PDFs, implementing custom CSV-splitting logic to isolate 5 distinct data tables, **reducing manual processing time by 70%**.

PROJECTS

AutoMend | Python, Airflow, Ray, Polars, PyTorch, QLoRA, HuggingFace, W&B, GCP, Docker

- Architected a production **ML infra remediation platform** that bridges observability stacks (Prometheus, Datadog) and execution layers (Kubernetes) to fully automate standard on-call incidents (**3rd Place**, Google Cambridge MLOps Expo).
- Fine-tuned Qwen2.5 LLM** using **QLoRA (4-bit PEFT)** and Bayesian sweeps to author JSON DAG playbooks, validating generations against a **54-metric evaluation pipeline** for strict Pydantic schema compliance.
- Implemented a decoupled execution pipeline where a **RoBERTa classifier** monitors real-time telemetry to instantly trigger fixes, routing structurally **validated workflows** to Temporal for cluster execution with **HITL governance**.

Store Sales Time Series Forecasting | Python, Pandas, XGBoost, LightGBM, CatBoost

- Developed a robust ML pipeline to forecast daily retail sales across **54 stores and 33 product families**, processing over **3 million transaction records** to identify complex demand patterns.
- Engineered **advanced temporal features** including Fourier series for seasonality, rolling window statistics, and Ecuador-specific socioeconomic indicators (wage days, oil prices) to capture non-linear market dynamics.
- Implemented a **stacked ensemble architecture** integrating XGBoost, LightGBM, and CatBoost with Ridge Regression meta-learner, achieving **Top 20%** on Kaggle leaderboard based on time-series cross-validation(RMSE).

GUARD-RAG | Python, GPT-4o, FAISS, Sentence-Transformers

- Engineered a **three-tier RAG** framework featuring a gatekeeper-controlled multi-agent debate system and Program-of-Thought (**PoT**) verification, **reducing calculation hallucinations** and achieving a **12.1% relative gain** over zero-shot baselines on the **financial TAT-QA benchmark**.

TECHNICAL SKILLS

Programming Languages: Python, R, SQL, Java, C, JavaScript, TypeScript

ML/DL Frameworks: PyTorch, TensorFlow, Keras, Transformers, scikit-learn, LangChain, LlamaIndex, RAG, LLMOps

Data & Analytics: Pandas, NumPy, Ray, Polars, DVC, EDA, Matplotlib, MySQL, MongoDB, Tableau, Power BI

Tools & Platforms: Git, CI/CD pipelines, Docker, Kubernetes, AWS, GCP, LLM APIs, Hugging Face, MLflow, Airflow, Vector Databases, LLM embeddings, Prompt Engineering, Cursor.

Web Technologies: React, Next.js, Node.js, Express.js, Redux, HTML, CSS, RESTful API, FastAPI, Vercel, Render.

Certifications: AI Practitioner (AWS), Machine Learning with Python (IBM), Text Retrieval and Search Engines (UIUC)