

Pranshu Acharya

Albany, NY • pranshuac1@gmail.com • (617) 952-3432 • [LinkedIn](#) • [Github](#)

Why Hire Me ←- Check out my work in this portfolio site
(Open to Relocate)

PROFESSIONAL SUMMARY

AI/ML Engineer with 3+ years of experience delivering end-to-end **ML, NLP and GenAI** solutions for Fortune 500 clients across **finance, automation, and information extraction**. Expertise in **ML modeling, fine-tuning LLMs, RAG, and agentic AI** workflows. Actively seeking opportunities to lead complex data science projects and deliver advanced ML and GenAI solution

PUBLICATIONS, CERTIFICATIONS

Papers published

- “Published research on **Offensive language identification** by benchmarking SVM, LSTM, and Multinomial Naive Bayes models on the SemEval-2019 OffensEval dataset, achieving 77% accuracy, and developed a Flask-based NLP application for real-time offensive content detection, demonstrating end-to-end model development and deployment.”

Certifications

- [Neural Networks & Deep Learning – DeepLearning.AI](#)
- [Big Data Foundations \(Hadoop, Spark\) – IBM](#)
- [Server-side Development \(Node.js, MongoDB\) – HKUST](#)

WORK EXPERIENCE

UMass Dartmouth Career Center Data & Automation Engineer

North Dartmouth, MA
August 2025 – July 2026

Developed end-to-end AI-powered automation, data pipelines, and business intelligence solutions that improved operational efficiency, reduced manual effort, and enabled self-service analytics.

- Architected and developed LLM-powered automation workflows that transformed unstructured model outputs into structured intents, enabling seamless backend task execution through **function calling** and workflow orchestration.
- Engineered scalable **ETL pipelines** integrating data from **10+** disparate internal and external sources, **standardizing schemas**, improving data quality, and eliminating **15+ hours** of manual processing per week.
- Designed and optimized **automated data transformation, validation, and ingestion** pipelines to deliver reliable, analytics-ready datasets for **downstream reporting** and **business intelligence**.
- Built **10+** interactive **Power BI dashboards** to monitor student-to-employer conversion funnels, hiring trends, and operational KPIs, enabling self-service **analytics** and **data-driven decision-making** for leadership.
- Automated end-to-end reporting and **data workflows** using **Python, SQL, and Power BI**, reducing operational overhead, improving reporting accuracy, and accelerating business insights.
- Collaborated with cross-functional teams to design scalable **automation** and analytics solutions, leveraging **AI, data engineering, and workflow automation** to improve operational efficiency and business outcomes.

UMass Dartmouth College of Engineering AI Research Engineer (LLM Privacy & Safety)

North Dartmouth, MA
May 2024 – May 2025

Owned the end-to-end development of LLM evaluation frameworks to benchmark foundation models on privacy, safety, and performance. Implemented ICV-based model steering for privacy-preserving model alignment and built real-time PII guardrails using SpaCy NER and custom risk scoring for secure LLM deployment.

- Designed and implemented a comprehensive evaluation framework to **benchmark** Llama 3.1 8B, Mistral 7B, GPT-4o, and Gemini on **privacy-preserving NLP tasks**, analyzing trade-offs between **PII leakage, semantic fidelity, and model utility**.

- Researched and implemented **In-Context Vector** (ICV) latent-state steering techniques using paired and unpaired demonstrations to improve **LLM alignment** toward privacy-safe, non-inferential responses while maintaining response quality.
- Developed a modular guardrail framework integrating **SpaCy NER**, custom PII detection, and real-time redaction to mitigate sensitive information leakage during LLM inference.
- Built **automated benchmarking pipelines** and custom evaluation metrics to systematically assess **LLM safety, privacy, and semantic performance**, enabling reproducible experimentation and **model comparison**.
- Conducted **comparative analysis** of open-source and proprietary LLMs across privacy, robustness, and response quality **benchmarks**, generating insights to guide model selection and **deployment**.
- Collaborated with research and engineering teams to translate AI safety and privacy-preserving **ML research** into scalable, production-ready LLM solutions.

Cognizant Technology Solutions

Programmer Analyst (Backend & Cloud)

Hyderabad, India

March 2022 – July 2023

Owled the end-to-end development of scalable Java/Spring Boot microservices, AWS cloud infrastructure, and CI/CD pipelines, delivering high-performance backend services, improving application reliability, and streamlining software delivery

- Architected and enhanced **Java/Spring Boot microservices** and **RESTful APIs** for a global retail platform, reducing API latency by **25%** through **performance optimization**, efficient **service orchestration**, and **database tuning**.
- Developed scalable, secure **backend services** using **microservices** architecture, enabling seamless integration with **enterprise applications** and **cloud-based services**.
- Managed **AWS** cloud infrastructure, utilizing **Auto Scaling**, **Elastic Load Balancing**, and **CloudWatch** to achieve **99.9% uptime** and ensure **high availability** under fluctuating workloads.
- Automated software delivery through **Jenkins CI/CD** pipelines, integrating **unit testing**, **build validation**, and **deployment automation** to improve release efficiency and reliability.
- Built robust **test automation** frameworks using **JUnit** and **Mockito**, increasing code coverage to **85%** and reducing production defects by **15%** through continuous **quality assurance**.
- Partnered with product managers, QA engineers, and DevOps teams to deliver high-quality, cloud-native backend solutions while driving performance, **scalability**, and operational excellence.

PROJECTS

Secure RAG QA System

- Architected a **RAG pipeline** with **LangChain** and **ChromaDB** using **hierarchical chunking** and **semantic indexing** (+40% retrieval relevance vs. keyword search), with an **in-chain PII redaction** guardrail for HIPAA/SOC2 compliance.

Repo Diagram Generator (Agentic)

- **Streamlit** app using **GPT** with recursive **AST parsing** to extract code structure and autogenerate **Mermaid.js** diagrams, an LLM tool-calling workflow over raw source.

AI Podcast Engine (YHack)

- **Go/Next.js** engine that scrapes trends, summarizes via **OpenAI**, and synthesizes audio with the **ElevenLabs API** — an end-to-end multi-step generative pipeline

Emotion & Genre CNN

- Multi-output **PyTorch CNN** on the **DEAM** dataset predicting **valence/arousal (0.107 RMSE)** while classifying genre.

SKILLS

- **LLM & RAG:** Retrieval-Augmented Generation (RAG), LangChain, LlamaIndex, LCEL, prompt engineering, embeddings, semantic search, chunking, fine-tuning, model alignment, OpenAI / Anthropic / Hugging Face APIs
- **Agents & Tooling:** Tool-calling / function calling, AI agents, orchestration, context & memory management, Model Context, Protocol (MCP), LangGraph
- **ML & NLP:** PyTorch, Scikit-learn, XGBoost, Transformers (BERT, RoBERTa), CNNs, SpaCy (NER), NLTK, Word2Vec, TF-IDF, Pandas, NumPy

- **Vector & Data Stores:** Pinecone, Chroma, Qdrant, Weaviate, PostgreSQL, MongoDB, DynamoDB, Oracle
- **Infra & MLOps:** AWS (S3, Lambda, ECS, Glue, Redshift, Athena), Docker, CI/CD (GitHub Actions, Jenkins), model evaluation, guardrails, PII redaction, monitoring
- **Languages:** Python, JavaScript (Node.js, React, Next.js), Java (Spring Boot), SQL / PL-SQL, Shell

EDUCATION

UNIVERSITY OF MASSACHUSETTS DARTMOUTH	Dartmouth, MA
<i>Master of Science in Data Science GPA(4/4)</i>	Aug 2023-May 2025
Vellore Institute of Technology (VIT)	Vellore, India
<i>Bachelor of Technology in Computer Science & Engineering</i>	Aug 2018 - May 2022