HARSH AGRAWAL

(857)-313-0855| Pleasanton, CA | agrawal.har@northeastern.edu | LinkedIn | GitHub | Google Scholar | Portfolio

EDUCATION

Northeastern University **Boston, MA**
Master of Science, Computer Science, ***GPA – 3.7*** *September 2022 - August 2024*
- **Relevant Courses:** Programming Design Paradigm, DBMS, Algorithms, Pattern Recognition and Computer Vision, ML

Narsee Monjee Institute of Management Studies **Mumbai, India**
Bachelor Of Technology (Hons.), Computer Engineering, ***GPA – 3.75*** *July 2018 - August 2022*
- **Relevant Courses:** Artificial Intelligence, Image processing, Soft Computing, Natural Language Processing

SKILLS

Languages: Python, Java, C++, SQL, R, JavaScript, MATLAB
Frameworks: TensorFlow, PyTorch, Scikit Learn, Keras, Hugging Face Transformers, PEFT, LangChain, NumPy, Pandas
Cloud & MLOps: AWS (S3, Lambda, SageMaker, Athena, Glue, Redshift), GCP, Databricks, Airflow, Docker, Jenkins, Kubernetes, Prometheus, FastAPI, Temporal, MCP, Triton Inference Server, CUDA, Git
Data & Vector Search: PostgreSQL, MinIO/S3, Pinecone, Weaviate, Apache Kafka
Publications: 10.1109/CONIT51480.2021.9498561, 10.1109/ICCCNT51525.2021.9579920, 10.1109/ICAIS50930.2021.9395895

PROFESSIONAL EXPERIENCE

Calfus Inc **Pleasanton, CA**
Software Engineer, Generative AI *May 2025 – Present*
- Architecting **Agent Foundry**, a **scalable generative-AI workflow platform**; designing core infrastructure with **Temporal, PostgreSQL, and S3-compatible MinIO** to support **fault-tolerant, multi-tenant agent workflows**, containerized and **executed across distributed Kubernetes clusters**
- Engineered a **Docker-Kubernetes** execution layer with a **custom pod controller to auto-scale worker pods** serving **LLM inference, vector search**, and **prompt-engineering/RLHF evaluation pipelines**, reducing **idle GPU time by 35%**

BulkMagic **Boston, MA**
Machine Learning Engineer *October 2024 – April 2025*
- Led the initiative for a collaborative filtering-based **recommender engine** and **prototyped transformer-based (BERT4Rec, SASRec)** and **graph-based (GraphSAGE)** recommender models achieving **20% higher NDCG@10** than **BPR model**
- Orchestrated **scalable data pipelines (Spark, Airflow)** and containerized **model deployment (Docker, Kubernetes),** with **Jenkins-driven CI/CD automation**, **cutting recommendation pipeline latency by 40%**
- **A/B tested** the system on **10K+ interactions** (incl. synthetic data), achieving **25% higher deal uptake**, **+18% CTR**, **+15% retention**, demonstrating strong user engagement and business impact

Amazon Robotics **Boston, MA**
Data Scientist Co-op *August 2023 – December 2023*
- Developed a system to **classify and categorize support tickets** based on complexity, addressing **the issue of ticket backlog** by employing **custom clustering algorithms** on integrated data from multiple sources, using **AWS SageMaker and Glue**
- Designed a **comprehensive downtime monitoring system for robotic arms**, using **AWS Lambda and Athena** to optimize operations**, identifying top contributors to downtime,** and **successfully mapping 70% of downtime occurrences**
- Conducted **extensive data analysis using AWS Data Lake, SQL, and PostgreSQL** to gather and process large datasets and applied ML techniques to solve operational challenges, **decreasing downtime for the robotic arm by 15%**

DosBro Infotech **Mumbai, India**
AI Developer *August 2020 – August 2022*
- **Engineered a BERT-/T5-based** content summarization pipeline for JioTV companion apps, achieving a **ROUGE-L score of 0.88** and **expediting editorial workflows** by **45%**, which boosted quick-turnaround news coverage and live event updates
- Implemented an **automated multi-lingual question-answering system** leveraging **PyTorch** and **attention-based architectures**, enabling dynamic content queries in **three Indian languages** and increasing **user engagement by 30%**
- **Developed a YOLOv4-based brand-detection framework for sponsor analytics**, **processing 300K+ social media images** monthly and delivering a **mean Average Precision (mAP) of 89%** while **cutting manual tagging efforts by 40%**
- Orchestrated a containerized **object tracking solution with Deep SORT** for real-time brand exposure insights, scaling to **1M+ video frames weekly** and maintaining **sub-200ms inference latency** with GPU acceleration

PROJECTS

Progress Note Understanding: Assessment and Plan Reasoning *May 2024 – August 2024*
- Engineered and fine-tuned **LLM-based transformer models (BERT, ClinicalBERT)** and **BiLSTM** to classify relationships in clinical notes, achieving a **Macro F1 score of 0.78**, with a focus on improving model generalization in healthcare tasks
- **Optimized Tiny-ClinicalBERT and Tiny-BioBERT** using **transformer-layer distillation**, **aligning the attention maps and hidden states to reduce model size by over 60% while retaining 95% of the original performance**

Personalized GIF-based Reply Recommendation System *January 2022 – May 2022*
- **Formulated** a **multi-modal transformer-based (VINVL)** approach to predict **relevant GIFs as text-message replies**, collecting **1.5M tweets** via Twitter API, and matching them with **115k GIFs**, exceeding **80%** overall precision
- **Engineered** a **collaborative filtering** framework on model responses, **combining sentiment analysis and user characteristics**, delivering personalized GIF replies and **slashing average response time by 50%** across chat platforms