ADITYA RATAN JANNALI

+19167767065 | jannali.a@northeastern.edu | LinkedIn | GitHub

EDUCATION

Northeastern University (GPA: 4.00) Boston, MA
MS in Artificial Intelligence, September 2023 – December 2025
Courses: Machine Learning, Foundations of Artificial Intelligence, Programming Design Paradigms, Algorithms, NLP, MLOps, Masters Project (Kolmogorov Arnold Networks & Reinforcement Learning), AI for HCI (Audit)

Vellore Institute of Technology (GPA: 9.11) Chennai, India
Bachelor of Technology in Electronics and Communications Engineering July 2017 – June 2021
Courses: Statistics, Calculus, Linear Algebra, Probability Theory, Data Structures, Machine Learning, Digital Image Processing.

EXPERIENCE

Institute for Experiential AI

Research Assistant *July 2025 – present. Boston, MA*

- Shaped research direction under Dr. Agata Lapedriza by synthesizing 35+ VLM Theory of Mind papers into novel taxonomy—designed experimental framework translating hypotheses into measurable methodology for A-tier conference publication.
- Accelerated experimentation by 80% through GPU-optimized inference pipeline—parallelized model execution to compress evaluation cycles from hours to under 30 minutes, enabling rapid iteration.

Data Scientist Co-op, Generative AI *Jan 2025 – Aug 2025. Portland, ME*

- Delivered a serverless RAG-based QA system for a Reference Management Service. Reduced query response time by 90% (from 1 sec to 100ms) for 10K userbase through optimized AWS system architecture. Engineered deployment with multi-format document ingestion, enabling instant answers from reference materials at scale.
- Improved retrieval accuracy by 40% by leading data quality strategy across 100K heterogeneous documents-analyzed corpus characteristics, designed filtering logic that reduced hallucinations and false positives for production deployment.
- Delivered 85% F1-score legal document classifier for a Legal Consulting Firm despite 10:1 class imbalance. Evaluated and benchmarked 5 ML approaches, implementing targeted solutions that enabled automated requirement extraction and classification at scale.

Amazon

Application Engineer III *August 2021 – August 2023. Chennai, India*

- Cut ticket volume by 35% and saved 20+ engineering hours monthly by leading cross-team automation initiative. Identified systemic DevOps issues, built remediation workflows, and scaled solutions organization-wide
- Enabled executive cost optimization by building real-time Redshift expense dashboard - architected ETL pipeline surfacing cluster utilization patterns, allowing leadership to eliminate waste and right size resources across teams
- Maintained 99.9% uptime for production data systems serving Amazon Exports Organization - owned operational reliability for Data Science/Engineering pipelines through proactive monitoring and rapid incident resolution

Software Support Engineer II Intern *January 2021 – July 2021. Chennai, India*

- Built and deployed an automated metrics distribution platform that reduced manual reporting overhead by 80% for service owners. Designed data collection and aggregated service-level metrics computation, and scaled solution across 5+ engineering teams, improving operational visibility for stakeholders.
- Owned production release cycle as designated release engineer for 6 critical pipelines, ensuring zero-downtime deployments and maintaining 99.9% service availability throughout tenure.

Antpod

System Development Engineer Intern *April 2020 – December 2020. Chennai, India*

- Involved in Research and Development of a proof of concept for an unmanned vehicle in ‘Land Stress Identification and Remote Sensing’.
- Prototyped an algorithm using Deep Fully Connected Convolution Network using Keras and TensorFlow to segment images from the dataset and perform classification to identify plant disease.

SKILLS

- **Machine Learning & AI:** PyTorch, TensorFlow, Keras, Scikit-learn, Transformers, Hugging Face, OpenAI Gymnasium, Reinforcement Learning (Q-Learning, DDQN), Deep Learning, Transfer Learning, Kolmogorov Arnold Networks (KANs)
- **NLP & Vision:** GPT-2, ALBERT, DeBERTa, Qwen, Gemma, Llava, Llama, RAG (Retrieval-Augmented Generation), OpenCV, Medical Image Analysis, Image Segmentation, Document Classification
- **MLOps & Data Engineering:** TensorFlow Data Validation (TFDV), DVC (Data Version Control), Model Versioning, CI/CD Pipelines, NumPy, Pandas, PostgreSQL, ETL/ETLM Pipelines, Data Preprocessing, Monitoring & Automation, Dashboard Development

- **Cloud & DevOps:** AWS (EC2, S3, Lambda, SageMaker, Redshift, DynamoDB, QuickSight, Athena, IAM, KMS, VPC, ECS), Google Cloud Platform (GCP), Serverless Architecture, Git, Release Engineering, End-to-End Testing, Container Orchestration
- **Software Engineering:** MVC Architecture, Design Patterns, Unit/Integration Testing, JSwing
- **Programming Languages:** Python, Java, C++, SQL, JavaScript, HTML/CSS, Shell/Bash

CERTIFICATIONS

- (CITI) Conflict of Interest, Human Subject Research, Responsible Conduct of Research for Engineers.
- (Coursera) Machine Learning, Deep learning using TF – CNN and NLP, Natural Language Processing – Classification And Vector Spaces, Probabilistic Models, Sequential Models, Digital Image Processing.

PROJECTS

- **Disease Prediction & Medical RAG system:** [GitHub], Team Lead, Sept 2025 – present
  Leading the team to build an end-to-end MLOps pipeline for disease prediction from radiological scan images with an integrated RAG system for medical report analysis and diagnosis assistance. Implementing data versioning with DVC, data validation using TFDV, and deploying on Google Cloud Platform.
  Skills: Google Cloud Platform**,** TensorFlow, TFDV, DVC, GCP, Medical Image Analysis, RAG, MLOps, Team Leadership

- **Research Project: Evaluating KANs for Reinforcement Learning Applications,** Sept 2024 – December 2024
  Implemented a DDQN using KANs Dr. Raj Venkat's supervision to evaluate performance on Atari game environments in OpenAI Gymnasium. This project is a study of comparative analysis between KAN-based and traditional MLP-based reinforcement learning architectures and training algorithms. Investigating KANs' potential as an alternative to Multilayer Perceptron models in Deep RL frameworks.
  Skills: PyTorch, Reinforcement Learning, Neural Network Design, Experimental Analysis

- **Language Model Interpretability** [GitHub]**,** Oct 2024 – Dec 2024
  Fine-tuned GPT-2 Small on a 6-label emotion dataset for sequence classification. Masked attention heads and replaced key tokens to analyze attention and token importance and interpretability of predictions.
  Skills: Transformers, NLP, Explainable AI, Model Evaluation, Attention Analysis

- **RL Tic-tac-toe** [GitHub], May 2024 – June 2024
  Implemented Q-Learning to train two agents with different reward schemes to play against each other. Each valid board configuration is a state in this environment which is modelled as a 9-digit ternary sequence eliminating the need to precompute the observation space.
  Skills: Reinforcement Learning, Python, NumPy, OpenAI Gym, Environment Design

- **LLM generated vs Human text classification** [GitHub], April 2024 – April 2024
  Finetuned and evaluated two transformer models, ALBERT and DeBERTa-XS on DAIGT Proper Train dataset to detect LLM generated essays. Achieved an accuracy of 96.57% for ALBERT, and 99.43% for DeBERTa.
  Skills: NLP, Transfer Learning, Transformer Fine-Tuning, Data Preprocessing, Model Evaluation

- **Image Processor** [GitHub], November 2023 – December 2023
  Designed and MVC, developed and wrote unit and integration tests for an image processing application that performs various image manipulations on custom images. The controller uses command design pattern to listen and process user interaction, and JSwing for the view.
  Skills: Java, Software Design Patterns, GUI Development, Testing & QA