ADITYA RATAN JANNALI

+19167767065 | jadityaratan@gmail.com | LinkedIn: [URL] | GitHub: [URL] | website: [URL]

Applied AI Researcher | Machine Learning | MLOps | NLP | Cloud Deployment (AWS/GCP)

EXPERIENCE

Institute for Experiential AI *"Jan 2025 – Present"*

AI Researcher *Boston, MA*

- **Optimized GPU-accelerated inference pipeline (PyTorch + vllm)** reducing evaluation latency by **80%** (3h to 30min) across **Gemma, LLaMA, Qwen, Llava, Bagel VLMs**, enabling faster multi-modal experimentation.
- **Synthesized 35+ VLM Theory-of-Mind papers** into a **taxonomy and experimental design framework**, operationalizing model cognition evaluation.

Data Scientist, Generative AI *Portland, ME*

- **Architected & deployed production-grade RAG system (AWS Lambda, EC2, FAISS, OpenSearch)** for a Reference Management Company, achieving **90% latency reduction** for **10K+ users**: improved retrieval precision by **40%** for user forum queries.
- **Performed statistical corpus analysis** across **100K+ heterogeneous documents** to identify noise patterns, designed and implemented filtering heuristics, and established evaluation framework measuring **hallucination rates and retrieval accuracy** in production environment.
- **Developed and deployed Text Extraction and Classification XGBoost model** achieving **85% F1-score** on **imbalanced datasets (10:1)**, automating extraction pipelines for a **Legal Consulting Firm**.
- Experimented on **Encoder, Decoder transformers** with **In-context Learning, ensemble methods, and SMOTE** to select best model.

Amazon *"Jan 2021 – Aug 2023"*

Application Engineer III *Chennai, India*

- **Reduced ticket volume by 35%** through strategic automation initiatives**, improving operational efficiency** and freeing up resources for high-priority tasks.
- **Directed a cross-team Operational Excellence initiative**, designing and implementing scalable solutions that enhanced workflow efficiency and **reduced engineering hours by 20+ hours monthly**.
- **Built ETL pipeline** processing cluster metrics, designed executive **QuickSight dashboard surfacing utilization patterns** and cost anomalies, enabling leadership to eliminate waste and **right-size resources across Data Science/Engineering teams**.
- **Owned operational reliability for production data systems** serving Amazon Exports Organization, maintaining 99.9% uptime. Implemented proactive monitoring and reporting pipelines **(Redshift, S3, SQL, QuickSight)** for Applied Science/Engineering teams, established incident response protocols, and **performed root cause analysis to prevent recurring failures**.

Software Support Engineer II *Chennai India*

- **Built and deployed an automated metrics aggregation utility** 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 team**s, improving operational visibility for stakeholders.
- Owned production release cycle as designated **release engineer** for **6 end customer production pipelines**, ensuring zero-downtime deployments and maintaining 99.9% service availability throughout tenure.

Antpod *"April 2020 – December 2020"*

AI Research Engineer *Chennai, India*

- Core researcher in the development of unmanned vehicle systems for remote soil health analysis.
- **Designed and built a Deep Neural Classification model (TensorFlow, Keras)** to segment images from the dataset and perform **classification on plant diseases**.

EDUCATION

Northeastern University (GPA: 4.00) *"September 2023 – December 2025"*

Master of Science, Artificial Intelligence Boston, MA

Courses: Machine Learning, Artificial Intelligence, Design Patterns, Algorithms, NLP, MLOps, Master's Research (KAN & RL), AI for HCI (Audit)

Vellore Institute of Technology (GPA: 9.11) *"July 2017 – June 2021"*

Bachelor of Technology, Electronics and Communications Engineering Chennai, India

Courses: Statistics, Calculus, Linear Algebra, Probability Theory, Data Structures, Machine Learning, Digital Image Processing.

SKILLS

- **Core AI/ML:** Machine Learning, Deep Learning, Transformers, Generative AI, Reinforcement Learning
- **Frameworks:** PyTorch, TensorFlow, Hugging Face, OpenAI Gymnasium, LangChain, vllm, ollama
- **Machine Learning Operations (MLOps):** MLFlow, DVC, TFDV, CI/CD, Docker, Apache Airflow
- **Data Engineering:** FAISS, OpenSearch, PostgreSQL, ETL Pipelines, Data Validation
- **Cloud & DevOps:** AWS, GCP, GitHub Actions
- **Programming:** Python, Java, C++, SQL, Shell
- **Specialties:** RAG Systems, Vision-Language Models, LLM Evaluation, Applied Research, Experimentation Frameworks

CERTIFICATIONS

- (Coursera) Machine Learning, Digital Image Processing, Deep learning using TensorFlow, Natural Language Processing
- (CITI) Conflict of Interest, Human Subject Research, Responsible Conduct of Research for Engineers.

PROJECTS

Disease Prediction & Medical RAG system: [GitHub], Team Lead, Sept 2025 – present

Leading end-to-end **MLOps pipeline** development for radiological scan-based disease prediction application. Implementing data versioning (**DVC**), validation (**TFDV**), experiment tracking (**MLFlow**) and **GCP** deployment. **Integrating a QA RAG system** for medical report analysis and diagnostic assistance. Orchestrating the pipeline using **Apache Airflow** DAGs.

Cite-your-Sources: [GitHub] May 2025 – June 2025

Developed and evaluated post-hoc **answer attribution models** for **long-document comprehension**, leveraging text span identification and **Vectara's HHEM model** to assess citation accuracy – advancing **trustworthy RAG and QA systems** through improved **source grounding**.

Research Project: Evaluating KANs for Reinforcement Learning Applications, Sept 2024 – December 2024

Implemented DDQN with **Kolmogorov-Arnold Networks** for **Atari** environments in OpenAI Gymnasium**.** Conducted comparative analysis between **KAN-based and MLP architectures.** Evaluated performance metrics and training efficiency under Dr. Raj Venkat's supervision..

Language Model Interpretability [GitHub]**,** Oct 2024 – Dec 2024

Fine-tuned GPT-2 Small on emotion classification dataset using **PyTorch and CUDA**. **Analyzed attention mechanisms** through head masking and token replacement techniques**.** Developed visualization methods for transformer attention patterns.

RL Tic-tac-toe [GitHub], May 2024 – June 2024

Implemented **Q-Learning with multi agent training** in a custom gymnasium environment using different reward schemes. **Optimized state representation** as ternary sequences for efficient training **minimizing memory requirements** to store states.

LLM generated vs Human text classification [GitHub], April 2024 – April 2024

Fine-tuned ALBERT and DeBERTa-XS transformers on the DAIGT Proper Train dataset. Achieved **more than 96% accuracy** in detecting AI-generated content. Implemented efficient data preprocessing pipeline and cross-validation framework.

Image Processor [GitHub] [Demo] October 2023 – December 2023

MVC architecture Image Processing **java** application with Controller based on **command design pattern**. Implemented comprehensive test suite including **unit and integration tests**. Built a **JSwing UI** for intuitive user experience and accepting user inputs.