Gayathri Akkinapalli

(413) 466-1166 | gayathri.akkinapalli2000@gmail.com | linkedin.com/in/gayathri-akkinapalli/ | github.com/akkina10gayu

Education

University of Massachusetts Amherst (UMASS) | MS in Computer Science | **CGPA: 4.0/4.0** — **Sep 2023 - May 2025**

Indian Institute of Information Technology (IIIT) | B.Tech in Computer Engineering | **CGPA: 9.07/10.0** — **Aug 2017 - May 2021**

Publications & Preprints

1. LS-GAN: Human Motion Synthesis with Latent-space GANs — **IEEE WACVW '25**
2. Iterative Critique-Refine Framework for Enhancing LLM Personalization — arXiv:2510.24469
3. Safe to Serve: Aligning Instruction-Tuned Models for Safety and Helpfulness — arXiv:2412.00074
4. Automated Model Selection for Tabular Data — arXiv:2401.00961

Technical Skills

Programming Languages: Python, R, C, HTML, SQL, MySQL, NoSQL | Familiar: Java, C++, PHP, CSS, JavaScript, C#

Tools/Libraries: LangChain, MCP, Unsloth, TRL, OpenAI, Wandb, Keras, PyTorch, TensorFlow, PySpark, SKLearn, NumPy, vLLM, Copilot

Software/Frameworks: Git, Docker, Kubernetes, Kafka, Flask, FastAPI, gRPC, AWS, Splunk, MongoDB, Tableau, CI/CD, Snowflake, FAISS

Professional Experience

Data Science Volunteer | UMass Amherst — **Aug 2025 – Present**

- Built **EduNotes,** a **multi-agent RAG** study assistant using a hybrid LLM setup (**Groq** Llama-3.1-70B API, local Flan-T5) with LangChain & **ChromaDB**, coordinating retriever, scraper, summarizer, note-maker agents via an async pipeline, query routing, with **85%** retrieval accuracy.
- Integrated AI-generated flashcards, quizzes, progress analytics with FastAPI backend & Streamlit UI, reducing note generation time by **70%**.

Graduate Student Researcher | Cisco — **Jan 2025 – May 2025**

- Developed **PerFine**, a training-free critique-refine **Agentic RAG** framework for enhancing personalization in long-text using **LangChain**
- Evaluated using **LLM-as-a-Judge** (G-Eval), improving personalization by **13%** and Meteor by **10%** over RAG-based baselines
- Enhanced outputs using profile-grounded critic feedback retrieved via **Pinecone** (Vector DB), **FAISS**, **MCP** to refine style and content relevance.

Research Assistant | UMass IESL Lab — **Aug 2024 – Dec 2024**

- Designed an **autoregressive** model that performs lookahead by decoding in superposition with just two forward passes using cross attention.
- Applied the approach to **machine translation**, improving **BLEU** score and generation quality of the **MT5** model by approximately **15%**.

Machine Learning Engineer | Carelon Global Solutions — **Jun 2021 – Jul 2023**

- Built **Recommendation Systems** using **NER**, SpaCy & ML models like **XGBoost**, LightGBM in **PySpark** improving NDCG@5 by **75%.**
- Deployed models into ENSO **ML pipeline** using RabbitMQ, **Kubernetes**, and Kafka, cutting deployment time by ~**70**% via **CI/CD** pipelines.
- Integrated **REST APIs** with **Flask** for end-to-end model automation using **Hive**, **MongoDB**, **Redis**, reduced care plan creation time by ~**2hrs**.
- Created **Splunk** Dashboard for user feedback KPIs & ran **A/B** testing on recommendations, driving a **60**% improvement in model performance.
- Designed Aspect-Based Sentiment Analysis on call transcripts using **RoBERTa**, BERT, SpaCy, achieved **85%** accuracy & **0.81** F1-score.
- Used AWS **SageMaker**, Kubeflow, S3, & GlueDB for model development and built **Conversational AI** bot resolving **65**% of patient queries.
- Integrated web-scraped healthcare articles into **Elasticsearch** with **ranking** optimization, reducing content retrieval time to **~120ms**.
- Developed an **IBM Watson** Assistant chatbot that handled **100+** queries daily, and automated real-time insights through metric generation.
- Built **Lambda** functions to extract conversational data from IBM Object Storage to **S3** and moved it into **DynamoDB** using **AWS Glue**.
- Deployed **ETL pipelines** and ML models using Google Cloud **Vertex AI** with **Airflow** for orchestration, and **Docker** containerization.

AI Engineer Intern | SensorDrops Networks (STEP at IIT Kharagpur) — **Aug 2020 – Sep 2020**

- Designed real-time Social Distance Monitoring system using **YOLOv3** with live feed, bounding boxes, and **90**% detection accuracy.
- Deployed the application on **AWS EC2** with **Docker**, enabling real-time analytics with ~**200ms** latency via socket-based data transfer.

AI Engineer Intern | Centre for Development of Advanced Computing (C-DAC) — **May 2020 – Aug 2020**

- Developed a prototype of a customized deep CNN model to identify COVID-infected chest X-rays with an **accuracy of around 92%**.
- Trained the model on **High Performance Computing (HPC)** for three chest X-ray classes, achieving a validation **F1-score of 0.9**.

Academic & Research Projects

Aligning LLMs towards safety and helpfulness | UMass | Github — **Feb 2024 - May 2024**

- Aligned **LLaMa-2** toward safety using **LoRA**, QLoRA on PKU-SafeRLHF benchmark with SFT, RAFT, **RLHF, DPO** in Unsloth & **TRL**.
- Scored **93% safe** on DPO (40% SFT) with Llama-Guard on I-CoNa. Implemented **LLM-as-a-judge** to evaluate safety and helpfulness.

Human Motion Synthesis with Latent-space GANs | UMass | Github — **Feb 2024 - May 2024**

- Generated text-to-motion sequences in latent space utilizing **GANS,** VAE, **CLIP** on HumanML3D with **Distributed training** in lightning.
- Achieved a **FID of 0.48** with GAN in the latent space with **91% in FLOPs reduction** compared to Latent **Diffusion Model** on HumanML.

Real time Stock Analysis | UMass | Github — **Sep 2024 - Dec 2024**

- Built **Kafka** & **PySpark** pipeline for Stock news analysis using LLaMa. Achieved 84% match with **GPT-4** reducing processing **time by 40%**.
- Integrated with a **RAG framework** for financial information retrieval (dense + keyword search), securing a further **5% increase** in accuracy.