

# Gayathri Akkinapalli

(413) 466-1166 | [gayathri.akkinapalli2000@gmail.com](mailto:gayathri.akkinapalli2000@gmail.com) | [Linkedin](#) | [Github](#) | [Google Scholar](#) | [Website](#)

## Summary

ML Engineer with 4+ years of experience building production ML systems, search and retrieval pipelines, recommendation models, and LLM based applications. Skilled in developing intelligent agents, with a strong background in NLP and scalable ML pipelines.

## 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

## Skills

Python, R, SQL; LangChain, MCP, Unsloth, Wandb, Keras, PyTorch, TensorFlow, PySpark, SKLearn, NumPy, vLLM, OpenAI, Copilot; Docker, Kubernetes, Kafka, Flask, FastAPI, AWS, Splunk, MongoDB, Tableau, CI/CD, Snowflake, FAISS, AWS, GCP

## Experience

**Machine Learning Engineer** | Oracle Aug 2025 - Present

- Built a hybrid search pipeline combining ColBERT-v2 dense retrieval over FAISS with BM25 sparse search, fused via Reciprocal Rank Fusion, improving Recall@20 by 26% over the keyword-only baseline.
- Fine-tuned MonoBERT re-ranker on 200K query-click pairs to reorder top-50 fused candidates, improving MRR by 18%.
- Implemented query normalization for noisy and verbose inputs, reducing zero-result rate from 7.2% to 3.1% across 30K daily queries.
- Evaluated retrieval quality using Recall, NDCG on a 3K human-ranked judgments, with weekly regression benchmarks on MLflow.
- Deployed with FastAPI, Kubernetes, and Redis caching, optimizing first-stage retrieval for p99 latency under 80ms.

**Graduate Researcher** | Cisco Jan 2025 - Jul 2025

- Developed PerFine, a training-free critique-refine framework for personalized long-text via iterative profile-grounded feedback using FAISS, MCP. Improving personalization by 13% and Meteor 10% over baselines with LLM-as-a-Judge (G-Eval) evaluation.

**Research Assistant** | UMass IESL Lab Aug 2024 - Dec 2024

- Implemented an autoregressive model performing lookahead via superposition decoding with cross attention in 2 forward passes, improving BLEU score and generation quality of the MT5 model by 15% on machine translation tasks.

**Data Scientist** | Caelon Global Solutions Jun 2021 - Jul 2023

- Built Recommendation Systems using NER, SpaCy & ML models like XGBoost in PySpark improving NDCG@5 by 75%.
- Designed and deployed end-to-end ML pipelines with Flask APIs, Hive, MongoDB, Redis, reducing careplan creation time by 2 hours.
- Performed Aspect-Based Sentiment Analysis on call transcripts with RoBERTa, SpaCy, achieved 85% accuracy & 0.81 F1-score.
- Engineered Conversational AI bot using DistilBERT for intent classification and BERT QA with RAG over S3, deployed on AWS SageMaker, resolving 65% of user queries across 1000+ daily interactions.
- Orchestrated model deployment with Kubernetes, Kafka, CI/CD pipelines, reducing deployment time by ~70%, improving scalability.
- Created Splunk dashboards and conducted A/B testing for feedback-driven model improvements, achieving 60% performance gain.

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

- Modeled real-time Social Distance Monitoring system leveraging YOLOv3 with live feed, bounding boxes, 90% detection accuracy.
- Deployed application on AWS EC2 and 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

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

## Projects

**EduNotes: AI Study Assistant** | UMass | [Github](#)

- Designed a multi-agent RAG study assistant, centralizing study materials into a queryable KB, using intelligent agents for search, content processing and adaptive summarization with Llama-70B/8B/Scout, Gemini-2.5 Flash, LangChain, and ChromaDB.
- Implemented flashcards, quizzes, notes, summaries, and conversational AI chat with 7 learning modes via FastAPI & Streamlit, supporting 6+ input types (PDF, URL, text, topic, web search) with Research Mode for academic references and PDF figure, table analysis.

**Aligning LLMs towards safety and helpfulness** | UMass | [Github](#)

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

## Publications & Preprints

- LS-GAN: Human Motion Synthesis with Latent-space GANs [IEEE WACVW '25](#)
- PerFine: Iterative Critique-Refine Framework for Enhancing LLM Personalization [arXiv:2510.24469](#)
- Safe to Serve: Aligning Instruction-Tuned Models for Safety and Helpfulness [arXiv:2412.00074](#)
- Automated Model Selection for Tabular Data [arXiv:2401.00961](#)