Praneeth Reddy Mallupalli

+1 (224) 536-1466 | praneeth.mallupalli@gmail.com | Evanston, IL | linkedin.com/in/praneethreddym | github.com/prm036

Education

Northwestern University, Evanston, IL — **Dec 2026 (Expected)**
M.S. in Computer Science (Concentration: Artificial Intelligence) — *GPA: 3.9/4.0*
- **Coursework:** Machine Learning, Deep Learning, Generative AI, Agent AI, NLP, Large Language Models

Osmania University, Hyderabad, India — **June 2021**
B.E. in Computer Science and Engineering — *GPA: 3.9/4.0*

Technical Skills

Languages	: Python, C/C++, Java, SQL.
ML Frameworks	: PyTorch, Keras, Scikit-learn, NumPy, SciPy, Pandas.
Generative AI	: LLMs, Video Foundation Models, Transformers, Diffusion, Fine-Tuning, RAG, AI Agents, Quantization.
Data & Infrastructure	: Docker, Kubernetes, Git, FastAPI, ETL Pipelines, PostgreSQL, Airflow, Elasticsearch, Distributed Systems.
Others	: Time-Series Modeling, Mechanistic Interpretability, SLURM.

Projects

Fine-Tuning LLMs to solve Math Problems without losing Safety Alignment | GitHub
- Fine-tuned Qwen2.5 on GSM8K dataset using LoRA, improving math accuracy to **0.81** while retaining a 0.88 safety alignment score.
- Conducted systematic ablations of LoRA rank, learning rate, and regularization (weight decay, dropout) to characterize and mitigate catastrophic forgetting.

Multimodal Human Activity Recognition(HAR) with IMU and Audio | GitHub
- Built an end-to-end multimodal ML pipeline for HAR, covering data preprocessing, feature extraction, model training, and evaluation using **15+** long-recording sessions.
- Extracted 48 statistical features from IMU time-series and generated mel-spectrograms from audio for activity classification.
- Trained and compared MobileNetV2 audio models, 1D CNN IMU models, with late-fusion multimodal architectures, achieving 0.73 F1 for activity detection and analyzing onset/offset delays for temporal localization.

Post-Training Quantization (PTQ) for Diffusion Transformers(DiTs) | GitHub
- Evaluated low-bit FP and INT quantization on the PixArt-Alpha (0.6B) **diffusion** model under equivalent precision budgets.
- Improved visual fidelity with FP over INT quantization, lowering **FID from 42.4 to 38.8** under **W4A8** with comparable CLIP and ImageReward scores.

Adaptive Hybrid RAG for Multi-Hop Question Answering | GitHub
- Built an adaptive RAG pipeline over 40K documents, combining keyword search and semantic retrieval with cross-encoder reranking.
- Implemented complexity-aware query routing to dynamically choose between single-hop and multi-hop retrieval before generation.
- Evaluated on 3K questions across 6 QA datasets, achieving 90.3% Hit@10 and 30% gain in answer accuracy over LLM-only baseline.

Professional Experience

Kilwa Technologies LLC — Chicago, IL
Machine Learning Engineer Intern — *June 2026–Aug 2026*
- Developed METI, a forecasting engine to helps investors time market entry across 54+ African markets over a 1–6 quarter horizon.
- Designed a medallion data pipeline to ingest, clean, and standardize 25+ macroeconomic metrics from 15+ data sources.
- Built a **hybrid time-series model** by fusing Bayesian-optimized sentiment signals with **ARIMAX** score and modeled uncertainty with Monte Carlo simulations.
- Benchmarked against a zero-shot **TimesFM** baseline, achieving **15%** lower RMSE and a **65%** 3-month forecast hit rate.

Arcesium (D.E. Shaw Group) — Hyderabad, India
Software Engineer (L1–L2) — *June 2021 – Aug 2025*
- Optimized the executor component of a distributed job system processing 1M+ jobs/day, cutting DynamoDB RCUs by 70% and removing a key bottleneck for horizontal autoscaling.
- Implemented HMAC-based API authentication for an enterprise alerting platform with sub-0.5 ms latency.
- Built a distributed ELK log-ingestion pipeline processing 40 TB+ logs/day across 350+ microservices for real-time analytics.
- Enhanced internal object storage SDK to support ephemeral and durable storage architectures, enabling adoption across 100+ apps.

Data Science Intern — *Feb 2021 – May 2021*
- Developed a **statistical model** to quantify the health of core infrastructure services by combining diverse operational telemetry data into a unified, interpretable health score.
- Applied exponential smoothing (ETS) time-series models for trend forecasting and anomaly detection, improving health-score accuracy by **30%** over heuristic baselines.

Leadership & Honors

- Teaching Assistant, Northwestern University — CS496: Machine Learning and Sensing.
- 1st Place, Arcesium Hackathon (2024) themed "Data Science & Analytics at Scale" — Won among 130+ teams.