# 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, Parallel Computing

**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, JavaScript, SQL.
**ML Frameworks** : PyTorch, Keras, Scikit-learn, NumPy, SciPy, Pandas.
**Software Engineering** : Spring Boot, Django, React, Datadog, Terraform, Kafka, Elasticsearch, AWS (S3, SQS, SNS, DynamoDB).
**Generative AI** : LLMs (Fine-Tuning, Evaluations, Agents), Transformers, Diffusion Models, RAG, AI Agents, Quantization.
**Data & Infrastructure** : Docker, Kubernetes, Git, FastAPI, ETL Pipelines, CI/CD, Linux, PostgreSQL, Airflow, Distributed Systems.

## Professional Experience

**Kilwa Technologies LLC** — Chicago, IL
*Machine Learning Intern* — *June 2026 – Aug 2026*
- Developed METI, a hybrid time-series model to help investors time market entry across 54+ African markets over 1–6 qtr. horizon.
- Designed a medallion data pipeline to ingest, clean, and standardize 25+ macroeconomic metrics from 15+ data sources.
- Factored in 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% hit rate on 3-month forecasts.

**Arcesium (D.E. Shaw Group)** — Hyderabad, India
*Software Engineer – L2* — *July 2023 – Aug 2025*
- Rearchitected the executor component of a distributed job execution engine processing 1M+ jobs per day, reducing DynamoDB read capacity consumption by **70%** and removing a key bottleneck to horizontal autoscaling.
- Implemented HMAC-based API authentication for an enterprise alerting platform with sub-0.5 ms latency.
- Developed an alert aggregation feature that grouped related alerts across multiple dimensions, improving alerting efficiency by 30%.

*Software Engineer – L1* — *June 2021 – June 2023*
- Built a distributed log ingestion pipeline (using ELK stack) processing 40 TB+ logs/day from 350+ microservices for real-time analytics.
- Enhanced the internal object storage SDK to support ephemeral and durable S3 buckets, enabling 100+ apps to adopt the firm's cell-based architecture in under a month.

*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.

## Projects

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

**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 across 15+ long-duration sessions.
- Extracted mel-spectrograms from audio and 48 statistical/time-series IMU features, then trained and compared MobileNetV2 audio models, 1D CNN IMU models, and late-fusion multimodal architectures.
- Achieved 0.73 F1 for activity detection and analyzed onset/offset delays to evaluate temporal localization accuracy across modalities.

**Post-Training Quantization (PTQ) for Diffusion Transformers(DiTs)** | GitHub
- Evaluated low-bit FP and INT quantization on PixArt-Alpha (0.6B) diffusion model under equivalent precision budgets.
- Observed FP yield better visual fidelity than 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 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 improving answer accuracy by 30% over LLM-only baseline.