Manush Kalwari

manushkalwari141@gmail.com | LinkedIn:manush-kalwari | GitHub:ManushKalwari | Portfolio Website

Technologies

Machine Learning & Generative AI: PyTorch, TensorFlow, HuggingFace, scikit-learn, LangChain, LangGraph, Optimizations (LoRA, GRPO, quantization, FlashAttention), Docker, CI/CD (GitHub Actions, GitLab), Comet ML, Weights & Biases
Cloud: AWS (S3, SageMaker, Lambda, EC2), GCP (Vertex AI, BigQuery, Cloud Run, GKE, Composer, Dataflow)
High Performance ML: CUDA, Sharding, multi-GPU training, profiling, distributed training

Education

Columbia University, MS in Electrical Engineering (Specialization in ML) Sept 2024 – Dec 2025 (Exp.)
- **Coursework:** Advanced Deep Learning, Scaling LLM Systems, High-Performance ML, Generative AI, Mathematics of ML

PROJECTS

ScaleRAG - Multi-Modal Hierarchical RAG [GitHub] Oct 2025 – Dec 2025
- Designed a hierarchical RAG framework for large scientific corpora integrating the **IBM Scientific Parser**, **RAPTOR-style summarization**, and **FAISS vector indexing**. Deployed using **vLLM + PagedAttention**, benchmarked latency, memory, and throughput—achieved 1.4× throughput improvement on single-GPU infrastructure.
- Built an end-to-end **evaluation dataset generation pipeline** using **OpenAI GPT-4.1, LangChain, and Python multiprocessing**, implementing **Self-Instruct-style QA synthesis**, dynamic chunk merging for scalable RAG benchmarking.
- Developed modular RAG evaluation workflows with **RAGAS, DeepEval, and FAISS metrics**, automating assessments for retrieval precision, grounding faithfulness, and latency-throughput trade-offs across configurations (Simple RAG, GraphRAG, RAPTOR).
- Conducted large-scale **scaling-law experiments** on multimodal corpora, analyzing grounding accuracy, context compression efficiency, and **GPU utilization profiling** to map quality–latency–cost curves and identify bottlenecks in LLM retrieval pipelines.
- Built a full-stack web interface for real-time RAG interaction using **FastAPI (Uvicorn backend)**, **Next.js/React frontend**, and **GCP model hosting**, integrating streaming inference with FAISS retrieval for production-grade deployment.

Resource-Efficient Model Optimization Pipeline [GitHub] March 2025 – May 2025
- Adapted DeepSeek & Phi-Mini models for math reasoning tasks(Countdown, AIME-2024) on a single T4 GPU. Designed **multi-objective GRPO reward functions** to improve output alignment and reduce output degeneracy.
- Applied LoRA + quantization, benchmarking vs FP runs using PyTorch Profiler; improved GPU utilization to 97%.
- Demonstrated a cost-efficient fine-tuning pipeline balancing VRAM constraints and model quality.

Few-Shot Traffic Scene Hazard Classification [GitHub] Feb 2025 – April 2025
- Collaborated in a 3-member team on multi-label deep learning pipeline for classifying images from 120th St Amsterdam Avenue intersection in NYC, aiming to foresee rare and dangerous traffic events and pedestrian hazards reliably.
- Tackled severe class imbalance with custom metrics and adaptive class thresholding. Experimented with EfficientNet, Vision Transformers, and CLIP using zero-shot and few-shot approaches to handle unseen hazard types.
- Achieved a few-shot accuracy of 60% by integrating patch-based CLIP embeddings, ranking among top 10 performers in the course Kaggle competition.

LLM-Guided Aesthetic Reward Modeling [GitHub] Oct 2025 – Dec 2025
- Built an agent-style loop around Anthropic Claude Sonnet (tool-use API) to iteratively write and refine a Aesthetic reward function over human-labeled image attributes (composition, subject focus, lighting, color palette).
- Implemented a Python evaluation harness that computes Spearman correlation on train/test splits and guards against degenerate solutions, feeding scores back into the LLM for further refinement.
- Achieved test correlation of $\rho \approx 0.82$ on unseen images, showing that an **LLM-guided symbolic reasoning** can approximate human aesthetic judgments without any gradient-based training.

Real-Time AI Monitoring Platform [GitHub] June 2025 – July 2025
- Built and deployed a containerized ML inference service using FastAPI and Docker Compose. Integrated Prometheus + Grafana dashboards for live latency and throughput tracking, plus Postgres-based logging.
- Implemented secure API-key authentication and a plug-and-play observability layer, enabling rapid instrumentation of new ML services.
- Delivered a modular stack that reduced setup time for new monitored endpoints by 50%.

RESEARCH

EEG-to-Text Translation [GitHub] Jan 2025 – May 2025
(Supervised by Prof. Nima Mesgarani, Zuckerman Institute, Columbia University)
- Built an end-to-end pipeline to convert raw brain signals into text, advancing research on non-invasive brain–computer interfaces and potential communication tools for speech-impaired patients.
- Engineered robust EEG preprocessing (artifact removal, normalization, word-level alignment) on ZuCo EEG-text dataset and designed self-supervised learning modules with contrastive alignment of embeddings.
- Achieved a **4% BLEU-1 improvement over state-of-the-art baseline** by aligning EEG and text embeddings with symmetric InfoNCE loss and integrating an EEG encoder with BART, demonstrating stronger generalization under realistic (non-teacher-forced) inference.