Bhuvan Nallamothu

bnallamo@andrew.cmu.edu • (857) 343-6774 • linkedin.com/in/bhuvan-nallamothu • github.com/iambhuvan

EDUCATION

CARNEGIE MELLON UNIVERSITY — **Pittsburgh, PA**
Master of Science in Internet of Things (Focus: Deep Learning)| GPA: 3.70/4.00 — Dec 2026

NORTHEASTERN UNIVERSITY — **Boston, MA**
Master of Science in Computer Science| GPA: 4.00/4.00 (Transferred to Carnegie Mellon University) — Aug 2025

SKILLS

Programming Languages: Python, C++, CUDA C/C++, Go, Java (SpringBoot), SQL, R, MATLAB, Bash, Makefile, Scala
ML & Deep Learning: PyTorch, TensorFlow, JAX, Hugging Face, NumPy, CUDA Kernels, ONNX Runtime, TensorRT, WandB
LLM & Generative AI: RLHF (PPO/DPO), LoRA, QLoRA, DeepSpeed ZeRO, FAISS, SGLang, vLLM, torch.distributed
Engineering & MLOps: Docker, Kubernetes, GitHub Actions, Airflow, Prometheus, Grafana, AWS, GCP, Git, Jenkins, MLflow, DVC
Databases: PostgreSQL, MongoDB, Kafka, Redis, Elasticsearch, Neo4j, Cassandra, Iceberg, JanusGraph, Apache Spark, Weaviate, Pinecone

PROFESSIONAL EXPERIENCE

CARNEGIE MELLON UNIVERSITY — **Pittsburgh, PA**
Lead Teaching Assistant, AI Model Development — Jan 2026 - Apr 2026

- Delivered 10+ lectures and live coding demos on BERT, NanoGPT, RAG, DeepSeek R1, DARE, and multi-agent AI to 50 graduate students across a 12-module LLM curriculum covering Transformers, LoRA/QLoRA fine-tuning, and agentic tool use
- Graded 150+ deliverables across 50 students on production RAG pipeline implementations, evaluating FAISS vector indexing, BM25+vector hybrid retrieval, cross-encoder reranking, multi-hop query decomposition, and chunk-level citation tracing
- Assessed full research portal deployments for hallucination detection rates, groundedness evaluation metrics, and trust-aligned output behavior patterns; verified exportable AI-generated artifacts met production grade RAG deployment standards

JAWAHARLAL NEHRU TECHNOLOGICAL UNIVERSITY — **Hyderabad, India**
Research Assistant | link — Mar 2023 - Apr 2024

- Built end-to-end IoT-ML agriculture decision system: ESP32 firmware with DHT11 and soil moisture sensors over WiFi feeding a Flask API serving ResNet50V2 seed defect classification, Random Forest soil quality (88.89%), and 22-crop recommendations
- Benchmarked 4 transfer learning architectures for agricultural seed defect detection, selecting ResNet50V2 with batch normalization and dropout, integrated with tabular soil models trained on 3 merged IoT and agricultural datasets for full yield prediction pipeline
- Trained a custom CNN on 5,856 chest X-rays with 10-fold Stratified K-Fold CV, class-weight balancing, and ReduceLROnPlateau, achieving 96.75% peak and 94.82% mean validation accuracy with Grad-CAM explainability for pneumonia detection

PROJECT EXPERIENCE

CUDA Accelerated LLM Training Engine - Kernels to Transformer (Carnegie Mellon University)

- Implemented CUDA C++ kernel primitives from scratch: stride-based map/zip operators, shared memory tree-reduce for parallel aggregation, and tiled matrix multiplication forming the low-level GPU compute backbone of a full LLM training stack
- Assembled miniTorch autograd engine (topological-sort DFS backprop, Linear/Dropout/LayerNorm1d) and Pre-LN decoder-only Transformer (Multi-Head Attention, causal masking, logsumexp loss), achieving BLEU 20 on IWSLT14 De-En in 10 epochs
- Accelerated transformer with LightSeq2-inspired fused CUDA kernels: numerically stable softmax (CUB BlockLoad, causal masking) and float4-SIMD LayerNorm, achieving 6.5x/5.5x softmax and 15.8x/3.7x LayerNorm forward/backward speedups

Distributed LLM Training and Inference Optimization (GPT-2, LLaMA-2) (Carnegie Mellon University)

- Developed data-parallel GPT-2 (rank-based DataPartitioner, torch.distributed all-reduce, 1.5x+ tokens/sec on 2 GPUs) and pipeline-parallel GPT-2 (clock-cycle microbatch scheduling, per-device worker thread queues), outperforming model parallelism
- Fine tuned LLaMA-2-7B on 2x V100 via DeepSpeed ZeRO sharding and LoRA within 32 GB GPU memory budget, demonstrating parameter-efficient adaptation of a 7B-scale model under tight multi-GPU memory constraints
- Deployed SGLang inference with RadixAttention (KV cache prefix reuse via radix tree, LRU leaf eviction, cache-aware scheduling prioritizing longer prefix matches) and compressed FSM constrained decoding via FlashInfer backend

RLHF Fine Tuning Pipeline - Reward Modeling and PPO via VERL (Carnegie Mellon University)

- Executed end-to-end PPO-based RLHF via VERL (HybridFlow: decoupled generation/training phases) fine tuning GPT-2, quantifiably shifting reward distribution from −0.5 to +0.5 post-training, validated against pre-RLHF reward baseline
- Constructed DistilBERT reward model with pairwise ranking loss on 10K Anthropic hh-rlhf preference pairs, achieving 60%+ accuracy

Deep Learning Primitives from Scratch - Autograd, CNN, RNN and CTC (Carnegie Mellon University)

- Engineered PyTorch-equivalent autograd engine in pure NumPy, deriving forward/backward passes for Linear, BatchNorm1d (EMA alpha=0.9, train/eval switching), GELU, learnable-beta Swish, SGD, Adam, and AdamW with decoupled weight decay.
- Extended autograd engine to CNN primitives: Conv1d/2d (tensordot sliding window, flipped-kernel backprop), ConvTranspose upsampling, and index-tracked MaxPool2d gradient scatter enabling full spatial feature learning with analytical gradients.
- Integrated GRUCell (reset/update/candidate gates, full BPTT through all gate interactions), CTC dynamic programming loss, and beam search decoding completing a full sequence modeling primitive stack from scratch.