Kaung Myat Kyaw

669-649-8171 | Rockville, MD | kaky2698@colorado.edu | LinkedIn | Github

EDUCATION

University of Colorado Boulder *Oct 2023 - Jan 2026*
Master of Science in Computer Science, GPA: 4/4

Dagon University *Jun 2015* - Dec 2019
Bachelor of Science in Geology, Minor in Computer Science

WORK EXPERIENCES

APS Asia PDA Enterprise **Rockville, MD**
Lead AI Engineer *Jan 2024 - Present*

- Designed and deployed production-grade GenAI and multi agentic systems using Python, PyTorch, Hugging Face**,** LangChain/LangGraph, AWS Bedrock, MCP, A2A, FastAPI, NodeJS (JavaScript) and RAG architectures, improving LLM application reliability, scalability, and response quality by 28%.
- Architected scalable high-throughput LLM inference services using Docker and Kubernetes on AWS EKS**,** leveraging vLLM, KV caching, quantization**,** FlashAttention**,** continuous batching**,** and speculative decoding to improve GPU utilization while reducing inference latency and serving costs by 68%.
- Fine-tuned Transformer-based LLMs using LoRA/PEFT adapters on domain-specific datasets, improving task accuracy by 15**%** while reducing trainable parameters and fine-tuning compute by 70%**.**
- Strengthened LLM application security by implementing prompt-injection detection**,** input/output sanitization**,** guardrails**,** PII filtering**,** and tool-access policies, reducing unsafe or unauthorized model interactions by 30%.
- Built MLOps/LLMOps pipelines incorporating CI/CD model and prompt versioning**,** automated evaluation**,** OpenTelemetry and LangSmith for LLM logging, tracing, and model inspection improving production observability and debugging efficiency by 25%**.**
- Designed scalable semantic-search infrastructure using embedding models, FAISS, ChromaDB, and OpenSearch, delivering low-latency similarity retrieval across large-scale unstructured datasets.

APS Asia PDA Enterprise **Rockville, MD**
Senior Machine Learning Engineer *Jan 2017 - Jul 2022*

- Engineered real-time personalized recommendation systems using TensorFlow, two-tower retrieval, sequence modeling, and Graph Neural Networks, improving recommendation relevance by 30% and user engagement by 43%**.**
- Developed computer vision models using YOLO and DenseNet with **TensorFlow** and **scikit-learn** for image detection and classification, improving model accuracy by 20% and reducing inference time by 25% and also built OCR and document-processing pipelines to automate information extraction and reduce manual processing time by **30%.**
- Led cross-functional engineering and product teams through the end-to-end AI/ML development lifecycle, driving technical decisions, architecture alignment, and delivery of production capabilities in a fast-paced Agile environment.

SKILLS

Programming & ML: Python, C++, TensorFlow, PyTorch, Hugging Face, Scikit-learn, Deep Learning, Machine Learning, Computer Vision, OCR, Recommendation Systems

Generative AI & LLMs: LLMs, Transformers, LoRA/PEFT, RAG, Agentic AI, LangChain, LangGraph, LlamaIndex, Prompt Engineering, LLM Fine-Tuning, LLM Inference, Guardrails, LLM Security, Claude Code, Cursor, Codex

AI Infrastructure & Retrieval: vLLM, Embeddings, Semantic Search, Vector Databases, FAISS, ChromaDB, OpenSearch, Similarity Search, Two-Tower Models, Sequence Models, Graph Neural Networks

Cloud & MLOps/LLMOps: AWS, Amazon EKS, SageMaker, Bedrock, Docker, CI/CD, Model Versioning, Prompt Versioning, Automated Evaluation, Model Monitoring, OpenTelemetry, LangSmith, Observability, Git

AI Systems: Model Serving, Inference Optimization, KV Cache, Quantization, FlashAttention, Continuous Batching, Speculative Decoding, GPU Optimization, Distributed Inference