KAI-YU LU

Seattle, WA | kaiyulu45@gmail.com | GitHub: github.com/K-YLMike | LinkedIn: linkedin.com/in/kaiyulu

Education

Northeastern University Seattle, WA

Master of Science in Computer Science (GPA 4.0) *Sep. 2024 to May 2026*

- Research Assistant, Generative AI Research Group | Teaching Assistant, CS6140 Machine Learning

Xi'an Jiaotong-Liverpool University Suzhou, China

Bachelor of Engineering with Honours in Computer Science and Technology *Sep. 2016 to Jul. 2020*

- National Scholarship (First Prize, Special Prize), Summer Undergraduate Research Fellowship, MoE/DoE Scholarship

Professional Experience

AI Application Engineer, Computer Vision, Quanta Storage Inc. (Taoyuan, Taiwan) *Dec. 2021 to Apr. 2023*

- Owned computer vision for enterprise HDD and NAS assembly lines as sole CV engineer, delivering three inspection stations and an action-recognition system from imaging design and labeling through **Python/PyTorch** training, deployment, and production support.
- Trained **YOLOv5** detectors for missing components and surface damage across three assembly stages, pairing each defect type with a dedicated camera and lighting setup to improve visual separability and cut defect escape from **10% to under 1%**.
- Deployed **SlowFast** across five assembly actions including screwdriving, head stack assembly, cover mounting, labeling, and cable routing, providing real time alerts, SOP audit reports, and cycle times for each action; reduced procedural errors by **30%**, corrected **80% of detected errors** on the line, and increased UPH by **24%**.
- Engineered a multi-camera inference pipeline on one workstation using per-camera workers, **FFmpeg** decode/encode, **OpenCV** preprocessing, and **ONNX/TensorRT**; sustained **six 1080p streams at 30 fps** with **40% lower end-to-end latency** than PyTorch.
- Automated the production data loop from line-image archiving and flagged-unit labeling queues to dataset versioning, one-command retraining, and redeployment, creating a closed-loop feedback system maintainable by a single engineer.
- Benchmarked C3D, R3D, R(2+1)D, CNN-LSTM, MediaPipe, ResNet18, MobileNetV3-Small, and EfficientNetV2-Small on recognition accuracy against inference cost and deployment constraints before committing to a production model.

Publications

[1] **Kai-Yu Lu**, Jayash Koshal, and Shanu Sushmita. "Fixing the Step, Missing the Answer: What Accuracy Cannot See in Reasoning Repair." *COLM 2026 Workshop on Efficient Reasoning*, 2026. [OpenReview]: openreview.net/forum?id=nMzRDHAz4f

[2] **Kai-Yu Lu**, Malhar Sham Ghogare, Zihan Su, and Shanu Sushmita. "LyricLens: An Interactive System for Multi-Label Music Content Rating." *NeurIPS 2025 Workshop on AI for Music*, 2025. [OpenReview]: openreview.net/forum?id=UzDOCp6R40

[3] **Kai-Yu Lu**, Yong Yue, and Jieming Ma. "Enhanced LoRaWAN RSSI Indoor Localization Based on BP Neural Network." *2021 IEEE 4th International Conference on Information Systems and Computer Aided Education (ICISCAE)*, 2021, pp. 190 to 195. [DOI]: doi.org/10.1109/ICISCAE52414.2021.9590790

Technical Skills

Languages: Python, C/C++, Java, SQL, R, MATLAB
Machine Learning & Deep Learning: PyTorch, TensorFlow, Hugging Face Transformers, scikit-learn, XGBoost, LightGBM
Large Language Models (LLM) & Natural Language Processing (NLP): Qwen / Qwen3, Llama, Gemma, Phi, Nemotron, Longformer
Fine-Tuning: supervised fine-tuning (SFT), instruction tuning, LoRA, parameter-efficient fine-tuning (PEFT); TRL
Reinforcement Learning (RL) & Alignment: DPO, RLHF, RLVR, PPO, GRPO, reward modeling, preference optimization
Agents & Retrieval-Augmented Generation (RAG): LangGraph, LangChain, LlamaIndex, tool calling; OpenAI, Anthropic APIs
Retrieval & Ranking: FAISS, approximate nearest neighbor (ANN) search, cross-encoder reranking, two-tower retrieval
Inference & Serving: vLLM, ONNX, TensorRT, llm-compressor; post-training quantization (INT8, INT4-GPTQ, W8A8)
Computer Vision & Video: YOLOv5, CLIP / OpenCLIP, SlowFast, Detectron2, MMDetection, OpenCV, FFmpeg
Systems & MLOps: Linux, Git, Docker, CUDA, Streamlit, HPC schedulers, NVIDIA Jetson, Snapdragon X Elite NPU

Projects

LLM Post-Training Benchmark (SFT, DPO, RLHF, RLVR) github.com/K-YLMike/llm-post-training

- Compared SFT, DPO, PPO-based RLHF, and GRPO-based RLVR on **Qwen3-1.7B**, branching all three post-SFT methods from one LoRA SFT checkpoint with identical adapter geometry, then scoring every variant on the same **500-question GSM8K** set through a single decoding path.
- Found DPO the strongest completed configuration at **0.732** GSM8K exact match against **0.676** for SFT and **0.460** for the base model, its **5.6-point** gain the only one in the study larger than the evaluation's roughly **2.1-point** standard error.
- Trained a separate **Qwen3-0.6B** reward model to **0.740** held-out pairwise accuracy on disjoint UltraFeedback partitions, then optimized the SFT policy against it with PPO under KL control; accurate preference ranking did not translate into stronger downstream accuracy.
- Identified vanishing within-group reward variance as a recurrent GRPO failure mode across **seven configurations**, reaching **92.5%** of prompt groups before collapse; disabling reward scaling stabilized training and lifted verifiable reward from **0.62** into the **0.66 to 0.75** range, with atomic checkpointing and validated resume keeping every crashed run recoverable under a one-hour job limit.

Agentic RAG Benchmark (Four Retrieval Architectures) github.com/K-YLMike/agentic-rag-benchmark

- Compared direct generation, vanilla RAG, RAG with cross-encoder reranking, and a **LangGraph** agentic loop on **1,500 HotpotQA** multi-hop questions, holding the model (**Qwen3-8B**), corpus, embeddings, FAISS index, and answer prompt constant; the agent judges its own retrieved evidence and reformulates the query, with its round budget enforced in graph routing rather than in a prompt.

- Measured retrieval as the dominant factor: attaching evidence raised exact match from **0.175 to 0.439**, a **26.4-point** gain, while cross-encoder reranking left Recall@5 unchanged at 0.845 and improved only the context reaching the model, **0.845 to 0.891**, worth **2.9** points at **12**× the latency.
- Found the agentic loop did not beat reranking, **0.465** against **0.468** exact match or four questions out of 1,500, for **3.6**× the LLM calls and **3.1**× the latency; attributed the shortfall to query rewriting by showing its first-round Recall@20 was already lower (**0.904** against 0.924).
- Built a ten-category failure taxonomy that split the agent's 802 errors into **225** evidence-judge false positives, **123** round-budget exhaustions, and **440** generation failures with correct evidence in context, against one undifferentiated bucket for the reranked baseline.

Master's Thesis: Stepwise Verification and First-Wrong-Step Repair for LLM Math Reasoning
- Developed a four-stage process-level framework for LLM mathematical reasoning that converts solutions into structured traces, removes computation-checkable noise through verifier-guided calibration, diagnoses residual failures with global error types and step-level labels, and repairs from the **First Wrong Step (FWS)** while preserving the stable correct prefix.
- Evaluated the framework on **500 MATH problems** with **Qwen2.5-14B-Instruct** and **Gemma-3-12b-it**; global error type matched the FWS error type in **94.26%** and **91.21%** of residual failures, respectively, supporting the FWS as an interpretable repair anchor across two model families.
- Compared No Hint, Answer-Free, and Answer-Aware repair conditions and separated local FWS correction from final-answer recovery; answer-free diagnostic guidance improved Full Repair Success from **23.3% to 32.5%** for Qwen and from **25.9% to 37.1%** for Gemma, while showing that successful local repair does not necessarily yield a correct final answer.

LLM Quantization: Accuracy/Latency/Memory Pareto Study github.com/K-YLMike/llm-quantization-pareto
- Quantized **Qwen2.5-14B-Instruct** with INT8 weight-only (W8A16), **INT4-GPTQ** (W4A16), and **W8A8 SmoothQuant** using **llm-compressor**; isolated WikiText-2 calibration data from the **MMLU** and **HumanEval** evaluation sets to prevent leakage.
- Benchmarked FP16 and all three quantized configurations on **vLLM** for throughput, first-token latency, and weight footprint over full **MMLU** (14,042 questions) and **HumanEval pass@1**, constructing an accuracy–latency–memory Pareto frontier on a single GPU.
- Reduced weight footprint by **3**×, from **29.5 GB to 9.9 GB** with INT4-GPTQ, at a 2.1-point MMLU and 1.2-point HumanEval cost; **W8A8** reached **3763 tok/s** and **1060 ms TTFT** while staying within 0.8 points of FP16 on MMLU and matching it on HumanEval.
- Tested whether low-bit quantization disproportionately harms multi-step code generation relative to multiple-choice accuracy; INT4 degraded MMLU slightly more than HumanEval, a negative result against that hypothesis.

Calibrated Multimodal AI Evaluation on Hateful Memes github.com/K-YLMike/multimodal-content-moderation
- Built a multimodal classifier from frozen **OpenCLIP ViT-B/32** image and text features and compared text-only, image-only, late-fusion, and cross-attention variants; fusion reached **AUROC 0.763** versus 0.675 and 0.629 for single modalities and 0.703 for cross-attention.
- Evaluated precision at fixed low false-positive rates and review-queue volume, then applied validation-only temperature scaling to reduce expected calibration error from **0.21 to 0.03** for reliable thresholding.
- Quantified failure modes: fusion recovered **235 of 560 cases (42%)** missed by at least one single modality, while the top 10% of modality-disagreement cases retrieved single-modality failures at **precision 1.0** as a targeted review and labeling signal.

LyricCert: Fine-Grained Lyric Content Classification and Age Rating
- Replaced a single binary explicit flag with four content categories, sexual, violence, substance, and language, under an auditable annotation protocol that records each label together with its supporting evidence spans.
- Combined category lexicons, repeated LLM judgments, deterministic consistency checks, and selective human review to distill a **31K-song dataset** from more than 40K Spotify tracks.
- Benchmarked transformer architectures for multi-label classification and selected **Longformer** at **macro-F1 0.85**, then tested external generalization on a cleaned **5.5K-song Billboard corpus**.
- Shipped a five-level content rating module mapping category scores to age-suitability tiers, with reproducible datasets, checkpoints, and audit templates.

Cross-Modal Visual Semantic Search with CLIP and FAISS github.com/K-YLMike/visual-semantic-search
- Built a cross-modal retrieval engine over **31,014 Flickr30K images** and **155,070 caption queries** using frozen **OpenCLIP ViT-B/32** embeddings and FAISS, supporting text-to-image and image-to-image search without model training.
- Compared Flat, IVF, HNSW, and IVFPQ indexes across recall, latency, and memory; **HNSW retained 98% of exact Recall@10** while reducing median query latency from **5.14 ms to 0.26 ms**, and IVFPQ compressed vectors from **2048 to 73.9 bytes**.
- Re-scored the top-100 IVFPQ candidates with full-precision embeddings, raising **Recall@10 from 0.388 to 0.546**; bucketed caption queries into counting, negation, and compositional/spatial categories to separate representation-level CLIP failures from index approximation.

Two-Stage Personalized Recommender System github.com/K-YLMike/recommender-system-movielens
- Built an end-to-end retrieval-and-ranking system with a PyTorch **two-tower** retriever, pretrained title features, **FAISS** IVF/HNSW search, and a **LightGBM LambdaRank** reranker, benchmarked against popularity and BPR-MF baselines.
- Found a **5× Recall@10 swing** from negative sampling alone; logQ reached 0.047 versus 0.043 for BPR-MF while increasing long-tail recall by a factor of **20** and catalog coverage by a factor of **26**.
- Evaluated with a leakage-aware temporal split, three seeds, and head/tail/cold-start segments; reranking improved **NDCG@10 by 27%**, while HNSW retained **97% recall at about 1/10 the latency** of exact search.