Valikhujaev Yakhyokhuja

github: https://github.com/yakhyo
linkedin: https://www.linkedin.com/in/y-valikhujaev

Email : yakhyo9696@gmail.com
Mobile : (+82) 010-9363-7088

Professional Summary

AI/ML & MLOps Engineer with 6+ years of industry and research experience focused on **Large Language Models (LLMs)**, **Retrieval-Augmented Generation (RAG)**, **Vision-Language Models (VLMs)**, and scalable **MLOps infrastructure**. Skilled in fine-tuning and deploying transformer architectures, building **Kubernetes-based ML platforms**, and automating model training and deployment across cloud and on-prem environments.

Currently working on designing and maintaining **private cloud and Kubernetes clusters** for distributed training and inference workloads. Building systems for resource discovery, policy enforcement, and workload scheduling using **Kubeflow**, **Kueue**, **Kai Scheduling**, and **Kyverno**, while focusing on scalable, reliable, and automated MLOps pipelines.

Experienced in developing and optimizing **LLM-based**, **multimodal**, and **retrieval-augmented systems** with strong background in model optimization (**TensorRT**, **Triton Inference Server**, **vLLM**), CI/CD integration, and production-grade deployment across **AWS** and **GCP**.

Skills Summary

- **Programming Languages**: Python, C/C++, Java
- **Data & Query Systems**: MySQL, PostgreSQL, PySpark; data wrangling with Pandas, Polars; visualization using Power BI
- **ML & DL Frameworks**: NumPy, Scikit-learn, PyTorch, PyTorch Lightning, TensorFlow, Keras, Hugging Face, LangChain, Unsloth
- **Retrieval & Indexing Systems**: FAISS, Pinecone, Elastic, Tantivy (lexical, semantic, hybrid search)
- **MLOps & Deployment**: Docker, FastAPI, Flask, gRPC, TorchServe, Triton Inference Server, TensorRT, vLLM, **Kubernetes**, Kubeflow, Helm, ArgoCD, Kueue, Kai Scheduling, Kyverno
- **Cloud Platforms**: AWS (EC2, S3, Lambda, SageMaker, EKS), Google Cloud Platform (Vertex AI, GKE, Compute Engine, Cloud Run)
- **Core Competencies**:
  - **LLMs & RAG**: LoRA/PEFT fine-tuning, retrieval pipeline design (Sentence Transformers, Tantivy), vLLM-based inference orchestration, knowledge base integration
  - **MLOps & Infrastructure**: Cluster orchestration, workload scheduling, resource management, model lifecycle automation, CI/CD, observability (Prometheus, Grafana, ELK)
  - **Conversational AI**: Voice-to-voice AI agents with ASR, LLMs, and TTS pipelines using LiveKit and related tools
  - **Multimodal AI**: Vision-Language Models, OCR, object detection, tracking, re-identification, video action recognition
  - **Generative AI**: sLLMs, Diffusion Models, GANs, DeepFakes, image-to-video, face restoration and enhancement
  - **Edge Deployment**: Model pruning, quantization, on-device inference, latency optimization
  - **ML Engineering**: End-to-end pipeline design, scalable deployment, model monitoring, MLOps best practices

Work Experience

- **MLOps Engineer** — Seoul, South Korea
  Thaki Cloud Co. Ltd — *August 2025 – Present*
  - **Kubernetes & Private Cloud Platform**: Designing, deploying, and maintaining **Kubernetes clusters** for a full-featured **private cloud platform** supporting distributed model training, inference, and MLOps workloads. Building multi-tenant infrastructure with automated resource provisioning, monitoring, and scaling for GPU-intensive pipelines.
  - **Resource Discovery & Policy Optimization**: Implementing dynamic resource discovery and scheduling strategies to maximize GPU and CPU utilization. Developing custom resource management operators for efficient allocation and preemption, and integrating **policy optimization** for fair and performance-aware workload balancing.

- **Policy Enforcement & Admission Control**: Establishing cluster-wide governance using **Kyverno** and custom **admission webhooks**. Automating security and compliance enforcement, applying mutation/validation logic, and standardizing deployment configurations across namespaces and teams.
- **Advanced Scheduling & Job Queueing**: Integrating and enhancing **Kueue** and **Kai Scheduling** frameworks to improve distributed job scheduling and workload queueing efficiency, achieving better throughput and fairness for model training and data processing jobs.
- **Kubeflow & MLOps Pipelines**: Managing the full lifecycle of ML pipelines using **Kubeflow**, including **Katib** for automated hyperparameter tuning, pipeline orchestration, and deployment. Leveraging **Helm** and **ArgoCD** for continuous delivery of ML workloads across staging and production environments.
- **Cluster Automation & Observability**: Automating cluster provisioning, logging, and monitoring pipelines using **Prometheus**, **Grafana**, and the **ELK stack**. Establishing CI/CD integration for infrastructure and ML workflow updates using GitOps principles.

- **AI/ML Research Engineer** — Seoul, South Korea
  Aria Studios Co. Ltd — *March 2024 – August 2025*
  - **LLM Fine-tuning & Adaptation**: Fine-tuned **Qwen-2.5-7B/3B-Instruct** models for the Korean language using **LoRA** and **DPO (Direct Preference Optimization)** for efficient on-device deployment. Also fine-tuned **GPT-3.5-turbo** on conversational data using custom augmentation workflows.
  - **AI Agent & Conversational Systems**: Built an end-to-end **voice-to-voice AI assistant** using **LiveKit** for real-time media streaming, **Whisper** for ASR, OpenAI-based LLMs for dialogue, and TTS for responses. Integrated multimodal understanding (face ID, age/gender, emotion). Enabled dynamic **function/tool calling** via LLMs to invoke external APIs using **MCP servers**, allowing real-time task execution and tool orchestration within conversation flow.
  - **Retrieval & Data Tooling for LLMs**: Built a simulated interaction tool to collect structured dialogue data for fine-tuning LLMs. Managed persistence with PostgreSQL and hosted the system on **GCP**.
  - **LLM Inference & API Optimization**: Designed scalable APIs using **vLLM** and FastAPI for real-time LLM inference. Applied **LangChain** and concurrency patterns to optimize API orchestration for latency-sensitive workflows.
  - **Generative Model Training & Deployment**: Fine-tuned **FLUX (Stable Diffusion)** using **LoRA** for stylized character generation. Deployed as a REST API on **GCP**.
  - **VLM Deployment for Context Awareness**: Developed a visual perception module for virtual characters using **Phi-3-Vision**, enabling multimodal awareness and interaction.
  - **Face Parsing & Enhancement**: Enhanced face-swapping pipeline by integrating a custom **face parsing** model and improving backbone efficiency for segmentation accuracy.
  - **DeepFake & Face Restoration**: Worked on face restoration and enhancement techniques to improve DeepFake video quality used in high-visibility media projects.
- **ML Engineer** — Seoul, South Korea
  Pyler Co. Ltd — *July 2022 – September 2023*
  - **Video-Based Visual Content Moderation**: Developed a robust pipeline for detecting inappropriate content in video streams using temporal action recognition models. Achieved a **10%+ improvement in accuracy** by optimizing model architecture and training strategy.
  - **Detection-Based Moderation Pipeline**: Implemented **real-time object detection and segmentation** models to flag unsafe visual elements for brand safety. Integrated **active learning** loops, improving precision and recall by **15%**. Built a scalable end-to-end training and deployment pipeline on **Kubeflow**.
  - **Classification-Based Moderation Framework**: Designed a **multi-label, multi-head classification system** combining self-supervised and supervised learning. Boosted precision by **20%** on hard samples and established this architecture as the standard for visual moderation across projects.
  - **Dataset Clustering with CLIP Embeddings**: Leveraged **CLIP vision-language embeddings** to perform unsupervised dataset analysis. Applied **PCA** for dimensionality reduction and used **KMeans** and **DBSCAN** for clustering to identify content groups and outliers. Enabled efficient dataset curation and weak supervision strategies.
  - **Model-Assisted Labeling System**: Built a feedback-driven pipeline that combined inference on labeled and unlabeled data to accelerate data annotation. Used **active learning** to improve labeling quality and reduce manual annotation time.
- **AI Research Engineer** — Seoul, South Korea
  D-Meta Co. Ltd — *November 2020 – July 2022*
  - **Slab Text Recognition**: Developed an OCR pipeline to recognize handwritten text on slab metals using **Spatial Transformer Networks (STN)** and sequential models. Built the full workflow from data preprocessing to training and evaluation. Achieved over **90% accuracy** by integrating state-of-the-art text detection and recognition techniques optimized for industrial scene images.
  - **Automatic Number Plate Recognition (ANPR)**: Designed an end-to-end pipeline for number plate detection and recognition. Improved performance by **15%** in precision and recall through **active learning**, synthetic data generation, and targeted augmentation strategies.
  - **Real-Time ANPR Inference**: Deployed ANPR models for **real-time video inference** from **RTSP streams**, enabling continuous monitoring and detection in live camera feeds. Handled frame capture, batching, and stream resilience for production environments.

- **Car Damage Detection**: Built and deployed a lightweight **car damage detection model** optimized for **Android devices** using TorchScript. Achieved a **10% improvement in precision** through hyperparameter tuning and efficient model design.
- **Shadow Removal using Pix2Pix GAN**: Applied **Pix2Pix GAN** to remove shadows cast on vehicles in captured images, enabling clearer downstream detection and damage assessment. Improved image quality and model robustness in low-light or occluded conditions.

Research Experience

- **Research Assistant** — Seongnam, South Korea
  AI and SC Lab — *Sep 2018 - Nov 2020*
  - **Computer Vision based Fire and Smoke Detection**: Designed and implemented a dilated CNN architecture for improved feature extraction and recognition in images/videos. Applied optimization techniques to reduce false positives and increase inference speed by 1.5× over baseline.
  - **Model Optimization for Edge Devices**: Improved the FPS on Edge device (Raspberry PI 2) by using hyper-parameter tuning and quantization for detection model.

Selected ML and AI project implementations: https://www.github.com/yakhyo

Education

- **Gachon University** — Seongnam, South Korea
  MSc in Computer Engineering; advised by Prof. Young Im Cho; CGPA: 4.0/4.5 — *Sep 2018 - Feb 2021*
- **Tashkent University of Information Technologies** — Tashkent, Uzbekistan
  BSc in Computer Engineering; CGPA(%): 85/100 or 3.72/4.0 — *Sep 2014 - June 2018*

Publications

- Muksimova S†, **Valikhujaev Y**†, Umirzakova S, Baltayev J, Cho YI. "GazeCapsNet: A Lightweight Gaze Estimation Framework". Sensors, 2025; 25(4):1224. Available at https://doi.org/10.3390/s25041224. [1]
- **Valikhujaev Y**, Abdusalomov A, Cho YI. "Automatic Fire and Smoke Detection Method for Surveillance Systems Based on Dilated CNNs". Atmosphere, **IF 2.9**. 2020; 11(11):1241. Available at https://doi.org/10.3390/atmos11111241.
- Muksimova Sh†, **Valikhujaev Y**†, Cho YI. "Automatic Fire and Smoke Detection System for Open Street CCTV Systems in Smart City Platforms". Korean Society of Information Scientists and Engineers, 412-414 pages, Domestic Conference.

Honors

Best paper award from Fire Investigation Society of Korea (FISK); (Domestic Conference, 2020)

Best presentation award from ISIS2019 & ICBAKE2019; (Domestic Conference, 2019)

Languages

English: Full Professional Proficiency (C1 Advanced);

Korean: Limited Working Proficiency (B1 Pre-Intermediate);

Russian: Limited Working Proficiency;

Uzbek: Native Proficiency;

[1] † These authors contributed equally to this work.