Yunmo Koo

Founding Engineer at FriendliAI | Optimizing Gen AI Inference

Redwood City, CA | mpbb03@gmail.com | yunmorning.me | github.com/kooyunmo

Summary

Founding engineer at FriendliAI with end-to-end experience building LLM inference, runtime, and distributed training platforms from the ground up. Specialize in production ML infrastructure that improves latency, reliability, and cost efficiency. Also led customer-facing engineering, product adoption, and technical sales growth in the US market.

Experience

FriendliAI San Mateo, CA

Founding Engineer *Feb 2021 – Present*

Inference Optimization (Speculative Decoding)

- Led end-to-end research and development of speculative decoding systems for production LLM inference.
- Architected and implemented online draft-model training, enabling continuous adaptation during production inference.
- Designed a hybrid speculator (model-based + model-free) with dynamic routing based on scoring mechanisms.
- Trained 50+ draft models across serverless and customer deployments, achieving 2-5x faster inference across diverse workloads.

Inference Optimization (Kernel-Level)

- Developed high-performance kernels for core LLM operations including attention, sampling, and decoding.
- Implemented specialized kernels for speculative decoding, improving end-to-end inference efficiency.

Inference Runtime Development

- Delivered inference support for major LLMs and multimodal models including Llama, Gemma, DeepSeek, Qwen, and GLM.
- Built core runtime systems including memory management, scheduling, and KV-cache optimization.

Distributed Training (PeriFlow)

- Led initial product development of PeriFlow, a distributed training platform for multi-cloud GPU environments.
- Architected fault-tolerant resource-management systems for reliable large-scale training.
- Led training and release of FAI-13B before Meta's Llama 2.
- Designed and implemented key LLMOps microservices for authentication, model registry, training, deployment, and monitoring.
- Developed and managed the SDK, CLI, documentation, and web interfaces across the full stack.

Lead Solutions Architecture & Forward-Deployed Engineering

- Led the US solutions architecture and forward-deployed engineering team, supporting 100+ customer PoCs.
- Managed strategic cloud partnerships, including AWS and Azure Marketplace launches and the AWS ISV Accelerate program.
- Delivered 30+ talks at industry events and developer communities.
- Built integrations with major open-source frameworks including LangChain and LlamaIndex.

Education & Research

Seoul National University

MS in Computer Science and Engineering *Aug 2020 – Aug 2022*

- "Terra: Imperative-Symbolic Co-Execution of Imperative Deep Learning Programs." *NeurIPS 2021*.
- Master's Thesis: *Cost-Efficient Machine Learning Training on Preemptible Cloud Clusters*.

BS in Computer Science and Engineering / Korean History (Double Major) *Mar 2014 – Aug 2020*

- Including 2 years of military service at the Korean Augmentation to the United States Army.

Skills

Languages: Python, C++, CUDA, TypeScript, Rust, Go

ML: PyTorch, TensorFlow, vLLM, SGLang, TensorRT-LLM

Kernel Programming: cuBLAS, cuDNN, CUTLASS, FlashInfer, CUB, NCCL, NIXL, Triton

Infrastructure: Kubernetes, Prometheus, Grafana, Terraform

Cloud Platforms: AWS, Azure, GCP, CoreWeave, Nebius

Specializations: Inference Optimization, Distributed Training, Multi-Cloud Infrastructure, LLMOps