

Jeonghun Kim

ML Engineer / AI Researcher

portfolio.jhkim3ispark.work



Senior Software Engineer, Inference Engine (Platform Software)

bandlagom0927@gmail.com

+82 10-9559-9471 · Seoul

github.com/thuygom



Cover Letter

Dear FuriosaAI Hiring Team,

I am applying for the Senior Software Engineer, Inference Engine (Platform Software) position (Job ID 4005788201). I bring practical experience operating model services, adapting task-specific models, and investigating GPU processing behavior. Working across these layers lets me connect a serving symptom to the workload and execution conditions beneath it. Before ISPARK, I served one year as an information-systems NCO in the Republic of Korea Air Force (Dec 2021–Dec 2022), maintaining Linux servers through shell scripting and log analysis.

I take responsibility for model deployment as well as training, building vLLM and FastAPI inference services in Docker with Prometheus and Grafana monitoring. Watching service failures alongside GPU usage helps me distinguish application problems from resource constraints. I ran distributed training on two RTX 4090 GPUs with BF16 precision and checkpoint recovery, and now also use two A100 GPUs through a national computing-support program. I track experiments in Weights & Biases so model changes can be compared across runs.

I redesigned long-audio processing and reduced server processing time by 36.9%, from 122.6 to 77.4 seconds on a 30-minute recording, using Silero voice activity detection and faster-whisper batching at batch size 8. The same workload had also exposed a delivery problem: the server finished after the synchronous HTTP request timed out. I separated live sessions from recoverable asynchronous jobs, with task polling and persisted progress. I treated recognition quality, execution time, and delivery as separate engineering problems, allowing each change to address a specific cause.

At FuriosaAI, I can contribute workload analysis and model validation to inference engine development. I would trace a request through execution, reproduce the limiting condition, and use reference outputs and timing measurements to assess implementation changes.

Sincerely,

Jeonghun Kim