Jeonghun Kim

ML Engineer / AI Researcher

portfolio.jhkim3ispark.work

bandlagom0927@gmail.com

+82 10-9559-9471 · Seoul

github.com/thuygom

FURIOSA **Algorithm - AI System Engineer**

Cover Letter

Dear FuriosaAI Hiring Team,

I am applying for the Algorithm - AI System Engineer position (Job ID 4005372201). I bring experience improving both model behavior and the system that delivers it. At TieMing, I develop core AI features through task-specific training, inference APIs, and workload optimization, using measured results to guide implementation choices. 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 improved a four-billion-parameter Text-to-SQL model from 26.6% to 76.6% normalized database-result agreement on a 64-question internal CRM development benchmark, using teacher-student training and rank-16 LoRA. The model matched 49 of 64 reference results, compared with 17 for the base model; scoring ignores row order and normalizes numbers and dates. I also verified the serving interface: missing schema input caused invented table names despite unchanged weights. I documented the required schema inputs alongside the adapter so the trained behavior could be reproduced in serving.

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 experiments that connect model requirements to execution and serving behavior. I would establish a representative workload, compare implementation choices against reference results, and use the findings to improve model support and performance.

Sincerely,
Jeonghun Kim