Jeonghun Kim

ML Engineer / AI Researcher

bandlagom0927@gmail.com

+82 10-9559-9471 · Seoul

portfolio.jhkim3ispark.work

github.com/thuygom

Senior Software Engineer, Inference Engine (Platform Software)

Profile

I am an ML engineer responsible for TieMing's core AI features, from model training to inference APIs and deployment. I bring a multimodal research background—five publications, four as first author—and turn experimental results into working products. I operate vLLM services and improve how GPU workloads progress, use resources, and deliver results.

Experience & Education

ISPARK R&D Center · AI Researcher Aug 2025–Present

Work-study intern Aug 2025–Feb 2026; full-time since Mar 2026

Develop TieMing's core AI features, owning model development, inference APIs, and deployment for document extraction, database queries, and meeting transcription.

Demonstrated at AI Summit Seoul 2026; Android private beta and Play Store registration in preparation

Republic of Korea Air Force · Information Systems NCO Dec 2021–Dec 2022 · 1 year

Maintained Linux servers through shell scripting and log analysis; earned Industrial Engineer Information Processing and Linux Master Level 2 during service.

Hansung University · B.S., Computer Engineering Mar 2019–Feb 2026

Department of Computer Engineering · GPA 3.95/4.5 · IRIS Lab undergraduate researcher, 2024–2026

Relevant Experience

- Built and operated vLLM/FastAPI inference services with GPU monitoring; ran distributed model training with checkpoint recovery.
- Reduced speech processing time by 36.9% on a 30-minute sample and designed asynchronous jobs with progress tracking and restart recovery.

Model Development & Service Operations

- Built and operated vLLM/FastAPI inference services in Docker, with Prometheus and Grafana monitoring to diagnose service behavior alongside GPU resource usage.
- Ran distributed training with checkpoint recovery on two RTX 4090 GPUs; now also using two A100 GPUs through NIPA and tracking experiments in Weights & Biases.

Awards, Certifications & Military Service

Selected Awards —Grand Prize, Hansung Engineering Competition (2023); Top Excellence Award, Hansung SW Festival (2024); Silver, AI Smart Workathon (2026). Research awards appear with publications.

Certifications —SW Development L5 (2026); Industrial Engineer Information Processing and Linux Master Level 2 (2022).

Military —Republic of Korea Air Force, information systems: 21 months as an enlisted airman.

Company Projects

CRM Text-to-SQL · Permission-aware database queries

Jun 2026–Present

- Developed a query workflow from schema retrieval through SQL generation and execution. Enriched database descriptions and combined semantic and keyword retrieval to resolve incorrect table selection.
- Enforced query execution controls: validated SELECT statements, common table expressions (CTEs), allowed schemas, and row limits; applied read-only transactions and timeouts.
- Designed and validated PostgreSQL Row Level Security (RLS) so row access is enforced independently of model output. Verified that document permission changes were reflected in available citation evidence.

Speech Transcription · Real-Time and Long-Form Processing

Mar 2026–Present

- Designed long-audio processing as recoverable asynchronous jobs with polling and persisted progress. Separated job completion from the HTTP request lifetime after a completed server task still produced a timeout.
- Reduced server processing time by **36.9% (122.6→77.4 seconds)** on one 30-minute recording, using Silero voice activity detection and faster-whisper batching, batch size 8.
- Integrated pretrained speech models through comparisons under the same conditions. Internal Korean meetings reached approximately **19% word error rate (WER)** and **23% diarization error rate (DER)**.
- Added a user-extensible glossary for domain terminology and per-user memory with Honcho. Implemented GPU job admission controls and tenant-scoped access for the service.

Web Search · Evidence collection

- Built evidence collection beyond search snippets: preserved full page content, source links, and collection status, with separate static, browser, and external crawler paths and per-host rate limits.

Coding-Session Search and Reuse

- Built searchable knowledge and wiki workflows from selected coding sessions, with source links, incremental ingestion, and tenant-, department-, and user-level access controls.

Model Training & Evaluation

Business-card OCR · Key Information Extraction Apr 2026–Present

- Developed business-card extraction for TieMing, reaching **94% accuracy across seven core fields** in a live workflow covering approximately 5,000 cards.
- Trained a smaller student model through teacher-student knowledge distillation and low-rank adaptation (LoRA), targeting incorrect field assignment that text recognition alone could not resolve.
- Improved field-value accuracy from **30.2% to 65.3%** in a separate difficult-case evaluation through further training and selective regeneration of malformed outputs.

Text-to-SQL · Research and model adaptation Dec 2025–Present

- Improved a four-billion-parameter model's database-result agreement from **26.6% to 76.6%** through teacher-student training and rank-16 LoRA: **17/64 → 49/64** on an internal CRM development benchmark, ignoring row order and normalizing numbers and dates.
- Validated the serving interface and documented required schema inputs alongside the adapter. This addressed invented table names when schema input was omitted despite unchanged weights.

Publications · Five papers, four as first author

Developed text-only comment analysis into multimodal research and implemented neural architecture search for choosing fusion paths. This work produced five publications, four as first author.

Multimodal Routing Optimization via Neural Architecture Search of Fusion Paths

IJIBC 17(3), 144–154 · KCI-indexed · Aug 2025 · First author

Multimodal routing optimization with implicit NAS policy learning

KSC, Dec 2025 · First author · Encouragement Award, Jan 2026

Attention-based user response analysis for multimodal influencer support estimation

KIPS, May 2025 · First author

SNS sentiment risk indicators for creator protection

KIPS, Nov 2025 · Coauthor · Silver Award

A Transformer-Based System for Predicting Influencer Popularity via Comment Sentiment and Topic Analysis

AICOMPS 2024 · First author

Patent Application · Examination requested

Text-to-SQL preprocessing with automatic database description generation and hallucination suppression

Korean application 10-2026-0119188 · Jun 30, 2026 · Inventor: Jeonghun Kim