Timur Khairulov

timurkhairulov@cau.ac.kr | github.com/KhrTim | linkedin.com/in/timur-khairulov

LLM Inference & Systems Engineer

Systems-focused engineer with a background in real-time performance engineering, specializing in LLM inference optimization and latency-critical deployment, with production experience in SIMD-accelerated C++ and end-to-end multimodal LLM pipelines.

Core Technical Stack

Inference & Systems	AVX SIMD, Linux, CMake, Profiling, Quantization, Pruning
Programming Languages	C++20, Python, C
Machine Learning	PyTorch, Hugging Face Transformers, ONNX, NumPy
Deployment & Tooling	Docker, Git, CLI/GUI Applications
Natural Languages	English (Advanced) · Russian (Native) · Korean (Basic)

Work Experience

C++ Software Engineer (L1 PHY), YADRO May 2023 - February 2024

- Optimized latency-critical Layer-1 signal processing pipelines using AVX SIMD intrinsics, improving throughput in a real-time execution environment with strict latency budgets
- Refactored memory layouts and access patterns in performance-sensitive code paths to reduce cache misses and stabilize low-latency execution
- Profiled and analyzed compute vs. memory bottlenecks in production telecom software using low-level profiling tools to guide optimization decisions
- Built an internal timing-visualization tool for L1–L2 communication analysis, improving performance debugging and cross-team diagnostics (Python, TypeScript, MongoDB)
- Applied real-time performance optimization principles to ML inference workloads

AI Optimization & Deployment Projects

Vision-Language Model Compression & Optimization GitHub | Blog

- Evaluated quantization and pruning strategies for vision-language models under latency and memory constraints
- Analyzed trade-offs between accuracy, latency, and memory footprint for downstream agent pipelines

Curriculum Learning Fine-Tuning for Mathematical Reasoning GitHub | Blog

- Fine-tuned LLMs using curriculum learning strategies to improve structured reasoning
- Achieved +1.37% accuracy improvement on GSM8K compared to baseline fine-tuning

BAGen: Multimodal Content Generation Pipeline GitHub

- Designed and implemented an end-to-end LLM-driven pipeline integrating language and diffusion models, delivered as a production-ready system with GUI and CLI interfaces

Education

M.S. in Artificial Intelligence, Chung-Ang University · GPA: 4.2/4.5 2024-Present
B.S. in Computer Science, ETU "LETI" · GPA: 4.3/5.0 2018-2023
Exchange Semester, Inha University · GPA: 3.9/4.5 2022-2023

Publications

2 Journal Papers (MDPI Symmetry, 2025) · 2 Conference Papers (IEEE ICCE 2025, IEIE 2024)
Google Scholar: Timur Khairulov

Last updated: December 24, 2025