

Minseo Choi

410-301-6057 | mchoi46@jh.edu | linkedin.com/in/minseoc03 | github.com/minseoc03 | minseoc03.github.io

EDUCATION

Johns Hopkins University

Bachelor of Science in Computer Science and Computer Engineering

Baltimore, MD

Expected May 2027

- **GPA:** 3.91/4.0
- **Relevant Coursework:** Machine Learning, Operating Systems, Deep Learning, Computer Systems Fundamentals, Data Structures, Algorithms, Information Retrieval, Web Agent
- **Self-Study:** Deep Learning Optimization, GPU Programming, Compiler Design, Performance Optimization, Computer Vision, Natural Language Processing, Linux Systems, OS Resource Management, Profiling & Debugging

TECHNICAL SKILLS

Languages: CUDA, Triton, Python, C/C++, Java, Assembly

Frameworks & Libraries: vLLM, KServe, Ray Serve, PyTorch, LLVM, Optuna, Hydra, NumPy, Pandas, Scikit-Learn

Developer Tools: Git, Docker, Kubernetes, Linux, Shell, Vim, VS Code

PROJECTS

Production Log Analysis LLM Agent | *Ray Serve, vLLM, Python, Performance Benchmarking* Jan. 2026 – Present

- Built an inference-focused LLM serving system for production log analysis, emphasizing request batching, KV-cache efficiency, long-context memory behavior, and tail-latency analysis
- Explored dynamic batching and request scheduling trade-offs under concurrent workloads, measuring TTFT, end-to-end latency, throughput, and GPU memory utilization, analyzing system behavior under bursty and distributed request patterns
- Characterized latency–throughput–memory trade-offs by varying batch size, context length, and concurrency, quantifying the impact of KV cache reuse and PagedAttention
- Documented system design decisions and benchmarking methodology to ensure reproducibility and maintainability

Medical Image Enhancement Acceleration | *Python, Triton, FastAPI, Docker* Oct. 2025 – Present

- Built GPU-accelerated CT/MRI enhancement pipeline combining DICOM preprocessing, DL denoising, and post-processing
- Developed custom Triton kernels and optimize UNet/DnCNN inference layers, achieving up to 10–30× faster runtime
- Implemented automatic image-quality evaluation (PSNR, SSIM, NIQE) to quantify fidelity, perceptual improvement
- Built a PyTorch + Triton zero-copy system with a FastAPI interface for real-time visualization and PACS-ready export

FlashAttention Implementation | *Python, C++, CUDA, Triton* Aug. 2025 - Oct. 2025

- Implemented FlashAttention v1 from scratch using Triton and CUDA, focusing on memory-efficient attention computation
- Designed tiled attention kernels to minimize HBM accesses and avoid materializing the full attention matrix
- Analyzed GPU performance characteristics including memory bandwidth utilization and kernel-level parallelism, comparing against naive attention baselines, achieving 2x kernel speed up and 5x memory reduction

COMPETITIONS

LG Aimers 5 Hackathon – Display Defect Detection

Aug. 2024 – Sep. 2024

LG AI Research

- Top 5% (62 / 1,123)
- Analyzed imbalanced manufacturing data and applied various tuning skills to mitigate severe class imbalance
- Developed reproducible training pipelines under strict time constraints—reflecting strong research iteration speed
- Implemented XGBoost with Optuna hyperparameter optimization, increasing F1-score by ~18% compared to baseline model

EXPERIENCE

Computer Systems Fundamentals – Course Assistant (CA)

Jan. 2026 – Present

Johns Hopkins University

Baltimore, MD

- Served as a Course Assistant for EN.601.229 Computer Systems Fundamentals, mentoring students on low-level systems topics such as cache behavior, memory hierarchy, and performance trade-offs in C/C++
- Debugged and reviewed student code involving pointer arithmetic, memory access patterns, and performance bottlenecks on Linux systems

Drill Instructor & Senior Squad Leader

Oct. 2023 – Apr. 2025

Republic of Korea Army

Nonsan, Republic of Korea

- Trained and mentored 2,000+ recruits, fostering teamwork, resilience, and discipline under high-pressure condition
- Led team operations as Senior Squad Leader, coordinating large-scale training programs
- Developed resilience and rapid decision-making—skills essential for ambiguous, research-driven environments
- Strengthened mentorship, communication, and accountability, directly applicable to collaboration and project execution