EESUN MOON

eesun.moon@columbia.edu | LinkedIn | GitHub | NY, United States

EDUCATION

Columbia University New York, NY
M.S. in Computer Science, GPA: 3.92/4.0 Expected Dec 2025
- Courses: Applied Machine Learning, Natural Language Processing, Spoken Language Processing, Computer Vision, Database

Sejong University Seoul, South Korea
B.S. in Intelligent Mechatronics Engineering, B.E. in Data Science, GPA: 4.4/4.5 Feb 2024
- Courses: Artificial Intelligence, Data Structures, Algorithms, Statistics, Computer Networks, Operating Systems, Web Programming
- Teaching Assistant: Algorithms using C programming, Python fundamentals

PROFESSIONAL EXPERIENCE

Samsung Research America Mountain View, CA
AI Algorithm/NPU Simulator Research Scientist Intern Jun 2025 – Present
- Built automated profiling pipeline on Exynos NPU for layer-wise (FlashAttention, CNN) and end-to-end (ViTs, diffusion LLMs), **scaling to 14.9B+ configurations** to identify system-level bottlenecks and improve latency and MAC efficiency
- Developed compiler-level analysis framework for layer-wise insight extraction, **accelerating** simulator validation by **over 10x**
- Optimized memory layout for LLM inference on NPU by applying PagedAttention-inspired KV cache paging and quantization

Humaner AI: AI Startup – Conversational Commerce & Recommendations [GitHub] Seoul, South Korea
Applied Machine Learning Engineer Intern Mar 2024 – May 2024
- Delivered AI-powered fan engagement product from prototype to production in **<2 months**, deployed at 500+ attendee live event
- Built Q&A chatbot with OpenAI API and LangChain on AWS EC2 to generate personalized athlete cheering messages
- Integrated post-event survey feedback into RAG pipeline to improve personalization, boosting **user satisfaction by 20%**

Mobile Intelligent Embedded System Laboratory, Sejong University [GitHub] Seoul, South Korea
Research Assistant Sep 2021 – Mar 2024
- Led multimodal emotion recognition project for On-device AI using TensorFlow and MongoDB on Linux for government initiatives
- Optimized ONNX-based deep models with score-based fusion of multimodal signals (heart rate, EEG, speech, image), achieving **99.68% classification accuracy** and reducing **power consumption by 3.12x** and **latency by 1.48x** for edge deployment [1], [2]

PUBLICATIONS

[1] **Eesun Moon**, A.S.M Sharifuzzaman Sugar, Hyung Seok Kim, "Multimodal Daily-life Emotional Recognition Using Heart Rate and Speech Data from Wearables," *IEEE Access*, vol. 12, pp. 96635-96648, 2024. DOI
[2] Taein Kim, **Eesun Moon**, Hoyeon Kang, Hyung Seok Kim, "OMER-NPU: On-device Multimodal Emotion Recognition on Neural Processing Unit for Low Latency and Power Consumption," *Neural Computing and Applications*, 2025. DOI.

PROJECTS

CS Advising Assistant Chatbot with LLM, RAG, and Agentic Flow [GitHub] Jan 2025 – May 2025
- Built DeepSeek-R1-based chatbot with Ollama, **eliminating API costs** via local inference and deploying to GCP for production use
- Integrated Agentic Flow into LangChain RAG via MCP server, improving multi-step retrieval and reasoning quality

Sentence Embedding Analysis in LLMs [GitHub] Jan 2025 – May 2025
- Analyzed interpretability of BERT and LLaMA through layer-/domain-specific embeddings, revealing semantic structure shifts
- Validated cluster coherence via Zipf slope steepening (from -0.87 to -1.42) in intermediate/domain-specific embeddings

Ranking-Based Spam Filtering on Social Networking Services [GitHub] Mar 2022 – Jun 2022
- Spearheaded project to prioritize organic user posts over likely ads from social media, earning **1st place** in graduation competition
- Automated data collection with Selenium and implemented unsupervised clustering with cosine similarity-based ranking, achieving **0.8 intra-cluster similarity** as coherence indicator

TECHNICAL SKILLS

Programming & Databases	Python, C, C++, R, Java, SQL (MySQL, PostgreSQL), NoSQL (MongoDB)
ML/AI Frameworks	PyTorch, TensorFlow, Keras, Scikit-Learn, XGBoost, Hugging Face, LangChain, OpenAI API, Ray
Model Optimization & Data Tools	GPU, NPU, ONNX, TFLite, TensorRT | Pandas, NumPy, Selenium, Matplotlib
DevOps	Git, Docker, Kubernetes, Linux, FastAPI, AWS EC2, Google Cloud Platform