

EESUN MOON

eesun.moon@columbia.edu | [LinkedIn](#) | [GitHub](#) | NY, United States

EDUCATION

Columbia University

M.S. in Computer Science, GPA: 3.92/4.0

New York, NY

Expected Dec 2025

- Courses: Applied Machine Learning, Natural Language Processing, Spoken Language Processing, Computer Vision, Database

Sejong University

B.S. in Intelligent Mechatronics Engineering, B.E. in Data Science, GPA: 4.4/4.5

Seoul, South Korea

Feb 2024

- Courses: Artificial Intelligence, Computer Networks, Operating Systems, Image Processing, Data Structures, Web Programming
- Teaching Assistant: Algorithms using C programming, Python fundamentals

PROFESSIONAL EXPERIENCE

Samsung Research America

AI Algorithm/NPU Simulator Research Scientist

Mountain View, CA

Jun 2025 – Expected Aug 2025

- Built automated profiling pipeline for ViTs and LLMs on Exynos NPU, enabling inference optimization by identifying system-level bottlenecks and improving latency and MAC Utilization on Galaxy AI's Next-GenAI Platform
- Developed compiler-level analysis framework for layer-wise insight extraction, **accelerating** simulator validation by **over 10x**
- Optimized LLaMA3 inference on NPU, leveraging vLLM design for KV cache quantization and memory-efficient attention

Humaner: Human-centered AI Software Development [\[GitHub\]](#)

Seoul, South Korea

Machine Learning Engineer

Mar 2024 – May 2024

- Built Q&A-based message generator using GPT-3.5 Turbo and LangChain, deployed on AWS EC2 at live event for 500+ fans
- Collected post-event surveys to extract domain-specific facts and integrated into RAG pipeline, enhancing message relevance
- Iteratively refined prompts based on feedback and RAG data, 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**, Hyeon 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 chatbot with DeepSeek-R1 on Ollama, **eliminating API costs** via local inference; deployed to GCP for production
- Integrated Agentic Flow into LangChain RAG via MCP server, enabling multi-step retrieval and dynamic tool-based reasoning

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, R, Java | MySQL, PostgreSQL, MongoDB

ML/AI Frameworks

PyTorch, TensorFlow, Keras, Scikit-Learn, Hugging Face, LangChain, OpenAI API

Model Optimization

GPU, NPU, ONNX, TFLite, TensorRT

DevOps & Data Tools

Git, Docker, Linux, FastAPI, AWS EC2, GCP | Pandas, NumPy, Selenium, Matplotlib