Minhyeong Yu (https://github.com/gogoymh) **Updated on Mar 4, 2025.**

SUMMARY

AI/ML Engineer specializing in LLM-driven agent development, with experience in building scalable LLM inference clusters and implementing RAG pipelines using LangChain, LlamaIndex, and Prompt Engineering techniques (ReAct, Chain of Thought, In-Context Learning). Skilled in end-to-end AI agent development, from conceptualization to engineering and deployment.

EDUCATION

Korea University Mar 2019 - Feb 2022
M.S. in Computer Science and Engineering (*Advisor: Prof. Seung Jun Baek*)

University of Seoul Mar 2012 - Feb 2019
B.S. in Statistics and Data Science

EXPERIENCE

Machine Learning Engineer — **Deeping Source Inc.** Jun 2022 - present

- Built and managed an LLM inference cluster utilizing multiple GPU servers, optimizing inference speed and model deployment.
- Developed RAG-based AI agents for retail analytics, integrating LangChain and LlamaIndex for efficient data retrieval.
- Applied Chain of Thought (CoT) and ReAct-based prompting techniques to enhance reasoning capabilities.
- Optimized vision-based AI models for store analytics, applying quantization and pruning techniques to achieve 2x speedup.
- Developed an internal LLM-powered automation tool for financial documentation and news summarization.

M.S. Candidate — **System INtelligence Group (SING) Lab, Korea University** Mar 2019 - Feb 2022

- Research on Federated Semi-Supervised Segmentation
- Research on AI system for Rehabilitation Medicine
- Research on Medical Image Segmentation

LLM PROJECTS

- **LLM-Powered Retail AI Agent**, Deeping Source Nov 2023 - Present
  Developed LLM-powered agents to generate sales strategies, integrating RAG for improved retrieval accuracy. Implemented prompt engineering techniques (ReAct, CoT, In-context learning) to enhance reasoning and response accuracy. Integrated LlamaIndex and LangChain, improving agent reasoning capabilities for sales-focused question answering.
- **LLM Inference Cluster Optimization**, Deeping Source oct 2024 - Present
  Designed and scaled an LLM inference cluster with 4 8-GPU servers, enabling low-latency AI-driven customer interactions. Integrated API-based orchestration for seamless LLM query processing, reducing system overhead.
- **LLM-Powered Company Internal Automation**, Deeping Source Aug 2023 - Present
  Developed an LLM-based expense report automation tool, streamlining financial documentation for internal use, reducing processing time from an average of 30 minutes to 5 minutes. Built a news summarization and reporting system for the retail industry, leveraging LLM-driven extraction and summarization techniques.
- **Chatbots for Storecare**, Deeping Source Nov 2023 - Aug 2024
  Developed a chatbot system integrated with LLM to assist store managers and employees in managing retail store operations via KakaoTalk API. This chatbot allows users to check real-time store status, inventory levels, and operational insights. Successfully deployed in collaboration with a major convenience store chain through an MoU.
- **AI-Generated YouTube Content**, Personal Jan 2023 - Jan 2024
  Developed an AI-driven & code-based video generation pipeline using LLM & Diffusion models to automate YouTube content creation. Operated 7 YouTube channels, reaching 3,870 subscribers within a year. Conducted user engagement analysis to evaluate audience responses to AI-generated educational content.
- **Research Assistant on Persona Extension**, Personal Sep 2023 - Oct 2023
  Assisted in NLP research on persona extension, focusing on resolving persona conflicts in multi-session conversations. Optimized experiment workflow, reducing runtime from 2 days to 2 hours. Research was accepted at EACL 2024, though not officially credited as an author due to unofficial contribution.

ML PROJECTS

- **Data Scarcity Resolution**, Deeping Source Nov 2023 - present
  Using NVIDIA Omniverse, a realistic retail store simulation was constructed to generate multi-camera tracking data by adjusting various camera angles. This data was then processed through Diffusion and NERF models to create high-fidelity synthetic images, closely resembling actual CCTV footage. The generated data was utilized to train object tracking algorithms, enhancing security and customer experience management in retail stores. Additionally, Semi-Supervised Learning (SSL) strategies were applied to improve model generalization without requiring labeled data, effectively addressing data scarcity in vision tasks.

- **Cumulative Model Compression**, Deeping Source Feb 2023 - Oct 2023
  Many model compression methods are often not compatible with each other for cumulative use. For instance, applying quantization may preclude the possibility of network pruning, and compressing weights to lower precision can impose additional burdens on activation functions. This incompatibility of techniques for cumulative application poses practical challenges in real-world development. We have developed a model compression method that is both accumulable and deployable, addressing these practical challenges. Applied quantization & network pruning, achieving 2x inference speed improvement and 50% memory reduction across 3 vision tasks.
- **Implementing Quantization on Multiple Hardware Systems**, Deeping Source Feb 2023 - Oct 2023
  We have performed quantization in various frameworks, including Pytorch, Onnx, TensorRT, Openvino, AIMET, and Furiosa SDK, to enable the use of models for different tasks (classification, object detection) across a range of hardware platforms, such as NVIDIA, Intel, Qualcomm, and Furiosa.
- **Quantization Aware Training for Object Detection**, Deeping Source Jun 2022 - Jan 2023
  Quantization is a promising technique for faster speed of inference. However, it often struggles to maintain its performance. To address this issue, we conducted a study on quantization aware training. Our findings suggest that the quantization bias between fake quantized activation and full precision one can be reduced when the interaction in matrix multiplication is taken into account. We have documented our observations in a paper for further reference.
- **Federated Semi-Supervised Segmentaton**, Korea University Feb 2022 - Mar 2022
  Medical Image Segmentation is challenging due to limited annotated data and privacy concerns. Federated Learning and Semi-Supervised Learning help train models in a private way. We introduce FedWeP, a Federated Semi-Supervised Segmentation method using Randomized Weight Perturbation, where the server adds Gaussian noise to model weights for client training.
- **AI system for Rehabilitation Medicine**, Korea University Sep 2020 - Dec 2021
  We have been developing an AI-based system for rehabilitation medicine, supported by the Ministry of Science and ICT (MSIT) of Korea and supervised by the Institute for Information and Communications Technology Planning and Evaluation (IITP). During the first year of the ICT Creative Consilience program, we developed a system to assess hemiplegic patients and recommend suitable exercises. In the second year, we created an automated system for the detection of videofluoroscopic swallowing studies in stroke patients.
- **Medical Image Segmentaton**, Korea University May 2020 - Nov 2021
  We collaborated with Korea Guro Hospital to study nerve segmentation on ultrasound imaging modality, for which we were awarded the Excellence Prize at the Korean Academy of Neuromusculoskeletal Sonography. Subsequent to this, we applied for a patent for this research and further studied it to propose a novel convolution, namely Scale Attentional Convolution, specialized in ultrasound nerve image segmentation.

PUBLICATION

- **Minhyeong Yu**, Federica Spinola, Myeongjun Kim, Philipp Benz, Tae-hoon Kim, "Rethinking of Straight-Through Estimator: Quantization-Bias Aware Training", (under revision), 2025.
- Federica Spinola, Philipp Benz, **Minhyeong Yu**, Tae-hoon Kim, "Knowledge Assembly: Semi-Supervised Multi-Task Learning from Multiple Datasets with Disjoint Labels", Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2023.
- **Minhyeong Yu**, Sunwoo Kim, Seungjun Baek. "Federated Semi-Supervised Segmentation with Randomized Weight Perturbation", International Symposium on Biomedical Imaging (ISBI), 2023.
- Beom Suk Kim*, **Minhyeong Yu***, Sunwoo Kim, Joon Shik Yoon, Seungjun Baek, "Scale-Attentional U-Net for the Segmentation of the Median Nerve in Ultrasound Images", Ultrasonography, 2022.
- Minki Kim* **Minhyeong Yu***, "Selection and Proposal of Vertical Building Forest Sites in preparation for the implementation of the Seoul Park Cancellation", Review of Korean Society for Internet Information, 2018.

PATENT & HONOR

- "Method and apparatus for automatically recognizing peripheral nerves and measuring nerve indicators in ultrasound images based on deep learning algorithms", 10-2020-0067199, Rep. of Korea Jun 2020
- Excellence Prize, Korean Academy of Neuromusculoskeletal Sonography Nov 2020
- Excellence Prize, Seoul Digital Foundation Nov 2018