# Sojin Lee

M.Sc. Student at Seoul National University
Phone: +82 10-3064-5846 ◇ Email ◇ GitHub ◇ LinkedIn

## Education

**Seoul National University** Sep. 2023 - Aug. 2025
M.Sc. student, Computer Science and Engineering
- Advisor: Prof. U Kang
- GPA: 3.89 / 4.3

**Seoul National University** Mar. 2018 - Feb. 2023
B.A., Linguistics
- Double major: Computer Science and Engineering
- GPA: 3.79 / 4.3

## Research Interests

- Large language model (LLM)
- Model compression, Quantization, Pruning

## Research Experience

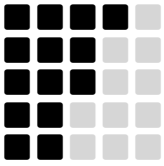
**Enhancing the Efficiency of Large Language Models via Extreme Compression** Jun. 2023 - Present
Youlchon AI Research Foundation
- Comprehensive survey on the state-of-the-art compression techniques for language models
- Currently working on devising a novel algorithm to reduce the memory usage of pre-trained LLMs via converting GQA to MLA and sensitivity-aware mixed-precision quantization

**Internship in NAVER LLM Solution** Jan.-Feb. 2025
NAVER | Tech.
- Developing the pipeline of automatic evaluation for LLMs based on Kubeflow framework
- Evaluating recently released LLMs (Phi-4, Mistral-24B, DeepSeek, etc.) using vLLM, an engine for accelerating LLM inference

**Devising a Novel Model Compression Algorithm for Large Language Models (PoC)** Feb.-Jun. 2024
LG AI Research
- Devising a mixed-precision quantization algorithm for large language models
- Sublayer-wise mixed-precision quantization based on the sensitivity difference between each sublayer
- Develop sublayer-wise pruning algorithm based on the findings from the project; the paper is submitted to IJCAI 2025 (Under review)

## Programming Skills

**Python/PyTorch**
**Java/C/C++**
**Kubernetes/Kubeflow/Docker**
**Cuda programming**
**Android/Kotlin**



## Language Skills

- TOEFL iBT (**99** / 120)
- OPIC (**IH**)

## Journal / Conference Papers

**A Comprehensive Survey of Compression Algorithms for Language Models** arXiv 2024
Seungcheol Park*, Jaehyeon Choi*, **Sojin Lee***, and U Kang *equal contribution.

**Accurate Sublayer Pruning for Large Language Models by Exploiting Latency and Tunability Information** IJCAI 2025 (under review)
Seungcheol Park*, **Sojin Lee***, Jongjin Kim*, Jinsik Lee, Hyunjik Jo, and U Kang *equal contribution.

## Teaching Experience

- Introduction to Data Mining @ SNU Sep.-Dec. 2024
- Data Structures @ SNU Mar.-Jun. 2024
- Model Compression @ HYUNDAI Nov. 2023