SUKMIN CHO

Republic of Korea
smcho@casys.kaist.ac.kr, nelllpic@kaist.ac.kr

RESEARCH STATEMENT

My research focuses on the **effective, efficient, and safe deployment of Large Language Models (LLMs)** for real-world applications. My previous work includes (1) enhancing LLM effectiveness by integrating them into Information Retrieval (IR), Question Answering (QA), and Retrieval-Augmented Generation (RAG), (2) assessing the robustness of LLMs against noisy real-world documents through adversarial attacks, and (3) developing efficient LLM solutions encompassing algorithmic and system-level optimizations, such as Speculative Decoding.

EDUCATION

Korea Advanced Institute of Technology and Science (KAIST) Daejeon, Korea
Ph.D. in School of Computing *September 2022 - Present*

Korea Advanced Institute of Technology and Science (KAIST) Daejeon, Korea
M.S. in School of Computing *March 2021 - August 2022*
Thesis: Template-based document labeling for dense retrieval

Korea Advanced Institute of Technology and Science (KAIST) Daejeon, Korea
B.S. in School of Computing *March 2016 - February 2021*
Minor in Mathematics

EMPLOYMENT

Graduate student, KAIST *March 2021 - Present*
Advisor: Prof. Youngjin Kwon

- Conducting research on developing an efficient LLM serving system with the search algorithm
- Conducted research on speculative decoding leveraging diverse data resources
- Conducted research on the robustness of Retrieval-Augmented Generation System
- Conducted research on adaptation of LLMs on IR, QA, and RAG

NAND Quality Assessment Intern, SK Hynix *January - February 2019*

- Developed visualization system for NAND testing.

PUBLICATIONS

C17 EXIT: Context-Aware Extractive Compression for Enhancing Retrieval-Augmented Generation
Taeho Hwang, Sukmin Cho, Soyeong Jeong, Hoyun Song, SeungYoon Han, and Jong C. Park
Findings of Annual Meeting of the Association for Computational Linguistics **(Findings of ACL)**, 2025.

C16 Does Rationale Quality Matter? Enhancing Mental Disorder Detection via Selective Reasoning Distillation
Hoyun Song, Huije Lee, Jisu Shin, Sukmin Cho, Changgeon Ko, and Jong C. Park
Findings of Annual Meeting of the Association for Computational Linguistics **(Findings of ACL)**, 2025.

C15 Lossless Acceleration of Large Language Models with Hierarchical Drafting based on Temporal Locality in Speculative Decoding
Sukmin Cho, Sangjin Choi, Taeho Hwang, Jeongyeon Seo, Soyeong Jeong, Huije Lee, Hoyun Song, Jong C. Park, and Youngjin Kwon

Findings of Nations of the Americas Chapter of the Association for Computational Linguistics **(Findings of NAACL)**, 2025.

C14 An Efficient Sign Language Translation Using Spatial Configuration and Motion Dynamics with LLMs
Eui Jun Hwang, Sukmin Cho, Junmyeong Lee, and Jong C. Park
Nations of the Americas Chapter of the Association for Computational Linguistics **(NAACL)**, 2025. **(Oral)**

C13 A Spatio-Temporal Representation Learning as an Alternative to Traditional Glosses in Sign Language Translation and Production
Eui Jun Hwang, Sukmin Cho, Huije Lee, Youngwoo Yoon, and Jong C Park
IEEE/CVF Winter Conference on Applications of Computer Vision **(WACV)**, 2025.

C12 PiLaMIM: Toward Richer Visual Representations by Integrating Pixel and Latent Masked Image Modeling
Junmyeong Lee, Euijun Hwang, Sukmin Cho, and Jong C. Park
Self-Supervised Learning - Theory and Practice **(SSL@NeurIPS)**, 2024.

C11 Typos that Broke the RAG's Back: Genetic Attack on RAG Pipeline by Simulating Documents in the Wild via Low-level Perturbations
Sukmin Cho, Soyeong Jeong, Jeongyeon Seo, Taeho Hwang, and Jong C. Park
Findings of Empirical Methods in Natural Language Processing **(Findings of EMNLP)**, 2024.

C10 Towards Effective Counter-Responses: Aligning Human Preferences with Strategies to Combat Online Trolling
Huije Lee, Hoyun Song, Jisu Shin, Sukmin Cho, SeungYoon Han, and Jong C. Park
Findings of Empirical Methods in Natural Language Processing **(Findings of EMNLP)**, 2024.

C9 DSLR: Document Refinement with Sentence-Level Re-ranking and Reconstruction to Enhance Retrieval-Augmented Generation
Taeho Hwang, Soyeong Jeong, Sukmin Cho, SeungYoon Han, and Jong C. Park
Knowledge Augmented Methods for NLP Workshop **(KnowledgeNLP@ACL)**, 2024.

C8 Preprocessing Mediapipe Keypoints with Keypoint Reconstruction and Anchors for Isolated Sign Language Recognition
Kyunggen Roh, Huije Lee, Eui Jun Hwang, Sukmin Cho, and Jong C. Park
Representation and Processing of Sign Languages: Evaluation of Sign Language Resources **(sign-lang@LREC-COLING)**, 2024.

C7 Adaptive-RAG: Learning to Adapt Retrieval-Augmented Large Language Models through Question Complexity
Soyeong Jeong, Jinheon Baek, Sukmin Cho, Sung Ju Hwang, and Jong C. Park
North American Chapter of the Association for Computational Linguistics **(NAACL)**, 2024.

C6 Improving Zero-shot Reader by Reducing Distractions from Irrelevant Documents in Open-Domain Question Answering
Sukmin Cho, Jeongyeon Seo, Soyeong Jeong and Jong C. Park
Findings of Empirical Methods in Natural Language Processing **(Findings of EMNLP)**, 2023.

C5 Test-Time Self-Adaptive Small Language Models for Question Answering
Soyeong Jeong, Jinheon Baek, Sukmin Cho, Sung Ju Hwang and Jong C. Park
Findings of Empirical Methods in Natural Language Processing **(Findings of EMNLP)**, 2023.

C4 Discrete Prompt Optimization via Constrained Generation for Zero-shot Re-ranker
Sukmin Cho, Soyeong Jeong, Jeongyeon Seo and Jong C. Park
Findings of Annual Meeting of the Association for Computational Linguistics **(Findings of ACL)**, 2023..

C3 Sign language production with avatar layering: A critical use case over rare words
Jung-Ho Kim, Eui Jun Hwang, Sukmin Cho, Du Hui Lee and Jong C. Park
International Conference on Language Resources and Evaluation **(LREC)**, 2022.

C2 Query generation with external knowledge for dense retrieval
Sukmin Cho, Soyeong Jeong, Wonsuk Yang and Jong C. Park
Deep Learning Inside Out **(DeeLIO@ACL)**, 2022.

C1 Augmenting Document Representations for Dense Retrieval with Interpolation and Perturbation
Soyeong Jeong, Jinheon Baek, Sukmin Cho, Sung Ju Hwang and Jong C. Park
Annual Meeting of the Association for Computational Linguistics **(ACL)**, 2022. **(Oral)**

ACADEMIC SERVICE

Reviewer of **ACL ARR 2025** February 2025
Reviewer of **ACL ARR 2024** February, April, June, October, December Reviewer 2024
Reviewer of **ACL ARR 2023** December Reviewer 2023
Reviewer of **IEEE Access** 2023

AWARD

Best Paper Award at Korea Computer Congress (KCC) 2024
Best Paper Award at Annual Conference on Human & Cognitive Language Technology (HCLT) 2023
Best Presentation Award at Korea Computer Congress (KCC) 2022

SKILLS

Language: Korean (mother tongue), English (fluent)
Programming: Python, C