Seogyeong Jeong

sg.jeong28@kaist.ac.kr | Homepage | LinkedIn | Google Scholar | GitHub

I am a Ph.D. student at KAIST School of Computing, advised by Prof. Alice Oh.

I study how reasoning works internally in language models and how that understanding can enable better intervention and control. My current research focuses on reasoning representations, mechanistic interpretability, and latent reasoning.

My research has evolved from evaluating language models through their observable behaviors [2, 3, 5], and gradually shifted toward **understanding the internal representations and mechanisms underlying those behaviors** [4, 6]. More recently, I have focused on **how reasoning unfolds beneath surface-level outputs** [1]. Building on these insights, I now aim to move beyond understanding reasoning toward **intervening in and controlling it**.

Broadly, I am interested in three connected questions:

- **How does reasoning work internally?**
- **Why does reasoning succeed or fail?**
- **How can we intervene to guide, improve, and control it?**

My long-term goal is to build language models capable of increasingly sophisticated reasoning while remaining **reliable**, **interpretable**, and **controllable**, enabling people to use AI safely and confidently.

Education

Korea Advanced Institute of Science and Technology (KAIST) Daejeon, Republic of Korea

- *Ph.D. in School of Computing* Sep 2026 – Present
  Advisor: Alice Oh
- *M.S. in School of Computing* Sep 2024 – Aug 2026
  Advisor: Alice Oh
- *B.S. in School of Computing* Feb 2019 – Aug 2024
  - Minor in *Mathematical Sciences*
  - Individually Designed Major in *Electrical Engineering* and *Bio and Brain Engineering*

Research Experience

NAVER AI Lab, Backbone Team Mar 2026 – Sep 2026
Research Intern

- Investigated **internal representations of LLM reasoning**, leading to a first-author paper at **EMNLP 2026 (Main Conference)** [1].
- Initiated research on **soft-token-based latent reasoning**, which is currently ongoing.

User & Information Lab Jun 2022 – Aug 2024
Research Intern — Graph neural networks, NLP

Emerging Nano Technology and Integrated Systems Lab Jun 2021 – Feb 2022
Research Intern — Neuromorphic computing

Publications

Conference Proceedings

1. **Seogyeong Jeong**, Dongyoon Han, Jaehui Hwang, Geonmo Gu, Alice Oh, Taekyung Kim. 2026. Beneath the Surface of Chains-of-Thought: A Mechanistic Interpretation of Reasoning Operations in LLMs. In *Main Conference of EMNLP 2026, Long*. [Demo]

2. **Seogyeong Jeong**[*], Kiwoong Park[*], Seyoung Song, Eunsu Kim, Ken E. Friedl, Jaeho Kim, Alice Oh. 2026. LoCar: Localization-Aware Evaluation of In-Vehicle Assistants through Fine-Grained Sociolinguistic Control. In *ACL 2026 Industry Track*.

3. Seyoung Song[*], **Seogyeong Jeong**[*], Eunsu Kim, Jiho Jin, Dongkwan Kim, Jay Shin, Alice Oh. 2025. MUG-Eval: A Proxy Evaluation Framework for Multilingual Generation Capabilities in Any Language. In *Findings of EMNLP 2025*.

Workshop Papers

4. Soyeon Park, **Seogyeong Jeong**[*], Sunwoo Kim[*], Alice Oh. 2026. The Role of Fine-grained Harm Signals in LLM Safety. In *NeurIPS 2026 Workshop – Interpretability for Science*.

5. Junyeong Park[*], **Seogyeong Jeong**[*], Seyoung Song[*], Yohan Lee, Alice Oh. 2025. LLM-C3MOD: A Human-LLM Collaborative System for Cross-Cultural Hate Speech Moderation. In *NAACL 2025 Workshop – C3NLP*.

Preprints

6. Haeun Yu, Arnav Arora, **Seogyeong Jeong**, Nadav Borenstein, Siddhesh Pawar, Jisu Shin, Jiho Jin, Junho Myung, Alice Oh, Isabelle Augenstein. 2025. Entangled in Representations: Mechanistic Investigation of Cultural Biases in Large Language Models. *arXiv preprint*.

7. Paul Röttger, Giuseppe Attanasio, Felix Friedrich, Janis Goldzycher, Alicia Parrish, Rishabh Bhardwaj, Chiara Di Bonaventura, Roman Eng, Gaia El Khoury Geagea, Sujata Goswami, Jieun Han, Dirk Hovy, **Seogyeong Jeong**, Paloma Jeretič, Flor Miriam Plaza-del-Arco, Donya Rooein, Patrick Schramowski, Anastassia Shaitarova, Xudong Shen, Richard Willats, Andrea Zugarini, Bertie Vidgen. 2025. MSTS: A Multimodal Safety Test Suite for Vision-Language Models. *arXiv preprint*.

** Indicates equal contribution.*

Awards & Honors

- **Excellence Award (2nd Place)**, 3rd AI Hackathon for Network Intelligence
  Open AI Networking Korea (ONK) 2025, hosted by AI Network Forum (**Top 2%** among 94 teams)

Activities

Academic Service 2026

- Reviewer, NeurIPS 2026 Workshop – Interpretability for Science
- Reviewer, EMNLP 2026 Industry Track
- Emergency Reviewer, NeurIPS 2026 Evaluations and Datasets Track

Teaching

- Teaching Assistant, *Introduction to Computer Programming*, KAIST Spring & Fall 2025