

Seungkyu Lee

✉ Email: sjl6932@psu.edu

🌐 Website: seungkyu-lee-01.github.io

RESEARCH INTERESTS

Personalized AI Agents: Developing agents that understand user intent and adapt their behavior to user preferences and context. *Keywords: agent memory, intent understanding, personalization.*

Efficient Language Models: Enabling language models to perform reliable and structured reasoning while minimizing inference cost and computational overhead. *Keywords: small language models, inference-time optimization.*

EDUCATION

The Pennsylvania State University

Ph.D. Student in Computer Science and Engineering

- Advisor: Prof. Wenpeng Yin

University Park, PA

Aug. 2026 – Present

Seoul National University (SNU)

B.S. in Industrial Engineering

Seoul, South Korea

Mar. 2019 – Feb. 2026

Stony Brook University (SUNY)

Exchange Student in Computer Science (Spring 2024)

Stony Brook, NY

Jan. 2024 – May 2024

HONORS AND AWARDS

Korean Government Scholarship Program for Study Overseas (KGSPSO)

National Institute for International Education (NIIED), Ministry of Education

Seoul, South Korea

Jul. 2026

HopperHackers 2024 - Best Beginner Hack (1st Place)

SUNY Stony Brook Hackathon

Stony Brook, NY

Feb. 2024

Lee Joonghan Award, Leadership Sector

Dept. of Industrial Engineering, Seoul National University

Seoul, South Korea

Dec. 2021

PUBLICATIONS

In-N-Out: A Parameter-Level API Graph Dataset for Tool Agents

Seungkyu Lee, Nalim Kim, Yohan Jo.

[TACL] Transactions of the Association for Computational Linguistics, 2026.

ThinkBrake: Efficient Reasoning via Log-Probability Margin Guided Decoding

Sangjun Song*, Minjae Oh*, Seungkyu Lee, Sungmin Jo, Yohan Jo.

[ACL] Findings of the Association for Computational Linguistics, 2026.

ThinkBrake: Mitigating Overthinking in Tool Reasoning

Minjae Oh*, Sangjun Song*, Seungkyu Lee*, Sungmin Jo, Yohan Jo.

* Equal contribution (co-first authors).

[NeurIPS] Workshop on Efficient Reasoning @NeurIPS 2025.

RESEARCH EXPERIENCE

Human-Oriented Language Intelligence Lab

Undergraduate Researcher | Advisor: Prof. Yohan Jo, SNU

Seoul, South Korea

Aug. 2024 – Jan. 2026

Structured Knowledge for Tool Agents

- Built *In-N-Out*, a parameter-level API graph dataset covering **550** real-world APIs; fine-tuned LLMs on *In-N-Out* to infer parameter-level connections from noisy API documentation
- Leveraged gold API graphs to improve LLM tool retrieval accuracy from **43.3%** to **84.3%** over documentation-only baselines; automatically generated graphs recovered **~90%** of that gain

Mitigation of Overthinking in Tool Reasoning

- Diagnosed overthinking in Small Reasoning Models (SRMs) for tool use; injecting a `</think>` token at sentence boundaries raised accuracy from **86%** to **94%** while reducing tokens by **87%**
- Proposed *ThinkBrake*, a training-free decoding rule that triggers an end-of-thought when the log-probability margin between `</think>` and the top token falls below a threshold; reduced tokens by up to **25%** without harming accuracy

Value Alignment in LLMs

- Investigated how vector steering can align LLMs' survey responses and behavioral outputs to achieve consistent value expression and reasoning

SNU Big Data AI Center

Research Assistant | Advisor: Prof. Sungzoon Cho, SNU

Seoul, South Korea

Jul. 2021 – Aug. 2021

- Implemented a Python-based Amazon product-review crawler (pagination, incremental refresh, rate-limit backoff) for the *Social Customer Value Factor (CVF) Exploration with Electronic Product Review Data* project
- Reviewed 10+ NLP papers (aspect extraction, sentiment classification, topical taxonomy) and assisted in shaping research directions and dataset design

PROFESSIONAL EXPERIENCE

SAP Labs Korea

Software Engineer Intern

Seoul, South Korea

Jul. 2025 – Jul. 2026

- Built an AI agent for *SAP HANA* that diagnoses current system states, recommends next best actions, and visualizes relevant data to help customers seamlessly resolve system issues
- Trained time-series models on database resource metrics (CPU, memory) to forecast trends; identified early precursor signals correlated with CPU spikes, achieving a **10%** improvement in prediction accuracy

Liner

Machine Learning Engineer

Seoul, South Korea

Nov. 2022 – Jan. 2024

- Designed and productionized an LLM-based autonomous agent that interprets user intent, decomposes requests into smaller subtasks, and delivers reliable answers to over **10k** user queries per day
- Developed a grounding copilot using retrieval-augmented generation (RAG) for content summarization and Q&A, improving factual accuracy and summary coherence, leading to a **20%** increase in daily active users

Nudge Healthcare (CashWalk)

Data Analyst

Seoul, South Korea

Jan. 2022 – Nov. 2022

- Conducted large-scale data analyses on over **3M** user records to identify critical bottlenecks in the onboarding process, introducing data-driven design changes that reduced user drop-off by **50%**
- Built and maintained over **20** Tableau dashboards to monitor key engagement metrics; discovered a strong correlation between referral rewards and first-open rates, optimizing reward policies and restoring growth momentum

SKILLS

Languages: Korean (Native), English (Fluent, TOEFL 104)

Programming: Python (Advanced), SQL (Proficient), Go, Java, R, C (Foundational)