

SEUNGJONG SUN

Ph.D. Candidate

tmdwhd406@g.skku.edu · +82-10-4057-3158

[Google Scholar](#) · [rss-researcher.github.io](#)

SUMMARY

I study how large language models imitate human personality, emotion, and interaction — what human values and behaviors they carry, and how those can be read and edited in the model's internals.

- **Human Values and Behavior in LLMs.** I ask how LLMs reproduce human personality, interaction, and social behavior, and how these are encoded in their internal representations [C1, C5, PP1].
- **Interpretability and Model Control.** I use activation steering and model diffing to locate the internal directions behind these behaviors [C1], and model merging to align models without additional training [C3].

I also work on AI safety, developing evaluation benchmarks for LLM safeguards in sarcasm understanding [C4] and jailbreak robustness [C2].

Keywords: Natural Language Processing, Affective Computing, Mechanistic Interpretability, Model Merging, AI Safety, Computational Social Science

EDUCATION

2023.09 – 2027.02	Sungkyunkwan University Ph.D. Candidate in Applied Artificial Intelligence · Expected February 2027 Advised by Prof. Jang Hyun Kim	Seoul, Korea
2021.03 – 2023.02	Sungkyunkwan University M.S. in Engineering, Human-Artificial Intelligence Interaction · Advised by Prof. Jang Hyun Kim	Seoul, Korea
2015.03 – 2021.02	Kyung Hee University B.A. in International Studies	Yongin, Korea

EXPERIENCE

2025.06 – 2025.09	Research Intern @ USC Institute for Creative Technologies • Developed soft-thinking LLM agents; investigated how LLMs process implicit (intrinsic) knowledge • Researched latent-space dialogue planning for task-oriented dialogue systems • Supervised by Dr. Gale M. Lucas	USC ICT
2021.03 – Present	Graduate Researcher @ Data Science and Social Analytics Lab (DSSAL), SKKU • LLM personality control, persona-conditioned behavior, and LLM-based social simulation • Advised by Prof. Jang Hyun Kim	DSSAL

PUBLICATION

** denotes equal contribution; ^ denotes corresponding author*

Conference

- [C1] **Seungjong Sun**, Hyunmin Jun, Minsik Yoon, Eungu Lee, Choongwon Kang, Young Seok Woo, Jang Hyun Kim. *No Matter Who You Are: Addressivity and Sycophancy in Large Language Models*. **EMNLP 2026**.
- [C2] Choongwon Kang, **Seungjong Sun***, Hyunmin Jun, Jang Hyun Kim. *Jailbreaking Multimodal Large Language Models using Multi-Clip Video*. **ACL 2026**.
- [C3] **Seungjong Sun**, Seo Yeon Baek, Jang Hyun Kim. *Personality Vector: Modulating Personality of Large Language Models by Model Merging*. **EMNLP 2025**.
- [C4] Choongwon Kang, Wonbyung Lee, Seunghyun Hwang, Sungho Tae, **Seungjong Sun^**, Jang Hyun Kim. *Sarcasm Subtype-Specific Reasoning in Dialogue with Multimodal Cues Using Large Language Models*. **CIKM 2025**.
- [C5] **Seungjong Sun**, Eungu Lee, Seo Yeon Baek, Seunghyun Hwang, Wonbyung Lee, Dongyan Nan, Bernard J. Jansen, Jang Hyun Kim. *Kiss up, Kick down: Exploring Behavioral Changes in Multi-modal Large Language Models with Assigned Visual Personas*. **EMNLP 2024**.

Journal

- [J1] Xiangying Zhao, Shunan Zhang, **Seungjong Sun**, Dongyan Nan, Jang Hyun Kim. *Exploring factors influencing word-of-mouth intentions for AI profile picture generation services*. **International Journal of Human-Computer Interaction**, 2025.

[J2] Dongyan Nan, Xiangying Zhao, Chaomei Chen, **Seungjong Sun**, Kyeo Re Lee, Jang Hyun Kim. *Bibliometric analysis on ChatGPT research with CiteSpace*. **Information**, 2025.

[J3] Dongyan Nan, **Seungjong Sun**, Shunan Zhang, Xiangying Zhao, Jang Hyun Kim. *Analyzing behavioral intentions toward generative artificial intelligence: The case of ChatGPT*. **Universal Access in the Information Society**, 2024.

[J4] **Seungjong Sun**, Jang Hyun Kim, Kwan Min Lee, Dongyan Nan. *Exploring the relationship between the Proteus effect and intention to play massive multiplayer online role-playing games*. **Internet Research**, 2023.

[J5] **Seungjong Sun**, Dongyan Nan, Jang Hyun Kim, Pim Kamphuis. *Exploring hype in the metaverse: Topic modeling analysis of Korean Twitter user data*. **Systems**, 2023.

[J6] Dongyan Nan, **Seungjong Sun**, Jang Hyun Kim. *Beyond avatar coolness: Exploring the impact of avatar coolness on continuance intention to play MMORPGs*. **International Journal of Human-Computer Interaction**, 2023.

[J7] Shunan Zhang, Dongyan Nan, **Seungjong Sun**, ShaoPeng Che, Jang Hyun Kim. *Teammate familiarity in distributed computer-supported collaborative learning: The mediating role of social presence*. **International Review of Research in Open and Distributed Learning**, 2023.

[J8] **Seungjong Sun**, Dongyan Nan, ShaoPeng Che, Jang Hyun Kim. *Investigating the knowledge structure of research on MMORPGs: A bibliometric analysis*. **Data and Information Management**, 2022.

Preprint

[PP1] **Seungjong Sun**, Eungu Lee, Dongyan Nan, Xiangying Zhao, Wonbyung Lee, Bernard J. Jansen, Jang Hyun Kim. *Random Silicon Sampling: Simulating Human Sub-Population Opinion Using a Large Language Model Based on Group-Level Demographic Information*. **arXiv:2402.18144**, 2024.

TEACHING EXPERIENCE

Fall 2026	Introduction to Artificial Intelligence Instructor · Dept. of Applied Artificial Intelligence	SKKU
Fall 2026	Natural Language Processing and AI Instructor · Dept. of Big Data	SKKU
Fall 2023	Metaverse Big Data Analysis Teaching Assistant	SKKU
Spring 2022	Creative Convergence Project Teaching Assistant	SKKU
Fall 2021	Data Mining Teaching Assistant	SKKU
Spring 2021	Data Analytics in Action with Python Teaching Assistant	SKKU

ACADEMIC & PROFESSIONAL SERVICE

Invited Talk

Jul. 2026	AI Transformation (AX) Strategy Technical advisor (one-time consultation), Korea Institute for Animal Products Quality Evaluation	KAPE
Mar. 2026	Expert Roundtable on AI Education Invited expert panelist, Seoul AI Hub	Seoul AI Hub
Nov. 2025	AI Health Chatbot for Older Adults Technical advisor (one-time consultation), Unimeo	Unimeo
Aug. 2025	Simulating Health and Medical Behaviors Using Large Language Models Invited talk at Samsung Advanced Institute for Health Sciences and Technology (SAIHST)	SAIHST
May 2024	Human Behavior Simulation Using Generative Artificial Intelligence Invited talk at COGNET 9, Inc.	COGNET 9

Reviewing

2026	Program Committee, CIKM 2026
2026	Secondary Reviewer, EMNLP 2026

Military Service

2017.01 – 2018.10	Republic of Korea Auxiliary Police • Served at 806th Company, 5th Mobile Division	ROK Police
-------------------	---	------------

AWARD & SCHOLARSHIP

Aug. 2026	Technical Achievement Award, UofT-SKKU International Entrepreneurship Bootcamp	UofT / SKKU
Aug. 2026	Encouragement Award, 1st Graduate Research Competition (Research Matters Plus)	SKKU
May 2023	1st Prize, PROMPTHON	Wrtm Technologies
May 2023	Outstanding Poster Award, 12th Conference of the Ergonomics Society of Korea	ESK
2021 – 2025	BK21 Scholarship (Ph.D. / M.S.)	NRF
2021 – 2025	Shim San Scholarship	SKKU

SKILL

Programming	Python, PyTorch
Language	Korean (native), English (fluent)