Injin Kong

M.S. Student, Graduate School of Data Science, Seoul National University
Advisor: Yohan Jo

Email: mtkong77@snu.ac.kr
Mobile: +82-10-6243-2892
Seoul, Republic of Korea

Personal Data

- **Nationality**: Republic of Korea
- **Language**: Native Korean, professional working proficiency in English

Education

- **Seoul National University** — Seoul, Korea
  M.S. in Data Science — March 2025 - Present
  Advisor: Yohan Jo
  Research Areas: Large Language Models, Mechanistic Interpretability, Diffusion Language Models, Multimodal Reasoning

- **Seoul National University** — Seoul, Korea
  B.S. in Mathematics — March 2020 - February 2025

Research Interests

- **Mechanistic Interpretability, Language Models, Diffusion Language Models, Multimodal Reasoning, and AI Agents**
  - Mechanistic interpretability of large language models
  - Autoregressive and diffusion-based language modeling
  - Representation learning and style extraction from text embeddings
  - Value expression, alignment, and internal mechanisms in LLMs
  - Multimodal reasoning, visual persuasion, and rationale faithfulness
  - AI agents and retrieval-augmented generation systems

Publications

*: Equal contribution

Peer-Reviewed Publications

- Jongwook Han, Jongwon Lim, **Injin Kong**, and Yohan Jo, "Dual Mechanisms of Value Expression: Intrinsic vs. Prompted Values in Large Language Models," International Conference on Machine Learning (ICML), 2026.

Preprints and Manuscripts

- **Injin Kong**, Hyoungjoon Lee, and Yohan Jo, "Where Should Diffusion Enter a Language Model? Geometry-Guided Hidden-State Replacement," arXiv preprint arXiv:2605.14368, 2026.
- Naeun Lee, Hyunjong Kim, Sunghwan Choi, **Injin Kong**, and Yohan Jo, "Can MLLMs Reason About Visual Persuasion? Evaluating the Efficacy and Faithfulness of Reasoning," arXiv preprint arXiv:2605.08965, 2026.
- **Injin Kong***, Hyoungjoon Lee*, and Yohan Jo, "Mechanism Shift During Post-training from Autoregressive to Masked Diffusion Language Models," arXiv preprint arXiv:2601.14758, 2026.
- **Injin Kong**, Shinyee Kang, Yuna Park, Sooyong Kim, and Sanghyun Park, "Style Extraction on Text Embeddings Using VAE and Parallel Dataset," arXiv preprint arXiv:2502.08668, 2025.

Research Experiences

- **Graduate Researcher, Seoul National University** — March 2025 - Present
  Graduate School of Data Science, advised by Yohan Jo
  Conducting research on mechanistic interpretability, value expression in LLMs, diffusion language models, and multimodal reasoning. Current projects analyze how internal representations and circuits change across prompting, post-training, and multimodal rationale generation.

- **Research on Diffusion Language Models** — 2025 - Present
  Mechanism analysis of autoregressive and masked diffusion language models
  Studied how post-training autoregressive models into masked diffusion models changes internal computation. Analyzed whether diffusion post-training preserves autoregressive circuits or induces new mechanisms for non-sequential and global planning tasks.

- **Research on Value Mechanisms in LLMs** 2025 - 2026
  Intrinsic and prompted value expression in large language models
  Investigated how LLMs express values through intrinsic model behavior and explicit prompting. Contributed to mechanistic analyses using value directions and value-related neurons to compare steerability, response diversity, and instruction-following effects.
- **Research on Visual Persuasion Reasoning in MLLMs** 2026 - Present
  Rationale supervision and faithfulness evaluation for multimodal reasoning
  Studied whether MLLMs can reason faithfully about visual persuasiveness. Contributed to a framework evaluating rationale-to-decision consistency, rationale-to-image groundedness, and rationale-to-decision sensitivity.
- **Research on Style Extraction from Text Embeddings** 2024 - 2025
  VAE-based analysis of style distributions in parallel text data
  Developed a VAE-based approach for separating and analyzing stylistic variation in text embeddings using parallel Bible translations. Studied how style can be represented as distributions in embedding space and applied to AI-based text generation and style analysis.

Entrepreneurial and Industry Experiences

- **Aardvark** November 2023 - August 2025
  Co-Founder and Chief Scientific Officer, AI EdTech Startup
  Built AI systems for generating English test questions, including data construction, retrieval, model prompting, and evaluation pipelines. Led research and product development for LLM-based educational content generation.
- **BLACKLABEL Geometry** 2020
  Mathematics Content Reviewer, Jinhaksa
  Reviewed advanced geometry problems for high-achieving students in *BLACKLABEL Geometry*, a mathematics workbook based on the 2015 revised Korean national curriculum.

Teaching Experience

- **Teaching Assistant, Computing 1** 2025
  Seoul National University

Academic Services

- Reviewer, ICML 2026 Workshop on Pluralistic Alignment

Honors and Awards

- **Second Place, CJ Logistics Future Technology Challenge** 2023
- **Academic Excellence Award (Highest Academic Achievement), Yonsei University** 2019
- **Gold Prize, Korean Mathematical Olympiad (KMO)** 2018
- **Grand Prize (1st Place), Earth Science Competition, Hana Academy Seoul** 2018
- **Gold Prize (2nd Place), Mathematics Research Presentation Contest, Hana Academy Seoul** 2018