Mengyu Ye

Ph.D. Student | **LLM Interpretability & Reasoning & Post-training & Diffusion Language Models**

Tohoku University, Sendai, Japan

Email: ye.mengyu.s1@dc.tohoku.ac.jp | URL: https://muyo8692.com

Education

Tohoku University — Sendai, Japan

Ph.D. Student in Information Science — *April 2024 - 2027 (expected)*
Advisor: Prof. Jun Suzuki

M.S. in Information Science — *April 2022 - 2024*
Advisor: Prof. Jun Suzuki

B.S. in Engineering — *April 2018 - 2022*
Advisor: Prof. Xiao Zhou

Awards & Grants

Google PhD Fellowship 2025 — *2025*

Gemma 2 Academic Research Program Grant — *2024*

BOOST Fellowship of JST — *2024*

Best Paper Award – ACL 2023 Student Research Workshop — *2023*

Publications

1. **Mengyu Ye**, Jun Suzuki, Tatsuro Inaba, Kuribayashi. Transformer Key-Value Memories Are Nearly as Interpretable as Sparse Autoencoders. *In The Thirty-ninth Annual Conference on Neural Information Processing Systems (NeurIPS 2025).* (to appear).

2. Tarek Naous, Anagha Savit, Carlos Rafael Catalan, Geyang Guo, Jaehyeok Lee, Kyungdon Lee, Lheane Marie Dizon, **Mengyu Ye**, Neel Kothari, Sahajpreet Singh, Sarah Masud, Tanish Patwa, Trung Thanh Tran, Zohaib Khan, Alan Ritter, JinYeong Bak, Keisuke Sakaguchi, Tanmoy Chakraborty, Yuki Arase, Wei Xu. Camellia: Benchmarking Cultural Biases in LLMs for Asian Languages. *arXiv preprint.* [pdf]

3. **Mengyu Ye**, Tatsuki Kuribayashi, Tatsuki Kuribayashi, Goro Kobayashi, Jun Suzuki. Can Input Attributions Explain Inductive Reasoning in In-Context Learning?. *In Findings of the Association for Computational Linguistics: ACL 2025 (Findings of ACL 2025).* [pdf]

4. **Mengyu Ye**, Tatsuki Kuribayashi, Jun Suzuki, Goro Kobayashi, Hiroaki Funayama. Assessing Step-by-Step Reasoning Against Lexical Negation: A Case Study on Syllogism. *In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP 2023).* [pdf]

    - **Mengyu Ye**, Tatsuki Kuribayashi, Jun Suzuki, Goro Kobayashi, Hiroaki Funayama. Assessing Chain-of-Thought Reasoning Against Lexical Negation: A Case Study on Syllogism. *Non-archival submission for ACL-SRW 2023.*
      🏆 Best Paper Award at ACL-SRW 2023

5. Hiroto Kurita, Ikumi Ito, Hiroaki Funayama, Shota Sasaki, Shoji Moriya, **Ye Mengyu**, et al. TohokuNLP at SemEval-2023 Task 5: Clickbait Spoiling via Simple Seq2Seq Generation and Ensembling. *In Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023).* [pdf]

Skills

Programming Languages: Python, MATLAB, Java, C/C++, SQL (MySQL), SML#

Languages: Languages: Chinese (Native), Japanese (Near-Native), English (Professional Working Proficiency)

Machine Learning Frameworks: Pytorch, JAX/Flax

Tools & Methods: ML pipeline development, High-Performance Computing (HPC) Clusters (ABCI etc.), Containerization (Docker, Singularity etc.), Synthetic data generation

Experience

Moonshot Research and Development Program *Sept. 2022 - 2024*

Research Assistant

- Conducted fine-tuning of LLaMA models to develop advanced Japanese LLMs, enhancing their linguistic and contextual performance.

- Designed and implemented a RAG system tailored for integration into prototype cybernetic avatar applications.

Tohoku University *April 2023 - Present*

Research Assistant

- Designed, implemented machine learning pipelines for large-scale model evaluation, involving dozens of LLMs, multiple tasks, feature analysis while ensuring efficient, reproducible workflows.

- Developed an end-to-end pipeline evaluating input attribution methods in LLMs within ICL settings through synthetic task generation, LLM fine-tuning, task accuracy assessment, attribution performance measurement.

- Collaborating with Georgia Tech researchers to develop evaluation frameworks for cultural bias detection in Asian languages. Leading data creation efforts for Japanese and Chinese, ensuring cross-lingual consistency.

- Benchmarked the interpretability of a large language model's internal feed-forward layers against state-of-the-art Sparse Autoencoders (SAEs), revealing them to be a surprisingly powerful and efficient baseline.

Projects

Post-Training and Inference Algorithm for Diffusion Language Models

- Investigating a lightweight post-training method combined with a novel inference algorithm to improve coherence and fluency in long-form text generation.
- Aiming to advance the applicability of masked diffusion LMs to real-world language tasks.

Deepresearch-based RAG agent perform in real-world conditions

- Developed an agentic RAG framework capable of deep research by performing iterative, multi-hop reasoning.
- Goal: enable robust information synthesis in real-world conditions.

Training Sparse Autoencoders for Japanese LLM

- Training and plan to releasing Sparse Autoencoders trained on Japanese LLMs to contribute valuable resources to the research community.

Teaching Experience

Teaching Assistant

2024 Spring - Seminar on System Information Sciences
2024 All year - Advanced Seminar on System Information Sciences B

Mentorship

2023 - Kazuki Yano, master's student researcher from Tohoku university GSIS department
2024 - Koichi Iwakawa, Haochen Zhu, master's student researcher from Tohoku university GSIS department
2025 - Wataru Ikeda, Hinata Sugimoto, master's student researcher from Tohoku university GSIS department

Service

2024 - ACL Rolling Review (ARR)