Mengyu Ye

Second Year Ph.D. Student | **Google PhD Fellow** | Tohoku University, Sendai, Japan

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

Research Interests: LLM interpretability, reasoning, post-training, diffusion LMs

Education

Tohoku University — Sendai, Japan

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

M.S. in Information Science — *April 2022 - 2024*

B.S. in Engineering — *April 2018 - 2022*

Awards & Grants

Google PhD Fellowship *2025*

Gemma 2 Academic Research Program Grant *2024*

BOOST Fellowship of JST *2024*

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

Selected Publications

- **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). [paper link]
- 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.* [paper link]
- **Mengyu Ye**, 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).* [paper link]
- **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).* [paper link]
  - **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
- Hiroto Kurita, Ikumi Ito, Hiroaki Funayama, Shota Sasaki, Shoji Moriya, **Ye Mengyu**, Kazuma Kokuta, Ryujin Hatakeyama, Shusaku Sone, Kentaro Inui. 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).* [paper link]

Skills

Core ML: Python, PyTorch, JAX/Flax, LLM post-training, interpretability analysis

Systems: HPC clusters, Linux, Docker/Singularity, distributed training workflows

Research Tooling: dataset construction, evaluation pipelines, reasoning evaluation

Additional Technical Exposure: Java, C/C++, MATLAB, SQL (MySQL), SML#

Research Focus Areas: machine learning research; deep learning; NLP/LLMs; generative models; diffusion language models; evaluation methodology; model analysis; alignment research

Selected Experience

Tohoku University *Apr 2023 – Present*
Research Assistant

- **General Pipeline Building**: Built automated LLM evaluation pipelines for HuggingFace-compatible models, enabling reproducible evaluation used across multiple research publications.
- **Interpretation Pipeline Building**: Developed ICL attribution-analysis pipelines by fine-tuning multiple LLMs (up to 27B) on synthetic tasks and evaluating their inductive-reasoning behaviors; supported ACL Findings 2025 experiments.
- **Dataset Construction**: Led creation of Japanese and Chinese subsets for a 19k+ cultural-bias dataset, in collaboration with researchers at Georgia Tech.
- **Deep Research Agent**: Built an agentic deep-research system based on an 80B open-weights LLM, capable of answering real-world, long-form questions through iterative evidence synthesis.
- **Diffusion LM Post-Training**: Developing lightweight post-training and inference methods to improve in-text coherence in diffusion-based LMs.

Moonshot Research and Development Program *Sept 2022 – 2024*
Research Assistant

- **Cross-Lingual Capability Transfer for LLMs**: Explored enhancing Japanese capabilities from English-centric LLM (LLaMA) through web-corpus translation and fine-tuning.
- **RAG Pipeline**: Built a low-latency RAG pipeline for a lab cybernetic-avatar prototype using Chroma DB, enabling real-time retrieval and domain-specific question answering.

Selected Projects

Optimizing LLM on Math Tasks
- Post-training the llm-jp model to improve mathematical problem-solving capabilities, with a focus on generating executable code directly as part of the solution process.

Bib Reference Auto Cleaner
- Developing a tool to clean and standardize raw BibTeX entries, supporting customizable key formats, automatic removal of redundant fields, and retrieval of the latest publication metadata from OpenReview. Planned release as a Python package.

Training Sparse Autoencoders for Japanese LLM
- Training Sparse Autoencoders on native Japanese LLMs (llm-jp models), with the aim of releasing the models and contributing high-quality interpretability 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)