Mengyu Ye

Second Year Ph.D. Student | **Google PhD Fellow** | **LLM Evaluation & Interpretability & Generative Models**

Tohoku University, Sendai, Japan

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

Research Summary

- Diagnose LLM failure modes through controlled experiments and causal interventions. Benchmarked FF key-value memories against Sparse Autoencoders, showing comparable interpretability at a fraction of the cost (NeurIPS 2025).
- Build automated evaluation harnesses and diagnostic protocols that convert qualitative failures into reproducible, actionable signals; won NeurIPS 2025 MMU-RAG Best Static Evaluation Award.
- End-to-end research spanning dataset construction, training, inference, and evaluation across interpretability, post-training, and diffusion language models.

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

Winner, Best Static Evaluation Award - NeurIPS 2025 MMU-RAG Competition	*2025*
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, Tatsuki 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).* [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).* [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).* [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
- **Mengyu Ye**, Ryosuke Takahashi, Keito Kudo, Jun Suzuki. Relaxing Positional Alignment in Masked Diffusion Language Models. *arXiv preprint.* [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.* [LINK]
- 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).* [LINK]

Skills

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

- *Evaluation & Diagnostics*: automated evaluation harnesses, error taxonomy, counterfactual testing, robustness analysis, regression tests, reasoning evaluation

- *Systems*: HPC environments (PJM, PBS), Linux, Docker/Singularity, distributed training workflows

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

Selected Experience

Tohoku University *Apr 2023 – Present*
Research Assistant

- **Automated Evaluation Harnesses**: Built automated LLM evaluation pipelines for HuggingFace-compatible models, enabling reproducible evaluation used across multiple research publications.
- **Mechanistic Diagnostics**: Developed diagnostic attribution pipelines for in-context learning by training/fine-tuning multiple LLMs (up to 27B) on controlled synthetic tasks (Findings of ACL 2025).
- **Deep Research Agent**: Trained an evidence-grounded deep-research agent for long-form question answering; designed and integrated a static-evaluation suite combining retrieval precision, citation faithfulness, and answer quality metrics; won NeurIPS 2025 MMU-RAG Best Static Evaluation Award. [LINK]
- **Post-training for Math Reasoning**: Post-trained llm-jp for math problem solving with executable-code generation; evaluated reliability across difficulty bands with systematic error analysis; won first place (open division) in the national llm-jp fine-tuning competition.
- **Cross-lingual Evaluation Data Curation**: Led construction and quality control of Japanese and Chinese subsets for a 19k+ cultural-bias dataset in collaboration with Georgia Tech.
- **Diffusion LM**: Identified a train–inference mismatch that degraded diffusion-LM generation quality; improved coherence by introducing an auxiliary training objective and tailored inference strategies to reduce error propagation.

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

- **Cross-Lingual Capability Transfer for LLMs**: Diagnosed cross-lingual capability gaps in English-centric LLMs (LLaMA) for Japanese; evaluated data and training interventions via controlled translation-corpus experiments and targeted 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

Bib Reference Auto Cleaner
- Developed a BibTeX normalization tool with automated metadata retrieval and validation against OpenReview; released as a Python package. [LINK]

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 – 2026	Reviewer, ACL Rolling Review (ARR)
2026	Reviewer, International Conference on Machine Learning (ICML 2026)