MENGYU YE Final Year Ph.D. Student | Google PhD Fellow

Tohoku University, Sendai, Japan | Email: ye.mengyu.s1@dc.tohoku.ac.jp | URL: https://muyo8692.com

Research Summary

I find where LLMs break and why, through controlled experiments.

- Simple negation breaks LLM step-by-step reasoning (**EMNLP 2023**, oral).
- Sparse autoencoders are only marginally more interpretable than the feed-forward memories they were meant to improve on (**NeurIPS 2025**).
- Input attributions only partly explain in-context learning (**Findings of ACL 2025**).
- *Separately, I build at scale.* Led **Sumi**, a fully open 7B diffusion LM pretrained from scratch on 1.5T tokens, matching autoregressive baselines and released for the community.
- *In collaborations I own the diagnostic layer.* Designed the static-evaluation suite behind a **NeurIPS 2025 MMU-RAG** award, and the reliability evaluation for a first-place **llm-jp** math system.

Education

Tohoku University Sendai, Japan
Ph.D. in Information Science – Advisor: Prof. Jun Suzuki *Apr 2024 – 2027 (expected)*

ETH Zürich Zürich, Switzerland
Visiting Researcher – Host: Prof. Mrinmaya Sachan *Sept 2026 - 2027 (expected)*

Selected Awards & Grants

Winner, Best Static Evaluation Award – NeurIPS 2025 MMU-RAG Competition *2025*
Google PhD Fellowship *2025*
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. *NeurIPS 2025.* [LINK]
- **Mengyu Ye**, Tatsuki Kuribayashi, Goro Kobayashi, Jun Suzuki. Can Input Attributions Explain Inductive Reasoning in In-Context Learning?. *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. *EMNLP 2023.* (🏆 Best Paper, ACL-SRW 2023) [LINK]
- **Mengyu Ye**, Keito Kudo, Ryosuke Takahashi, Jun Suzuki. Reconsidering Positional Supervision in Masked Diffusion Language Model Training. *arXiv preprint. SPIGM@ICML2026.* [LINK]

Selected Projects

Transformer FF-KV vs. Sparse Autoencoders *(first author, NeurIPS 2025)* *2025*

- Benchmarked transformer feed-forward key-value memories against sparse autoencoders on auto-eval and human annotation.
- Showed the FF-KV are nearly as interpretable at a fraction of the compute, suggesting a clean baseline for feature-based interpretability research.

Sumi: Open 7B Diffusion Language Model *(project lead)* *2025 – Present*

- Led end-to-end pretraining of a 7B uniform diffusion LM from scratch on 1.5T tokens on 80 H100 nodes; first open model at this scale to match autoregressive baselines.
- Early analysis: although the model can revise its own tokens, it rarely does so under confidence sampling, the strategy that also drives most of its gains on generation tasks (GSM8K, HumanEval).

Positional Supervision in Masked Diffusion LM Training *(first author, SPIGM@ICML 2026)* *2025 – 2026*

- Identified a train–inference mismatch that degraded masked diffusion LM generation quality.
- Proposed an auxiliary training objective and tailored inference strategy that recovered the gap; a full diagnosis-to-fix result on LLaDA-8B-Instruct.

Math Post-training for llm-jp *(diagnostic lead)* *2024*

- Diagnosed that the base model was weak at text-based math and redirected it to solve via executable Python rather than natural-language reasoning.
- In post-training, localized the model's weak math domains and found that correct answers were failing the automatic verifier on answer format rather than correctness, a scoring artifact that masked true capability.