MENGYU YE

Ph.D. Student in Information Science | **Multilingual AI & LLM Interpretability**
Email: ye.mengyu.s1@dc.tohoku.ac.jp | URL: https://muyo8692.github.io

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

Gemma 2 Academic Research Program Grant (JP/KR 2024) *2024*
BOOST Fellowship *2024*
Best Paper Award – ACL 2023 Student Research Workshop *2023*

Publications

1. **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
2. 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]

Experience

Moonshot Research and Development Program *Sept. 2022 - Present*
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.
- Collaborated with international teams to create multilingual evaluation frameworks for LLMs, ensuring cross-lingual consistency and robust evaluation metrics.

Projects

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.

Cross-Cultural AI Safety Initiative
- 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.

Mechanistic Interpretability for Multilingual Foundation Models
- Exploring techniques to enhance reliability and explainability in multilingual AI systems, currently in the conceptual stage.