Yunzhen He

Email: heyunzhenwhat@gmail.com

Research interests: LLM Hallucinations, Sparse Autoencoder, Mechanistic Interpretability.

Education

Kyoto University

Graduate School of Informatics, Data Science Course

- **Master's Degree:** Expected Completion, 2025
- Current **Ph.D. Candidate**
- **Adviser:** Hidetoshi Shimodaira

2023 – Present

Chuo University

Faculty of Science and Engineering, Department of Information and System Engineering

- **Bachelor of Science in Information and System Engineering**

2018 – 2022

Professional Experience

Spiral.AI, *R&D Intern, LLM*

April 2024 – August 2024

- Contributed to research and development projects focusing on multi-turn dialogues from large language models (LLMs).

Kyoto University, *Research Assistant*

April 2024 – Present

- Contributed to research projects focusing on mitigating hallucination in large language models (LLMs) .

Awards and Honors

Sponsor Award (Recruit)

- Awarded at the Association for Natural Language Processing (ANLP, Japan) 2024 conference for outstanding contributions in the field of natural language processing.

DoGS Fellow Candidate

- Recognized as a candidate for the prestigious DoGS fellowship at Kyoto University.

Papers

1. 大規模言語モデルにおける幻覚緩和のための単語確率の外挿. **(The Extrapolation of Word Probabilities for Mitigating Hallucinations in Large Language Models)**
   Yunzhen He, Yusuke Takase, Yoichi Ishibashi, Hidetoshi Shimodaira. *The Association for Natural Language Processing*, 2024.
2. **Shimo Lab at "Discharge Me!": Discharge Summarization by Prompt-Driven Concatenation of Electronic Health Record Sections.**
   Yunzhen He, Hiroaki Yamagiwa and Hidetoshi Shimodaira. The 62nd Annual Meeting of the Association for Computational Linguistics. 2024

Skills

- **Programming Languages:** Python, C
- **Tools & Technologies:** PyTorch, Transformers, Transformer lens.