Daking Rai

Fairfax, Virginia, USA

✉ drai2@gmu.edu | ⌂ https://dakingrai.github.io/ | https://github.com/Dakingrai

Education

George Mason University *Fairfax, VA*

Ph.D. student in Computer Science *Aug 2021 – November 2026 (Expected)*

Research Advisor: Dr. Ziyu Yao

Tribhuvan University *Kathmandu, Nepal*

Bachelors in Computer Science and Information Techology *Aug. 2012 – Aug. 2016*

Awarded **“Outstanding Student of the Batch 2012”** Award

Publications

2025

- Daking Rai, Samuel Miller, Kevin Moran, and Ziyu Yao. **Failure by Interference: Language Models Make Balanced Parentheses Errors When Faulty Mechanisms Overshadow Sound Ones**. *NeurIPS*, 2025.
- Daking Rai, Yilun Zhou, Shi Feng, Abulhair Saparov, and Ziyu Yao. **A Practical Review of Mechanistic Interpretability for Transformer-Based Language Models**. *Arxiv pre-print*, 2025.
- Dong Shu, Xuansheng Wu, Haiyan Zhao, Daking Rai, Ziyu Yao, Ninghao Liu, Mengnan Du. **A Survey on Sparse Autoencoders: Interpreting the Internal Mechanisms of Large Language Models**. *EMNLP Findings*, 2025.
- Siddarth Mamidanna, Daking Rai, Ziyu Yao, Yilun Zhou. **All for One: LLMs Solve Mental Math at the Last Token With Information Transferred From Other Tokens**. *EMNLP*, 2025.
- Samuel Miller, Daking Rai* and Ziyu Yao. **Mechanistic Understanding of Language Models in Syntactic Code Completion**. *AAAI KnowFM Workshop*, 2025.

2024

- Daking Rai, and Ziyu Yao. **An Investigation of Neuron Activation as a Unified Lens to Explain Chain-of-Thought Eliciting Arithmetic Reasoning of LLMs**. *ACL*, 2024.
- Daking Rai, Rydia R. Weiland, Kayla Margaret Gabriella, Tyler H. Shaw, and Ziyu Yao. **Understanding the Effect of Algorithm Transparency of Model Explanations in Text-to-SQL Semantic Parsing**. *Arxiv pre-print*, 2024.

2023

- Daking Rai, Bailin Wang, Yilun Zhou and Ziyu Yao. **Improving Generalization in Language Model-based Text-to-SQL Semantic Parsing: Two Simple Semantic Boundary-Based Techniques**. *ACL*, 2023.
- Daking Rai, Yilun Zhou, Bailin Wang and Ziyu Yao. **Explaining Large Language Model-Based Neural Semantic Parsers** . *AAAI Student Abstract and Poster Program*, 2023.

Research & Work Experience

Graduate Research Assistant *Fairfax, VA, USA*

NLP Lab, George Mason University *May 2022 - Present*

- Advisor: Dr. Ziyu Yao
- My research focuses on the **interpretability of language models**, spanning behavioral interpretability (explaining behavior through input–output analysis) and mechanistic interpretability (understanding model internals). I am also equally interested in applying these insights to practical domains such as output steering, performance improvement, and AI safety.
- We recently conducted Tutorial on Mechanistic Interpretability for Language Models, **ICML 2025**.
- We recently organized The First Workshop on the Application of LLM Explainability to Reasoning and Planning @ **COLM 2025**

Graduate Teaching Assistant

Fairfax, VA, USA

George Mason University

Aug 2021 - April 2022

- Teaching Assistant for Computer Systems and System Programming (CS531) & Essentials of Computer Science (CS110).

Machine Learning Engineer

Kathmandu, Nepal

Infodev Pvt Ltd

April 2019 - June 2020

- Led a research and development team for the integration of semantic search enhanced by named entity recognition (NER) within an e-commerce platform. Additionally, worked on projects involving face recognition, facial aliveness detection, object detection, and localization for various prototypes.

Services

- Organized The First Workshop on the Application of LLM Explainability to Reasoning and Planning @ **COLM 2025**
- Conducted Tutorial on **ICML 2025** — Tutorial on Mechanistic Interpretability for Language Models
- Served as a reviewer for **NeurIPS'25 MI Workshop, AAAI'25, ARR May'25, ARR June'24, AAAI'24**.
- Served as a secondary reviewer for **NeurIPS'23, NeurIPS'25**.
- Invited talk on the topic of Mechanistic Interpretability of Language Models at The George Washington University.
- Volunteer for "The 10th annual Mid-Atlantic Student Colloquium on Speech, Language and Learning".
- Volunteer for "The third ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO '23)".
- Executive committee member (volunteer) for Rose Foundation Nepal, an NGO working to raise awareness about different types of cancers, especially breast cancer in Nepal.
- Organizer and mentor on 10+ machine learning and deep learning workshops conducted by AI Developer Nepal, an AI community based in Kathmandu.

Awards

- NeurIPS 2025 Scholar Award
- Graduate Student Travel Fund (GSTF) for ICML'25, ACL'24 and ACL'23.
- Honorary mention, GMU CS Research Symposium (Poster Presentation).
- "Outstanding Student of the Batch 2012" Award. Bachelors in Computer Science and Information Technology, Tribhuvan University (TU), Kathmandu.

References

Ziyu Yao (ziyuyao@gmu.edu)

Assistant Professor
Dept. of Computer Science (CS)
George Mason University
4400 University Dr, Fairfax, VA 22030