

Zirui He

East Orange, NJ, USA

zh296@njit.edu | (201) 539-1663 | Personal Website | Google Scholar

RESEARCH INTERESTS

My research focuses on **LLM Interpretability** and **Representation Space Analysis**. I aim to move LLMs from being usable to being understandable, diagnosable, and controllable, and my work explores this potential by improving transparency and behavior alignment in multiple real world tasks.

EDUCATION

New Jersey Institute of Technology (NJIT)

Newark, NJ, USA

Ph.D. in Data Science

Sep. 2024 – Present

Stevens Institute of Technology

Hoboken, NJ, USA

M.S. in Computer Science

Sep. 2021 – Jun. 2023

China University of Geosciences, Wuhan

Wuhan, China

B.S. in Computer Science

Sep. 2017 – Jun. 2021

RESEARCH EXPERIENCE

New Jersey Institute of Technology (NJIT)

Newark, NJ, USA

Ph.D. Candidate, Data Science (Advisor: Mengnan Du)

Oct. 2025 – Present

- Developed activation-space methods for LLM interpretability, including memorization analysis, activation inversion, and cross-model representation explanation.
- Developed inference time intervention(ITI) methods for LLM behavior calibration via LM-head editing and logit space steering.
- Built domain-specific LLM systems for financial NLP, RAG, and evidence-grounded reasoning.

New Jersey Institute of Technology (NJIT)

Newark, NJ, USA

Ph.D. Student, Data Science (Advisor: Mengnan Du)

Sep. 2024 – Oct. 2025

- Developed algorithms to identify and mitigate shortcut learning in language models, including extensions to reward models.
- Applied sparse autoencoders to analyze instruction-following behavior and steer controllable open-ended generation in LLMs.

PUBLICATIONS

Reliable and Controllable LLMs

1. **SAE-SSV: Supervised Steering in Sparse Representation Spaces for Reliable Control of Language Models**

Zirui He, Mingyu Jin, Bo Shen, Ali Payani, Yongfeng Zhang, Mengnan Du

Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 2025

2. **SAIF: A Sparse Autoencoder Framework for Interpreting and Steering Instruction Following of Language Models**

Zirui He, Haiyan Zhao, Yiran Qiao, Fan Yang, Ali Payani, Jing Ma, Mengnan Du

European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML-PKDD), 2026

3. **HeadEdit: Calibrating Language Model Behavior Through the Frozen Unembedding Matrix**
Zirui He, Haiyan Zhao, Jingyu Hu, Yinghao Wu, Chenxi Yuan, Yingcong Li, Yandong Bai, Mengnan Du
 Under review at ICLR, 2027
4. **DynaCF: Mitigating Shortcut Learning in Reward Models via Dynamic Counterfactual Sensitivity**
 Fengyuan Liu, Yongliang Miao, **Zirui He**, Yanguang Liu, Fei Sun, Mengnan Du
 Under review at AAAI Phase 2, 2026
5. **Mitigating Shortcuts in Language Models with Soft Label Encoding**
Zirui He, Huiqi Deng, Haiyan Zhao, Ninghao Liu, Mengnan Du
Proceedings of the 31st International Conference on Computational Linguistics (COLING), 2024

Mechanistic Interpretability and Activation-Space Analysis

1. **LogitTrace: Detecting Benchmark Contamination via Layerwise Logit Trajectories**
Zirui He, Haiyan Zhao, Yingcong Li, Ali Payani, Mengnan Du
 Under review at NAACL, 2026
2. **Rep2Text: Decoding Full Text from a Single LLM Token Representation**
 Haiyan Zhao, **Zirui He**, Yingming Tang, Fan Yang, Ali Payani, Dianbo Liu, Mengnan Du
 Advances in Neural Information Processing Systems (NeurIPS), 2026
3. **Universal Activation Verbalizer: A Unified Framework for Cross-Model Activation Explanation**
 Haiyan Zhao, **Zirui He**, Guanchu Wang, Ali Payani, Yingcong Li, Mengnan Du
Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 2026
4. **Faithful Activation Verbalization: Reducing Hallucinations in LLM Representation Interpretation**
 Haiyan Zhao, **Zirui He**, Wei Shi, Huiqi Deng, Na Zou, Mengnan Du
 Under review at ICLR, 2027

Knowledge-Grounded and Domain-Specific Reasoning

1. **SAE-FiRE: Enhancing Earnings Surprise Predictions Through Sparse Autoencoder Feature Selection**
 Huopu Zhang, Yanguang Liu, Miao Zhang, **Zirui He**, Mengnan Du
Findings of the Association for Computational Linguistics (ACL), 2026
2. **FinAnchor: Anchor-Based Representation Alignment for Financial Prediction**
Zirui He, Huopu Zhang, Yanguang Liu, Yingcong Li, Mengnan Du
 Under review at ACL ARR Oct, 2026
3. **HyperPathsRAG: Multi-Target Path Approximation over Hypergraph Knowledge Graphs for Multi-Hop QA**
 Guanhua Zhu, **Zirui He**, Mengnan Du, Chenxi Yuan
Findings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 2026
4. **RealRoute: Dynamic Query Routing System via Retrieve-then-Verify Paradigm**
 Jiahe Liu, Qinkai Yu, Jinmen Zhao, Xi Zhu, **Zirui He**, Zhen Xiang, Fan Yang, Jingcheng Niu
 Under review at ACL ARR Oct, 2026

TECHNICAL SKILLS

- **ML / LLM:** Python, PyTorch, CUDA, Transformers, RAG, financial NLP, LLM evaluation
- **Interpretability:** Sparse Autoencoders (SAEs), LLM steering, probing, activation inversion
- **Post-training / Alignment:** Supervised fine-tuning (SFT), LoRA-based instruction tuning, reward modeling, inference-time intervention, LLM behavior calibration, knowledge distillation

TEACHING EXPERIENCE

- Teaching Assistant, DS677: Deep Learning, NJIT Spring 2026
- Teaching Assistant, DS675: Machine Learning, NJIT Fall 2025
- Teaching Assistant, DS680: Natural Language Processing, NJIT Spring 2025

ACADEMIC SERVICE

Conference Reviewer

- ICML (International Conference on Machine Learning), 2026
- NeurIPS (Conference on Neural Information Processing Systems), 2026
- ICLR (International Conference on Learning Representations), 2027
- EMNLP (Conference on Empirical Methods in Natural Language Processing), 2026
- AISTATS (International Conference on Artificial Intelligence and Statistics), 2026
- ECML-PKDD (European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases), 2026

Journal Reviewer

- Applied AI Letters (AIL), 2026
- Transactions on Machine Learning Research (TMLR), 2026