Haiyan Zhao

hz54@njit.edu | hy-zhao23.github.io | Google Scholar | GitHub

Research

My research broadly focuses on large language models, with interests in post-training, alignment, interpretability, and trustworthy AI. I develop methods for understanding, adapting, and improving LLM behavior through representation analysis, concept-subspace modeling, activation-based explanation, and sparse-autoencoder-based steering. More broadly, I am interested in building reliable and controllable AI systems, including LLM agents and post-trained models for real-world applications.

Interests: Trustworthy AI · LLMs · Post-training · Alignment

Education

2023 – Present	**Ph.D.** in Computer Science, New Jersey Institute of Technology Advisor: Dr. Mengnan Du
2017 – 2019	**M.E.** in Control Engineering, Huazhong University of Science and Technology Advisor: Dr. Xiaolong Shi
2013 – 2017	**B.E.** in Automation, Wuhan University of Technology

Publications

Publications: 17+ papers, including 9 first/co-first-author papers.
Selected venues: ACL, EMNLP, ICLR, COLING, EACL, ECML-PKDD, ACM TKDD.
$*$ denotes equal contribution. Full list on Google Scholar.

H. Zhao, Z. He, G. Wang, A. Payani, Y. Li, and M. Du. Universal Activation Verbalizer: A Unified Framework for Cross-Model Activation Explanation. *EMNLP*, 2026. *Explain activations in natural lanuage.*

J. He, **H. Zhao**, R. Shi, Y. Liu, X. Wang, F. Sun, and M. Du. SAEExplainer: Interpreting SAE Features with Activation-Guided Preference Optimization. *EMNLP*, 2026. *Refines SAE-feature explanations via iterative DPO.*

H. Zhao, X. Wu, F. Yang, B. Shen, N. Liu, and M. Du. Denoising Concept Vectors with Sparse Autoencoders for Improved Language Model Steering. *EACL Findings*, 2026. *Uses SAEs to denoise concept vectors.*

Z. He*, **H. Zhao***, Y. Qiao, F. Yang, A. Payani, J. Ma, and M. Du. Saif: A sparse autoencoder framework for interpreting and steering instruction following of language

models. *ECML PKDD*, 2026

X. Wu*, **H. Zhao***, Y. Zhu*, Y. Shi*, F. Yang, T. Liu, X. Zhai, W. Yao, J. Li, M. Du, et al. Usable XAI: 10 strategies towards exploiting explainability in the LLM era. *TKDD*, 2026

H. Zhao, H. Zhao, B. Shen, A. Payani, F. Yang, and M. Du. Beyond Single Concept Vector: Modeling Concept Subspace in LLMs with Gaussian Distribution. *ICLR*, 2025. *Models a concept as a Gaussian subspace to describe variance.*

H. Zhao, F. Yang, S. Bo, H. Lakkaraju, and M. Du. Towards Uncovering How Large Language Model Works: An Explainability Perspective. *SIGKDD Explorations*, 2025

D. Shu, X. Wu, **H. Zhao**, M. Du, and N. Liu. Beyond Input Activations: Identifying Influential Latents by Gradient Sparse Autoencoders. *EMNLP*, 2025

M. Jin, Q. Yu, J. Huang, Q. Zeng, Z. Wang, W. Hua, **H. Zhao**, K. Mei, Y. Meng, K. Ding, et al. Exploring Concept Depth: How Large Language Models Acquire Knowledge at Different Layers? *COLING*, 2025

S. Liu, Z. Li, R. Ma, H. Zhao, and M. Du. Contracteval: Benchmarking llms for clause-level legal risk identification in commercial contracts. *EMNLP Workshop on NLLP*, 2025. **Best Presentation Award (Top 1 out of 40 presentations)**

D. Li, M. Jin, Q. Zeng, **H. Zhao**, and M. Du. Exploring Multilingual Probing in Large Language Models: A Cross-Language Analysis. *XLLM Workshop*, 2025

H. Zhao, H. Chen, F. Yang, N. Liu, H. Deng, H. Cai, S. Wang, D. Yin, and M. Du. Explainability for large language models: A survey. *ACM TIST*, 2024

M. Jin, Q. Yu, **H. Zhao**, W. Hua, Y. Meng, Y. Zhang, M. Du, et al. The impact of reasoning step length on large language models. *ACL Findings*, 2024

Preprints

H. Zhao, Z. He, F. Yang, A. Payani, and M. Du. Rep2Text: Decoding Full Text from a Single LLM Token Representation. *Under review*, 2025. *Reinverse activation into input text using adapters.*

Z. He, **H. Zhao**, A. Payani, and M. Du. MemLens: Uncovering Memorization in LLMs with Activation Trajectories. *Under review*, 2025. *Detects memorization from layer-wise activation trajectories.*

Experience

May – Aug 2026 — Research Scientist Intern, Cloudflare
Tool-calling algorithms for LLM agents with dynamic activation steering; extending interpretability methods to agent behavior control.

Feb 2022 – Feb 2023	Research Intern, Deakin University *Error-Aware Probabilistic Logical Reasoning Model for Entity Alignment*.
Sep 2019 – Dec 2021	Technical Specialist, Tianjin Weiman Information Technology Co. *Design LaTeXtemplate for academic journals.*

Talks & Presentations

June 2026	Towards Natural Language Explanation of LLM Internal Representations @MLNLP
April 2024	Explaining training samples and enhancing trustworthiness with LLM @IEEE ICHI
Oct 2023	Explainability on LLMs @MLLM-AI
Oct 2023	Explainability on LLMs @UberAI

Professional Service

Conference Reviewer
ACL, EMNLP, ICML, ICLR, NeurIPS, AAAI, AISTATS, ICHI

Journal Reviewer
TIST, Web of Science, Scientific Report, and others.

Teaching

2023 - Now	Teaching Assistant, DS680 Natural Language Processing, NJIT
Fall 2024/2025	Teaching Assistant, DS675 Machine Learning, NJIT
Spring 2025	DS636 Data Analytics with R Program, NJIT

Technical Skills

ML / LLM: PyTorch, HuggingFace Transformers & PEFT, vLLM, DeepSpeed / FSDP
Interpretability: Linear probing, Logit / Tuned Lens, Patchscopes, sparse autoencoders, activation patching, representation steering
Post-training: Supervised fine-tuning, knowledge distillation, DPO

Honors & Awards

2025	**Best Presentation Award** (NNLP 2025)

2018 – 2019	**National Scholarship**(Ministry of Education of P.R. China); **Outstanding Graduate, Merit Postgraduate**(HUST)
2017 – 2018	**First-class academic scholarships for postgraduates**(HUST)
2016 – 2017	**Outstanding Graduate Thesis**(Hubei)
2015 – 2016	**First-class University Scholarship**(WUT)
2014 – 2015	**National Scholarship**(Ministry of Education of P.R. China)
2013 – 2014	**Third-class University Scholarship**(WUT)