Zining Zhu

✉ zzhu41@stevens.edu   🐦 @zhuzining   🌐 https://ecailab.org

Employment

2024 - present   **Assistant Professor,** Stevens Institute of Technology, Hoboken, New Jersey, US

Employment (Internship)

June 2022 - Sep 2022   **Applied Scientist Intern,** Amazon Search, Palo Alto, California
Advisor: Haoming Jiang, Applied Research Scientist.

May 2019 - Aug 2019   **Research Intern,** Tencent Jarvis Lab, Shenzhen, Guangdong, China.
Advisor: Zachary Zhao, Senior Researcher.

Sep 2017 - Aug 2018   **Software Engineering Intern,** Winterlight Labs, Toronto, Ontario, Canada.
Advisor: Jekaterina Novikova, Director of Machine Learning.
Source of research funding: Winterlight Labs.

June 2017 - Aug 2017   **Software Engineering Intern,** TripAdvisor Inc., Needham, Massachussetts.
Advisor: Raksik Kim, Software Engineer.

May 2016 - Aug 2016   **Research Assistant,** Dynamic Systems Lab, Toronto, Ontario.
Advisor: Angela Schoellig, Assistant Professor.
Source of research funding: Engineering Science (ESROP) at University of Toronto

Education

2019 - 2024   **PhD, University of Toronto** Computer Science.
Advisor: Frank Rudzicz, Associate Professor at University of Toronto. Now Associate Professor at Dalhousie University.
Thesis: *Methods and Applications for Probing Deep Neural Networks*.

2014 - 2019   **BASc., University of Toronto** Engineering Science, Robotics Option.

Grants & Awards

Grants

- *NAIRR: Generating Feasible Follow-up Research Questions*, National Artificial Intelligence Research Resource Pilot Program, National, ≈$15,000 (in forms of credits and node hours), 2025
- *NAIRR: Accelerating Neural Network Circuits with Software-Hardware Co-Design*, National Artificial Intelligence Research Resource Pilot Program, National, ≈$15,000 (in forms of computation credits and node hours), 2025
- *Cloud Computing Program Grant*, Google, Institutional, $5,000, 2025
- *Toward Safer Mixture-of-Expert Models*, Stevens Institute for Artificial Intelligence, Institutional, $5,000, 2025
- *An LLM-enhanced Framework for Targeted Practices*, Teaching and Learning Center at Stevens, Institutional, $5,000, 2025
- *Vector Institute PhD Research Grant*, Institutional, $6,000 each, 2020-2024

Awards

- Outstanding paper award, NAACL. 2025
- Top Reviewer Award, NeurIPS. 2023
- Ontario Graduate Scholarship, Provincial, $15,000. 2022-2023

Publications

Refereed Conference Proceedings

1. Wang H, **Zhu Z**, and Shi HF. Distribution Prompting: Understanding the Expressivity of Language Models Through the Next-Token Distributions They Can Produce. *EMNLP*. 2025

2. Lei Y, Niu J, **Zhu Z**, Chen X, and Penn G. Dynamic Granularity in the Wild: Differentiable Sheaf Discovery with Joint Computation Graph Pruning. *EMNLP*. 2025

3. Li H, Cao Y, Yu Y, Javaji SR, Deng Z, He Y, Jiang Y, **Zhu Z**, Subbalakshmi K, Xiong G, Huang J, Qian L, Peng X, Xie Q, and Suchow JW. INVESTORBENCH: A Benchmark for Financial Decision-Making Tasks with LLM-based Agent. *ACL*. 2025

4. Roewer-Despres F, Feng J, **Zhu Z**, and Rudzicz F. ACCORD: Closing the Commonsense Measurability Gap. *NAACL*. Outstanding paper award. 2025

5. **Zhu Z**, Chen H, Ye X, Lyu VQ, Marasović A, Tan C, and Wiegreffe S. Explanation in the Era of Large Language Models. *NAACL Tutorial Abstracts*. 2024

6. Jingcheng N, Wang A, **Zhu Z**, and Penn G. What does the Knowledge Neuron Thesis Have to do with Knowledge? *ICLR*. Spotlight (Top 5%). 2024

7. Sahak E, **Zhu Z**, and Rudzicz F. A State-Vector Framework for Dataset Effects. *EMNLP*. 2023

8. **Zhu Z**, Shahtalebi S, and Rudzicz F. Predicting fine-tuning performance with probing. *EMNLP*. 2022

9. **Zhu Z**, Wang J, Li B, and Rudzicz F. On the data requirements of probing. *Findings of ACL*. 2022

10. Li B, **Zhu Z**, Thomas G, Rudzicz F, and Xu Y. Neural reality of argument structure constructions. *ACL*. 2022

11. Ramezani A, **Zhu Z**, Rudzicz F, and Xu Y. An unsupervised framework for tracing textual sources of moral change. *Findings of EMNLP*. 2021

12. Li B, **Zhu Z**, Thomas G, Xu Y, and Rudzicz F. How is BERT surprised? Layerwise detection of linguistic anomalies. *ACL-IJCNLP*. 2021

13. **Zhu Z** and Rudzicz F. An information-theoretic view on selecting linguistic probes. *EMNLP*. 2020

14. **Zhu Z**, Novikova J, and Rudzicz F. Detecting cognitive impairments by agreeing on interpretations of linguistic features. *NAACL*. 2019

15. Li Q, Qian J, **Zhu Z**, Bao X, Helwa M, and Schoellig A. Deep neural networks for improved, impromptu trajectory tracking of quadrotors. *ICRA*. 2017

Refereed Workshop Proceedings

16. Li H, Cao Y, Yu Y, Suchow JW, and **Zhu Z**. Truth Neurons. *ACL Knowlegable Foundational Models Workshop*. 2025

17. Javaji SR and **Zhu Z**. What Would You Ask When You First Saw $a^2 + b^2 = c^2$? Evaluating LLM on Curiosity-Driven Questioning. *AAAI workshop on AI4Science*. 2025

18. Jaipersaud B, **Zhu Z**, Rudzicz F, and Creager E. Show, Don't Tell: Uncovering Implicit Character Portrayal using LLMs. *NeurIPS Workshop on Creativity and GenAI*. 2024

19. Ajwani RD, Javaji SR, Rudzicz F, and **Zhu Z**. LLM-Generated Explanations May be Adversarially Helpful. *NeurIPS Regulatable ML Workshop*. Featured in Stevens School of Engineering & Science Newsletter. 2024

20. Ajwani RD, **Zhu Z**, Rose J, and Rudzicz F. Plug and Play with Prompts: A Prompt Tuning Approach for Controlling Text Generation. *AAAI ReLM*. 2024

21. **Zhu Z**, Jiang H, Yang J, Nag S, Zhang C, Jie H, Gao Y, Rudzicz F, and Yin B. Situated Natural Languages Explanations. *ACL NLRSE*. 2023

22. **Zhu Z**, Shahtalebi S, and Rudzicz F. OOD-Probe: A Neural Explanation of Out-of-Domain Generalizations. *ICML SCIS Workshop*. 2022

23. Shahtalebi S, **Zhu Z**, and Rudzicz F. Out-of-Distribution Failure through the Lens of Labeling Mechanisms. *ICML SCIS Workshop*. 2022

24. **Zhu Z**, Pan C, Abdalla M, and Rudzicz F. Examining the rhetorical capacities of neural language models. *EMNLP BlackboxNLP Workshop*. 2020

25. Hsu YT, **Zhu Z**, Wang CT, Fang SH, Rudzicz F, and Tsao Y. Robustness against the channel effect in pathological voice detection. *NeurIPS ML4H Workshop*. 2018

26. **Zhu Z**, Novikova J, and Rudzicz F. Semi-supervised classification by reaching consensus among modalities. *NeurIPS IRASL Workshop*. 2018

Preprints

27. Mi Z, Tan Q, Yu X, **Zhu Z**, Yuan G, and Huang S. KerZOO: Kernel Function Informed Zeroth-Order Optimization for Accurate and Accelerated LLM Fine-Tuning. 2025

28. Javaji SR, Cao Y, Li H, Yu Y, Muralidhar N, and **Zhu Z**. Can AI Validate Science? Benchmarking LLMs for Scientific Claim → Evidence Reasoning. 2025

29. Doda S, Javaji SR, and **Zhu Z**. VERBA: Verbalizing Model Differences Using Large Language Models. 2025

30. Vinden N, Saqur R, **Zhu Z**, and Rudzicz F. Contrastive Similarity Learning for Market Forecasting: The ContraSim Framework. 2025

31. Cao Y, Li H, Yu Y, Javaji SR, He Y, Huang J, **Zhu Z**, Xie Q, Liu Xy, Subbalakshmi K, et al. FinAudio: A Benchmark for Audio Large Language Models in Financial Applications. 2025

32. Qiu P, Rudzicz F, and **Zhu Z**. Scenarios and Approaches for Situated Natural Language Explanations. 2024

33. **Zhu Z** and Rudzicz F. Measuring Information in Text Explanations. 2023

34. Huang J, Gao Y, Li Z, Yang J, Song Y, Zhang C, **Zhu Z**, Jiang H, Yin B, and Chang KCC. CCGen: Explainable Complementary Concept Generation in E-Commerce. 2023

35. **Zhu Z**, Balagopalan A, Ghassemi M, and Rudzicz F. Quantifying the Task-Specific Information in Text-Based Classifications. 2021

36. **Zhu Z**, Li B, Xu Y, and Rudzicz F. What do writing features tell us about AI papers? 2021

37. **Zhu Z**, Xu Y, and Rudzicz F. Semantic coordinates analysis reveal language changes in AI research. 2020

38. **Zhu Z**, Novikova J, and Rudzicz F. Deconfounding age effects with fair representation learning when assessing dementia. Featured in Medical Xpress. 2019

Press coverage

- Stevens News: Zining Zhu Helps Developers and Users Navigate 'Adversarial Helpfulness' in Artificial Intelligence (December 18, 2024)
- Medical Xpress: A new machine learning model to isolate the effects of age in predicting dementia (July 27, 2018)

Teaching

Instructor

Courses at Stevens Institute of Technology:

- CS 810A Topics in Explainable Natural Language Processing (2025 spring). New course.
- CS 584C Natural Language Processing (2024 fall, 2025 fall).

Courses at University of Toronto:

- CSC401 / 2511 Natural Language Computing (2023 winter). Co-instructing with En-Shiun Lee and Raeid Saqur
- CSC401 / 2511 Natural Language Computing (2022 winter). Co-instructing with Frank Rudzicz and Raeid Saqur

Teaching Assistant (at University of Toronto)

- CSC108 Introduction to Computer Programming (2023 summer)
- ECE1786 Creative Applications for NLP (course prep TA in 2022 summer and TA in 2022 fall)
- CSC2515 Introduction to Machine Learning (2021 fall)
- CSCC24 Principles of Programming Languages (2021 summer)
- CSC148 Introduction to Computer Science (2021 summer)
- CSC401/2511 Natural Language Computing (2021 winter)
- CSC309 Web Programming (2020 fall)
- CSC401/2511 Natural Language Computing (2020 winter)
- ECE324 Introduction to Machine Intelligence (2019 fall)
- CSC180 Introduction to Computer Programming (2016 fall)

Seminars

- Interpretable NLP seminar at UofT CompLing (2021 winter)
- Introduction to ML seminar at UTADA (2017 fall)

Advising

Current:

- Weijin Liu – 2025-present – PhD at Stevens
- Ziruo Zhao – 2025-present – PhD at Stevens
- Shashidhar Reddy Javaji – 2024-present – PhD at Stevens
- Preet Jhanglani – 2024-present – PhD at Stevens
- Haohang Li – 2024-present – PhD at Stevens (co-advised with Jordan Suchow).
- Molly DiCampli – 2025 Undergraduate research project (Pinnacle Scholar program) Project: Sports performance data analysis.
- Yitao Xu – 2025 – Master research project. Topic: Expert preferences in Mixture-of-expert (MOE) language models.
- Dev Arpan Desai – 2025 – Master research project (SIAI AIRS program). Topic: Logit lens and steering vectors on multi-GPU LLMs.
- Onkar Mahamuni – 2025 – Master research project (TLC grant project)
- Xingyue Qiu – 2025 – Master research project. Topic: Steering vector model steering.
- Muhammad Shahzaib Chaudhury – 2025 – Master research project. Topic: Sparse Autoencoders model steering.

Alumni:

- Shravan Doda – 2024-present – Master research project. Topic: Natural language explanation of model differences.
- Samaksh Bhargav – 2024 – Edison Academy Magnet School. Topic: Sparse autoencoder (SAE) model steering.
- Rohit Sandadi – 2024 – Dougherty Valley High School. Topic: Natural language explanation of neurons.
- Patrick Wierzbicki – 2024 – Undergraduate research project. Topic: Retrieval language model.
- Paul Gao – 2024 – Undergraduate research project. Topic: Graph-based LLM reasoning.
- Pengshuo Qiu – 2024 – Undergraduate research project: Situated natural language explanations.
- Sean Wang – 2024 – Undergraduate research project. Topic: Analyzing the distribution in next-token prediction.
- Sam Pan – 2024 – Undergraduate research project. Topic: Use steering vector to control model honesty.
- Huaizhi Ge – 2023 – Master research project. Topic: mechanistic interpretability of language models.

- Sanika Mhadgut – 2024 – Master research project. Topic: neural network circuits.
- Jim Yang – 2023 – Summer research project: Natural language explanation
- Jason Zuo – 2023 – Summer research project: Probing and explanation
- Rohan Deepak Ajwani – 2022-2023 – Summer research project & ECE MEng projects. Topic: adversarial helpfulness explanations, controllable generation.
- Philipp Eibl – 2021 – Undergraduate research project: Information estimators
- Esmat Sahak – 2021 – Undergraduate research project: Multitask learning and probing.

Services

Organizing Workshops, Tutorials, etc.

- Explanation in the Era of Large Language Models, tutorial at NAACL 2024
- Machine Learning for Cognitive Mental Health (ML4CMH), workshop at AAAI 2024
- ACL Student Research Seminar at UofT Computational Linguistics, 2023

Area Chair or Action Editor

- 2025: ACL Rolling Review (NAACL), ICML, NeurIPS
- 2024: ACL Rolling Review (EMNLP), NeurIPS

Reviewing

- 2025: ICLR, AAAI, ACL Rolling Review, COLM, LLMSEC@ACL
- 2024: ICLR, ICML, ACL Rolling Review, COLM, AAAI
- 2023: ICLR, ICML, IJCAI, ACL Rolling Review, FAccT, NeurIPS, EMNLP, R2HCAI@AAAI
- 2022: ACL Rolling Review, EMNLP, NeurIPS, RobustSeq@NeurIPS, LT-EDI@ACL, CMCL@ACL
- 2021: ACL, EMNLP, NAACL, AAAI
- 2020: ACL, IEEE Journal of Biomedical and Health Informatics
- 2018: Computer Methods & Programs in Biomedicine

Selected Talks

- *Artificial Intelligence for Teaching and Learning*, panelist at Stevens Teaching and Learning Center Symposium. May 15, 2025
- *Mechanistic Interpretability and AI Safety*, USC CSCI699 guest lecture, Online. March 31, 2025.
- *Chatting with GPT about a PDF*, Stevens CS188 guest lecture, March 26, 2025.
- *Mechanistic Interpretability and AI Safety*, Rutgers University New Brunswick, New Jersey. October 15, 2024.
- *A Communication Channel of Explanations*, AI camp event, Toronto. February 7, 2024. Vector Institute NLP Workshop, Toronto. February 16, 2024.
- *Challenge and Opportunities in AI Safety and Alignment*, TGO Virtual Panel Talk. Nov 11, 2023.
- *Towards Interpretable and Controllable AI*, Stevens Institute of Technology (May 18, 2023), UW-Madison iSchool, (February 16, 2023)
- *Torwards Interpretable and Controllable AI*, NEC Laboratories Europe, February 2, 2023
- *Interpretable and Controllable Pretrained Language Models*, UT Computational Linguistics talk, Nov 15, 2022
- *Incorporating probing in the development of large language models*, Vector Institute Endless Summer School (ESS) invited talk, March 1, 2022
- *On the data requirements of probing*, Vector Institute Research Symposium, Virtual poster presentation, Feb 22, 2022
- *Predicting fine-tuning performance with probes*, UT Computational Linguistics, virtual presentation, Feb 15, 2022
- *Quantifying the task-specific information in text-based classifications*, UT Language Research Day, Virtual presentation, Nov 12, 2021
- *Probing neural language models*, AISC Recent Trends in NLP discussion, Video talk, Aug 15, 2021
- *Writing can predict AI papers acceptance, but not their impact*, Vector Institute Research Symposium, Virtual poster presentation, Feb 16, 2020
- *Improving the neural NLP model performances with linguistic probes*, Zhi-Yi NLP Open Course, Video talk, Nov 20, 2020

- *An information-theoretic view on selecting linguistic probes*, TsingHua University AI TIME, Video talk, Oct 30, 2020
- *Examining the rhetorical capacities of neural language models*, Vector Institute NLP Symposium spotlight presentation, Video talk, Sep 16, 2020.
- *Speeding up computation with GPU and Google Cloud*, ECE324 tutorial, Toronto, Canada, Oct 31, 2019
- *Efficient pre-training methods for language modeling*, Tencent Jarvis Lab, Shenzhen, China, Aug 5, 2019
- *Automatic assessment of cognitive impairments*, UTMIST tech talk, Toronto, Canada, Nov 20, 2018
- *Probabilistic graphical models*, UTADA tech talk, Toronto, Canada, Oct 21, 2017

Affiliation

Association of Computational Linguistics (ACL), International Electrical and Electronics Engineers (IEEE)

Last updated: August, 2025