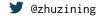
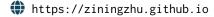
Zining Zhu







Employment (Academia)

2024 - present

Assistant Professor, Stevens Institute of Technology, Hoboken, New Jersey, US

Employment (Internship)

June 2022 - Sep	2022 App	lied Scientist Inte	rn, Amazon S	Search, Palo Alto	o, California

Advisor: Haoming Jiang, Applied Research Scientist.

Source of research funding: Amazon.

May 2019 - Aug 2019 **Research Intern,** Tencent Jarvis Lab, Shenzhen, Guangdong, China.

Advisor: Zachary Zhao, Senior Researcher. Source of research funding: Tencent.

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 Profes-

sor at Dalhousie University.

Thesis: Methods and Applications for Probing Deep Neural Networks.

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

Publications

Refereed Conference Proceedings

- 1. **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
- 2. Jingcheng N, Wang A, **Zhu Z**, and Penn G. What does the Knowledge Neuron Thesis Have to do with Knowledge? *ICLR*. 2024
- 3. Sahak E, **Zhu Z**, and Rudzicz F. A State-Vector Framework for Dataset Effects. *EMNLP*. 2023
- 4. **Zhu Z**, Shahtalebi S, and Rudzicz F. Predicting fine-tuning performance with probing. *EMNLP*. 2022
- 5. **Zhu Z**, Wang J, Li B, and Rudzicz F. On the data requirements of probing. *Findings of ACL*. 2022
- 6. Li B, **Zhu Z**, Thomas G, Rudzicz F, and Xu Y. Neural reality of argument structure constructions. *ACL*. 2022

- 7. Ramezani A, **Zhu Z**, Rudzicz F, and Xu Y. An unsupervised framework for tracing textual sources of moral change. *Findings of EMNLP*. 2021
- 8. Li B, **Zhu Z**, Thomas G, Xu Y, and Rudzicz F. How is BERT surprised? Layerwise detection of linguistic anomalies. *ACL-IJCNLP*. 2021
- 9. **Zhu Z** and Rudzicz F. An information-theoretic view on selecting linguistic probes. *EMNLP*. 2020
- 10. **Zhu Z**, Novikova J, and Rudzicz F. Detecting cognitive impairments by agreeing on interpretations of linguistic features. *NAACL*. 2019
- 11. 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

- 12. 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
- 13. **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
- 14. **Zhu Z**, Shahtalebi S, and Rudzicz F. OOD-Probe: A Neural Explanation of Out-of-Domain Generalizations. *ICML SCIS Workshop.* 2022
- 15. Shahtalebi S, **Zhu Z**, and Rudzicz F. Out-of-Distribution Failure through the Lens of Labeling Mechanisms. *ICML SCIS Workshop.* 2022
- 16. **Zhu Z**, Pan C, Abdalla M, and Rudzicz F. Examining the rhetorical capacities of neural language models. *EMNLP BlackboxNLP Workshop*. 2020
- 17. 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
- 18. **Zhu Z**, Novikova J, and Rudzicz F. Semi-supervised classification by reaching consensus among modalities. *NeurIPS IRASL Workshop*. 2018

Preprints

- 19. Qiu P, Rudzicz F, and **Zhu Z**. Scenarios and Approaches for Situated Natural Language Explanations. 2024
- 20. Ge H, Rudzicz F, and Zhu Z. What Do Circuits Mean? A Model Edit View. 2024
- 21. Lei Y, Niu J, Zhu Z, and Penn G. Circuit discovery by differentiable masks. 2024
- 22. Roewer-Despres F, Feng J, **Zhu Z**, and Rudzicz F. ACCORD: Closing the Commonsense Measurability Gap. 2024
- 23. Ge H, Rudzicz F, and **Zhu Z**. How Well Can Knowledge Edit Methods Edit Perplexing Knowledge?
- 24. Ajwani R, Javaji SR, Rudzicz F, and **Zhu Z**. LLM-Generated Explanations May be Adversarially Helpful. 2024
- 25. **Zhu Z** and Rudzicz F. Measuring Information in Text Explanations. 2023
- 26. 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
- 27. **Zhu Z**, Balagopalan A, Ghassemi M, and Rudzicz F. Quantifying the Task-Specific Information in Text-Based Classifications. 2021

- 28. **Zhu Z**, Li B, Xu Y, and Rudzicz F. What do writing features tell us about AI papers? 2021
- 29. **Zhu Z**, Xu Y, and Rudzicz F. Semantic coordinates analysis reveal language changes in AI research. 2020
- 30. **Zhu Z**, Novikova J, and Rudzicz F. Deconfounding age effects with fair representation learning when assessing dementia. 2019

Press coverage

• 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 584 Natural Language Processing (2024 fall, Thursday section). *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:

- Shashidhar Reddy Javaji 2024-present PhD at Stevens
- Preet Jhanglani 2024-present PhD at Stevens
- Haohang Li 2024-present PhD at Stevens
- Shravan Doda 2024 Master research project
- Paul Gao 2024 Undergraduate research project
- Patrick Wierzbicki 2024 Undergraduate research project
- Rohit Sandadi 2024 Dougherty Valley High School

Alumni:

- Paul (Yuzhi) Tang 2024 Master research project
- Pengshuo Qiu 2024 Undergraduate research project
- Arijit Chowdhury 2024 Undergraduate research project
- Chu Wu 2024 Undergraduate research project
- Sean Wang 2024 Undergraduate research project
- Sam Pan 2024 Undergraduate research project
- Huaizhi Ge 2023 Master research project
- Sanika Mhadgut 2024 Master research project

- 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 project: Controllable language 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

Reviewing

- 2024: ICLR, ICML, ACL Rolling Review February, COLM, NeurIPS (as area chair), ACL Rolling Review June (as action editor)
- 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

Volunteering

- Toronto Graduate Application Assistance Program (2021, 2022)
- NSight Mentorship Program (2016)

Selected Talks

- *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

Awards

- Top Reviewer Award, NeurIPS. 2023
- Ontario Graduate Scholarship, Provincial, \$15,000. 2022-2023
- Vector Institute PhD Research Grant, Institutional, \$6,000 each. 2020-2023
- ICRA RAS Travel Grant, Institutional, \$500. 2017
- Engineering Science Research Opportunity Program (ESROP) fellowship, Departmental, \$3000. 2016
- Dean's List, Institutional. 2014-2019
- UofT Entrance Scholarship, Institutional, \$5,000. 2014
- Chinese Physics Olympics (CPhO) Bronze medal, National. 2013

Affiliation

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

T , 1 , 1	0 1	
Last updated:	September	2024