Zhengxuan Wu
Phone: 216-551-7046
wuzhengx@cs.stanford.edu
https://nlp.stanford.edu/~wuzhengx

Education

Stanford University *2022 -*
Ph.D. in Computer Science
Advised by Chris Manning and Chris Potts

Stanford University *2020 - 2022*
M.S. in Symbolic Systems Program

University of Pennsylvania *2015 - 2017*
M.S. in Computer Science

Case Western Reserve University *2012 - 2015*
B.S. in Aerospace Engineering

Manuscripts and Publications[1]

preprint Causal Abstraction: A Theoretical Foundation for Mechanistic Interpretability
Atticus Geiger, Duligur Ibeling, Amir Zur, Maheep Chaudhary, Sonakshi Chauhan, Jing Huang, Aryaman Arora, **Zhengxuan Wu**, Noah Goodman, Christopher Potts, Thomas Icard https://arxiv.org/abs/2301.04709.

preprint A Reply to Makelov et al.(2023)'s "Interpretability Illusion" Arguments
Zhengxuan Wu, Atticus Geiger, Jing Huang, Aryaman Arora, Thomas Icard, Christopher Potts, Noah D. Goodman https://arxiv.org/abs/2401.12631.

EMNLP '24 Dancing in Chains: Reconciling Instruction Following and Faithfulness in Language Models
Zhengxuan Wu, Yuhao Zhang*, Peng Qi*, Yumo Xu*, Rujun Han, Yian Zhang, Jifan Chen, Bonan Min, Zhiheng Huang https://arxiv.org/pdf/2407.21417.

NeurIPS '24 *spotlight* ReFT: Representation Finetuning for Language Models
Zhengxuan Wu*, Aryaman Arora*, Zheng Wang, Atticus Geiger, Dan Jurafsky, Christopher D. Manning, Christopher Potts https://arxiv.org/abs/2404.03592.

CoLM '24 Mapping the Increasing Use of LLMs in Scientific Papers
Weixin Liang*, Yaohui Zhang*, **Zhengxuan Wu***, Haley Lepp, Wenlong Ji, Xuandong Zhao, Haocheng Cao, Sheng Liu, Siyu He, Zhi Huang, Diyi Yang, Christopher Potts, Christopher D. Manning, James Y. Zou https://arxiv.org/abs/2404.01268.

NAACL '24 pyvene: A library for understanding and improving PyTorch models via interventions
Zhengxuan Wu, Atticus Geiger, Aryaman Arora, Jing Huang, Zheng Wang, Noah Goodman, Christopher D. Manning, Christopher Potts
https://arxiv.org/abs/2403.07809.

[1] *equal contribution

CogSci ’24 Symbolic Variables in Distributed Networks that Count
Satchel Grant, **Zhengxuan Wu**, James Lloyd McClelland, Noah Goodman
https://escholarship.org/uc/item/7gm9d3hp.

ICML ’24 In-Context Sharpness as Alerts: An Inner Representation Perspective for Hallucination Mitigation
Shiqi Chen*, Miao Xiong*, Junteng Liu, **Zhengxuan Wu**, Teng Xiao, Siyang Gao, Junxian He https://arxiv.org/abs/2403.01548.

ACL ’24 RAVEL: Evaluating Interpretability Methods on Disentangling Language Model Representations
Jing Huang, **Zhengxuan Wu**, Christopher Potts, Mor Geva, Atticus Geiger
https://arxiv.org/abs/2402.17700.

CleaR ’24 Finding Alignments Between Interpretable Causal Variables and Distributed Neural Representations
Atticus Geiger*, **Zhengxuan Wu***, Christopher Potts, Thomas Icard, Noah D. Goodman, https://arxiv.org/abs/2303.02536.

BlackboxNLP ’23 Rigorously Assessing Natural Language Explanations of Neurons [**Best Paper Award**]
Jing Huang, Atticus Geiger, Karel D’Oosterlinck, **Zhengxuan Wu**, Christopher Potts, https://arxiv.org/abs/2309.10312.

EMNLP ’23 MQuAKE: Assessing Knowledge Editing in Language Models via Multi-Hop Questions
Zexuan Zhong*, **Zhengxuan Wu***, Christopher D. Manning, Christopher Potts, Danqi Chen, https://arxiv.org/abs/2202.12312.

EMNLP ’23 Oolong: Investigating What Makes Crosslingual Transfer Hard with Controlled Studies
Zhengxuan Wu*, Isabel Papadimitriou*, Alex Tamkin*,
https://arxiv.org/abs/2202.12312.

TACL ’23 ReCOGS: How Incidental Details of a Logical Form Overshadow an Evaluation of Semantic Interpretation
Zhengxuan Wu, Christopher D. Manning, Christopher Potts,
https://arxiv.org/abs/2303.13716.

NeurIPS ’23 Interpretability at Scale: Identifying Causal Mechanisms in Alpaca
Zhengxuan Wu*, Atticus Geiger*, Thomas Icard, Christopher Potts, Noah D. Goodman, https://arxiv.org/abs/2305.08809.

Findings ACL ’23 Inducing Character-level Structure in Subword-based Language Models with Type-level Interchange Intervention Training
Jing Huang, **Zhengxuan Wu**, Kyle Mahowald, Christopher Potts,
https://arxiv.org/abs/2209.14279.

ICML ’23 Causal Proxy Models for Concept-Based Model Explanations
Zhengxuan Wu*, Karel D’Oosterlinck*, Atticus Geiger*, Amir Zur, Christopher Potts, M.s., Stanford University, https://arxiv.org/abs/2209.14279.

NeurIPS ’22 ZeroC: A Neuro-Symbolic Model for Zero-shot Concept Recognition and Acquisition at Inference Time

Tailin Wu, Megan Tjandrasuwita, **Zhengxuan Wu**, Xuelin Yang, Kevin Liu, Rok Sosic, Jure Leskovec,
https://arxiv.org/abs/2206.15049.

NeurIPS '22 CEBaB: Estimating the Causal Effects of Real-World Concepts on NLP Model Behavior
Eldar David Abraham*, Karel D'Oosterlinck*, Amir Feder*, Yair Ori Gat*, Atticus Geiger*, Christopher Potts*, Roi Reichart*, **Zhengxuan Wu***,
https://arxiv.org/abs/2205.14140.

ICML '22 Inducing Causal Structure for Interpretable Neural Networks
Atticus Geiger*, **Zhengxuan Wu***, Hanson Lu*, Josh Rozner, Elisa Kreiss, Thomas Icard, Noah D. Goodman, Christopher Potts, https://arxiv.org/abs/2112.00826.

NAACL '22 Causal Distillation for Language Models
Zhengxuan Wu*, Atticus Geiger*, Josh Rozner, Elisa Kreiss, Hanson Lu, Thomas Icard, Christopher Potts, Noah D. Goodman, https://arxiv.org/abs/2112.02505.

RepL4NLP '22 Identifying the Limits of Cross-Domain Knowledge Transfer for Pre-trained Models [**Best Paper Award**]
Zhengxuan Wu, Nelson F. Liu, Christopher Potts,
https://arxiv.org/abs/2104.08410.

NeurIPS '21 ReaSCAN: Compositional Reasoning in Language Grounding
Zhengxuan Wu*, Elisa Kreiss*, Desmond C. Ong, Christopher Potts,
https://arxiv.org/abs/2109.08994.

ACL '21 DynaSent: A Dynamic Benchmark for Sentiment Analysis
Christopher Potts*, **Zhengxuan Wu***, Atticus Geiger, Douwe Kiela,
https://arxiv.org/abs/2012.15349.

NAACL '21 Dynabench: Rethinking Benchmarking in NLP'
Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, **Zhengxuan Wu**, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Zhiyi Ma, Tristan Thrush, Sebastian Riedel, Zeerak Waseem, Pontus Stenetorp, Robin Jia, Mohit Bansal, Christopher Potts and Adina Williams, https://arxiv.org/abs/2104.14337.

AAAI '21 Context-Guided BERT for Targeted Aspect-Based Sentiment Analysis
Zhengxuan Wu, Desmond C. Ong, https://arxiv.org/abs/2010.07523.

CHI '21 Not Now, Ask Later: Users Weaken Their Behavior Change Regimen Over Time, But Believe They Will Imminently Re-Strengthen It
Geza Kovacs, **Zhengxuan Wu** and Michael S. Bernstein,
https://arxiv.org/abs/2101.11743..

SCiL '21 Pragmatically Informative Color Generation by Grounding Contextual Modifiers
Zhengxuan Wu, Desmond C. Ong, https://arxiv.org/abs/2010.04372.

BlackboxNLP '20 Structured Self-Attention Weights Encode Semantics in Sentiment Analysis
Zhengxuan Wu, Thanh-Son Nguyen and Desmond C. Ong,
https://arxiv.org/abs/2010.04922.

ACII '19 Attending to Emotional Narratives
Zhengxuan Wu, Xiyu Zhang, Zhi-Xuan Tan, Jamil Zaki, Desmond C. Ong, https://arxiv.org/abs/1907.04197.

TAC '19 Modeling emotion in complex stories: the Stanford Emotional Narratives Dataset
Desmond C. Ong, **Zhengxuan Wu**, Zhi-Xuan Tan, Marianne Reddan, Isabella Kahhale, Alison Mattek and Jamil Zaki, https://arxiv.org/abs/1912.05008.

CHI '19 Conservation of Procrastination: Do Productivity Interventions Save Time or Just Redistribute It?
Geza Kovacs, Drew Mylander Gregory, Zilin Ma, **Zhengxuan Wu**, Golrokh Emami, Jacob Ray and Michael S. Bernstein, https://dl.acm.org/doi/10.1145/3290605.3300560.

CSCW '18 Rotating Online Behavior Change Interventions Increases Effectiveness But Also Increases Attrition
Geza Kovacs, **Zhengxuan Wu** and Michael S. Bernstein, https://dl.acm.org/doi/10.1145/3274364.

Other Professional Experience

Meta, Inc. - Research Scientist Intern *2024 - 2024*
- Developing interpretability tools for LLaMA models.

Amazon, Inc. - Research Scientist Intern *2023 - 2023*
- Investigated faithfulness and instruction following of language models.

VMware, Inc. - Senior Member of Technical Staff *2017 - 2022*
- Developed scalable data-center management platform.

Swift Capital (Paypal, Inc.) - Machine Learning Intern *2016 - 2016*
- Developed machine learning systems to predict the credit scores of loan applicants.

Academic Experience

- Reviewer for CHI19, *CL22-24, ICML22-24, NeurIPS22-23, COLM24
- Invited Abstract Presentation in IC2S2 2019, University of Amsterdam, Netherlands

Technical Strengths

- **Program Languages**: Python, C++/C, C#, Java, R, Matlab, Haskell, Bash.
- **Machine Learning**: Discriminative and Generative Models; Reinforcement Learning; Multi-task Learning; Graph Neural Networks.
- **AI + Big Data**: PyTorch, scikit-learn, Keras, TensorFlow, NumPy, Pandas, H2O, MapReduce (Hadoop).
- **Data Mining**: PyData, SciPy, SNAP, SQL, NoSQL (Mongo), NetworkX, Jupyter.
- **Data Science**: Mixed Linear Model, Hierarchical Logistic Regression, A/B Testings, Crowdsourcing (MTurk).
- **Server + Database**: Node.js, Flask, MongoDB, PostgreSQL, Kubernetes, Docker, Google Cloud, AWS EC2, Heroku, Azure, Jenkins CICD.
- **Web + Mobile**: HTML/CSS/JS, Polymer, React, Webpack, Apache, Android (Java), Xcode.