

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

MANUSCRIPTS AND PUBLICATIONS¹

- preprint* 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>.
- preprint* MAPPING THE INCREASING USE OF LLMs IN SCIENTIFIC PAPERS
Weixin Liang*, Yaohui Zhang*, **Zhengxuan Wu***, Haley Lepp, Wenlong Ji, Xuan-dong 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>.
- preprint* 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>.
- preprint* 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>.
- preprint* 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>.
- 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>.
- CleaR ’24 FINDING ALIGNMENTS BETWEEN INTERPRETABLE CAUSAL VARIABLES AND DISTRIBUTED NEURAL REPRESENTATIONS

¹*equal contribution

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 ZERO-C: 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 Susic, 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 IMMEDIATELY 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

RESEARCH EXPERIENCE

Stanford NLP Group - Graduate Researcher *2020 - Present*

- Working with Chris Manning and Chris Potts.

Amazon AWS AI - Research Scientist Intern *June 2023 - Sept. 2023*

- Worked with Yuhao Zhang, Peng Qi and Yumo Xu.

OTHER PROFESSIONAL EXPERIENCE

VMware, Inc. - Senior Software Engineer *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.