Xinpeng Wang

xinpeng@cis.lmu.de xinpeng-wang.github.io Xinpeng-Wang Google Scholar

Research Experience

New York University **Jan. 2025 – May. 2025 (Expected)**
Visiting Scholar at NYU Alignment Research Group - Advised by Prof. He He *New York, US*

- LLM Safety Alignment

Education

Ludwig Maximilian University of Munich **Dec. 2022 – present**
PhD student - Advised by Prof. Barbara Plank *Munich, Germany*

- Trustworthiness and Alignment of Language Model

Technical University of Munich **Oct. 2019 – Oct. 2022**
M.Sc. in Robotics, Cognition, Intelligence *Munich, Germany*

Technical University of Munich **Oct. 2016 – Aug. 2017**
Exchange program in Mechanical Engineering *Munich, Germany*

Dalian University of Technology **Sep. 2014 – Jul. 2018**
B.Eng. in Mechanical Engineering *Dalian, China*

Selected Publication

Surgical, Cheap, and Flexible: Mitigating False Refusal in Language Models via Single Vector Ablation.
***Wang, X.**, Hu, C., Röttger, P., Plank, B.* **ICLR 2025**
Look at the Text: Instruction-Tuned Language Models are More Robust Multiple Choice Selectors than You Think.
***Wang, X.**, Hu, C., Ma, B., Röttger, P. & Plank, B.* **COLM 24'**
"My Answer is C": First-Token Probabilities Do Not Match Text Answers in Instruction-Tuned Language Models.
***Wang, X.**, Ma, B., Hu, C., Weber-Genzel, L., Röttger, P., Kreuter, F., Hovy, D., Plank, B* **ACL 24' Findings**
"Seeing the Big through the Small": Can LLMs Approximate Human Judgment Distributions on NLI from a Few Explanations? Chen, B., ***Wang, X.**, Peng, S., Litschko, R., Korhonen, A., Plank, B.* **EMNLP 24' Findings**
The Potential and Challenges of Evaluating Attitudes, Opinions, and Values in Large Language Models. *Ma B.*, **Wang X.***, Hu T., Haensch A, Hedderich M., Plank B., Kreuter F.* **EMNLP 24' Findings**
FinerCut: Finer-grained Interpretable Layer Pruning for Large Language Models.
*Zhang Y., Li Y., **Wang X.**, Shen Q., Plank B., Bischl B., Rezaei M., Kawaguchi K.* **Compression@NeurIPS 24'**
ACTOR: Active Learning with Annotator-specific Classification Heads to Embrace Human Label Variation.
***Wang, X.**, Plank, B.* **EMNLP 23'**
How to Distill your BERT: An Empirical Study on the Impact of Weight Initialisation and Distillation Objectives.
***Wang, X.**,Weissweiler L., Schütze H. & Plank, B.* **ACL 2023**
SceneFormer: Indoor Scene Generation with Transformers.
Wang, X., Yeshwanth, C.*, & Nießner, M.* **3DV 21', Oral Presentation** *Equal contribution

Talks

Workshop Tutorial: How to Query LLMs / Pitfalls in Prompting **Oct. 2024**
Social Data Science Center (SoDa) and the Artificial Intelligence Interdisciplinary Institute (AIM) *University of Maryland*

Bias and Robustness of LLM Evaluation **Apr. 2024**
Social Data Science and AI Lab *LMU Munich*

Teaching Experience

Human-Centric NLP *Guest Lecture, LMU Munich* **SS24**

Introduction to Deep Learning *Teaching Assistant, Technical University of Munich* **SS20, WS20/21**

Service

Program Committee/ Review: ARR 2023 Dec, EACL UncertaiNLP, ARR 2024 June - Oct, ICLR 2025

Technical Skills

Languages: Python, HTML, C, MATLAB
Frameworks: PyTorch, TensorFlow, HuggingFace, Fairseq, DeepSpeed