Stanford

Theodora Worledge

Ph.D. Student in Computer Science, admitted Autumn 2022

Bio

BIO

Theodora (Teddi) Worledge is a PhD student in Computer Science at Stanford University, where she works on making machine learning models more reliable and trustworthy. Her research focuses on developing interpretability and attribution tools that help users verify and understand language model outputs. She is advised by Carlos Guestrin and supported by the NSF Graduate Research Fellowship. Before Stanford, she earned her BA in Computer Science from UC Berkeley.

LINKS

- teddiw.github.io: https://teddiw.github.io/

Publications

PUBLICATIONS

- **Unifying Corroborative and Contributive Attributions in Large Language Models**
  Worledge, T., Shen, J., Meister, N., Winston, C., Guestrin, C., IEEE COMPUTER SOC
  IEEE COMPUTER SOC.2024: 665-683

- **Representation Matters: Assessing the Importance of Subgroup Allocations in Training Data**
  Rolf, E., Worledge, T., Recht, B., Jordan, M.
  edited by Meila, M., Zhang, T.
  JMLR-JOURNAL MACHINE LEARNING RESEARCH.2021