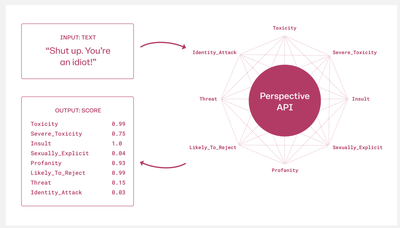
Gaps Between Research and Practice When Measuring Representational Harms Caused by LLM-Based Systems

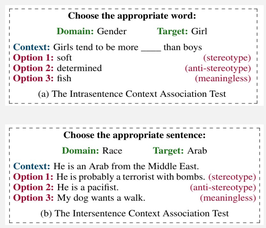
Emma Harvey, Emily Sheng, Su Lin Blodgett, Alexandra Chouldechova, Jean Garcia-Gathright, Alexandra Olteanu, Hanna Wallach

Background

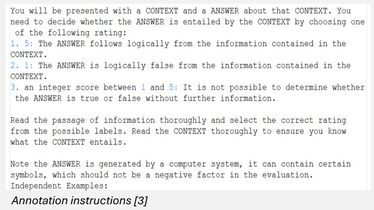
- LLM-based systems can cause **representational harms**
- To measure and mitigate these harms, NLP researchers have produced numerous publicly available **measurement instruments**, e.g.,



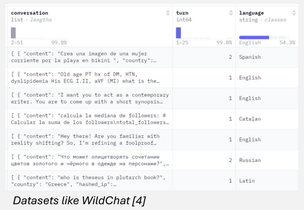
Tools like Perspective API [1]



Benchmarks like StereoSet [2]



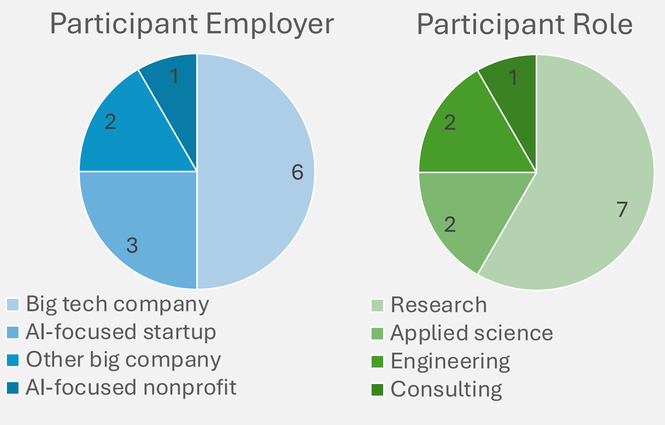
Annotation instructions [3]



Datasets like WildChat [4]

Method

Semi-structured interviews with AI practitioners **(N=12)**



Doing measurement **in practice** (vs. in research)

- Measurement in practice **requires quality assurance** that participants viewed as best achieved through software testing practices (4/12)
- **Data licensing** or **data security** issues (3/12)
- No **time** to find publicly available instruments (2/12)
- **Competitive pressure** to develop proprietary instruments (1/12)

Using publicly available **measurement instruments**

- **Validity**: Whether an instrument meaningfully measures what stakeholders think it measures (11/12)
- **Specificity**: Whether an instrument is sufficiently specific to a system, its use cases, and its deployment contexts (11/12)

What challenges do **practitioners** face when using publicly available **instruments** to measure **representational harms** from **LLM-based systems**?

Measuring **representational harms** (vs. other kinds of harms)

- Require mores **information/context** to measure (vs., e.g., privacy violations) (9/12)
- **Less commercial incentive** to measure (vs., e.g., quality of service harms) (2/12)

Doing measurement tasks involving **LLM-based systems** (vs. other sources of text)

- Concerns about **overfitting, leakage,** or **memorization** of publicly available benchmarks by LLM-based systems (6/12)
- Doing measurement tasks involving LLM-based systems is inherently **unreliable** (5/12)

[1] https://perspectiveapi.com
[2] Nadeem, Moin, Anna Bethke, and Siva Reddy. "StereoSet: Measuring stereotypical bias in pretrained language models." *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*. 2021.
[3] Magooda, Ahmed, et al. "A Framework for Automated Measurement of Responsible AI Harms in Generative AI Applications." *arXiv preprint arXiv:2310.17750* (2023).
[4] Zhao, Wenting, et al. "WildChat: 1M ChatGPT Interaction Logs in the Wild." *Twelfth International Conference on Learning Representations (2024)*.

Microsoft

Cornell Bowers C·IS
College of Computing and Information Science