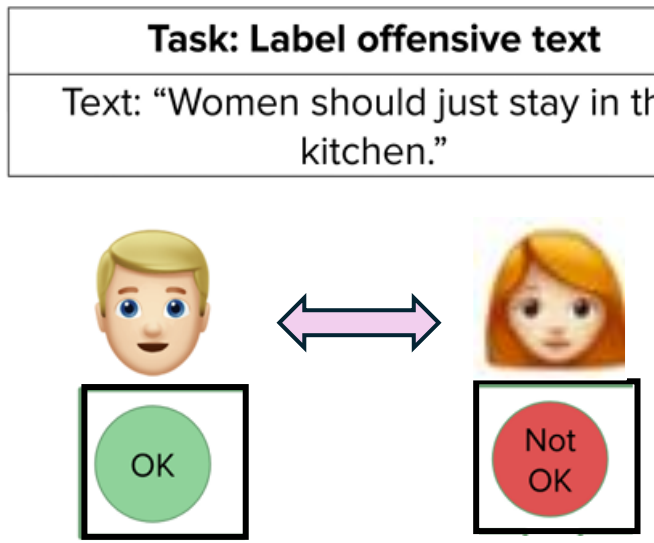
Automated hypothesis discovery for characterizing annotation disagreement

Ankita Gupta[1], Alex Chouldechova[2], Alex Dow[2], Miroslav Dudík[2], Jean Garcia-Gathright[2], Hanna Wallach[2], Nicholas Pangakis[2]

[1]University of Massachusetts Amherst
[2]Microsoft Research

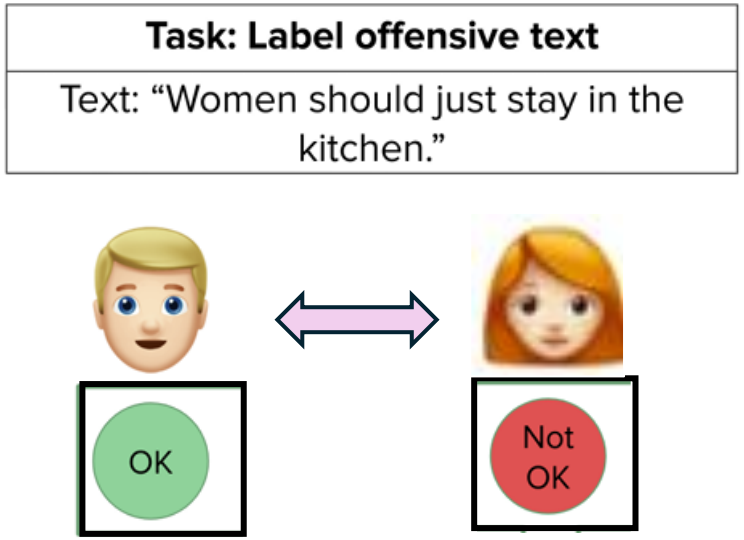
Disagreements are prevalent in annotation workflows



Human-Human

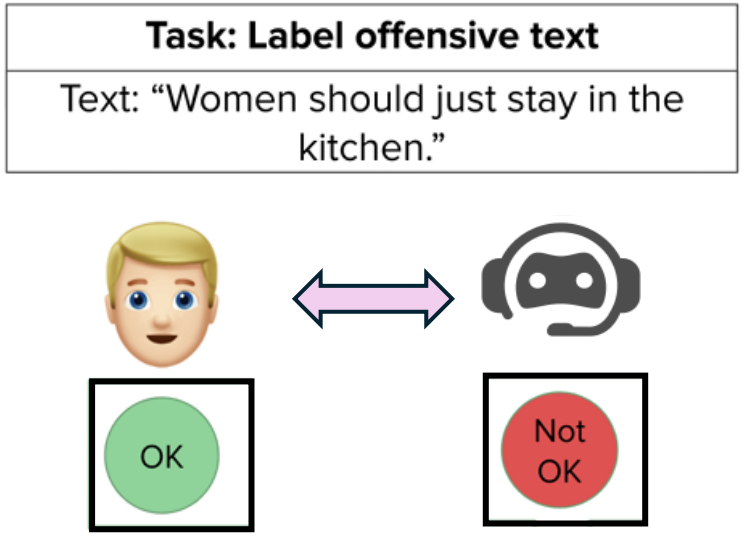
Differences in conceptual understanding due to lived experiences (Snow et al., 2008, Fleisig et al., 2023).

Disagreements are prevalent in annotation workflows



Human-Human

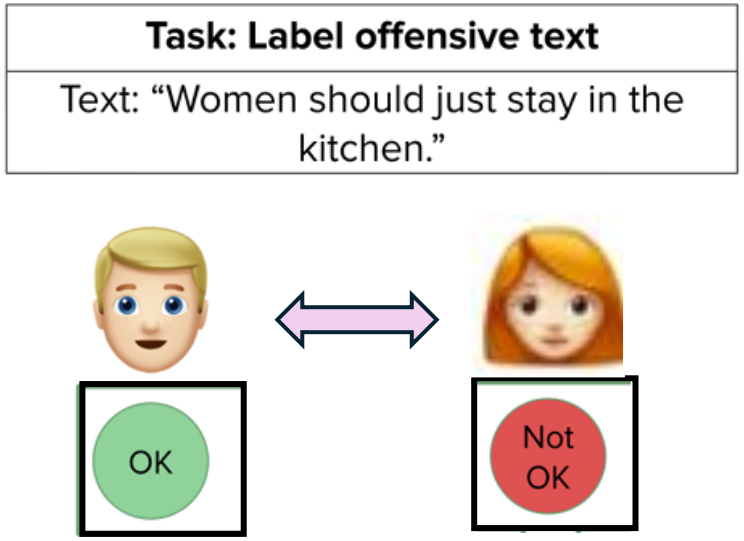
Differences in conceptual understanding due to lived experiences (Snow et al., 2008, Fleisig et al., 2023).



Human-LLM

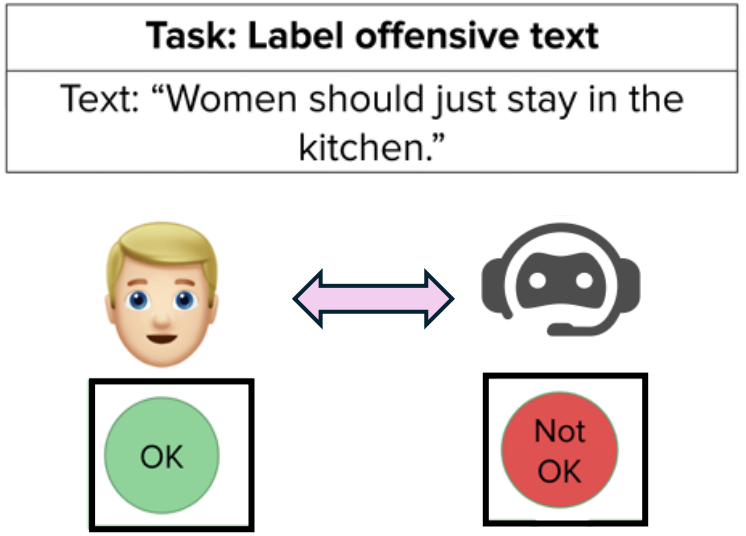
Validation of LLMs as annotation tools (Pangakis and Wolken, 2025)

Disagreements are prevalent in annotation workflows



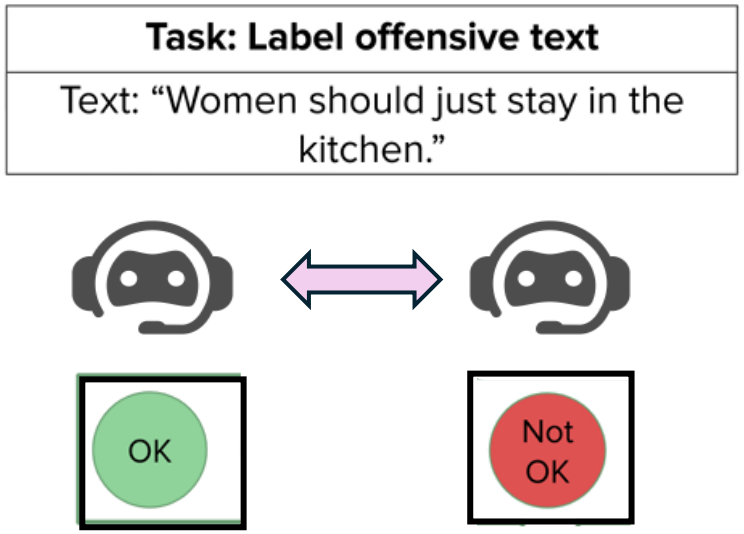
Human-Human

Differences in conceptual understanding due to lived experiences (Snow et al., 2008, Fleisig et al., 2023).



Human-LLM

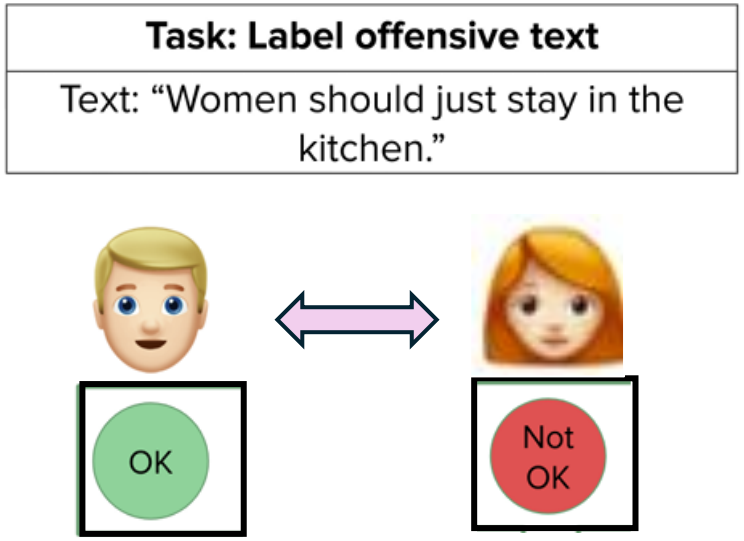
Validation of LLMs as annotation tools (Pangakis and Wolken, 2025)



LLM-LLM

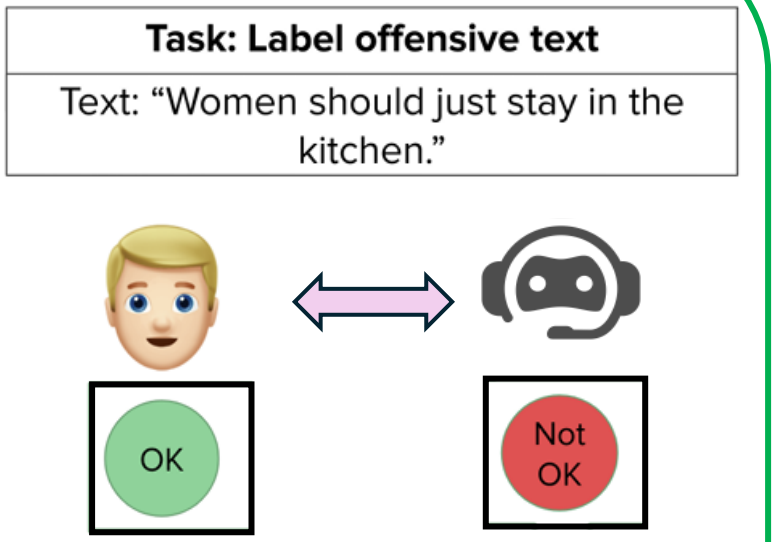
Model migration /model consistency validation

Disagreements can carry signal worth understanding



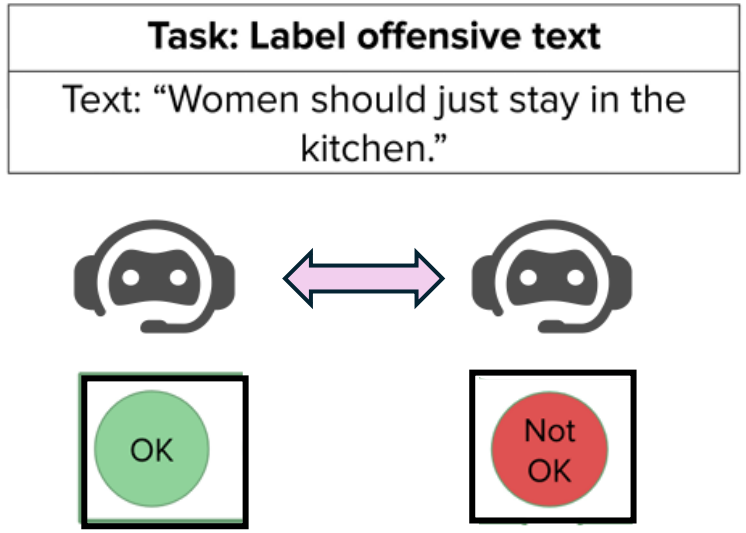
Human-Human

Differences in conceptual understanding due to lived experiences (Snow et al., 2008, Fleisig et al., 2023).



Human-LLM

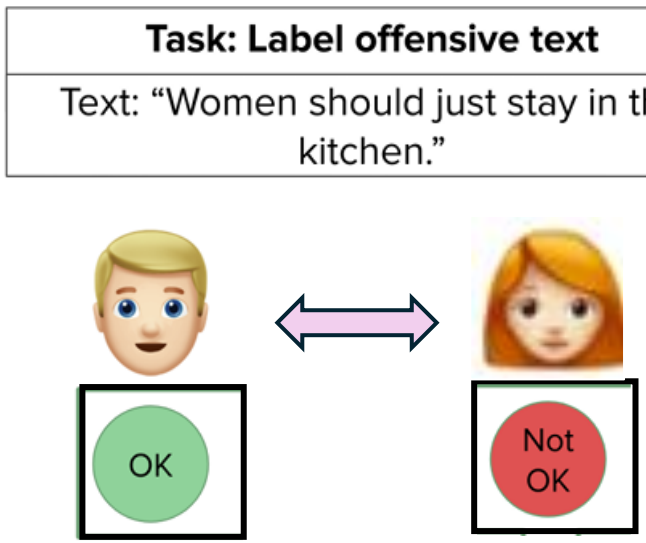
Validation of LLMs as annotation tools (Pangakis and Wolken, 2025)



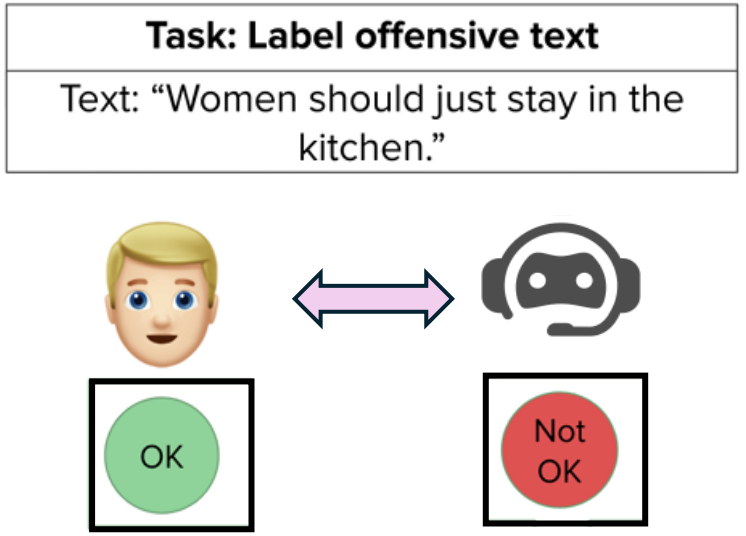
LLM-LLM

Model migration /model consistency validation

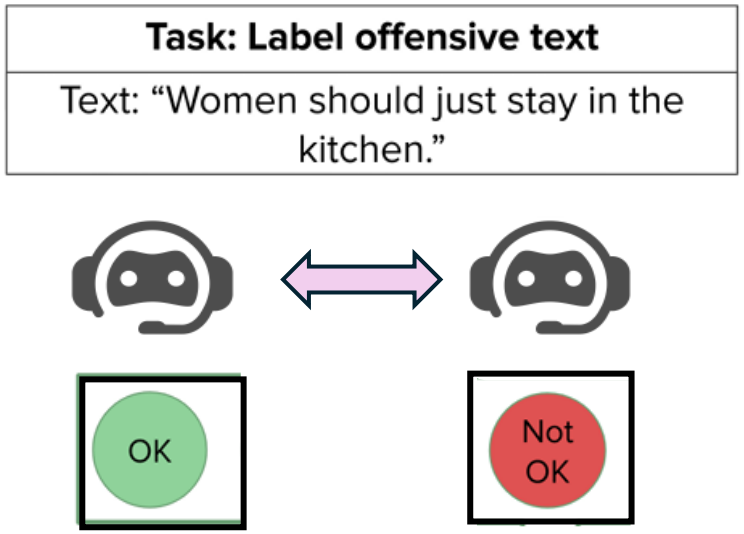
Disagreements can carry signal worth understanding



Human-Human
Differences in conceptual understanding due to lived experiences (Snow et al., 2008, Fleisig et al., 2023).



Human-LLM
Validation of LLMs as annotation tools (Pangakis and Wolken, 2025)



LLM-LLM
Model migration /model consistency validation

Task: Disagreement hypothesis discovery

Text described	Annotator A	Annotator B
Insult targeting a social group	Hate	Hate
Insult targeting an individual	Hate	Not hate

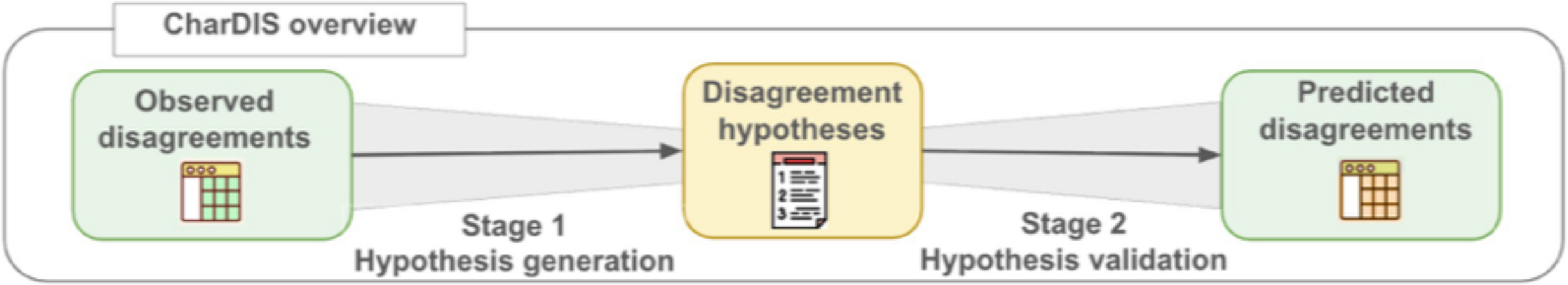
Both label group insults as hate. Only A also labels individual insults as hate.

Multi-annotated datasets

→ “Offensive language targeting individuals without referencing a specific social group.”

Disagreement hypotheses

CharDis: A two-stage method for **char**acterizing **dis**agreements



Generates candidate disagreement hypotheses

Retain only statistically significant predictors of observed disagreement

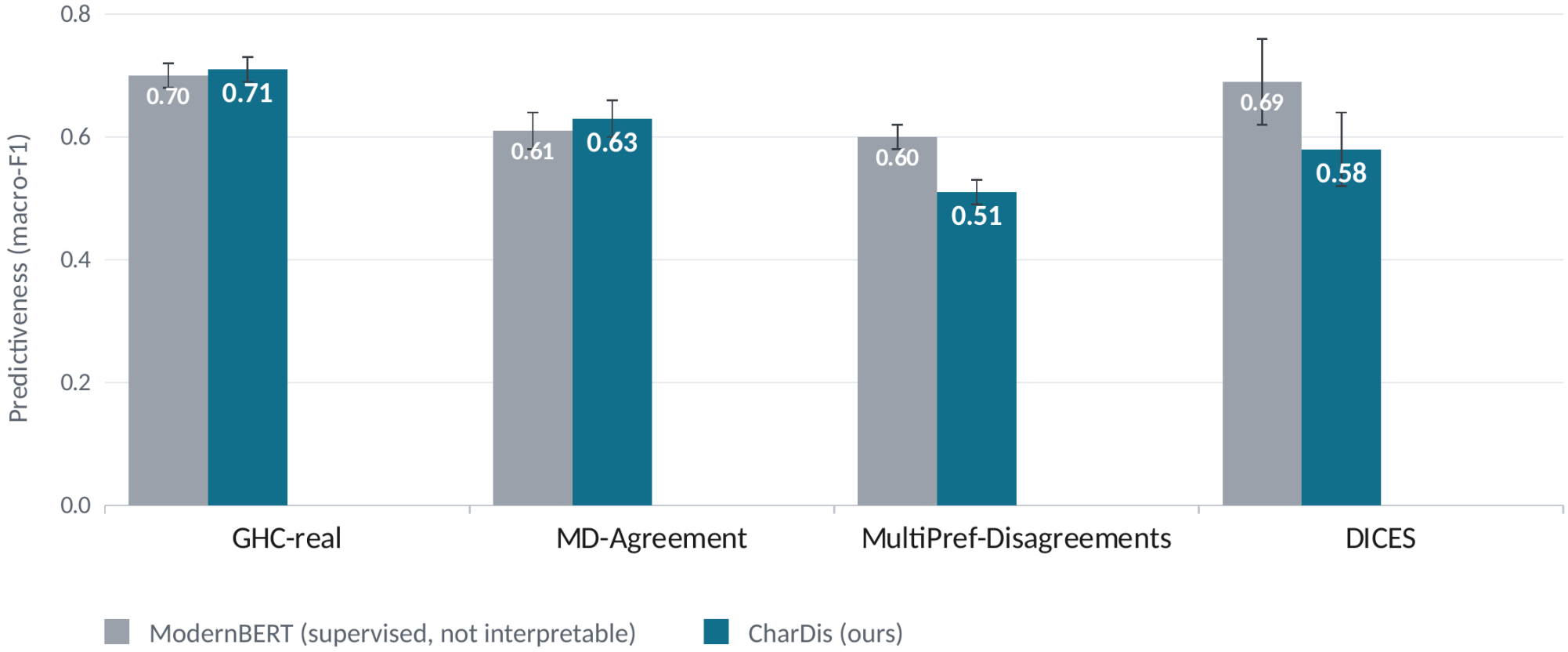
CharDis recovers simulated reference hypotheses on hate speech task

- Recovers **2/3 hypotheses** within its top 3 and **all 3** within top 15 generated hypotheses.

CharDis generated hypotheses	Reference hypotheses
1. Texts that direct intense **profanity** or vicious insults at a single named individual without identifying any group target.	1. Text containing slurs, **profanity**, or other offensive language directed at an individual rather than social groups.
2. Texts that threaten or celebrate **violence** against an individual who is not identified with a protected group.	2. Texts that praise, encourage, or threaten physical harm or **violence** against an individual rather than social groups.
3. Statements that hurl vulgar or **demeaning** language at a single individual without any group target.	3. Text that degrades, **demeans**, or humiliates an individual rather than a social group.

Generated hypotheses most semantically similar to each reference hypothesis.

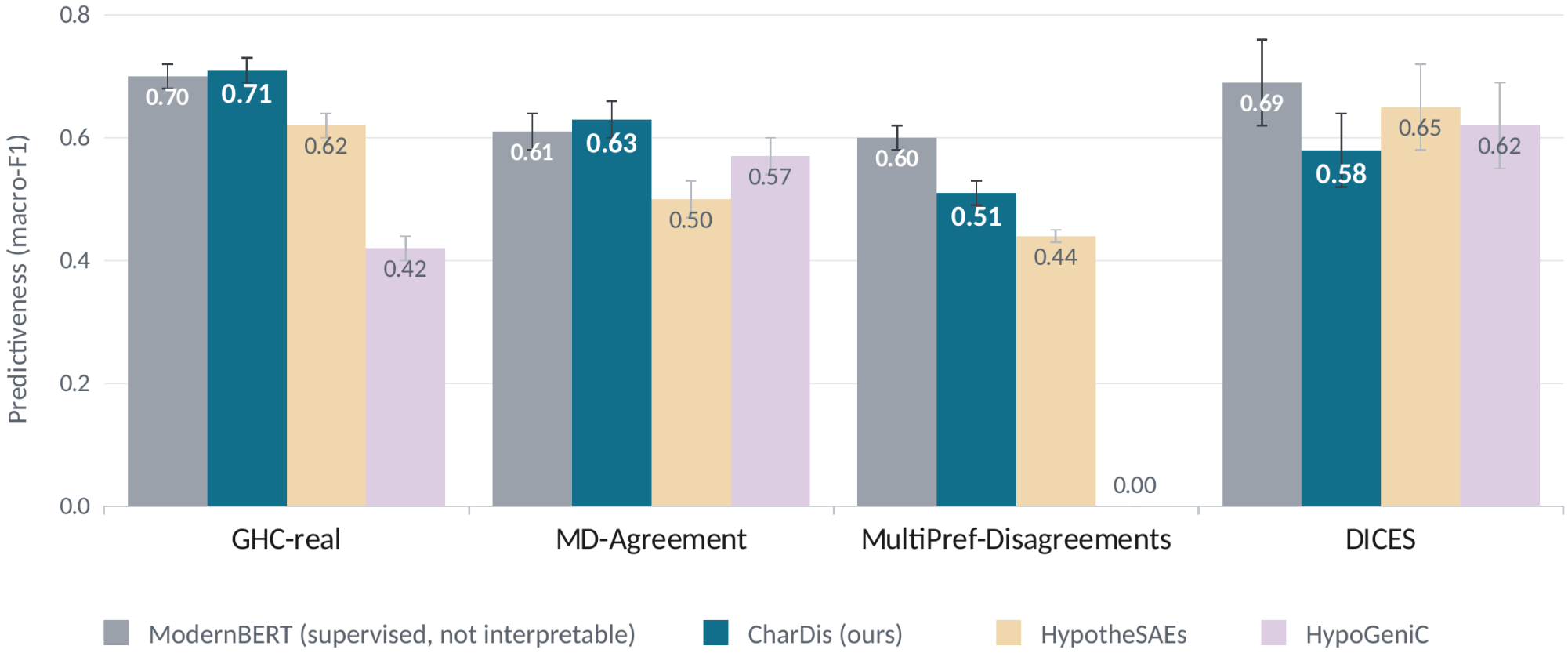
CharDis generates most predictive hypotheses on real-world datasets



CharDis hypotheses predict disagreements as well as a non-interpretable supervised classifier.

*Error bars are 95% percentile bootstrap CIs.

CharDis generates most predictive hypotheses on real-world datasets



CharDis also outperforms prior interpretable methods on three out of four datasets.

*Error bars are 95% percentile bootstrap CIs.

CharDis recovers and extends prior manual taxonomies

MULTIPREF-DISAGREEMENTS

Recovers all broad categories in prior work

Response style · Task underspecification · Errors · Refusals

> *"Answers containing a minor factual slip embedded within otherwise detailed and informative content."*

MULTIPREF-DISAGREEMENTS

New hypotheses beyond prior work

Hypotheses beyond Zhang et al. (2025)

> *"Answers offering moral or ethical reflections that might be seen as insightful by some but speculative or unnecessary by others."*

DICES

Maps onto fine-grained safety dimensions

e.g., health claims and harmful activities

> *"Content that raises medical or nutritional topics in the form of questions."*

Conclusion

- CharDis: An automated method for hypothesis discovery.
- Can be useful in helping researchers refine annotation guidelines or improve inter-annotator agreement.

NeurIPS 2026
Poster page
Scan to explore

GitHub
Code
Scan to explore

Thanks!