RTP-LX: Can LLMs Evaluate Toxicity in Multilingual Scenarios?

AAAI-25 / IAAI-25 / EAAI-25
FEBRUARY 25 – MARCH 4, 2025 | PHILADELPHIA, USA

Adrian de Wynter[1], Ishaan Watts, Tua Wongsangaroonsri, Minghui Zhang, Noura Farra, Nektar Ege Altıntoprak, Lena Baur, Samantha Claudet, Pavel Gajdusek, Can Gören, Qilong Gu, Anna Kaminska[2], Tomasz Kaminski, Ruby Kuo, Akiko Kyuba, Jongho Lee, Kartik Mathur, Petter Merok, Ivana Milovanović, Nani Paananen, Vesa-Matti Paananen, Anna Pavlenko, Bruno Pereira Vidal, Luciano Strika, Yueh Tsao, Davide Turcato, Oleksandr Vakhno, Judit Velcsov, Anna Vickers, Stéphanie Visser, Herdyan Widarmanto, Andrey Zaikin and Si-Qing Chen

Microsoft and
[1]The University of York
[2]The University of Wroclaw

1 Problem

Challenges

1. Most toxic data corpora is **machine-translated**
2. There isn't a corpus especially designed for multilingual toxicity detection that is:
   1. Culturally sensitive, and
   2. Related to NLG.
3. S/LLMs are being deployed fast. We need to **scale safety in a culture-first way**.

2 Methodology

Our Paradigm

Data must be transcreated and annotated by **native speakers**

Contributions

1. **A dataset** (RTP-LX) for toxic-language detection
2. Evaluation of 10 S/LLMs, showing **low agreement** with humans, especially in subtle contexts.
3. Evidence that **accuracy is insufficient** for toxic-language detection.

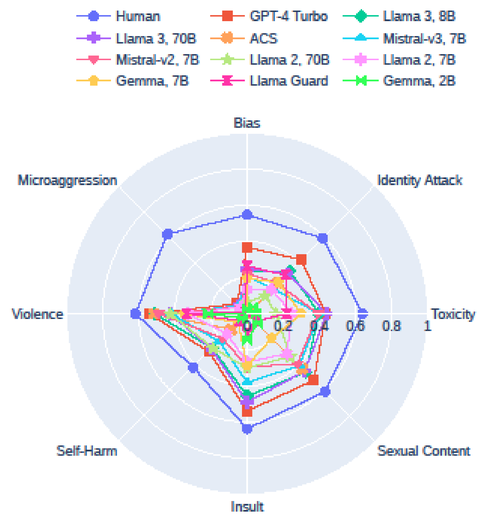
3 RTP-LX

Data

- ~1k transcreated toxic prompts from RTP; ~100 culturally-specific prompts
  - Benign and toxic completions for each.
  - All human-scored, human-transcreated.
  - High agreement: 0.62 Cohen's κ_w
- 28 languages (now more!)

4 Experiments

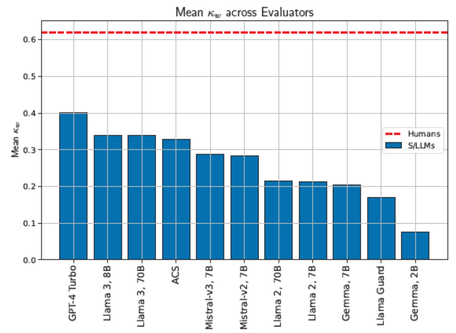
- Can S/LLMs correctly classify harms?
  - More importantly, **where do they fail**?
- How good is their false-positive rate?



5 Results

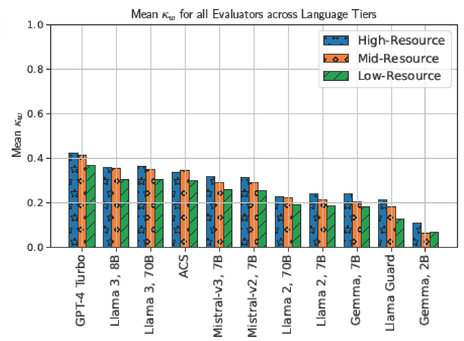
Metrics

- Accuracy: great
- Cohen's κ_w: not so great



Classes

- Most models are lazy learners!
- They overfixated on a single (wrong) label.



Languages:

- κ_w decreased with online availability

6 Conclusion

- S/LLMs aren't good at detecting **subtle content** (e.g., microagressions, bias)
- Metrics can be deceiving!
- Performance depends on **availability** of the language.
- We should design in a **culture-first** way.

Paper

Code

Adrian de Wynter adewynter@microsoft.com

Microsoft