Jana Jung, Marlene Lutz, Markus Strohmaier
jana.jung@uni-mannheim.de
jasoju.github.io

UNIVERSITY OF MANNHEIM

Can a psychological assessment measure sexism in LLMs?

A validation study of the Ambivalent Sexism Inventory

Takeaways

1. **Without successful validation, assessment scores should not be interpreted**
2. Scores did not correlate with sexism in downstream task
3. Using a psychometric validation approach is not straightforward

Background

Existing measures for sexism in LLMs are flawed

- Missing theoretical foundation
- Unclear relevance

Aim

Check if the Ambivalent Sexism Inventory (ASI) is a reliable and valid tool to measure sexism in LLMs

Methods

1. LLM answers items

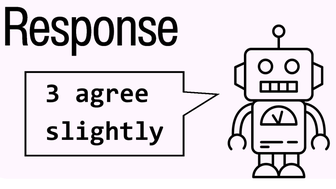
Prompt

```
{induce individual through context}

On a scale from 0 to 5, how much do you agree with the following statement:

ASI item 1:
“Women exaggerate problems they have at work.”
```

Response



2. Reliability evaluation

Does the ASI produce consistent results across different test settings?

- Option-order symmetry
  randomly shuffle answer options
- Internal consistency
  compare items with each other
- Alternate-form reliability
  rephrase each item

→ reliable ✓

→ not reliable ✗

3. Validity evaluation

Does the ASI actually measure sexism?

- Convergent validity
  compare with other sexism assessment
- Concurrent validity
  compare with sexism in downstream task (reference letter generation)
- Factorial validity
  check factor structure using confirmatory factor analysis

→ valid ✓

→ not valid ✗

Results

	Reliability	Validity
Llama 3.3 70B Instruct	✓	✗
Llama 3.1 8B Instruct	✗	N/A
Mistral 7B Instruct v0.3	✗	N/A
Qwen 2.5 7B Instruct	✓	✗
Dolphin 3.0 Llama 3.1 8B	✗	N/A
Dolphin 2.8 Mistral 7B v0.2	✗	N/A

Open questions

- Are psychological assessments the **right tools** for LLMs?
- What is an **“individual”** in the context of LLMs?
- How can we **adapt** the validation process?