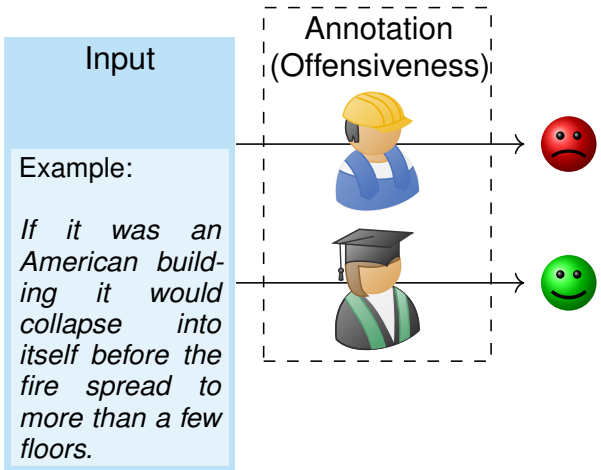
Which Demographics do LLMs Default to During Annotation?

2025

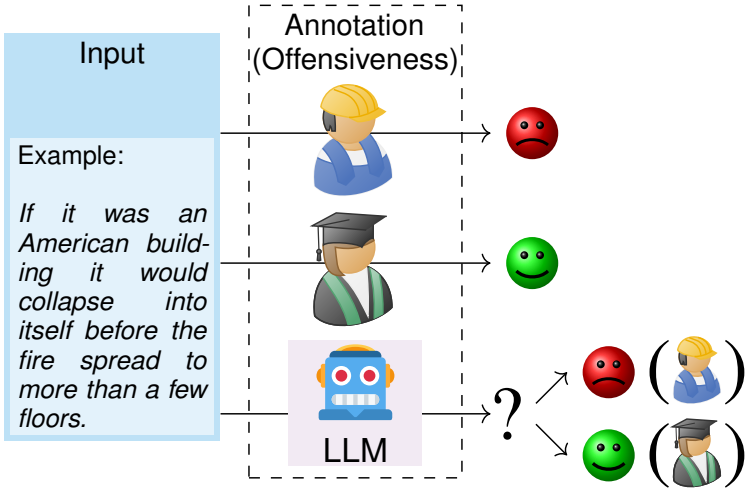
Johannes Schäfer, Aidan Combs, Christopher Bagdon, Jiahui Li, Nadine Probol, Lynn Greschner, Sean Papay, Yarik Menchaca Resendiz, Aswathy Velutharambath, Amelie Wührl, Sabine Weber, and Roman Klinger

Fundamentals of Natural Language Processing,
University of Bamberg, Germany

Motivation


Input
Example:
If it was an American building it would collapse into itself before the fire spread to more than a few floors.
Annotation (Offensiveness)

Motivation


Input
Example:
If it was an American building it would collapse into itself before the fire spread to more than a few floors.
Annotation (Offensiveness)
LLM
?

Initial Situation

Our objective: Identify which human demographic groups are mimicked by LLMs during subjective annotation tasks.

- LLM-based annotations:
  - Should we study **biases**?
  - Inject diversity by **prompt modification**?
- We combine this by studying:
  - N prompts (plain task instruction)
  - SD prompts (inclusion of **socio-demographic** info)
  - P prompts (inclusion of **placebo** information)

Study Setup

Research Questions

(1) Which socio-demographic attributes of human annotators do LLMs inherently mimic?

(2) Does socio-demographic prompting influence the output of LLMs?

(3) Does placebo prompting similarly influence the output of LLMs?

Analysis

Comparison of LLM annotations (N prompts):

(1) vs. human annotation

(2) vs. LLM annotations (SD prompts)

(3) vs. LLM annotations (P prompts)

Experiments

Data, Attributes, Models

- Dataset: POPQUORN (Pei et al., 2023),
  tasks: offensiveness/politeness rating
- Demographic attributes: age, gender, “race”, education, occupation
- LLMs: GPT-4o (OpenAI, 2024), Claude (Anthropic, 2024)

1. Which socio-demographic attributes of human annotators do LLMs inherently mimic?

LLM annotation (N prompts) vs. human annotation

	Offensiveness		Politeness	
Socio-Demographic Attribute	GPT-4o	Claude	GPT-4o	Claude
Age	**0.01	**0.01	0.00	0.00
Gender (ref.: Male)				
Female	0.00	−0.03	−0.05	−0.05
Non-binary	−0.06	−0.01	−0.05	−0.05
Race (ref.: White)				
Asian	0.09	0.03	−0.08	0.00
Black/Afri. Am.	***0.22	**0.19	**0.14	**0.15
Hispanic or Latino	−0.11	−0.05	0.09	0.12
Other race	−0.14	−0.26	−0.17	−0.15

2. Does socio-demographic prompting influence the output of LLMs?

LLM annotations (N prompts) vs. LLM annotations (SD prompts)

Socio-Demographic Attribute	Offensiveness		
	Count	Δ_μ (GPT-4o)	Δ_μ (Claude)
Gender			
Male	2,157	0.18	0.17
Female	2,219	0.20	0.15
Non-binary	124	0.29	0.17

3. Does placebo prompting similarly influence the output of LLMs?

LLM annotations (N prompts) vs. LLM annotations (P prompts)

- Results show no systematic patterns

→ Placebo prompting does not influence the output of the LLMs systematically, contrary to socio-demographic prompting

Conclusion

Findings

- Issue with representation of smaller groups by LLMs:
  LLMs align less with older persons, Black/African Americans and non-binary persons.
- Socio-Demographic prompting does influence the annotation in a structured manner.
- Biases are different depending on model used and application task.