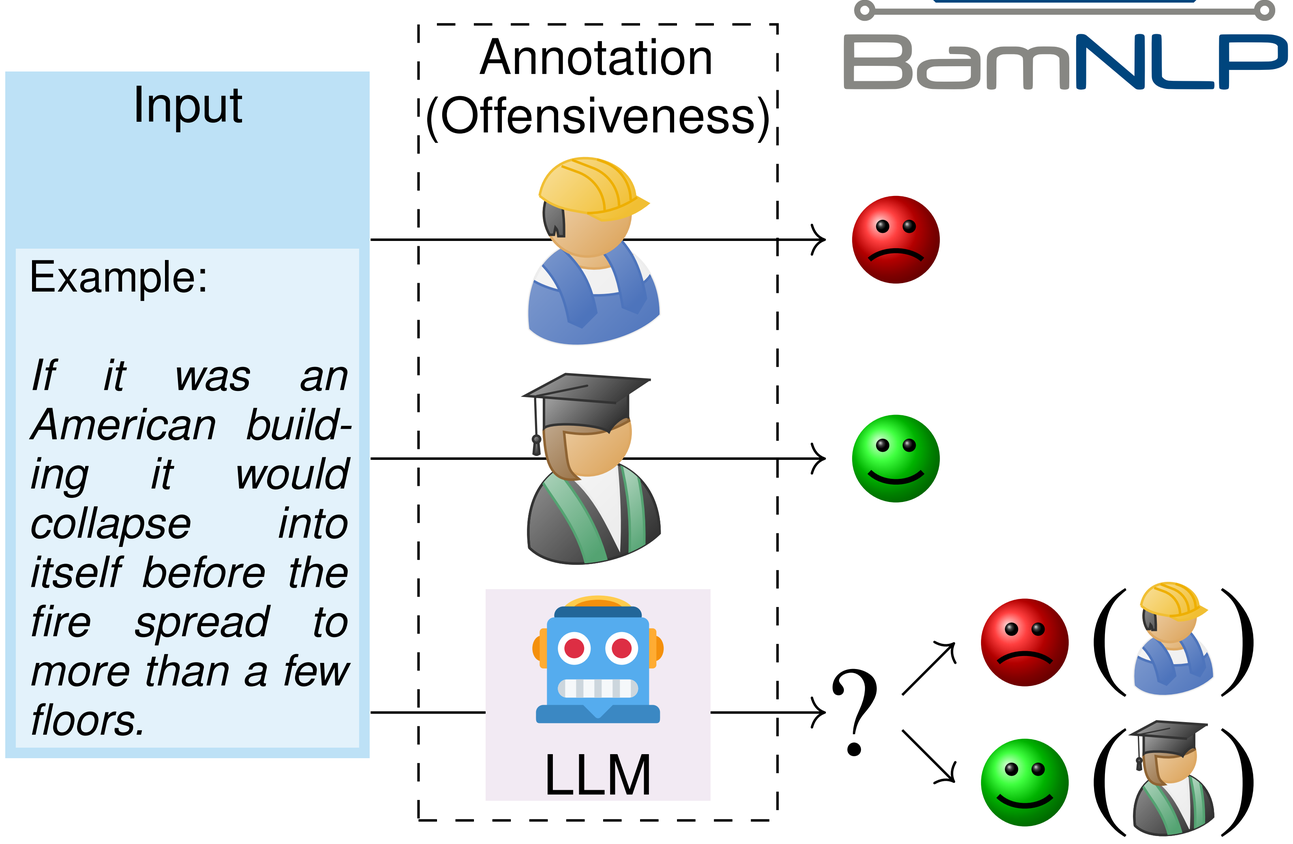
Which Demographics do LLMs Default to During Annotation?

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

Initial Situation

- **Human analysis** of text data shows variety of **perspectives** during annotation, i.e. label variations; these can be explained by **demographics**.
- LLM-based annotations: Should we study biases? Inject diversity by prompt modification?
- Our objective: Identify which human demographic groups are mimicked by LLMs during subjective annotation **tasks (offensiveness/politeness rating)**.
- LLM-based annotation methods:
  - **N prompts** (plain task instruction)
  - **SD prompts** (inclusion of socio-demographic info)
  - **P prompts** (inclusion of placebo information)



Research Questions

1. Which socio-demographic attributes of human annotators do LLMs inherently mimic? → LLM annotation with N prompts vs. human annotation
2. Does socio-demographic prompting influence the output of the LLMs? → LLM annotation with N prompts vs. with SD prompts
3. Does placebo prompting similarly influence the output of the LLMs? → LLM annotation with N prompts vs. with P prompts

Analysis Setup

Results

1. Regression analysis, N prompts vs. human annot.

	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

⇒ Overall: some significant differences, biases for age (offensiveness rating task) and "race" (both tasks)

2. Inclusion of socio-demographic info in the prompt

	Offensiveness		
Socio-Demographic Attribute	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

⇒ GPT-4o's annotation influenced by SD prompt with gender non-binary.

3. Placebo prompting

⇒ Does not show systematic patterns.

Conclusion

- **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.**