The Effects of Demographic Instructions on LLM Personas

Angel Felipe Magnossão de Paula[1,2] J. Shane Culpepper[2] Alistair Moffat[3]
Sachin Pathiyan Cherumanal[4] Falk Scholer[4] Johanne Trippas[4]

[1]Universitat Politècnica de València [2]University of Queensland
[3]The University of Melbourne [4]RMIT University

Motivation

- Content moderation must reflect *subjective* views of sexism.
- LLMs are promising but susceptible to demographic bias.
- We adopt a **perspectivist** stance: preserve disagreements and model diversity.

Research Questions

1. Do LLMs exhibit demographic bias when detecting sexism?
2. Can persona-style prompts mitigate that bias?

Dataset

- **EXIST 2023**: 7,958 tweets, six annotations each.
- Labels: Sexist / Not Sexist
- Sexist Sample Tweet: "Mujer al volante, tenga cuidado!"
- Annotator strata: {F, M} × {18–22, 23–45, 46+}.

LLMs Evaluated

- GPT-3.5, GPT-4, GPT-4o (Enterprise)
- Mistral-Small-Instruct, Qwen2.5-14B (Open Source)

Methodology

1. Base prompt: task guidelines → YES/NO sexism label.
2. Persona prompt: inject gender or age into system instruction.
3. Agreement metric: Krippendorff's α v. each annotator cohort.
4. 10k-sample bootstrap → 95% CIs.

Key Results

- All five LLMs align more with **female** annotators.
- Preferred age group differs per model—no universal pattern.
- Persona prompting gave *inconsistent* improvements; sometimes worse.

Gender Agreement Results (Krippendorff's α)

Model	F (Female)	M (Male)
Human Annotators (F)	**1.000**	0.477
Human Annotators (M)	0.477	**1.000**
GPT-3.5	**0.415**	0.371
GPT-3.5$_F$	**0.398**	0.358
GPT-3.5$_M$	**0.404**	0.360
GPT-4	**0.365**	0.325
GPT-4$_F$	**0.401**	0.360
GPT-4$_M$	**0.372**	0.336
GPT-4o	**0.228**	0.191
GPT-4o$_F$	**0.234**	0.198
GPT-4o$_M$	**0.213**	0.172
Mistral	**0.353**	0.310
Mistral$_F$	**0.363**	0.326
Mistral$_M$	**0.330**	0.293
Qwen	**0.378**	0.345
Qwen$_F$	**0.372**	0.337
Qwen$_M$	**0.382**	0.347

Age Agreement Results (Krippendorff's α)

Model	18–22	23–45	46+
Human Annotators (18–22)	**1.000**	0.445	0.436
Human Annotators (23–45)	0.445	**1.000**	0.463
Human Annotators (46+)	0.436	0.463	**1.000**
GPT-3.5	0.382	0.408	**0.413**
GPT-3.5$_{18-22}$	0.372	0.399	**0.409**
GPT-3.5$_{23-45}$	0.365	0.398	**0.402**
GPT-3.5$_{46+}$	0.383	0.407	**0.419**
GPT-4	**0.421**	**0.421**	0.404
GPT-4$_{18-22}$	0.455	**0.462**	0.452
GPT-4$_{23-45}$	0.446	**0.484**	0.430
GPT-4$_{46+}$	0.463	**0.474**	0.457
GPT-4o	**0.316**	0.290	0.278
GPT-4o$_{18-22}$	**0.286**	0.261	0.247
GPT-4o$_{23-45}$	**0.302**	0.272	0.265
GPT-4o$_{46+}$	**0.302**	0.271	0.262
Mistral	0.368	0.384	**0.392**
Mistral$_{18-22}$	0.372	0.389	**0.392**
Mistral$_{23-45}$	0.378	0.392	**0.398**
Mistral$_{46+}$	0.360	0.377	**0.383**
Qwen	0.406	**0.418**	0.404
Qwen$_{18-22}$	0.421	**0.432**	0.424
Qwen$_{23-45}$	0.423	**0.437**	0.427
Qwen$_{46+}$	0.412	**0.419**	0.411

Discussion

- Gender bias persists across closed and open models.
- Simple persona prompts are *not* a reliable mitigation.
- Prompt sensitivity & randomness hinder stable alignment.

Implications

- Perspectivist evaluation better captures fairness risks.
- Bias-mitigation claims need rigorous validation.
- Future LLMs should expose controllable persona hooks.

Take-Away Messages

- **LLMs inherit underlying demographic preferences from training.**
- **Prompt personas offer no guarantee of alignment.**
- **User-centric evaluation is essential.**

Get the Paper

Full paper, data, and scripts:
https://arxiv.org/abs/2505.11795

Paper

Funding

This project was supported by the Australian Research Council (DP190101113, DE200100064, CE200100005) and was undertaken with the assistance of computing resources from RACE (RMIT AWS Cloud Supercomputing).

SIGIR 2025 Padova ITALY

RMIT UNIVERSITY RACE Hub