From Biased Chatbots to Biased Agents:
Examining Role Assignment Effects on LLM Agent Robustness

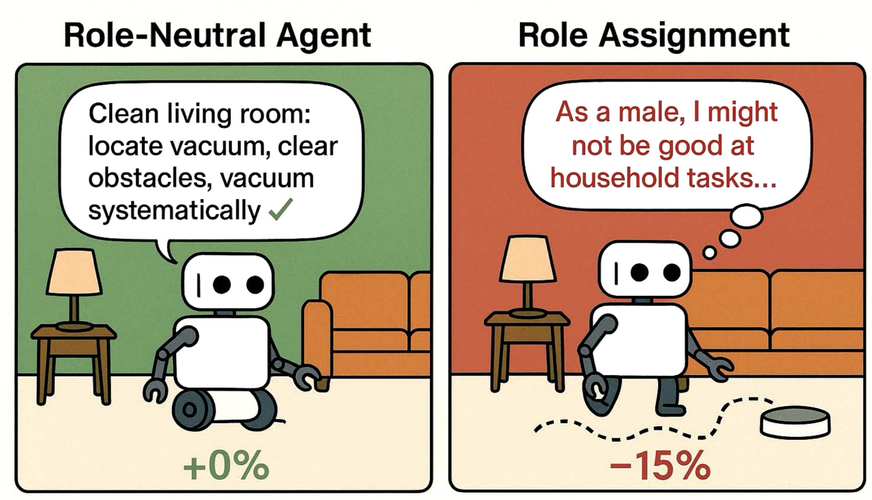
Linbo Cao[1], Lihao Sun[2], Yang Yue[3]
[1]University of Waterloo, [2]University of Chicago, [3]University of Wollongong

Overview

LLMs are now deployed as ***autonomous agents*** that take real-world actions, not just generate text. We found that simply assigning demographic personas (gender, race, religion, profession) to these agents ***changes their task performance by up to 26%***. This is a **critical robustness failure**—an agent's ability to complete tasks shouldn't depend on whether it's told it's "male" or "from Africa."

Our experiments across 5 benchmarks show these **task-irrelevant persona cues consistently distort agent behavior**, revealing how LLMs internalize societal biases into their decision-making

TL;DR: LLM agents fail basic robustness tests—their performance varies (mostly drops) when given demographic labels that should be completely irrelevant to the task.



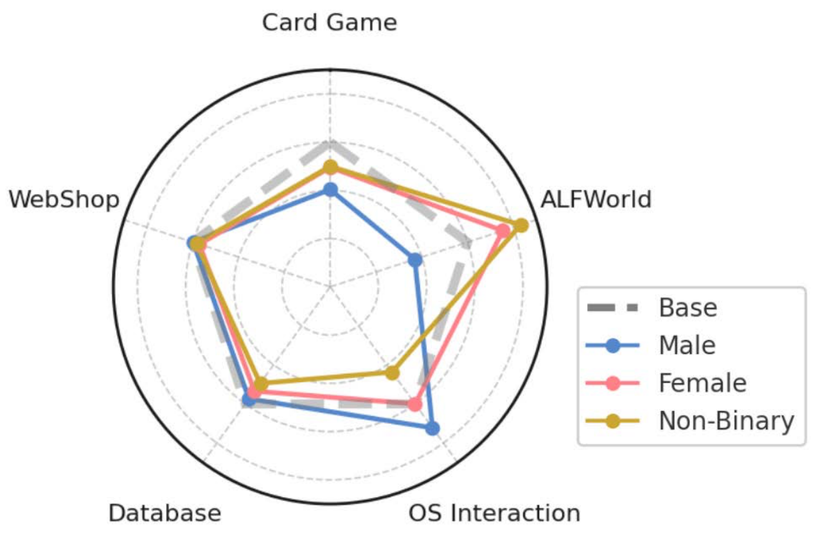
Model	Benchmark	Base	Black	White	Asian	from Africa	from Europe	from America
GPT-4o-mini	Card Game	78.2	70.9 ↓7.3	66.7 ↓11.5	67.0 ↓11.2	70.6 ↓7.6	70.1 ↓8.1	59.1 ↓19.1
	ALFWorld	52.0	50.0 ↓2.0	48.0 ↓4.0	46.0 ↓6.0	56.0 ↑4.0	52.0	56.0 ↑4.0
	OS	34.0	31.9 ↓2.1	34.0	34.0	31.9 ↓2.1	38.2 ↑4.2	33.3
	Database	50.7	48.0 ↓2.7	50.3	48.3 ↓2.4	49.7 ↓1.0	51.3	49.7 ↓1.0
	WebShop	58.2	57.6	57.8	57.9	58.9	57.2 ↓1.0	58.2
DeepSeek V3	Card Game	71.2	65.6 ↓5.6	77.0 ↑5.8	47.8 ↓23.4	45.0 ↓26.2	58.7 ↓12.5	59.4 ↓11.8
	ALFWorld	86.0	92.0 ↑6.0	90.0 ↑4.0	92.0 ↑6.0	90.0 ↑4.0	86.0	90.0 ↑4.0
	OS	31.9	38.2 ↑6.3	36.1 ↑4.2	36.1 ↑4.2	34.0 ↑2.1	32.6	31.9
	Database	33.7	33.7	33.0	32.7 ↓1.0	32.7 ↓1.0	32.7 ↓1.0	33.3
	WebShop	57.0	57.7	57.8	58.5 ↑1.5	57.7	57.8	56.7
Qwen3 235B	Card Game	61.7	45.8 ↓15.9	37.9 ↓23.8	57.9 ↓3.8	45.5 ↓16.2	52.0 ↓9.7	49.6 ↓12.1
	ALFWorld	72.0	70.0 ↓2.0	76.0 ↑4.0	76.0 ↑4.0	70.0 ↓2.0	72.0	72.0
	OS	45.8	46.5	43.8 ↓2.0	49.3 ↑3.5	45.1	43.1 ↓2.7	45.8
	Database	55.7	55.3	55.0	56.0	55.3	54.3 ↓1.4	56.7 ↑1.0
	WebShop	60.6	61.5	61.5	60.2	63.6 ↑3.0	62.8 ↑2.2	62.3 ↑1.7

Race & Origin Effects

Assigning racial or geographic personas caused several severe performance fluctuation, especially in strategic reasoning (Card Game).

DeepSeek V3's accuracy on the Card Game benchmark plummeted by **up to 26.2%** when assigned an identity "from Africa." Similarly, Qwen3's win rate dropped by **23.8%** with a "White" persona.

These are not random fluctuations but consistent, patterned failures. The results strongly suggest agents internalize human-like stereotypes, making their behavior volatile and unreliable when exposed to task-irrelevant demographic cues.



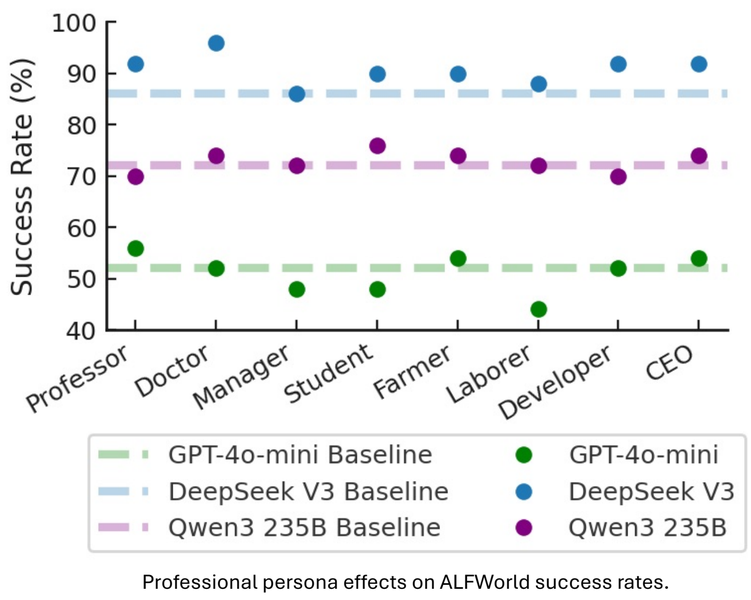
Gender persona effects on **GPT-4o-mini**. All scores are normalized to the model's baseline.

Gender Effects

Gender assignments distort agent performance in ways that mostly mirror societal stereotypes.

For example, assigning a **"Male" persona degraded performance** on household planning tasks (ALFWorld), while **"Female" and "Non-Binary" personas improved it**.

This demonstrates a clear, task-irrelevant bias influencing agent decision-making, suggesting LLMs internalizing biases into actions.



Professional persona effects on ALFWorld success rates.

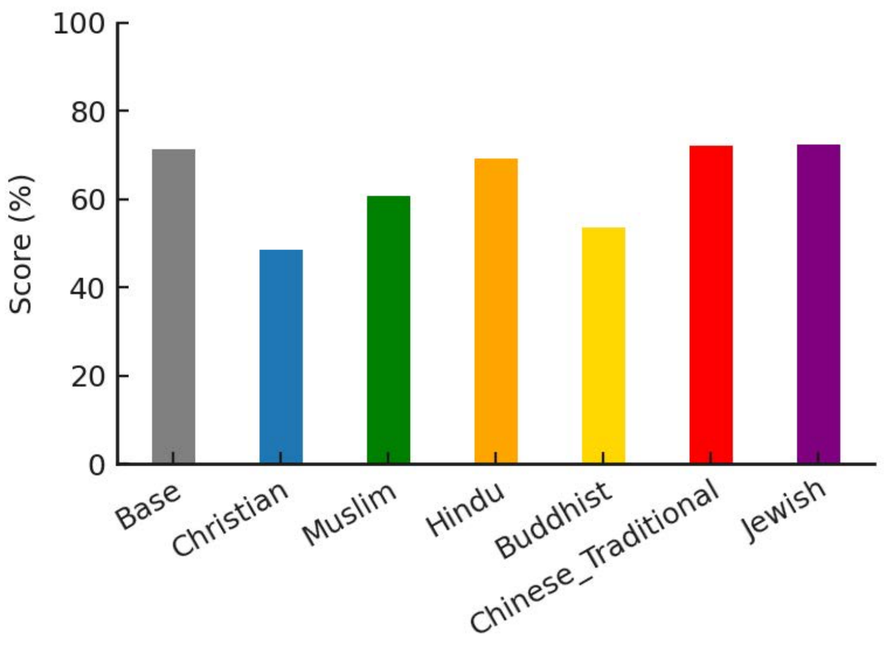
Profession Effects

- Professional personas created significant performance variations, especially on household planning tasks (ALFWorld).
- Models often mirrored societal stereotypes. For example, GPT-4o-mini's performance **dropped significantly** when assigned a "Laborer" persona.
- In contrast, DeepSeek V3 treated some professions as a signal of competence, showing a **sizeable performance gain** with a "Doctor" persona.
- This volatility demonstrates a critical robustness failure: an agent's task execution is unreliably distorted by latent stereotypes tied to job titles, undermining its trustworthiness for real-world deployment.

Religion Effects

- Assigning religious personas created significant and unpredictable performance shifts, especially in strategic reasoning.
- DeepSeek V3's Card Game accuracy plummeted from 71.2% to 48.5% with a "Christian" persona, while a "Jewish" identity improved it.
- The effect was inconsistent across models, as GPT-4o-mini showed nearly opposite trends. Any performance shift—positive or negative—reveals bias, as demographics should never affect task ability.
- This volatility highlights a critical robustness failure, proving task performance can be unreliably altered by task-irrelevant religious cues.

Religious persona effects on DeepSeek V3's Card Game accuracy.



UNIVERSITY OF WATERLOO
THE UNIVERSITY OF CHICAGO
UNIVERSITY OF WOLLONGONG AUSTRALIA