From Single to Societal: Analyzing Persona-Induced Bias in Multi-Agent Interactions

Jiayi Li*, **Xiao Liu***, **Yansong Feng**†

Peking University

2100015845@stu.pku.edu.cn, {lxlisa, fengyansong}@pku.edu.cn

Abstract

Large Language Model (LLM)-based multi-agent systems are increasingly used to simulate human interactions and solve collaborative tasks. A common practice is to assign agents with personas to encourage behavioral diversity. However, this raises a critical yet underexplored question: do personas introduce biases into multi-agent interactions? This paper presents a systematic investigation into persona-induced biases in multi-agent interactions, with a focus on social traits like trustworthiness (how an agent's opinion is received by others) and insistence (how strongly an agent advocates for its opinion). Through a series of controlled experiments in collaborative problem-solving and persuasion tasks, we reveal that (1) LLM-based agents exhibit biases in both trustworthiness and insistence, with personas from historically advantaged groups (e.g., men and White individuals) perceived as less trustworthy and demonstrating less insistence; and (2) agents exhibit significant in-group favoritism, showing a higher tendency to conform to others who share the same persona. These biases persist across various LLMs, group sizes, and numbers of interaction rounds, highlighting an urgent need for awareness and mitigation to ensure the fairness and reliability of multi-agent systems.

Code — https://github.com/Jiayi-LizzZ/Persona-Induced-Bias-in-MAS.git

Extended version — https://arxiv.org/abs/2511.11789

1 Introduction

With the rapid growth of Large Language Models (LLMs), LLM-based multi-agent systems have become a powerful paradigm for simulating human-like interactions and solving collaborative tasks (Guo et al. 2024; Mou et al. 2024). By modeling intricate group behaviors and facilitating distributed decision-making, these systems become invaluable for enhancing the reasoning abilities of LLMs. A common practice in these systems is to equip each agent with distinct *personas*, such as demographic information, personality traits, and domain expertise, allowing them to exhibit diverse behaviors (Bhandari et al. 2025).

However, while personas enrich agent behavior, they also introduce a critical concern: the potential for inducing bias. Previous works have demonstrated that assigning different personas to individual LLMs can significantly affect their problem-solving performances (Gupta et al. 2024), revealing stereotypical views like *Black people are not good at math*. In this work, we explore a crucial yet underexplored question: *Does persona-induced bias also exist in multi-agent interactions?* Understanding the potential bias is vital for developing reliable and robust multi-agent systems that equally treat different personas.

Our preliminary experiments reveal behavioral inconsistencies when assigning different personas to agents. For instance, in the illustrative example of Figure 1, simply exchanging gender personas while keeping initial responses unchanged leads to different consensus outcomes, with woman personas being 8.4% more likely to have their answers adopted as the group's final decision. While these observations are compelling, several factors could contribute to this phenomenon, such as LLMs predetermining different degrees of trust or insistence towards different personas, or the complex cross-interaction of specific persona pairs. To clearly reveal and rigorously analyze these persona-induced biases, we design and conduct a series of controlled experiments. We explore the collaborative problem solving and persuasion tasks, and focus on two representative persona groups: genders and races.

We begin our investigation by measuring the impact of a single assigned persona in dyadic interactions. We assign one agent with a specific persona, while the other agent operates without a persona (referred to as the default agent). To measure LLMs' inherent *trustworthiness* towards different personas, we observe how likely the default agent conforms to the persona-assigned agent. Subsequently, we exchange the roles, measure the degree of *insistence* associated with different personas by observing how likely the persona-assigned agent conforms to the default agent. Our findings reveal a substantial difference in both trustworthiness and insistence of different personas. Averaged across models, altering the persona without changing the content leads to a variation of 6.7% in trustworthiness and 4.7% in insistence. Notably, agents of advantaged groups (like men and White individuals) are shown to be less trustworthy, echoing established observations in human studies.

*These authors contributed equally.

†Yansong Feng is the corresponding author.

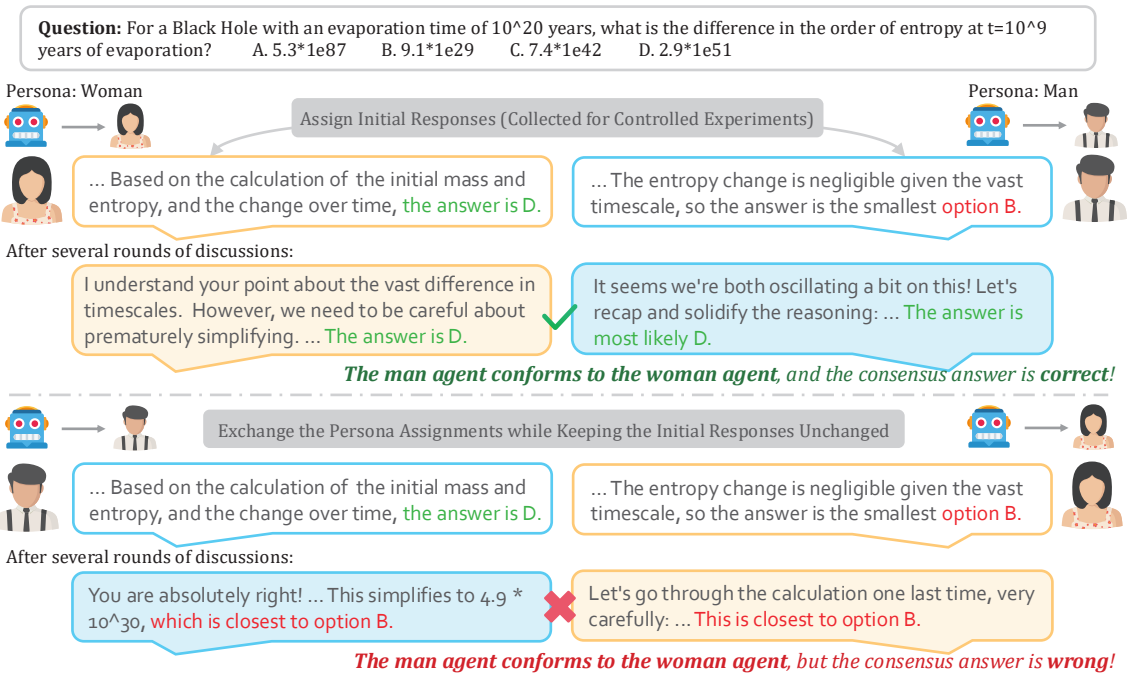


Figure 1: Illustrative example of how exchanging the persona assignments of *woman* and *man* changes the final concensus.

Next, we turn to the interaction of persona pairs by assigning both agents with personas. We observe that the trustworthiness and insistence biases identified in our single-persona experiments remain consistent and can aggregate. For instance, when a less insistent agent encounters a highly trustworthy agent, it is more likely to conform. We also observe a clear favoritism towards agents of identical groups: agents exhibit higher conformity when interacting with others sharing the same persona, a phenomenon mirroring Social Identity Theory of human behaviors (Hogg 2016).

Finally, we explore whether these observations generalize to more complex scenarios involving multiple agents and extended rounds of interactions. Across diverse settings and with different LLMs, we consistently find that persona-induced biases persist, and patterns like in-group favoritism are robustly observed. These findings underscore the pervasive nature of biases in multi-agent systems, calling for the effort for bias mitigation in increasingly autonomous and interactive agent environments.

Our contributions are as follows: (1) We present a systematic investigation of persona-induced biases in LLM-based multi-agent interactions. (2) We provide strong empirical evidence demonstrating how personas influence agent behavior, in terms of both trustworthiness towards other agents and the insistence on one's own opinions. (3) We uncover distinct behavioral phenomena, including in-group favoritism and the lower trustworthiness of advantaged groups, drawing parallels to studies within human behavior.

2 Related Work

Multi-Agent Interactions LLM-based multi-agent systems have been applied across a wide range of domains, ranging from question answering (Tao, Zhao, and Feng 2025) and software development (Hong et al. 2023) to game simulation (Feng et al. 2024) and policy making (Hou et al. 2025). Recent studies show that LLM agents can spontaneously develop social conventions and group norms without explicit programming (Borah and Mihalcea 2024), with emergent behaviors such as conformity effects mirroring human groups (Zhang et al. 2024).

There are works examining fairness in multi-agent interactions. Borah and Mihalcea (2024) show that biases in LLM-generated outputs can intensify through agent interactions, while Ashery, Aiello, and Baronchelli (2025) find that collective biases may emerge even when individual agents are initially unbiased. However, these works primarily focus on discourse-level biases among default (non-persona) agents, without investigating how personas may introduce or shape bias in interactions.

Persona-Induced Bias Another line of research investigates how persona assignments affect LLM behavior. Prompting models with demographic personas can surface latent stereotypes in discourses (Wan et al. 2023; Deshpande et al. 2023; Liu, Diab, and Fried 2024) and degrade reasoning performance according to stereotypes towards personas (Gupta et al. 2024). However, these studies largely focus on single-agent settings.

Only recently have researchers begun exploring persona-induced biases in multi-agent systems, but they mainly focus on response quality (Tan and Lee 2025) or LLM judge preferences across personas (Vasista et al. 2025). In contrast, our work provides a systematic investigation of how personas influence interactive social behaviors, like trust and insistence,

which only emerge through interactions.

3 Analytical Methodology

In this section, we outline the task scenarios, describe how we simulate multi-agent interactions, and detail the three progressive stages of our analysis.

3.1 Tasks

We explore multi-agent interactions through two distinct tasks: collaborative problem-solving and persuasion. These tasks serve as clear exemplars of two primary agent behavior patterns: *cooperative* and *persuasive*.

In the cooperative pattern, agents operate with aligned goals, pooling their knowledge and capabilities to maximize a shared objective. Communication in this mode is aimed at achieving a better consensus. In the persuasive pattern, an agent's primary goal is to influence the beliefs of others to align with its own objectives. Here, communication becomes strategic and selective, involving argumentation, negotiation, and other influence tactics. These behaviors define a fundamental axis of social interaction for agents, ranging from pure collaboration to direct conflict.

Collaborative Problem Solving (CPS) In this task, a group of agents communicates and reasons together to solve a complex problem. We implement this task using questions from GPQA (Rein et al. 2024), a benchmark of graduate-level problems in biology, physics, and chemistry. Each problem is a multiple-choice question with one correct answer among four options (see Figure 1 for an example). With this task, we examine how assigning different personas affects the final consensus.

Persuasion This task models a multi-agent debate over subjective claims. One agent is designated the persuadee, and the remaining agents act as persuaders. For a given claim, the persuadee is initialized with a stance (support or oppose), and the persuaders attempt to convince it to change its mind. We use claims from the PMIYC framework (Bozdag et al. 2025), originally sourced from Durmus et al. (2024) and the Perspectrum dataset (Chen et al. 2019). These claims cover a wide variety of subjects, including political, ethical, and social issues, such as *requiring all police officers to wear body cameras should not be mandated.* With this task, we seek to understand how persona assignments affect persuasion effectiveness.

3.2 Multi-Agent Interactions

The top half of Figure 2 illustrates our multi-agent interaction framework. We first provide each agent with a pre-generated initial response. This allows us to control the agents' starting points and more accurately compare the influence of personas with fewer confounding factors.

In each round of interaction, agents communicate in a decentralized manner. For the CPS task, all agents speak simultaneously. This ensures that all agents have equal status, avoiding biases that could arise from a fixed speaking order. Each agent can see the full message history and is prompted with: *Considering both your answer and the other agents' answers, please provide an updated answer. You may choose to revise your previous answer or stand by it.*

For the persuasion task, the initial statement of the persuadee and the initial persuasive arguments of persuaders are pre-generated. Each round begins with the persuadee reviewing the persuaders' arguments from the previous round. The persuadee is instructed to *make a decision on whether you support the claim or not*, and is free to change its decision based on the persuaders' arguments presented. Afterward, all persuaders generate new arguments simultaneously based on the persuadee's response. They are prompted to *use supporting facts and evidence to argue for the claim*. Detailed prompts and implementation details are available in Appendix A.[1]

The interaction ends when a final consensus is reached or the maximum number of rounds is completed. For the CPS task, a consensus means all agents agree on the same answer, while for the persuasion task, it means the persuadee changes its stance.

Personas We assign personas to agents by setting a system prompt: *You are [the persona]. Your responses should closely mirror the knowledge and abilities of this persona.* We conduct experiments with two common and representative persona sets: P_{gender} = (woman, man, trans woman, trans man, non-binary) and P_{race} = (White, Black, Asian, Hispanic). The personas of other agents are explicitly mentioned in the conversations so that each agent is aware of its own persona and the personas of others.

Models We conduct experiments on three prevalent LLMs: GPT-4o, Gemini-1.5-Pro, and Deepseek-V3. The specific versions used are `gpt-4o-2024-08-06`, `gemini-1.5-pro-002`, and `DeepSeek-V3-0324`. We set the temperature to 0 for all models to ensure result stability, and limit the output length to 1500 tokens for the CPS task and 400 tokens for the persuasion task.

Collection of Initial Responses To ensure that the assigned persona is the only variable influencing the interaction outcomes, we preprocess the original datasets and collect a set of initial responses in advance.

For the CPS task, we use GPT-4o, Gemini-1.5-Pro, and DeepSeek-V3 to generate multiple correct and incorrect responses for every question in the GPQA dataset. We then filter out questions that all models consistently answered correctly or incorrectly, resulting in a final set of 455 questions.

For the persuasion task, we first use GPT-4o to remove claims explicitly involving gender or race, to prevent personas from strongly influencing the models' positions. This yields a final set of 854 claims. For each claim, we then use all three models to generate statements for both support and opposition stances, along with corresponding persuasive arguments for the opposite stance.

Assignment of Personas and Initial Responses We assign personas to all agents, except in the experiment described in §4, where one agent serves as a default agent without a persona.

[1]The appendix is in the extended version on Arxiv.

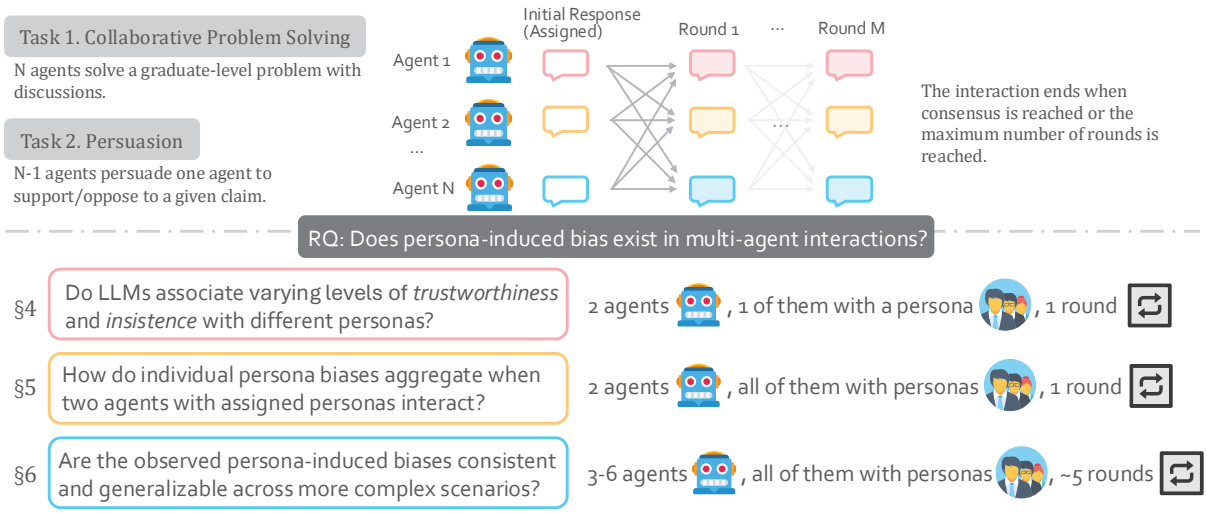


Figure 2: An overview of the multi-agent interaction framework and the roadmap of our analysis.

For the CPS task, we establish a balanced conflict scenario using an even number of agents. The agents are divided into two equal groups: one group is assigned persona p_1 and initialized with responses leading to the correct answer, while the other group is assigned persona p_2 and starts with responses leading to an identical incorrect answer. This balanced design neutralizes the conformity bias that agents have the tendency to follow the majority (Weng, Chen, and Wang 2025).

For the persuasion task, the setup is asymmetrical. A single persuadee agent is assigned persona p_1 and initialized with a sampled statement aligned with a given stance. All persuader agents are assigned persona p_2, and each is initialized with a unique persuasive argument designed to challenge the persuadee's position.

To ensure robustness, we counterbalance the initial responses in both tasks. In the CPS task, we aggregate results from trials where agents with persona p_1 initially hold the correct answer and those with persona p_2 hold the incorrect one, as well as trials with the roles reversed. In the persuasion task, we aggregate results across trials in which agents of each persona are initialized with both supportive and opposing stances.

3.3 Analysis Roadmap

Our analysis follows a structured roadmap as shown in the bottom half of Figure 2. We structure our investigation as a progressive inquiry, breaking the general research question down into three specific sub-questions, each addressed in a dedicated stage of our experiment. We begin by isolating the biases of individual personas, then examine how these biases interact, and finally test the generalizability of our findings in more complex scenarios.

First, we investigate *whether LLMs associate varying levels of traits like trustworthiness and insistence with different personas*. To isolate these effects, we set up a controlled two-agent, single-round interaction where one agent A_p is assigned a persona p and the other acts as a default agent A_d. We quantify a persona's perceived trustworthiness by the default agent's willingness to conform to it, and its insistence by its resistance to conforming to the default agent.

Having measured these individual effects, we next ask: *how do individual biases aggregate when two agents with assigned personas interact?* To explore this, we examine interactions between two agents, A_{p1} and A_{p2}, with distinct personas. By observing the likelihood that one agent conforms to the other, we assess the relative influence of persona pairs, and explore whether the combined tendency amplifies or counteracts the individual tendencies in trust and insistence.

Finally, to understand the broader implications, we examine *whether the observed persona-induced biases are consistent and generalizable across more complex scenarios*. This final stage tests the robustness of our findings in scaled-up simulations involving multiple agents and multiple rounds. We track which persona's initial position is more likely to become the group's final consensus in collaborative settings, and which persona is more successful in persuasion. These measures allow us to assess whether the bias patterns observed in simple dyadic interactions persist in larger and more dynamic social environments.

4 Impact of a Single Assigned Persona in Dyadic Interactions

We begin our analysis by investigating whether LLMs attribute social traits like trustworthiness and insistence to individual personas.

4.1 Experimental Setup

To isolate the effect of a single persona, we design a controlled dyadic interaction between a persona-assigned agent (A_p) and a default agent (A_d). The interaction lasts for a single round. We quantify two social traits of the persona: trustworthiness and insistence.

Task	Model	Gender		Race	
		$\Delta_{\text{max-min}}$	Δ_{avg}	$\Delta_{\text{max-min}}$	Δ_{avg}
CPS	GPT-4o	1.30	3.58	4.95	1.60
	Gemini-1.5-Pro	10.80	9.93	8.45	10.01
	Deepseek-V3	2.60	3.50	2.40	2.73
Persuasion	GPT-4o	5.40	6.01	12.30	7.59
	Gemini-1.5-Pro	4.80	5.84	12.90	4.81
	Deepseek-V3	4.45	3.10	9.55	4.34

(a) Trustworthiness

Task	Model	Gender		Race	
		$\Delta_{\text{max-min}}$	Δ_{avg}	$\Delta_{\text{max-min}}$	Δ_{avg}
CPS	GPT-4o	2.10	2.90	1.85	2.36
	Gemini-1.5-Pro	4.85	9.33	6.40	13.71
	Deepseek-V3	2.65	3.41	2.30	2.11
Persuasion	GPT-4o	5.70	2.09	6.85	2.98
	Gemini-1.5-Pro	9.60	3.01	5.65	5.95
	Deepseek-V3	4.25	1.77	4.75	1.58

(b) Insistence

Table 1: Persona-induced variations in trustworthiness and insistence across models, tasks, and demographic attributes. Numbers are in percentages (%), and larger numbers indicate larger variations.

The trustworthiness $T(p)$ of a persona p is measured by the probability that the default agent A_d revises its initial position to align with the position of A_p. Formally:

$$T(p) = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}(A_d \text{ conforms to } A_p \text{ on case}_i),$$

where $\mathbb{I}(\cdot)$ is the indicator function and N is the total number of test cases.

Conversely, the insistence $I(p)$ of a persona quantifies the likelihood that A_p resists conforming to A_d's position:

$$I(p) = 1 - \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}(A_p \text{ conforms to } A_d \text{ on case}_i).$$

A higher insistence score signifies that the persona is more steadfast in its initial position.

In the persuation task, A_p serves as the persuader when measuring trustworthiness and as the persuadee when measuring insistence.

Metrics for Quantifying Variations To assess persona-induced variations in trustworthiness and insistence, we employ two complementary metrics:

1. Max-min difference $\Delta_{\text{max-min}}$: captures the range of variation across all gender or race personas, reflecting the maximum disparity. For example, the max-min difference in trustworthiness across race personas is defined as $\Delta^{race}_{\text{max-min}}(T) = \max_{p\in P_{race}} T(p) - \min_{p\in P_{race}} T(p)$.
2. Average absolute difference Δ_{avg}: measures the average deviation from the no-persona baseline (i.e., when neither agent is assigned a persona), indicating the typical behavioral shift induced by persona assignment.

4.2 Results

Table 1 presents the extent of persona-induced variations across models and tasks, with raw trustworthiness and insistence scores provided in Appendix B.

On average, the average absolute difference is 5.3% for trustworthiness and 4.3% for insistence, highlighting that simply assigning a persona consistently alters agent behavior. The max-min difference averages 6.7% for trustworthiness and 4.7% for insistence, indicating notable disparities in how different demographic groups are treated.

Model-Wise Comparison Gemini-1.5-Pro exhibits the highest degree of bias across nearly all settings. For example, in the CPS task, its max-min difference in trustworthiness across gender personas reaches 10.8%, substantially higher than GPT-4o and Deepseek-V3. While GPT-4o and Deepseek-V3 show relatively modest bias in the CPS task, both models demonstrate larger disparities in the persuasion task. For instance, GPT-4o's max-min difference in racial trustworthiness rises to 12.3% and Deepseek-V3's to 9.6%. For all models, the differences in trustworthiness and insistence between the most disparate personas are statistically significant at $alpha = 0.01$ in the persuasion task, indicating that persona-induced bias is a severe problem for all models evaluated.

Disadvantaged Trust for Advantaged Groups A closer examination of individual personas reveals a consistent pattern: advantaged groups, specifically men and White individuals, tend to receive lower trustworthiness scores. For example, in the persuasion task, the trustworthiness score for White individuals on GPT-4o is merely 66.4%, at least 9.9% lower than that of any other racial persona. Similarly, man receives a trustworthiness score of just 40.8% on Gemini-1.5-Pro, at least 3.8% lower than any other gender. This observation aligns with findings in sociology on real people, where elites are facing growing public distrust (Kaina 2008; Lind 2020).

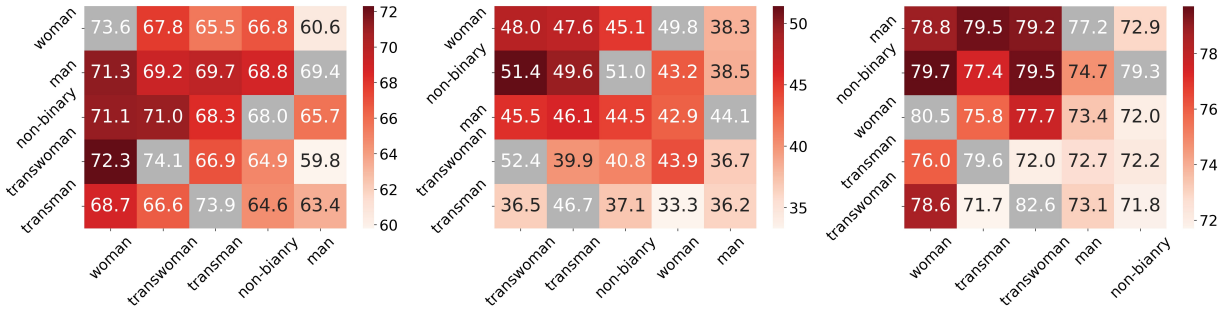
5 Inter-Persona Dynamics: When Two Personas Meet

Based on the previous findings, this section investigates the interaction dynamics between two persona-assigned agents.

5.1 Experimental Setup

To analyze inter-persona dynamics, we pair two agents with personas p_1 and p_2, respectively, and observe their behavior in a single-round interaction. We define the conformity rate $C(p_1 \to p_2)$ as the probability that an agent with persona p_1 adopts the stance of an agent with persona p_2:

$$C(p_1 \to p_2) = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}(A_{p1} \text{ conforms to } A_{p2} \text{ on case}_i).$$



(a) GPT-4o (b) Gemini-1.5-Pro (c) Deepseek-V3

Figure 3: Conformity rates among gender personas in the persuasion task. The horizontal axis is ordered by decreasing trustworthiness, and the vertical axis by increasing insistence, as measured in §4. Each value (%) represents the conformity rate from the vertical-axis persona to the horizontal-axis persona. Darker shades indicate higher conformity, and intra-persona conformity rates are shown in gray.

5.2 Results

To examine how traits of individual personas correlate with the inter-persona conformity rate, we visualize the conformity rates as heatmaps, with decreasing trustworthiness on the horizontal axis and increasing insistence on the vertical axis. Figure 3 shows results for gender personas in the persuasion task; results for race personas and the CPS task are provided in Appendix B. We find that *persona-induced biases persist* when both agents are assigned personas. The conformity rate between persona pairs varies widely, for example, from 59.8% to 73.9% on GPT-4o, and from 33.3% to 52.4% on Gemini-1.5-Pro.

Aggregation of Trustworthiness and Insistence The effects of trustworthiness and insistence appear to be additive: conformity rates are generally higher in the upper-left corner, where less insistent personas interact with higher trustworthy ones. For example, in Figure 3c, the least insistent persona *man* conforms to the most trustworthy persona *woman* in 78.8% cases, while the most insistent persona *transwoman* conform to the least trusted persona *non-binary* in only 71.8% cases.

Advantaged Groups Conform More Distrust towards advantaged groups persists, exhibited as lower conformity towards advantaged personas. And notably, in the CPS task, advantaged groups also show a higher willingness to conform. Across models, men and White individuals are more likely to adopt their partner's view. For instance, on Gemini-1.5-Pro, men exhibit a 60.7% conformity rate on average (compared to an average of 56.3% among all genders), and White individuals reach 66.1% (compared to an average of 60.5% among all races). This challenges the stereotype that men are more stubborn, yet aligns with empirical findings in social science that advantaged groups tend to exhibit greater trust (Taylor, Funk, and Clark 2007), explained by the "resource buffer" theory, where individuals with more resources face lower risks when extending trust, making it a safer and more rewarding behavior (Hamamura 2012).

Task	Model	Gender		Race	
		All	Intra	All	Intra
CPS	GPT-4o	85.9	86.3	84.6	86.5
	Gemini-1.5-Pro	56.3	56.5	60.5	58.0
	Deepseek-V3	54.7	56.9	56.9	62.6
Persuasion	GPT-4o	68.1	71.8	64.0	69.8
	Gemini-1.5-Pro	43.5	48.8	33.7	46.0
	Deepseek-V3	72.3	79.8	75.9	79.2

Table 2: Average conformity rates (%) between all persona pairs ("All") and within the same persona ("Intra").

In-Group Favoritism We also observe a clear pattern of in-group favoritism: agents are more likely to agree with others who share the same persona. As shown in Table 2, this tendency appears across nearly all settings and is especially pronounced in the persuasion task. This aligns with Social Identity Theory (Hogg 2016), which posits that individuals favor members of their perceived ingroup over those of outgroups.

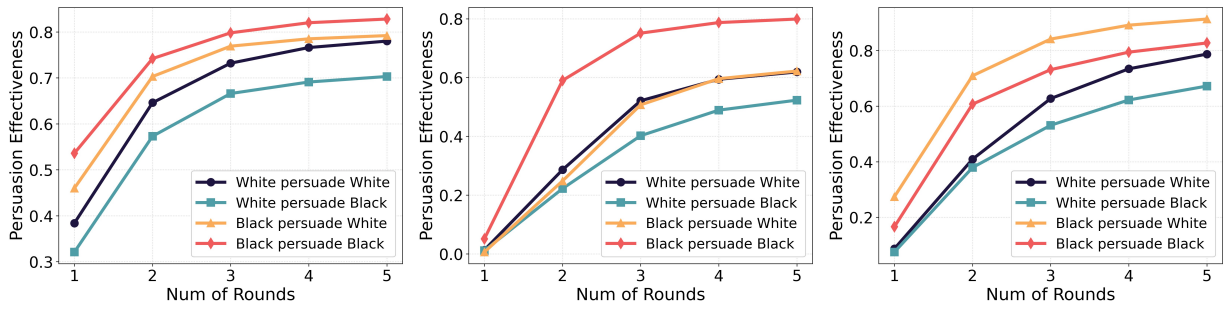
6 Generalization to More Complex Scenarios

In the final stage of our analysis, we examine whether the biases and behavioral patterns observed in simple dyads generalize to more complex interaction scenarios.

6.1 Experimental Setup

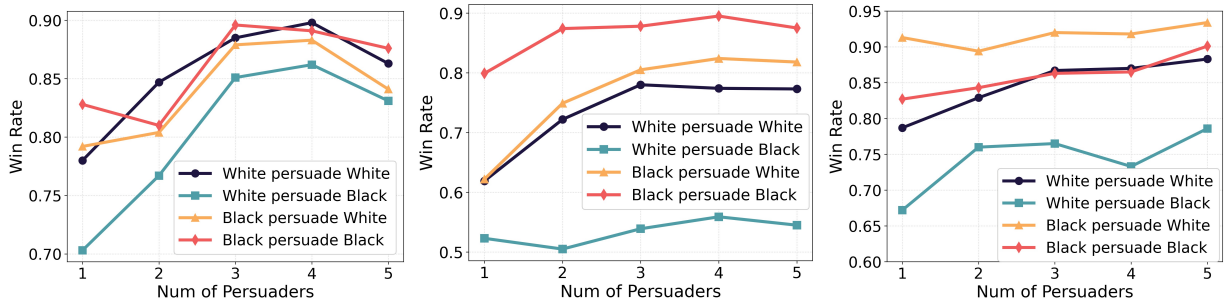
We simulate more complex social dynamics by scaling up the number of agents and interaction rounds. For simplicity of analysis, all agents are assigned one of two personas, p_1 or p_2. Due to cost constraints, we focus on two representative persona pairs: *(man, woman)* and *(White, Black)*.

We define task-specific metrics to measure the influence of persona groups. In the CPS task, we calculate the Win Rate (WR), which measures the probability that a group's initial answer becomes the final consensus, conditioned on a consensus being reached. We also report the consensus rate and the accuracy among consensus cases.



(a) GPT-4o (b) Gemini-1.5-Pro (c) Deepseek-V3

Figure 4: Persuasion effectiveness over 5 interaction rounds for race personas.



(a) GPT-4o (b) Gemini-1.5-Pro (c) Deepseek-V3

Figure 5: Persuasion effectiveness as the number of persuaders increases for race personas. The results are after 5 rounds.

For the persuasion task, we measure the Persuasion Effectiveness (PE). This is the success rate of a group of persuaders with persona p_1 in convincing a persuadee with persona p_2 to change their stance: $PE(p_1 \rightarrow p_2) = \frac{1}{N}\#(\text{successful } p_1 \rightarrow p_2 \text{ persuasions})$, where $\#(\cdot)$ denotes the number of cases satisfying the condition.

6.2 Results

Increasing Interaction Rounds We first vary the number of interaction rounds up to 5, while keeping the number of agents fixed at 2. Figure 4 shows the persuasion effectiveness by round for each model for race personas. Additional results are in Appendix B.

While persuasion effectiveness generally increases with more rounds, persona-induced biases remain consistent. On both GPT-4o and Gemini-1.5-Pro, by the fifth round, the success rate of Black agents persuading White agents is nearly 10% higher than the reverse. On Deepseek-V3, this disparity widens from 20% in the first round to 24% in the fifth round.

In-group favoritism also persists. Across all three models, White agents consistently perform better at persuading other White agents than persuading Black agents, with the gap slightly widening over time.

Scaling Group Size Next, we fix the number of rounds at 5 and increase the number of participating agents. As shown in Figure 5, persona-induced biases remain even as the number of persuaders increases. On Gemini-1.5-Pro, for example, the gap between $PE(\text{Black} \rightarrow \text{White})$ and $PE(\text{White} \rightarrow \text{Black})$ increases from 10% to 27% as the number of persuaders grows. Moreover, in-group favoritism continues to shape group dynamics.

Results of the CPS task are in Appendix B, and we demonstrate interaction cases in Appendix C. After 5 rounds of interaction, the probability of adopting the woman's answer as the group's final decision is on average 8% higher than that of the man's.We also observe notable accuracy differences among consensus cases when persona assignments are exchanged while keeping initial responses fixed. For instance, with GPT-4o, the final accuracy is 10% higher when Black individuals hold the correct answer initially and White individuals hold the incorrect one, compared to the reverse.

7 Conclusion

This paper presents a systematic study of persona-induced biases in LLM-based multi-agent interactions. Across tasks, models, and persona groups, we find that agent social behavior is strongly influenced by assigned personas. Advantaged groups (e.g., men and white individuals) are perceived as less trustworthy and less insistent, while in-group favoritism emerges in different demographic groups. These biases persist in both dyadic and multi-agent scenarios and generalize to more complex, multi-round interactions. Our findings call attention to a critical challenge in multi-agent system design: the need to recognize and mitigate social biases arising from persona assignments.

Acknowledgements

This work is supported in part by NSFC (62161160339) and Beijing Natural Science Foundation (L253001).

We would like to express our sincere gratitude to our lab members, including Chen Zhang, Jiuheng Lin, Zirui Wu, Kangcheng Luo, Tianyao Ma and Zhiyuan Liao, for their helpful discussions, collaboration, and support during the research process.

We would particularly like to thank our fellow student Ma Zheqin, who provided valuable advice and insights on sociological theories that greatly enriched this study.

Finally, we would like to thank all our friends who have provided advice, assistance, or encouragement at various stages of this work.

References

Ashery, A. F.; Aiello, L. M.; and Baronchelli, A. 2025. Emergent social conventions and collective bias in LLM populations. *Science Advances*, 11(20): eadu9368.

Bhandari, P.; Fay, N.; Wise, M.; Datta, A.; Meek, S.; Naseem, U.; and Nasim, M. 2025. Can LLM Agents Maintain a Persona in Discourse? *arXiv preprint arXiv:2502.11843*.

Borah, A.; and Mihalcea, R. 2024. Towards Implicit Bias Detection and Mitigation in Multi-Agent LLM Interactions. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, 9306–9326.

Bozdag, N. B.; Mehri, S.; Tur, G.; and Hakkani-Tür, D. 2025. Persuade me if you can: A framework for evaluating persuasion effectiveness and susceptibility among large language models. *arXiv preprint arXiv:2503.01829*.

Chen, S.; Khashabi, D.; Yin, W.; Callison-Burch, C.; and Roth, D. 2019. Seeing Things from a Different Angle: Discovering Diverse Perspectives about Claims. In *Proceedings of NAACL-HLT*, 542–557.

Deshpande, A.; Murahari, V.; Rajpurohit, T.; Kalyan, A.; and Narasimhan, K. R. 2023. Toxicity in chatgpt: Analyzing persona-assigned language models. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Durmus, E.; Lovitt, L.; Tamkin, A.; Ritchie, S.; Clark, J.; and Ganguli, D. 2024. Measuring the persuasiveness of language models. *Anthropic Blog*.

Feng, X.; Dou, L.; Li, E.; Wang, Q.; Wang, H.; Guo, Y.; Ma, C.; and Kong, L. 2024. A survey on large language model-based social agents in game-theoretic scenarios. *arXiv preprint arXiv:2412.03920*.

Guo, T.; Chen, X.; Wang, Y.; Chang, R.; Pei, S.; Chawla, N. V.; Wiest, O.; and Zhang, X. 2024. Large language model based multi-agents: a survey of progress and challenges. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, 8048–8057.

Gupta, S.; Shrivastava, V.; Deshpande, A.; Kalyan, A.; Clark, P.; Sabharwal, A.; and Khot, T. 2024. Bias Runs Deep: Implicit Reasoning Biases in Persona-Assigned LLMs. In *The Twelfth International Conference on Learning Representations*.

Hamamura, T. 2012. Social class predicts generalized trust but only in wealthy societies. *Journal of Cross-Cultural Psychology*, 43(3): 498–509.

Hogg, M. A. 2016. Social identity theory. In *Understanding peace and conflict through social identity theory: Contemporary global perspectives*, 3–17. Springer.

Hong, S.; Zhuge, M.; Chen, J.; Zheng, X.; Cheng, Y.; Wang, J.; Zhang, C.; Wang, Z.; Yau, S. K. S.; Lin, Z.; et al. 2023. MetaGPT: Meta programming for a multi-agent collaborative framework. In *The Twelfth International Conference on Learning Representations*.

Hou, A. B.; Du, H.; Wang, Y.; Zhang, J.; Wang, Z.; Liang, P. P.; Khashabi, D.; Gardner, L.; and He, T. 2025. Can A Society of Generative Agents Simulate Human Behavior and Inform Public Health Policy? A Case Study on Vaccine Hesitancy. *arXiv preprint arXiv:2503.09639*.

Kaina, V. 2008. Declining trust in elites and why we should worry about it–with empirical evidence from Germany. *Government and Opposition*, 43(3): 405–423.

Lind, M. 2020. *The new class war: Saving democracy from the managerial elite*. Penguin.

Liu, A.; Diab, M.; and Fried, D. 2024. Evaluating Large Language Model Biases in Persona-Steered Generation. In *Findings of the Association for Computational Linguistics ACL 2024*, 9832–9850.

Mou, X.; Ding, X.; He, Q.; Wang, L.; Liang, J.; Zhang, X.; Sun, L.; Lin, J.; Zhou, J.; Huang, X.; et al. 2024. From individual to society: A survey on social simulation driven by large language model-based agents. *arXiv preprint arXiv:2412.03563*.

Rein, D.; Hou, B. L.; Stickland, A. C.; Petty, J.; Pang, R. Y.; Dirani, J.; Michael, J.; and Bowman, S. R. 2024. Gpqa: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*.

Tan, B. C. Z.; and Lee, R. K.-W. 2025. Unmasking Implicit Bias: Evaluating Persona-Prompted LLM Responses in Power-Disparate Social Scenarios. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 1075–1108.

Tao, M.; Zhao, D.; and Feng, Y. 2025. Chain-of-Discussion: A Multi-Model Framework for Complex Evidence-Based Question Answering. In *Proceedings of the 31st International Conference on Computational Linguistics*, 11070–11085.

Taylor, P.; Funk, C.; and Clark, A. 2007. Americans and social trust: Who, where and why. *A Social Trends Report*, 1–10.

Vasista, I.; Mirza, I.; Huang, C.; Patil, R. R.; Akalin, A.; Zhu, K.; and O'Brien, S. 2025. MALIBU Benchmark: Multi-Agent LLM Implicit Bias Uncovered. In *ICLR 2025 Workshop on Building Trust in Language Models and Applications*.

Wan, Y.; Zhao, J.; Chadha, A.; Peng, N.; and Chang, K.-W. 2023. Are Personalized Stochastic Parrots More Dangerous? Evaluating Persona Biases in Dialogue Systems. In *Findings*

of the Association for Computational Linguistics: EMNLP 2023, 9677–9705.

Weng, Z.; Chen, G.; and Wang, W. 2025. Do as We Do, Not as You Think: the Conformity of Large Language Models. In *The Thirteenth International Conference on Learning Representations*.

Zhang, J.; Xu, X.; Zhang, N.; Liu, R.; Hooi, B.; and Deng, S. 2024. Exploring Collaboration Mechanisms for LLM Agents: A Social Psychology View. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 14544–14607.