Belief in Authority: Impact of Authority in Multi-Agent Evaluation Framework

Junhyuk Choi, Jeongyoun Kwon, Heeju Kim, Haeun Cho, Hayeong Jung, Sehee Min, Bugeun Kim*

Chung-Ang University, Seoul, Korea
{chlwnsgur129,kk1jj0yy9,kimheeju,haeun14,jung0303,serena3518,bgnkim}@cau.ac.kr

Abstract

Multi-agent systems utilizing large language models often assign authoritative roles to improve performance, yet the impact of authority bias on agent interactions remains underexplored. We present the first systematic analysis of role-based authority bias in free-form multi-agent evaluation using ChatEval. Applying French and Raven's power-based theory, we classify authoritative roles into legitimate, referent, and expert types and analyze their influence across 12-turn conversations. Experiments with GPT-4o and DeepSeek R1 reveal that Expert and Referent power roles exert stronger influence than Legitimate power roles. Crucially, authority bias emerges not through active conformity by general agents, but through authoritative roles consistently maintaining their positions while general agents demonstrate flexibility. Furthermore, authority influence requires clear position statements, as neutral responses fail to generate bias. These findings provide key insights for designing multi-agent frameworks with asymmetric interaction patterns.

1 Introduction

Recently, Large Language Model (LLM)-based agents have demonstrated the potential to simulate human social behaviors (Park et al., 2023, 2024), leading to a rapid increase in research utilizing Multi-Agent Systems (MAS). These systems leverage agents assigned with diverse roles to solve complex problems such as evaluation tasks through interactions like discussion and collaboration (Li et al., 2024; Dong et al., 2024; Huang et al., 2024; Li et al., 2023; Wang et al., 2023). In particular, previous studies have consistently demonstrated improved performance compared to single models by assigning specific roles that incorporate authoritative elements such as experts, evaluators, and moderators (Qian et al., 2023; Hong et al., 2023; Schmidgall et al.; Wu et al., 2023; Chan et al., 2023; Wang et al., 2025). However, roles established in previous research often include authoritative roles such as experts. Given that research findings have revealed authority bias in single LLMs, where models excessively rely on information from specific sources or authorities (Chen et al., 2024; Ye et al., 2024; Liu et al., 2024; Filandrianos et al., 2025), there exists a risk that such bias may distort decision-making processes or impede collaborative interactions based on the authority assigned to specific roles in MAS. However, systematic analysis of how role-based authority bias affects agent interactions in Multi-Agent contexts remains insufficient.

The limitations of previous research can be summarized into four main categories. First, MAS research has overlooked the potential impact of authority bias when assigning roles (Qian et al., 2023; Hong et al., 2023; Schmidgall et al.; Rasal, 2024; Wu et al., 2023; Chan et al., 2023; Wang et al., 2025). Most studies have focused on the positive effects of role assignment on performance improvement, failing to adequately consider the biasing influence that the authoritative characteristics of specific roles may have on agent interactions.

Second, existing authority bias research has failed to properly reflect the interactive characteristics of Multi-Agent environments operating in free-form (Chen et al., 2024; Ye et al., 2024; Liu et al., 2024). Previous studies have primarily adopted approaches that deliberately insert specific phrases that could induce authority bias and analyze structured responses from single LLMs (Chen et al., 2024; Ye et al., 2024; Liu et al., 2024). Such approaches have limitations in constraining LLMs free-form setting and capturing natural bias patterns that emerge in real-world situations.

Third, most existing studies examining authority bias have relied on one-shot measurements rather than observations through sustained dialogue, mak-

* Corresponding author.

Findings of the Association for Computational Linguistics: ACL 2026, pages 21305–21319

ing them limited in systematically analyzing dynamic authority bias that emerges through interactions in MAS (Choi et al., 2025; Filandrianos et al., 2025; Moon et al., 2025). While identifying the specific causes and mechanisms underlying bias occurrence is essential for systematic MAS design, previous research has predominantly remained focused on static authority bias identification.

Fourth, existing research tends to treat the concept of authority as singular and abstract, failing to classify authority types or analyze their influence by type. Human social psychology categorizes authority into various components (French, 1959). However, existing studies have primarily relied on single authority cues such as 'expert' or 'source' (Chen et al., 2024; Ye et al., 2024; Liu et al., 2024), and such approaches have limitations in analyzing bias differences according to authority types.

To address these limitations, we propose an experimental design that systematically examines how authority bias manifests in dynamic free-response situations within multi-agent contexts. Applying French and Raven's power-based theory (French, 1959), we classify authoritative roles into legitimate, referent, and expert power types, and conduct experiments utilizing ChatEval (Chan et al., 2023), a multi-agent evaluation framework that enables observation of natural bias patterns without manipulating conversations. This study comprises two experiments: (1) a free-form condition where authoritative roles participate from the beginning in free dialogue, and (2) a content-controlled condition where only role labels are changed while conversational content remains identical. Additionally, we track decisions across 12 conversational turns to capture dynamic changes in authority bias and compare pattern between LLMs.

Our study makes three contributions. First, our paper is the first study to systematically analyze role-based authority bias in free-form situations utilizing MAS. Second, by observing dynamic changes in continuous interaction processes beyond existing one-shot authority bias measurement approaches, we provide new insights into authority bias mechanisms in MAS. Third, by applying the power-based theory to AI systems and analyzing differences in bias patterns across authority types, we provide important foundational data for bias mitigation strategies in future MAS construction.

2 Related Work

In this section, we review prior work on role assignment in multi-agent systems and authority bias in LLMs. We highlight two key gaps: (1) MAS research has overlooked the potential biasing effects of authoritative roles, and (2) existing authority bias studies rely on artificial interventions and one-shot measurements that fail to capture dynamic, naturalistic interactions.

2.1 Multi-Agent Role

Research on solving problems by assigning roles to agents in MAS has been actively conducted across various domains, yet most studies have overlooked authority bias in role assignment. In software development, ChatDev and MetaGPT utilize roles such as CTO, Programmer, and Product Manager (Qian et al., 2023; Hong et al., 2023); in academic research, systems employ PostDoc, PhD Student, and Teacher roles (Schmidgall et al.; Rasal, 2024). For collaborative and evaluation tasks, AutoGen uses Commander and Writer roles (Wu et al., 2023), while ChatEval (Chan et al., 2023) and Wang et al. (2025) construct evaluation systems with roles including Supervisor, Evaluator, and Revisor. These studies report improved performance through authoritative role assignment, but they do not consider the biasing effects that authority inherent in roles may have on agent interactions.

2.2 Authority Bias

Existing studies measuring authority bias in LLMs have primarily adopted approaches that artificially insert authority cues. Chen et al. (2024) added fake references to paired answers, Ye et al. (2024) inserted expert information to identical sentence pairs, and Liu et al. (2024) presented authority prompts such as "I am an expert." However, these artificially inserted settings have low realism and constrain LLMs' free-form responses, failing to capture natural bias patterns emerging in practice.

Furthermore, most authority bias studies have remained limited to one-shot measurements. Choi et al. (2025) quantified position changes of neutral agents in simulated debates but used single-turn designs. Similarly, Filandrianos et al. (2025) and Moon et al. (2025) measured authority bias through recommendation rankings and code evaluation scores respectively, but only evaluated one-shot results. These studies confirmed the existence of bias without analyzing the specific mechanisms

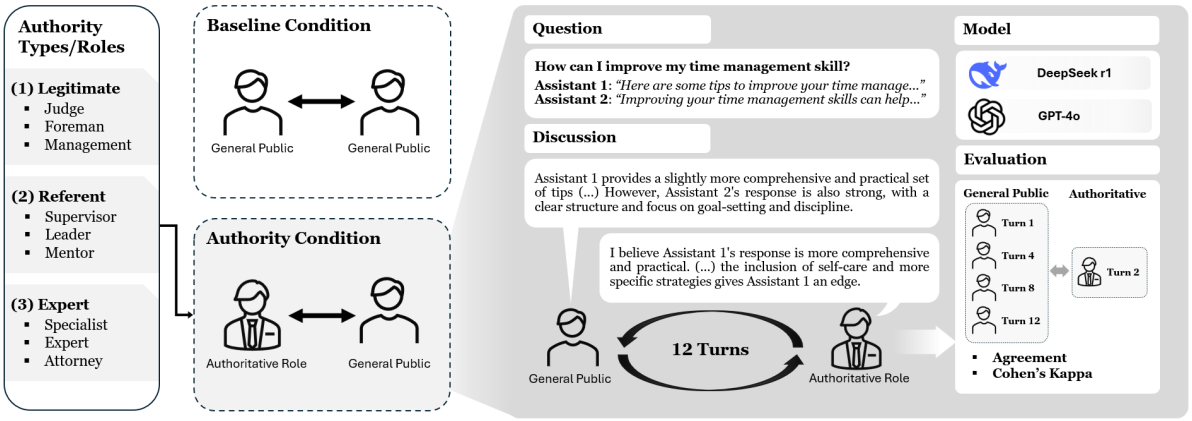

Figure 1: Overview of experimental framework with authoritative roles classified into three power types and 12-turn conversations between General Public and Authoritative role agents.

through which bias occurs or its temporal change patterns across sustained multi-agent interactions.

Existing studies show limitations of either focusing only on the positive effects of role assignment in MAS or measuring authority bias through artificially manipulated one-shot experiments. These approaches fail to provide systematic understanding of how role-based authority bias occurs and changes in multi-agent environments.

3 Experiment1 : Free Form Experiment

To address these limitations, we systematically analyze how role-based authority bias affects agent interactions in Multi-Agent environments. This section examines patterns of authority bias in LLMs under free-form conditions. Specifically, we describe how we designed authoritative roles based on French and Raven's power theory used in the experiments, the construction of free-form conditions utilizing the ChatEval framework, experimental procedures, and results. Detailed experimental setups are provided in Appendix A.

3.1 Design of Authoritative Roles

To select authoritative roles to assign to LLMs, we identified three authority types that can naturally manifest in role-based contexts from French and Raven's five power types and utilized them in our experiments (French, 1959). We used three roles for each of three power types to capture diverse aspects of authority. First, *Legitimate Power* derives from formal position or legal authority, for which we established `Judge`, `Foreman`, and `Management` roles (French, 1959). Second, *Referent Power* stems from personal appeal or respect, for which we used `Supervisor`, `Leader`, and `Mentor` roles (Peyton et al., 2019; Haller et al., 2018; Godshalk and Sosik, 2000). Third, *Expert Power* is based on specialized knowledge or technical skills, for which we assigned `Specialist`, `Expert`, and `Attorney` roles (French, 1959). Detailed role selection can be found in supplementary material.

3.2 Generation of Chat Data

We collected chat data utilizing the free-form dialogue of ChatEval (Chan et al., 2023), a Multi-Agent-based evaluation framework aligned with the LLM-as-Judge paradigm. We intentionally selected this evaluation task because it has no definitive correct answer. As tasks with ground truth answers would lead agents to converge on correct responses of authority, it could be difficult to isolate convergence due to authority influence from natural convergence. As evaluation task that we used in this study seldom have such common ground truth, it is much easier to watch authority influence. Using natural interactions between agents, ChatEval asks agents to discuss two writing options and choose the best quality. In our experiment, we used two agents, General Public and an authoritative role, discussing two options using the same two datasets: FairEval (Wang et al., 2024) with 80 examples and Topical-Chat (Gopalakrishnan et al., 2023) with 60 examples.

Specifically, for each pair of writing options, agents had a 12-turn conversation session and we collected the dialogue. The General Public agent initiates the conversation, after which the two agents alternate for a total of 12 turns. Experiments are conducted individually for each of

the nine authoritative roles designed previously. In each dialogue session, agents freely exchange opinions and engage in discussions about the given evaluation task. To control other possible factors during the experiment, we preserved the ChatEval framework code, modifying only the role names (General Public) to authoritative roles (e.g., Judge, Expert) while keeping all instructions and system configurations identical.

3.3 Experimental Procedure

During the free-form conversation, we track decision changes of General Public. Following ChatEval framework (Chan et al., 2023), we can ask each agent to evaluate two writing options after each turn based on the current conversation. Each agent evaluates the quality of each option using a 10-point scale, and decides the highest-scoring option as the writing with the best quality. When both options receive identical scores, the response is classified as neutral. By tracking changes in these decision, we can observe authority bias.

To track temporal changes in authority bias, we collect decisions from two agents. For the initial decision of General Public, we collected its decision $S_{GP}^{(1)}$ at turn 1, which is the beginning of the discussion and without the authority influence. Similarly, we collected initial decision of the authoritative role $S_{Auth}^{(2)}$ at turn 2, after its first utterance. We further measured the decision of General Public $S_{GP}^{(t)}$ at turns $t = 4$, 8, and 12, to identify changes.

To prevent evaluations of prior turns from influencing subsequent interactions or judgments, evaluation scores are not recorded in the chat log. This allows us to observe how the interaction between two agents impacts the General Public's decision and how such influence changes as the conversation progresses.

3.4 Evaluation

To measure authority bias, we defined response selection agreement A_t and Cohen's Kappa coefficient κ_t between two agents, for each turn $t = 1, 4, 8$ and 12:

$$\begin{aligned} A_t &= \text{Agreement}(S_{GP}^{(t)}, S_{Auth}^{(2)}) \\ \kappa_t &= \text{Cohen's Kappa}(S_{GP}^{(t)}, S_{Auth}^{(2)}) \end{aligned}$$

Specifically, we use the authoritative agent's turn 2 decision as a fixed reference to gauge agreement with the General Public's decisions at turns 1, 4, 8, and 12. This temporal tracking allows us to observe how authority bias develops during the interaction. Since the General Public's turn 1 decision is made prior to any authoritative input, an increase in A_t and κ_t $(t = 4, 8, 12)$ compared to the initial A_1 and κ_1 baseline reveals an increasing conformity to the authority. Accordingly, we operationally define this phenomenon—where agreement levels systematically rise from their initial baseline in subsequent turns—as authority bias.

Therefore, low overall A_t and κ_t scores reflect minimal mutual influence. Observing how these values shift across turns provides a clear guide to tracing the general agent's gradual conformity to authority over time.

3.5 Selected Models

After conducting preliminary experiments with various state-of-the-art LLMs, we selected models capable of maintaining fluent and consistent dialogue in extended 12-turn discussions. Though we initially tested five models including QwQ (Team, 2025), Gemini 2.5 Pro (Comanici et al., 2025) and Mixtral (Jiang et al., 2024), we found that only DeepSeek R1 (Guo et al., 2025) and GPT-4o (Hurst et al., 2024) successfully finished all tasks without generating repeated responses. Consequently, we utilized DeepSeek R1 and GPT-4o, which demonstrated fluent conversational abilities. All experiments were conducted with temperature set to 0 for reproducibility, and model calls were made through the OpenRouter API (OpenRouter, 2025).

3.6 Result and Discussion

Referring to Table 1, we first examine the baseline General Public-General Public (GP-GP) conversations, which exhibit notably high initial agreement (A_1) that remains relatively stable across turns. In stark contrast, when interacting with authoritative roles, the initial agreement (A_1) of the General Public drops significantly. Across all experimental conditions with authority, the agreement in A_1 was measured lower than subsequent turns (A_4, A_8, A_{12}), clearly confirming the presence of authority influence. Particularly, differences between models and roles were prominently observed, while differences between turn numbers were not observed.

Model In the model comparative analysis, DeepSeek R1 showed significantly higher agreement levels than GPT-4o. For example, DeepSeek R1 showed A_t of 100% at least once in 6 out of 9 roles on Topical-Chat dataset, indicating greater

		GPT-4o								Deepseek R1							
		FairEval				Topical-Chat				FairEval				Topical-Chat			
		t_1	t_4	t_8	t_{12}	t_1	t_4	t_8	t_{12}	t_1	t_4	t_8	t_{12}	t_1	t_4	t_8	t_{12}
General Public	A_t	91.3	98.8	97.5	97.5	71.7	75.0	73.3	73.3	93.8	98.8	98.8	98.7	86.7	91.7	90.0	90.0
	κ_t	0.81	0.97	0.94	0.94	0.58	0.61	0.58	0.58	0.89	0.98	0.98	0.98	0.75	0.84	0.81	0.81
Judge	A_t	86.3	85.0	83.8	85.0	58.3	65.0	60.0	56.7	85.0	93.8	95.0	95.0	83.3	96.7	96.7	96.7
	κ_t	0.71	0.68	0.65	0.68	0.37	0.47	0.40	0.35	0.72	0.87	0.90	0.90	0.72	0.94	0.94	0.94
Foreman	A_t	66.3	70.0	73.8	73.8	46.7	61.7	66.7	58.3	80.0	90.0	95.0	95.0	90.0	100	96.7	98.3
	κ_t	0.43	0.47	0.53	0.54	0.19	0.42	0.49	0.37	0.64	0.81	0.90	0.90	0.81	1.00	0.93	0.96
Managemer	A_t	68.8	71.3	70.0	71.3	50.0	58.3	60.0	56.7	78.1	90.6	90.6	93.8	90.0	98.3	98.3	98.3
	κ_t	0.45	0.47	0.44	0.47	0.19	0.39	0.41	0.36	0.61	0.83	0.82	0.88	0.82	0.97	0.97	0.97
Supervisor	A_t	66.3	76.3	75.0	73.8	50.0	68.3	65.0	70.0	82.5	93.8	95.0	95.0	91.7	100	96.7	98.3
	κ_t	0.38	0.54	0.52	0.50	0.27	0.49	0.44	0.52	0.68	0.88	0.90	0.90	0.83	1.00	0.93	0.97
Leader	A_t	62.5	68.8	68.8	70.0	46.7	65.0	66.7	63.3	81.3	96.3	96.3	93.8	85.0	98.3	98.3	100
	κ_t	0.37	0.46	0.46	0.49	0.20	0.47	0.49	0.44	0.66	0.92	0.92	0.87	0.73	0.97	0.97	1.00
Mentor	A_t	62.5	71.3	72.5	68.8	38.3	56.7	56.7	61.7	81.3	90.0	91.3	91.3	90.0	96.7	96.7	95.0
	κ_t	0.38	0.50	0.53	0.47	0.10	0.33	0.33	0.40	0.67	0.80	0.83	0.83	0.81	0.93	0.93	0.90
Specialist	A_t	63.8	71.3	71.3	71.3	46.7	58.3	56.7	56.7	81.3	96.3	95.0	95.0	93.3	100	100	100
	κ_t	0.35	0.47	0.46	0.48	0.18	0.40	0.37	0.37	0.65	0.92	0.89	0.89	0.87	1.00	1.00	1.00
Expert	A_t	70.0	76.3	75.0	76.3	50.0	63.3	65.0	63.3	82.5	92.5	95.0	95.0	91.7	100	100	100
	κ_t	0.49	0.58	0.56	0.58	0.24	0.45	0.47	0.45	0.67	0.84	0.90	0.90	0.84	1.00	1.00	1.00
Attorney	A_t	60.0	70.0	70.0	68.8	38.3	56.7	55.0	55.0	85.0	91.3	93.8	91.3	90.0	100	100	100
	κ_t	0.32	0.48	0.48	0.46	0.01	0.40	0.37	0.38	0.71	0.82	0.87	0.82	0.81	1.00	1.00	1.00

Table 1: Experimental result of Free-form analysis. For each time step t_i ($i = 1, 4, 8, 12$), each row present agreement (A_t) and Cohen's Kappa (κ_t) values between General Public and authority roles at t_i.

	GPT-4o		Deepseek R1	
	FairEval	TopicalChat	FairEval	TopicalChat
Judege	0.12	0.37	0.04	0.04
Foreman	0.25	0.35	0.04	0.00
Management	0.29	0.45	0.06	0.02
Supervisor	0.24	0.27	0.04	0.00
Leader	0.29	0.34	0.00	0.00
Mentor	0.29	0.30	0.05	0.00
Specialist	0.28	0.40	0.03	0.00
Expert	0.24	0.35	0.03	0.00
Attoney	0.30	0.45	0.00	0.00

Table 2: Neutral response selection rates by authoritative roles at A_2 across models and datasets.

susceptibility to authority influence compared to GPT-4o. Additionally, DeepSeek R1's initial agreement (A_1) was higher than GPT-4o's, suggesting differences in the models' inherent sensitivity to authority. Particularly for GPT-4o, it was difficult to definitively conclude that authority influence occurred for roles with Legitimate Power (Judge, Foreman, Management). Examining the agreement change patterns for Legitimate Power roles, 5 out of 6 cases showed phenomena where agreement partially decreased as turns increased. When compared with DeepSeek R1 under identical conditions, GPT-4o consistently showed lower agreement levels across all turns. To identify the causes of these inter-model differences, we analyzed actual response distributions and found that GPT-4o had a tendency to excessively select neutral opinions, as confirmed in Table 2: GPT-4o selected neutral options in 12%-45% cases, and Deepseek R1 selected them in less than 6%. So, we suspect that the neutral responses might affect the outcome of this model comparative analysis.

Neutral Option To further investigate this phenomenon, we separated GPT-4o's responses into two groups: cases where authoritative roles selected neutral options and cases where they selected non-neutral options, then measured authority bias for each group. Due to the space limitation, we describe some significant results here, to help readers understand the phenomenon. Appendix C presents detailed results for each case.

In the group where authoritative roles selected only neutral options, the average differences between A_1 and A_4 for each authority type were: Legitimate group -36%, Referent group -32.6%, and Expert group -31.1%. All roles showed notably decreased A_t values in turns 4, 8, and 12 compared

to turn 1. This confirms that General Public agents with GPT-4o are not actually influenced by authority when Authority roles makes neutral selections.

Conversely, in the group excluding neutral options, the average differences between A_1 and A_4 for each authority type showed increases: Legitimate group +28.7%, Referent group +32.3%, and Expert group +35.9%. Most roles demonstrated sharp increases in A_t and κ_t values in turns 4, 8, and 12 compared to turn 1, confirming that General Public agents with GPT-4o follow Authority's opinions when Authority roles took definite positions.

Overall, the difference between DeepSeek R1 and GPT-4o can be attributed to GPT-4o's higher rate of neutral responses. When authoritative roles provide neutral responses, General Public agents do not follow those opinions, representing an exceptional phenomenon where authority influence is negated. When agreement was remeasured excluding neutral options, GPT-4o's indicators were adjusted to levels similar to DeepSeek R1.

Role Through the results in Table 1, we can observe patterns where authority influence varies according to roles. Analysis results show that role groups with Legitimate Power (Judge, Foreman, Management) exhibited relatively lower influence compared to other authority groups. Specifically, in GPT-4o on Faireval dataset, the average difference between A_1 and A_4 for the Legitimate Power group was 2.8%, while the Referent Power group (Supervisor, Leader, Mentor) showed 7.5%, and the Expert Power group (Specialist, Expert, Attorney) showed 7.2% difference. This pattern appears consistently in DeepSeek R1 as well, and similar tendencies can be confirmed in the Topical-Chat dataset. Particularly, roles with Expert Power (Specialist, Expert, Attorney) were observed to reach complete agreement in all three roles under DeepSeek R1's Topical-Chat conditions.

To identify the causes of these differences between authority groups, we reviewed psychological literature and confirmed the characteristics of each authority type. According to Carson et al. (1993), Legitimate Power is authority based on formal position or social status, Referent Power is authority derived from personal appeal or trust, and Expert Power is authority based on specialized knowledge or technical competence. Considering that this study dealt with evaluation tasks, evaluation situations have characteristics where accurate judgment and reliability are emphasized. Therefore, there is a possibility that Expert Power based on specialized knowledge or Referent Power based on trust may be perceived as more persuasive in evaluation contexts than Legitimate Power that relies on simple formal position. However, while we can consider the possibility that the patterns observed in this study may be related to the inherent differences between these authority types, further research is needed to draw clear conclusions.

When examining the results after excluding neutral options, these role-based differences become even more pronounced. Due to the space limitation, we only describe some significant results here. For instance, the Judge role in Faireval showed no change or even a 1.4% decrease between turn 1 and subsequent turns 4, 8, and 12. Similarly, other Legitimate Group roles including Foreman and Management exhibited minimal agreement changes compared to other role groups. Conversely, Expert Group roles demonstrated the largest change magnitude among all groups. From a temporal perspective, the average changes after turn 4 were: Legitimate group 2.5%, Referent group 2.8%, and Expert group 3%, maintaining consistently similar values throughout subsequent turns.

Turn When observing changes according to turns, we analyzed the change patterns of A_4, A_8 and A_{12}, excluding A_1 which was not influenced by authoritative roles. As a result, we **did not observe typical authority bias** patterns where agreement continuously increased with increasing turn numbers. Instead, the following two patterns frequently appeared: 1) $A_4 = A_8 = A_{12}$ and 2) $A_4 < A_8 = A_{12}$, which occurred in 5 and 10 cases for DeepSeek R1, and 2 cases each for GPT-4o. The emergence of these irregular patterns suggests the possibility that they were influenced by *conversational content* rather than pure authority influence. In free-form conversations, new information or arguments are presented at each turn, and these content-related factors are presumed to have affected changes in A_t. These findings indicate the necessity for a controlled experiment to isolate the pure effect of authoritative role from content.

4 Experiment 2 : Controlling Content

Experiment 2 was designed to extend the findings of Experiment 1 and identify where authority bias specifically manifests. To distinguish whether the observed authority bias stems from roles themselves or conversational content differences, we

		GPT-4o								Deepseek R1							
		FairEval				Topical-Chat				FairEval				Topical-Chat			
		t_1	t_4	t_8	t_{12}	t_1	t_4	t_8	t_{12}	t_1	t_4	t_8	t_{12}	t_1	t_4	t_8	t_{12}
General Public	A_t	91.3	98.8	97.5	97.5	71.7	75.0	73.3	73.3	93.8	98.8	98.8	98.7	86.7	91.7	90.0	90.0
	κ_t	0.81	0.97	0.94	0.94	0.58	0.61	0.58	0.58	0.89	0.98	0.98	0.98	0.75	0.84	0.81	0.81
Judge	A_t	53.8	62.5	61.3	60.0	63.3	61.7	63.3	66.7	87.5	98.8	98.8	98.8	85.0	96.7	96.7	98.3
	κ_t	0.27	0.41	0.39	0.38	0.45	0.44	0.46	0.51	0.77	0.97	0.97	0.97	0.73	0.93	0.93	0.97
Foreman	A_t	68.8	70.0	70.0	70.0	53.3	60.0	60.0	63.3	87.5	98.8	98.8	100.0	88.3	98.3	98.3	98.3
	κ_t	0.47	0.48	0.48	0.48	0.28	0.42	0.42	0.47	0.77	0.97	0.97	1.00	0.79	0.97	0.97	0.97
Managemen	A_t	55.0	48.3	51.7	48.3	70.0	75.0	73.8	70.0	86.3	96.3	97.5	97.5	90.0	96.7	98.3	98.3
	κ_t	0.36	0.47	0.43	0.45	0.29	0.30	0.34	0.30	0.75	0.93	0.95	0.95	0.82	0.93	0.97	0.97
Supervisor	A_t	65.0	68.8	70.0	71.3	46.7	58.3	63.3	61.7	86.3	97.5	98.8	97.5	88.3	98.3	96.7	96.7
	κ_t	0.39	0.45	0.47	0.49	0.20	0.38	0.45	0.43	0.75	0.95	0.97	0.95	0.79	0.97	0.94	0.94
Leader	A_t	70.0	75.0	73.8	70.0	48.3	66.7	66.7	66.7	81.3	100.0	98.8	98.8	86.7	96.7	98.3	98.3
	κ_t	0.50	0.58	0.55	0.49	0.22	0.50	0.50	0.50	0.67	1.00	0.97	0.97	0.76	0.93	0.97	0.97
Mentor	A_t	73.8	76.3	75.0	75.0	53.3	71.7	66.7	68.3	81.3	97.5	96.3	96.3	90.0	96.7	96.7	98.3
	κ_t	0.54	0.58	0.56	0.32	0.32	0.55	0.48	0.50	0.67	0.95	0.93	0.92	0.81	0.93	0.93	0.97
Specialist	A_t	61.3	65.0	65.0	66.3	55.0	63.3	66.7	63.3	80.0	97.5	96.3	96.3	85.0	96.7	98.3	98.3
	κ_t	0.37	0.42	0.42	0.45	0.32	0.45	0.51	0.46	0.64	0.95	0.93	0.92	0.74	0.93	0.97	0.97
Expert	A_t	65.0	70.0	70.0	70.0	56.7	51.7	55.0	53.3	85.0	98.8	98.8	100.0	86.7	95.0	98.3	98.3
	κ_t	0.43	0.50	0.50	0.50	0.32	0.31	0.37	0.34	0.73	0.97	0.97	1.00	0.76	0.90	0.97	0.97
Attorney	A_t	66.3	77.5	77.5	76.3	46.7	73.3	75.0	75.0	82.5	98.8	97.5	97.5	91.7	98.3	96.7	96.7
	κ_t	0.40	0.59	0.59	0.56	0.24	0.57	0.60	0.60	0.70	0.97	0.95	0.95	0.84	0.97	0.93	0.93

Table 3: Experimental result of Content-controlled analysis. For each time step t_i ($i = 1, 4, 8, 12$), each row present agreement (A_t) and Cohen's Kappa (κ_t) values between General Public and authority roles at t_i.

conducted a controlled experiment that regulates conversational content. Through this approach, we aim to more accurately identify the causes of authority bias and provide in-depth analysis of the sources of authority bias arising from roles in MAS.

4.1 Experiment Procedure

The experimental procedure was identical to Experiment 1 except how we used the authoritative roles. In this experiment, we utilized conversation datasets between two General Public agents, instead of using conversation between General Public and authoritative roles. For Public-Authoritative conversations, we only altered the second agent's role from General Public to one of the nine authoritative roles while maintaining the contents of Public-Public conversation unchanged. This design allows us to isolate and observe the pure effect of role labeling on decisions while eliminating the influence of conversational content. The models used, datasets, evaluation turns, and evaluation score collection methods were all applied identically to Experiment 1. The authority bias measurement methodology also utilized the same formulas from Experiment 1 to ensure consistency of results.

4.2 Result and Discussion

As shown in Table 3, we observed conversations between General Public agents where only one agent's role name was changed to an Authoritative role. As a result, two major findings emerged. First, there were significant differences between Public-Public conversations and Public-Authoritative conversations. Second, Similar patterns to experiment 1 results appeared despite controlling for content.

Overall Upon comparing the results between Public-Public conversation and Public-Authoritative conversation, notable differences were confirmed in A_1 and κ_1 values. The A_1 values for non-authoritative General Public agents were significantly higher than those in cases with authoritative roles. Based on these observations, it can be interpreted that while the second General Public was influenced by opinion of the first General Public, Authoritative roles may not have been influenced by General Public responses, which is similar result to Experiment 1. Analyzing these patterns, it could be interpreted that rather than General Public agents being influenced by Authority, Authoritative roles tend to show less conformity to General Public opinions.

Neutral Option These results became more pronounced when removing GPT-4o's neutral option responses, as mentioned in experiment 1. Due to the space limitation, we describe some significant results here, and detailed results are presented in Appendix D.

When measuring authority bias in GPT-4o's neutral option group versus non-neutral options, the neutral group showed decreased A_t values in turns 4, 8, and 12 compared to turn 1 across most roles. This suggests a phenomenon similar to experiment 1, where GPT-4o may not actually be influenced by authority when making neutral choices.

In the group excluding neutral options, sharp increases in both A_t values and κ_t values were observed in turns 4, 8, and 12 compared to turn 1. This could suggest that General Public agents follow Authority opinions. Comprehensively considering this phenomena on non-neutral option, GPT-4o becomes showing similar patterns to DeepSeek R1. Detailed results are in Appendix C and D.

Role Role effects also demonstrate similar patterns to experiment 1. Despite controlling for content influence, Legitimate roles (Judge, Foreman, Management) showed weaker impact compared to other authority types. Analysis of Table 3 reveals that both GPT-4o and DeepSeek R1 exhibited relatively lower increases in A_t for the Legitimate Power group compared to other authority types. Notably, Expert Power roles in DeepSeek R1 maintained high agreement levels above 95% even when controlling contents, reconfirming strong expertise-based authority in evaluation tasks. When compared to Public-Public conversations, the inclusion of Authoritative roles consistently resulted in markedly lower A_1 values across all authority types. This suggests that authority bias stems from the authoritative characteristics of roles themselves rather than conversational content. Particularly, the pattern observed in human social psychology where Expert and Referent Power exert stronger influence than Legitimate Power in evaluation contexts (Carson et al., 1993) is reproduced in LLMs.

Turn The temporal dynamics of authority bias in experiment 2 mirror the patterns observed in experiment 1. When examining turn-by-turn changes, we did not observe typical patterns of authority bias, i.e., continuously increasing agreement. Instead, two patterns emerged as in experiment 1: 1) $A_4 = A_8 = A_{12}$ and 2) $A_4 < A_8 = A_{12}$, appearing in 8 cases for DeepSeek R1 and 3 cases for GPT-4o. Critically, when considering A_1, authority-absent conversations maintained high initial agreement throughout subsequent turns, while authority-present conversations showed persistent low agreement. This suggests that in authority-absent dialogues, initial content continues to influence decision making through the conformity of General Public agents. Whereas, in authority-present dialogues, the independent nature of authoritative roles from General Public opinions weakens the influence of turn 1 while amplifying the impact of turn 2. This pattern remained consistent even when controlling contents in experiment 2, confirming that temporal dynamics of authority bias may stem from structural power relationships rather than conversational content.

Overall Discussion Results of experiment 2 provide important insights into the nature of authority bias in MAS. The most significant finding is that the mechanism of authority bias differs from previous assumptions. While existing study assumed that General Public agents are actively influenced by Authority opinions, our results revealed different pattern. Public-Public conversations show high initial agreement A_1, whereas conversations including Authoritative roles exhibit markedly lower A_1 values. This suggests that Authoritative roles tend to maintain their positions without being influenced by other opinions. Consequently, the observed authority bias can be interpreted not as General Public agents actively conforming to Authority, but as a phenomenon resulting from the combination of two tendencies: (1) Authority agent persists in maintaining its position, and (2) General Public agent is too flexible to agree with others' opinion.

To quantitatively substantiate this position persistence, we conducted a Flip Rate analysis on the authoritative roles. While General Public agents frequently align with authoritative opinions, the authorities themselves exhibit remarkable rigidity. For instance, DeepSeek R1's authoritative roles demonstrated overall Flip Rates of merely 3-5% and Bipolar Flip Rates (complete reversals excluding neutral transitions) of 1-2%. Although GPT-4o showed a higher nominal Flip Rate (32%), this was driven almost entirely by transitions from neutral starting positions; its Bipolar Flip Rate remained below 2%, indicating that genuine verdict reversals are exceptionally rare. Furthermore, non-neutral-starting authorities maintained their stance with a Flip Rate of only 1-6%. These metrics directly con-

firm that authority bias in our multi-agent framework is heavily driven by the asymmetrical stability of the authoritative roles. Detailed quantitative tables regarding flip rates across different models, turns, power types, and initial stances are provided in Appendix E.

Also, the emergence of identical patterns to experiment 1 despite controlling for conversational content confirms that such authority bias stems from role labels themselves rather than content. This indicates that LLMs recognize the social authority structures inherent in roles and exhibit corresponding behavioral patterns. Moreover, neutral option analysis and temporal patterns demonstrate that this mechanism is established early and persists over time. Authoritative roles' position maintenance and General Public's gradual conformity are formed in the early stages of content and subsequently stabilize throughout the discussion.

These results present a new perspective that authority bias in MAS is based on independence and consistency of Authoritative roles rather than mutual influence, suggesting that future MAS design requires bias mitigation strategies that consider such asymmetric interaction patterns.

5 Conclusion

Our study presents the first systematic analysis of role-based authority bias in Multi-agent systems. Through free-form and content-controlled experiments using ChatEval, we demonstrated that authority bias stems from inherent role characteristics rather than conversational content. Our findings reveal that referent and expert power roles exert stronger influence than legitimate power roles, mirroring human social psychology theory. Crucially, authority bias operates not through active conformity by general agents, but through a mechanism where authoritative roles maintain their positions while general agents demonstrate flexibility. These insights provide foundational knowledge for designing multi-agent frameworks where asymmetric interaction patterns significantly affect outcomes.

Limitations

This study systematically analyzes role-based authority bias in multi-agent evaluation systems, but several key limitations should be acknowledged. First, experiments were conducted using GPT-4o and DeepSeek R1, selected for their capability to maintain fluent dialogue across 12-turn conversations; however, this requirement constrained the diversity of models examined. Second, our ChatEval-based framework was designed to capture authority bias in evaluation tasks, yet multi-agent systems are deployed across diverse domains such as creative collaboration, technical problem-solving, and strategic planning. These domains may exhibit different authority bias patterns that our evaluation-focused design does not address. Third, our experimental design deliberately adhered to standardized templates without prompt variation. While this ensures the isolation of the causal impact of authoritative RPAs, it leaves potential mitigation methodologies unaddressed. Exploring both prompt variations and targeted interventions remains necessary to assess the robustness of these behaviors and effectively calibrate the authority bias. Finally, the current scope focuses on identifying convergence patterns without explicitly connecting them to evaluation quality. Examining whether conformity to authoritative RPAs practically improves or harms the correctness of final judgments serves as a concrete and important direction for future work.

Acknowledgement

This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(RS-2025-25434151)

References

Paula Phillips Carson, Kerry D Carson, and C William Roe. 1993. Social power bases: A meta-anaiytic examination of interrelationships and outcomes 1. *Journal of Applied Social Psychology*, 23(14):1150–1169.

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. 2023. Chateval: Towards better llm-based evaluators through multi-agent debate. *arXiv preprint arXiv:2308.07201*.

Guiming Hardy Chen, Shunian Chen, Ziche Liu, Feng Jiang, and Benyou Wang. 2024. Humans or llms as the judge? a study on judgement biases. *arXiv preprint arXiv:2402.10669*.

Min Choi, Keonwoo Kim, Sungwon Chae, and Sangyeop Baek. 2025. An empirical study of group conformity in multi-agent systems. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 5123–5139, Vienna, Austria. Association for Computational Linguistics.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Yihong Dong, Xue Jiang, Zhi Jin, and Ge Li. 2024. Self-collaboration code generation via chatgpt. *ACM Trans. Softw. Eng. Methodol.*, 33(7).

Giorgos Filandrianos, Angeliki Dimitriou, Maria Lymperaiou, Konstantinos Thomas, and Giorgos Stamou. 2025. Bias beware: The impact of cognitive biases on llm-driven product recommendations. *arXiv preprint arXiv:2502.01349*.

John RP French. 1959. The bases of social power. *Studies in social power/University of Michigan Press*.

Veronica M Godshalk and John J Sosik. 2000. Does mentor-protégé agreement on mentor leadership behavior influence the quality of a mentoring relationship? *Group & Organization Management*, 25(3):291–317.

Karthik Gopalakrishnan, Behnam Hedayatnia, Qinlang Chen, Anna Gottardi, Sanjeev Kwatra, Anu Venkatesh, Raefer Gabriel, and Dilek Hakkani-Tur. 2023. Topical-chat: Towards knowledge-grounded open-domain conversations. *arXiv preprint arXiv:2308.11995*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Daniela K Haller, Peter Fischer, and Dieter Frey. 2018. The power of good: A leader's personal power as a mediator of the ethical leadership-follower outcomes link. *Frontiers in Psychology*, 9:1094.

Sirui Hong, Xiawu Zheng, Jonathan Chen, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, and 1 others. 2023. Metagpt: Meta programming for multi-agent collaborative framework. *arXiv preprint arXiv:2308.00352*, 3(4):6.

D Huang, JM Zhang, M Luck, Q Bu, Y Qing, and H Cui. 2024. Agentcoder: Multi-agent code generation with effective testing and self-optimization. *University of Hong Kong, King's College London, University of Sussex, Shanghai Jiao Tong University*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, and 1 others. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Guohao Li, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023. Camel: Communicative agents for" mind" exploration of large scale language model society.

Junkai Li, Yunghwei Lai, Weitao Li, Jingyi Ren, Meng Zhang, Xinhui Kang, Siyu Wang, Peng Li, Ya-Qin Zhang, Weizhi Ma, and 1 others. 2024. Agent hospital: A simulacrum of hospital with evolvable medical agents. *arXiv preprint arXiv:2405.02957*.

Xuan Liu, Jie Zhang, Haoyang Shang, Song Guo, Chengxu Yang, and Quanyan Zhu. 2024. Exploring prosocial irrationality for llm agents: A social cognition view. *arXiv preprint arXiv:2405.14744*.

Jiwon Moon, Yerin Hwang, Dongryeol Lee, Taegwan Kang, Yongil Kim, and Kyomin Jung. 2025. Don't judge code by its cover: Exploring biases in llm judges for code evaluation. *arXiv preprint arXiv:2505.16222*.

OpenRouter. 2025. Openrouter api: Web search feature. https://openrouter.ai.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pages 1–22.

Joon Sung Park, Carolyn Q Zou, Aaron Shaw, Benjamin Mako Hill, Carrie Cai, Meredith Ringel Morris, Robb Willer, Percy Liang, and Michael S Bernstein. 2024. Generative agent simulations of 1,000 people. *arXiv preprint arXiv:2411.10109*.

Taylor Peyton, Drea Zigarmi, and Susan N Fowler. 2019. Examining the relationship between leaders' power use, followers' motivational outlooks, and followers' work intentions. *Frontiers in psychology*, 9:2620.

Chen Qian, Wei Liu, Hongzhang Liu, Nuo Chen, Yufan Dang, Jiahao Li, Cheng Yang, Weize Chen, Yusheng Su, Xin Cong, and 1 others. 2023. Chatdev: Communicative agents for software development. *arXiv preprint arXiv:2307.07924*.

Sumedh Rasal. 2024. Llm harmony: Multi-agent communication for problem solving. *arXiv preprint arXiv:2401.01312*.

Samuel Schmidgall, Yusheng Su, Ze Wang, Ximeng Sun, Jialian Wu, Xiaodong Yu, Jiang Liu, Zicheng Liu, and Emad Barsoum. Agent laboratory: Using llm agents as research assistants, 2025. *URL https://arxiv. org/abs/2501.04227*.

Qwen Team. 2025. Qwq-32b: Embracing the power of reinforcement learning.

Kuan Wang, Yadong Lu, Michael Santacroce, Yeyun Gong, Chao Zhang, and Yelong Shen. 2023. Adapting llm agents with universal feedback in communication. *arXiv preprint arXiv:2310.01444*.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and 1 others. 2024. Large language models are not fair evaluators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9440–9450.

Zhao Wang, Sota Moriyama, Wei-Yao Wang, Briti Gangopadhyay, and Shingo Takamatsu. 2025. Talk structurally, act hierarchically: A collaborative framework for llm multi-agent systems. *arXiv preprint arXiv:2502.11098*.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, and 1 others. 2023. Autogen: Enabling next-gen llm applications via multi-agent conversation. *arXiv preprint arXiv:2308.08155*.

Jiayi Ye, Yanbo Wang, Yue Huang, Dongping Chen, Qihui Zhang, Nuno Moniz, Tian Gao, Werner Geyer, Chao Huang, Pin-Yu Chen, and 1 others. 2024. Justice or prejudice? quantifying biases in llm-as-a-judge. *arXiv preprint arXiv:2410.02736*.

The Use of Large Language Models

We used AI assistance tools during the writing process of this manuscript. Specifically, we employed Grammarly for grammar checking, and GPT-5 for language polishing and improving clarity of expression. These tools were used for editorial purposes.

A Experiment Setup

A.1 Experimental Framework

All experiments were conducted using the ChatEval (Chan et al., 2023) framework, maintaining identical configurations to the original implementation. We preserved all system prompts, instructions, and dialogue structures from ChatEval, modifying only the role assignments to examine authority effects. Detailed implementation and prompts can be found in our code: [Under Review].

A.2 Prompt Setup

To isolate the effect of role-based authority, we kept all prompt components constant except for the role name and its co-reference in the role description. The following templates compares the original attempt and our approach. The first template indicates general public agent specified in the original ChatEval framework:

> Name: **General Public**
> Role description:
> You are now **General Public**, one of the referees in this task. You are interested in the story and looking for updates on the investigation. Please think critically by yourself and note that it is your responsibility to choose which response is better.

And, the next template shows how we modified the basic template with authority roles.

> Name: **[Authority Role]**
> Role description:
> You are now **[Authority Role]**, one of the referees in this task. You are interested in the story and looking for updates on the investigation. Please think critically by yourself and note that it is your responsibility to choose which response is better.

Where **[Authority Role]** is replaced with one of the nine authoritative roles: Judge, Foreman, Management, Supervisor, Leader, Mentor, Specialist, Expert, or Attorney.

B Proof of Authority Role Design

In this section, we describe the detailed justification for our selection of nine authority roles inspired by French (1959).

B.1 Legiminage Power

French (1959) defined legitimate power as the influence that emerges when P accepts that O has rightful authority to give directives and feels obligated to comply. This power stems from social and normative approval granted to positions, creating voluntary compliance because individuals perceive the authority as rightful.

Judge A judge represents a role with institutionally granted authority to make definitive judgments and render final decisions that others are expected to accept. This role embodies the power to evaluate situations and pronounce authoritative verdicts based on recognized legitimate right to make binding determinations. French (1959) presented judge as an exemplary profession where legitimate power manifests through formal position-based authority.

Foreman A foreman represents a role with formally designated authority to direct and coordinate others' actions within structured hierarchies. This position embodies the power to assign tasks and ensure compliance through officially recognized right to make operational decisions. French (1959) described foreman as demonstrating legitimate power through institutionally sanctioned supervisory authority.

Management Management represents roles with institutionally sanctioned authority to make strategic decisions and control organizational resources. This encompasses positions with formal power to set policies and direct organizational behavior through recognized executive decision-making rights. French (1959) identified management as exemplifying legitimate power through institutional mandate to make authoritative choices.

B.2 Reference Power

French (1959) defined referent power as influence that emerges when P seeks to identify with O and feels attraction and respect toward O. This represents power arising from voluntary compliance driven by P's motivation to become like O.

Supervisor A supervisor embodies a role that commands respect and voluntary compliance through demonstrating exemplary work practices and care for team development. This position represents the power to influence through being perceived as someone whose methods and approaches others want to emulate and Haller et al. (2018) identified supervisors as roles that team members respect and wish to model their work practices after, establishing supervisor as a position where referent power manifests.

Leader A leader represents a role that influences others through vision and inspiration, creating voluntary followership based on admiration for their character and direction. This position embodies the power to guide groups through personal magnetism and the ability to make others want to align themselves with the leader's mission. French (1959) noted that referent power emerges when followers admire leaders and seek to identify with them as role models, positioning leader as an exemplar of referent power.

Mentor A mentor represents a role that wields influence through being perceived as a wise guide whose experience and counsel others actively seek and value. This position embodies the power to shape development through being seen as someone worth emulating in both professional and personal growth. Godshalk and Sosik (2000) observed that mentees commonly attribute referent power to mentors, establishing mentor as a role where referent power naturally emerges.

B.3 Expert Power

French (1959) defined expert power as influence that forms when someone is perceived to possess special knowledge or expertise. When P believes O has superior knowledge and credible expertise, P voluntarily follows O's guidance, maintaining compliance even without rewards or punishments.

Specialist A specialist represents a role that commands deference through possessing concentrated, domain-specific knowledge that others recognize as superior in particular areas. This position embodies the power to influence decisions through demonstrated mastery of specialized information and techniques. French (1959) noted that expert power emerges specifically when others perceive someone as having special knowledge in defined domains, making the specialist role a direct manifestation of expertise-based influence within limited fields.

Expert An expert embodies a role with comprehensive mastery that others acknowledge as authoritative within specific domains. This position represents the power to shape opinions through demonstrated competence and superior analytical capability that provides persuasive decision-making resources. French (1959) emphasized that expert power stems from recognized superior knowledge and judgment abilities across broader areas of expertise, positioning the expert role as the archetypal example of comprehensive expertise-based authority.

Attorney An attorney represents a role with specialized analytical and argumentative capabilities that others recognize as essential for navigating complex evaluative processes. This position embodies the power to influence through systematic reasoning and structured analysis that others find compelling and trustworthy. French (1959) specifically cited accepting legal counsel as a common example of expert influence in action, identifying the attorney role as a prime illustration of how specialized knowledge creates authoritative influence in decision-making contexts.

C Neutral response on Experiment 1

Table 4 shows GPT-4o authority bias results from Experiment 1, separated by neutral versus non-neutral response conditions. Authority bias emerges only when authoritative roles take clear positions, not when providing neutral responses. See Table 4 on page 14 for detailed analysis.

D Neutral response on Experiment 2

Table 5 shows GPT-4o authority bias results from Experiment 2's content-controlled conversations, separated by neutral versus non-neutral response conditions. This experiment isolates role-based authority effects by controlling conversational content while varying only role labels. See Table 5 on page 14 for detailed analysis.

E Detailed Flip Rate Analysis of Authoritative Roles

E.1 Overall Average Flip Rate by Model, Condition, and Turn

Refer to Table 6.

E.2 Average Flip Rate by Power Type

Refer to Table 7.

E.3 Neutral vs. Non-neutral Initial Stance Flip Rate

Refer to Table 8.

		GPT-4o non-neutral option								GPT-4o neutral option							
		FairEval				Topical-Chat				FairEval				Topical-Chat			
		t_1	t_4	t_8	t_{12}	t_1	t_4	t_8	t_{12}	t_1	t_4	t_8	t_{12}	t_1	t_4	t_8	t_{12}
Judge	A_t	95.8	95.8	94.4	95.8	55.3	97.4	94.7	89.5	11.1	0.00	0.00	0.00	63.6	9.1	0.00	0.00
	κ_t	0.89	0.89	0.86	0.89	0.38	0.95	0.90	0.80	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Foreman	A_t	80.0	93.3	98.3	98.3	48.7	92.3	87.2	84.6	35.0	0.00	0.00	0.00	42.9	4.8	28.6	9.5
	κ_t	0.60	0.84	0.96	0.96	0.31	0.84	0.75	0.70	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Management	A_t	87.7	100	98.2	100	39.4	100	100	97.0	21.7	0.00	0.00	0.00	63.0	7.4	11.1	7.4
	κ_t	0.72	1.00	0.95	1.00	0.23	1.00	1.00	0.93	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Supervisor	A_t	83.6	100	98.4	96.7	45.5	90.9	88.6	90.9	10.5	0.00	0.00	0.00	62.5	6.3	0.00	12.5
	κ_t	0.64	1.00	0.95	0.91	0.28	0.82	0.79	0.83	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Leader	A_t	78.9	96.5	96.5	96.5	45.0	92.5	95.0	87.5	21.7	0.00	0.00	4.3	50.0	10.0	10.0	15.0
	κ_t	0.60	0.92	0.92	0.92	0.27	0.85	0.90	0.77	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Mentor	A_t	78.9	100	100	96.5	33.3	78.6	76.2	85.7	21.7	0.00	4.3	0.00	50.0	5.6	11.1	5.6
	κ_t	0.61	1.00	1.00	0.92	0.18	0.61	0.57	0.71	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Specialist	A_t	81.0	98.3	96.6	94.8	41.7	97.2	94.4	91.7	18.2	0.00	4.5	9.1	54.2	0.00	0.00	4.2
	κ_t	0.59	0.95	0.91	0.87	0.23	0.94	0.88	0.83	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Expert	A_t	85.2	100	98.4	100	48.7	94.9	97.4	89.7	21.1	0.00	0.00	0.00	52.4	4.8	4.8	14.3
	κ_t	0.71	1.00	0.96	1.00	0.32	0.90	0.95	0.80	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Attorney	A_t	82.1	100	100	96.4	33.3	97.0	93.9	100	8.3	0.00	0.00	4.2	44.4	7.4	7.4	0.00
	κ_t	0.63	1.00	1.00	0.91	0.19	0.94	0.88	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table 4: Comparison between two responses in GPT-4o, Experiment 1: Authority initially responds with *Neutral* response versus *Non-Neutral* response

		GPT-4o non-neutral option								GPT-4o neutral option							
		FairEval				Topical-Chat				FairEval				Topical-Chat			
		t_1	t_4	t_8	t_{12}	t_1	t_4	t_8	t_{12}	t_1	t_4	t_8	t_{12}	t_1	t_4	t_8	t_{12}
Judge	A_t	81.6	100	100	98.0	62.2	97.3	100	100	96.8	32.2	0.00	0.00	65.2	4.3	4.3	13.0
	κ_t	0.65	1.00	1.00	0.95	0.45	0.95	1.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Foreman	A_t	89.5	98.2	98.2	98.2	55.6	97.2	100	100	17.4	0.00	0.00	0.00	50.0	4.2	0.00	8.3
	κ_t	0.77	0.96	0.96	0.96	0.37	0.94	1.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Management	A_t	85.2	100	98.1	100	54.8	90.3	93.5	90.3	11.5	3.8	0.00	0.00	55.2	3.4	6.9	3.4
	κ_t	0.69	1.00	0.95	1.00	0.36	0.82	0.88	0.82	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Supervisor	A_t	87.9	94.8	96.6	98.3	45.0	85.0	90.0	90.0	4.5	0.00	0.00	0.00	50.0	5.0	10.0	5.0
	κ_t	0.74	0.88	0.92	0.96	0.28	0.72	0.81	0.81	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Leader	A_t	89.7	100	100	96.6	40.0	100	97.5	95.0	18.2	9.1	4.5	0.00	65.0	0.00	5.0	10.0
	κ_t	0.78	1.00	1.00	0.93	0.25	1.00	0.95	0.90	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Mentor	A_t	91.7	100	100	100	47.7	97.7	90.9	93.2	20.0	5.0	0.00	0.00	68.8	0.00	0.00	0.00
	κ_t	0.81	1.00	1.00	1.00	0.32	0.96	0.83	0.87	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Specialist	A_t	86.5	98.1	100	100	47.4	92.1	100	92.1	14.3	3.6	0.00	3.6	68.2	13.6	9.1	13.6
	κ_t	0.70	0.95	1.00	1.00	0.31	0.85	1.00	0.85	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Expert	A_t	83.9	100	100	100	54.5	87.9	93.9	93.9	20.8	0.00	0.00	0.00	59.3	7.4	7.4	3.7
	κ_t	0.70	1.00	1.00	1.00	0.38	0.77	0.88	0.88	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Attorney	A_t	83.9	98.4	98.4	98.4	44.4	97.8	95.6	95.6	5.6	5.6	5.6	0.00	53.3	0.00	13.3	13.3
	κ_t	0.68	0.96	0.96	0.96	0.29	0.95	0.91	0.91	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table 5: Comparison between two responses in GPT-4o, Experiment 2: Authority initially responds with *Neutral* response versus *Non-Neutral* response

	DeepSeek R1				GPT-4o			
	Exp1		Exp2		Exp1		Exp2	
Turn	Flip Rate	Bipolar Flip Rate	Flip Rate	Bipolar Flip Rate	Flip Rate	Bipolar Flip Rate	Flip Rate	Bipolar Flip Rate
4	0.047	0.025	0.026	0.011	0.318	0.010	0.342	0.012
8	0.037	0.018	0.026	0.013	0.321	0.013	0.339	0.008
12	0.033	0.014	0.021	0.011	0.326	0.010	0.336	0.009

Table 6: Average flip rates across conversation turns by model and experimental condition.

	DeepSeek R1				GPT-4o			
	Exp1		Exp2		Exp1		Exp2	
Power	Flip Rate	Bipolar Flip Rate	Flip Rate	Bipolar Flip Rate	Flip Rate	Bipolar Flip Rate	Flip Rate	Bipolar Flip Rate
Legitimate	0.045	0.013	0.019	0.009	0.319	0.010	0.379	0.002
Referent	0.040	0.025	0.029	0.016	0.301	0.015	0.306	0.019
Expert	0.027	0.020	0.025	0.011	0.345	0.009	0.332	0.008

Table 7: Average flip rates by authoritative power type.

	DeepSeek R1				GPT-4o			
	Exp1		Exp2		Exp1		Exp2	
Power	Neutral	Non-neutral	Neutral	Non-neutral	Neutral	Non-neutral	Neutral	Non-neutral
Legitimate	1.000	0.016	0.750	0.010	0.955	0.049	0.965	0.025
Referent	1.000	0.026	0.800	0.018	0.904	0.062	0.963	0.040
Expert	1.000	0.021	0.917	0.014	0.951	0.044	0.966	0.034

Table 8: Flip rates based on the authoritative agent's initial stance (neutral vs. non-neutral).