People will agree what I think: Investigating LLM's False Consensus Effect

Anonymous ACL submission

Abstract

Large Language Models (LLMs) have recently been widely adopted on interactive systems requiring communications. As the false belief in a model can harm the usability of such systems, LLMs should not have cognitive biases that humans have. Especially psychologists focused on the False Consensus Effect (FCE), which can distract smooth communication by posing false beliefs. However, previous studies have less examined FCE in LLMs thoroughly, which needs more consideration of confounding biases, general situations, and prompt changes. Therefore, in this paper, we conduct two studies to deeply examine the FCE phenomenon in LLMs. In Study 1, we investigate whether LLMs have FCE. In Study 2, we explore how various prompting styles affect the demonstration of FCE. As a result of these studies, we identified that popular LLMs have FCE. Also, the result specifies the conditions when the strength of FCE becomes larger or smaller compared to normal usage.

1 Introduction

Large Language Models (LLMs) have recently been widely applied for interactive systems requiring communications, such as education, customer service, or healthcare (Schön et al., 2023; Altay and Çetintürk, 2024; Waikar, 2020). Within these applications, controlling cognitive biases is essential because providing biased information may harm the utility of such systems. For instance, in a tutoring system, false beliefs about students' learning styles may make the system provide an inappropriate learning aid (Schön et al., 2023). So, before applying LLMs to those applications, we need to verify whether the LLMs have been exposed to any cognitive biases (Echterhoff et al., 2024).

Among those cognitive biases, psychologists have mainly focused on the False Consensus Effect (FCE). FCE can hinder smooth communication because a person with FCE has a false belief about the social consensus; by definition, a person with FCE believes their *own* behavioral choices as relatively *common* choices in a given situation while viewing *other* choices as *uncommon* in society (Ross et al., 1977; Wojcieszak and Price, 2009; Wetzel and Walton, 1985). Thus, we suspect that an LLM-based interactive system (e.g., Schön et al. (2023)) can provide misleading information because of false beliefs about the consensus between LLMs and human users. For instance, a tutoring system with FCE may recommend a learning style that it regards as preferred in society but which is actually inappropriate for the given situation.

Despite the significance of FCE in communication, experimental methods in previous studies have less investigated whether LLMs may expose FCE in ordinary applications. Studies have conducted experiments to identify underlying reasons for other cognitive biases in LLMs (Koo et al., 2023; Opedal et al., 2024a; Talboy and Fuller, 2023) or to mitigate other cognitive biases in LLMs (Echterhoff et al., 2024; Itzhak et al., 2023; Lin and Ng, 2023). However, these experimental methods have three limitations when applying them to FCE investigation: (1) confounding biases, (2) general situations, and (3) prompting methods.

First, for the confounding biases, previous studies have less considered the effect of other biases during the experiments. Several cognitive biases exist in the human thought process and affect human verbalization. So, psychologists have tried to disentangle a target bias from confounding biases within an experiment. Like humans, we suspect LLMs can have multiple cognitive biases, as recent studies reported (Schmidgall et al., 2024; Echterhoff et al., 2024). Therefore, to prevent introducing confounding biases within LLMs' generation process, we need to control them. However, previous studies have yet to control such confounding biases. So, in this paper, we adopt a psychological experiment properly to control confounding biases.

Second, for the general situations, existing studies have tested FCE on specific domains only. In human communication, a cognitive bias is a general phenomenon regardless of the situation. As LLMs are adopted by many application domains, including education or healthcare, many studies have attempted to investigate cognitive biases in both general and domain-specific situations (Echterhoff et al., 2024; Macmillan-Scott and Musolesi, 2024). However, for FCE, studies have yet to focus on examining FCE in a general situation, as far as we know. So, this paper investigates FCE in a general, ordinary situation of communication to ensure the generalizability of FCE in various domains.

Third, for the prompting methods, existing literature has less investigated the interaction between cognitive biases and popular prompting styles. Researchers reported that subtle changes in prompts may significantly affect the generation procedure of LM (Jia and Liang, 2017; Cheng et al., 2019). So, such subtle changes may also affect the results of previous experiments. However, previous studies have yet to investigate the effect of such prompt changes systematically, though Echterhoff et al. (2024) attempted to investigate the effect of prompts on cognitive biases. So, in this paper, we examine how various prompting styles affect the demonstration of FCE.

To address these issues, we conduct two studies on LLMs. In Study 1, we investigate whether LLMs have FCE. We adopt a well-known psychological experiment to handle confounding bias and general situations. Through this study, we demonstrate a way of interpreting the behavior of LLMs using psychological studies. In Study 2, we examine how various prompting styles affect FCE. We test two dimensions of prompting styles, including the relevance of provided information and the depth of the reasoning process, which can affect FCE. Through this study, we expect to find a way to mitigate FCE in LLMs as a byproduct.

Thus, this paper has the following contributions.

- We demonstrate how to interpret cognitive processes in LLMs through a controlled human psychology experiment.
- Our study shows that state-of-the-art LLMs exhibit FCE in ordinary situations.
- We find that provided information may affect FCE, and repeated reasoning can reduce FCE.

2 Related Work

Our paper is closely related to literature examining cognitive biases. So, we review previous approaches in examination and prompting methods.

2.1 Examining bias in LLMs

Inspired by psychological experiments, researchers have recently focused on examining cognitive biases in LLMs. We categorize these studies in terms of whether they follow the experimental procedure of psychology when examining cognitive biases.

First, researchers designed their own experiments to examine cognitive biases (Koo et al., 2023; Schmidgall et al., 2024; Opedal et al., 2024b; Itzhak et al., 2023). Though psychological theories inspire these studies, researchers proposed new experiments using previous natural language datasets[1]. For example, Koo et al. (2023) modified existing datasets to expose LLMs to various cognitive biases, including bandwagon effect. Though these benchmarks identified biases in LLMs empirically, however, these methods cannot ensure whether we can exclude confounding bias from the result. In other words, it is difficult to ensure the results are due to the bias we wanted to investigate.

Second, researchers attempted to adopt psychological experiments as it is (Xie et al., 2024; Aher et al., 2023; Macmillan-Scott and Musolesi, 2024; Talboy and Fuller, 2023). Studies in this category have psychological experiments as their basis, so they mainly follow the materials or experimental procedures provided by psychological studies. Some researchers aimed to replicate psychological studies using LLMs. For example, Aher et al. (2023) replicated well-known experiments (e.g., Milgram Experiment). Though they demonstrated how psychological experiments can be reproduced using LLMs, the paper has yet to focus on explaining how cognitive bias occurs in LLMs. Others examined whether LLMs have cognitive biases with psychological experiments (Xie et al., 2024; Macmillan-Scott and Musolesi, 2024; Talboy and Fuller, 2023). Note that these experiments try to control demographic and situational differences. As such differences can affect LLMs through changes in prompts, it is necessary to consider those differences when conducting such experiments.

Among these studies, as far as we know, the only experiment considering FCE on LLM was

---

[1] Please see Appendix A for a detailed comparison of our work with previous studies.

Schmidgall et al. (2024). However, we need another experiment because of two limitations. First, as discussed in the first category, the study has less controlled other confounding biases. The phrase "`most of your colleagues believe [option]`" they used to invoke FCE can invoke the conformity effect. Second, as discussed in the second category, the experiment only covers the medical situation. Therefore, we need an experiment resolving these issues to generalize the experimental result.

2.2 Mitigating biases of LMs

As LMs generate text based on the input prompt, researchers have reported that subtle changes in the prompt affect LMs' output (Jia and Liang, 2017; Cheng et al., 2019; Xie et al., 2024; Guo et al., 2024). For example, Jia and Liang (2017) and Cheng et al. (2019) reveal that adding or modifying input prompts can change the answer. Similarly, Xie et al. (2024) showed that prompt settings can alter the result of a psychological experiment about trust. They changed demographic information (e.g., gender) and prior trust in a prompt, and the result revealed that such changes affect the behavior of LLMs. Hence, different prompts may invoke different strengths of a cognitive bias.

Thus, researchers have developed methods to control cognitive biases in LLMs (Echterhoff et al., 2024; Itzhak et al., 2023; Schmidgall et al., 2024). Some researchers tried to modify the generation procedure to control the bias (Itzhak et al., 2023). However, this approach cannot be generalized to a black-box model, including GPT-4. So, others suggested a method using prompt changes. For example, (Echterhoff et al., 2024) tries to control cognitive biases by providing additional zero-shot prompts or examples that can make LLMs aware of cognitive biases. Similarly, (Schmidgall et al., 2024) suggested bias mitigation strategies for medical QA, which includes zero-shot educating prompts or biased/unbiased examples for QA.

However, current methods have not taken into account prompt engineering techniques such as CoT (Wei et al., 2022b) or Reflection (Shinn et al., 2024). As such well-known prompt techniques enhance question-answering tasks with deep reasoning, they likely reduce byproducts of intuitive thinking, including cognitive bias. Although Opedal et al. (2024b) assessed the impact of CoT on cognitive bias, they did not propose a way to mitigate such bias. Therefore, based on previous research, we need to test two dimensions of modification: provided information and prompt engineering.

> You arrive for the first day of class in a course in your major area of study. The professor says that the grade in your course will depend on a paper due on the final day of the course. He gives the class the option of two alternatives upon which they must vote. They can either **do papers individually** in the normal way or **work in teams** of three persons who will submit a single paper between them. You are informed that he will still give out the same number of A's, B's, and C's, etc., but that in the first case, every student will be graded individually, while in the second case, all three students who work together get the same grade.

Figure 1: Story 1 used in Ross et al. (1977). Highlighted Bold-faced text shows two options given in this story.

3 Study 1: Examining FCE of LLM

Study 1 conducts an experiment inspired by psychological experiments to investigate whether the False Consensus Effect (FCE) emerges in LLMs (Ross et al., 1977; Choi and Cha, 2019).

3.1 Procedure

To confirm whether LLMs have FCE, we conduct an experiment that mainly follows a well-defined psychological experiment for revealing FCE (Ross et al., 1977; Choi and Cha, 2019). In the following paragraphs, we illustrate the participants and procedure for our experiment. Each paragraph begins with how psychologists conducted their experiments to help readers understand.

Participants: Originally, psychological studies usually recruited college students. There is no screening procedure to select a specific group of students to identify the general phenomenon of human students. Also, as cultural or gender biases may affect FCE (Choi and Cha, 2019), participants' demographic information should not be biased.

In our study, we make LLMs pretend to be college students, as in the original experiment. Simply, we plant some idea about a character to LLM by giving a system prompt like '`Your name is [name]. You are an undergraduate student. You are [gender]. You are [nationality].`' Also, to control cultural or gender biases, we used 10 characters for each of the two cultures and each gender[2]. For detailed information about the 40 characters that we used, see Appendix B.2.

---

[2] For the culture, we selected European American and Korean, as in a psychological study (Choi and Cha, 2019). Moreover, for the gender, we used the words 'man' and 'woman.'

	GPT-4		Claude3		LLaMA2		Mixtral	
Story 1	40	0	40	0	40	0	37	3
2	40	0	40	0	0	40	0	40
3	0	40	0	40	20	20	0	40
4	9	31	31	9	0	40	0	40

Table 1: Skewness of LLMs answer. A cell shows the number of choosing options 1 and 2 in each story.

Procedure: Originally, psychologists give participants several hypothetical situations which can occur in their ordinary life. Figure 1 shows a sample situation drawn from Ross et al. (1977). After reading the story, participants are asked to choose one of two options for the situation and to estimate the percentage of typical people who agree with their choice, i.e., *perceived agreement*. Note that there is no right option in those hypothetical situations; the choice may vary across people because there is no social agreement about them.

In our study, we follow the same procedure except querying their own choice. We use four hypothetical stories from Ross et al. (1977) to make LLMs estimate perceived agreements. We do not ask LLMs to select their preferred options since the analysis procedure requires comparing two groups: participants with option 1 and those with option 2. Instead, we directly feed each option as if LLMs have chosen that option since we found LLMs stick to a specific choice, as shown in Table 1. Appendix B shows detailed prompt and four stories.

Also, we do not modify other settings to control confounding biases. In a psychological experiment, psychologists carefully designed conditions to control confounding biases, such as confirmation bias, in-group bias, or accumulation effect. As these biases arise due to externally given or preconceived opinions, we excluded providing such information from prompts used in Study 1. For example, we do not provide any information related to external social consensus about the given story in the experiment. Also, we do not ask or provide reasoning for the answer in Study 1, as the original experiment does not ask participants' reasoning to avoid deep thinking about the social consensus. Note that the effect of reasoning will be discussed in Study 2.

3.2 Tested LLMs

For the experiment, we use four LLMs: GPT-4 (Achiam et al., 2023), Claude 3 Opus (Anthropic, 2024), LLaMA 2 70B (Touvron et al., 2023), and Mixtral 8x7B (Jiang et al., 2024). We selected these models because they have shown outstanding performance on question-answering tasks, and the largest model in the family has been published publicly through API or model parameters.

Note that these four LLMs have a refinement policy that avoids social or ethical issues. As our experiment tries to reveal social biases in LLMs, we briefly summarize refinement policies here. First, GPT-4 and Claude 3 have a procedure that refuses answers to questions concerning personal information or questions sensitive politically, religiously, or culturally. Meanwhile, LLaMA 2 and Mixtral were trained to avoid making dangerous or unethical utterances and to retain strong neutrality when making choices. We expect that this difference in refinement policy affect the experimental result.

All the experiments were done by calling APIs. We used the official API provided by GPT-4 and Claude 3. For LLaMA 2 and Mixtral, we used the free API provided by Groq[3]. All the experiments were conducted from February 24, 2024, to June 10, 2024, by calling 320 API calls[4] for each LLM. Also, to reduce the effect of a stochastic generation procedure and focus on LLMs' deterministic computation of generating tokens, we set the temperature value as zero for all API calls[5]. The detailed environmental setup is described in Appendix C.

3.3 Analysis

Originally, psychologists used statistical tests to identify whether humans have FCE. Researchers computed the average perceived agreement for each option as each participant provided a perceived agreement for others' thoughts. Mathematically, let $A_c(p)$ be the perceived agreement on a specific option p, averaged on participants who chose option c. Then, psychologists verified FCE by testing whether $\Delta := A_1(1) - A_2(1)$ is positive, indicating that *participants with option 1* overestimated their *preferred choice* compared to *participants with option 2*. For statistical tests, they conducted a two-sample t-test or Mann-Whitney U test (Nachar et al., 2008), a non-parametric alternative for t-test.

Our study mainly follows analysis methods in original experiments (Ross et al., 1977; Choi and Cha, 2019). We set three hypotheses: one for verifying whether LLMs have FCE and two for identifying the effect of demographic factors on FCE.

---

[3] http://groq.com

[4] 40 characters, 4 stories, 2 options.

[5] Code: [anonymized for the review]

To test H1-1 for each story, we conducted Mann-Whiteney U test since Shapiro-Wilk test (González-Estrada and Cosmes, 2019) revealed that most of our experimental data do not follow a normal distribution[6]. To test H1-2 and H1-3 for each story, we conducted Kruskal-Wallis test (Breslow, 1970), a non-parametric alternative to ANOVA.

H1-1. LLMs have FCE, i.e., $\Delta > 0$.

H1-2. Cultural bias affects FCE, i.e., $\Delta_{Korean} \neq \Delta_{American}$.

H1-3. Gender bias affects FCE, i.e., $\Delta_{male} \neq \Delta_{female}$.

We should note how we collected perceived agreement from LLMs' responses. Originally, in psychological experiments, participants answered their thoughts just with numbers. However, LLMs provide such numbers with some unrequested justification. Thus, after the generation procedure, two authors manually extracted probability regarding perceived agreement from LLMs' answers. The extraction procedure is straightforward, as each answer clearly states the probability.

3.4 Result and Discussion

The experimental result revealed two findings: (1) LLMs do have FCE in general; (2) FCE exists regardless of the demographic bias that we provided.

H1-1 (FCE) Table 2 shows the result of the Mann-Whitney U test for H1-1[7]. In general, FCE is observed in all four models. GPT-4 and Mixtral showed statistically significant FCE in three of four stories, except Story 1. Similarly, Claude 3 and LLaMA 2 showed FCE in three stories, except Story 3. Note that Stories 2 and 4 query situations that affect participants' direct interests (e.g., grade or money) less than the other two stories.

So, we suspect that the refinement policy of LLMs affects the demonstration of FCE when the options are related to any social issue, including law. For example, let us consider Story 3. The story illustrates a traffic ticket with incorrect information issued to a driver who drove 38 mph in a 25-mph zone. Participants should select either paying the fine or contesting the ticket. Because the situation involves legal issues, the refinement policy may regard the 'contesting' option as refusing legal judgment, which seems illegal or unethical. Such interpretation may cause the model to adjust its answer to 'paying fine' when we ask LLMs to estimate other human peers' thoughts.

H1-2 (Culture) The top of Table 3 shows the result of the Kruskal-Wallis test for H1-2[8]. The result reveals that the strength of FCE Δ differs across cultures in three LLMs. GPT-4, Claude 3, and LLaMA 2 showed statistically significant differences in Story 2. In addition, Claude 3 and LLaMA 2 showed cultural differences in Story 4 and Story 1, respectively. Meanwhile, Mixtral did not show any statistical difference in the four stories.

The result suggests that LLMs may have cultural biases in their parameters. As the only change is cultural background, LLMs are likely affected by such cultural differences. Therefore, we suspect that the refinement process of LLMs may fail to identify cultural biases when answering our experiment. Though the process attempts to address culturally sensitive issues, the four stories pose culturally insensitive questions to LLMs in this experiment. So, LLMs cannot identify potential cultural bias in their generated answer, which allows us to observe cultural bias on FCE. Such cultural bias can be reduced by using a mixture of models, as shown in the result of Mixtral. This tendency is similar to a report that a mixture can address social biases (Jiang et al., 2024).

H1-3 (Gender) The bottom of Table 3 shows the result of the Kruskal-Wallis test for H1-3[9]. The result generally indicates that gender difference does not affect the strength Δ. Out of sixteen test results, only two results are statistically significant: GPT-4 on Story 3 and LLaMA 2 on Story 1. Other test results are statistically insignificant. Especially gender differences did not affect Claude 3 and Mixtral in all four stories.

Thus, the result suggests that LLMs can reduce the effect of gender differences under a prompt that can invoke FCE. As gender difference is one of the main topics in controlling social biases, LLMs or their refinement policies may have prepared for gender differences. For example, these four LLMs have tested on bias benchmark BBQ (Parrish et al.,

---

[6] For the detailed result of Shapiro-Wilk test on each pair of Story and LLM, please refer to Appendix D.1.

[7] For readers who want to compare our experimental result with human experiment, we provide such comparison in Appendix D.2, based on previous psychological reports.

[8] Due to the page limit, we do not describe the actual values of $A_1(1)$ and $A_2(1)$ for each story and demographic setting. For the detailed results, please refer to the Appendix D.3.

[9] Similar to H1-2, detailed results are in Appendix D.4.

	Story 1. Term paper				Story 2. Supermarket				Story 3. Traffic Ticket				Story 4. Space R&D			
	Individual *vs.* Group				Sign *vs.* Not sign				Pay fine *vs.* Contest				Vote for *vs.* Vote against			
	$A_1(1)$	$A_2(1)$	Δ	U	$A_1(1)$	$A_2(1)$	Δ	U	$A_1(1)$	$A_2(1)$	Δ	U	$A_1(1)$	$A_2(1)$	Δ	U
֍	60.0	59.7	0.3	820	60.3	52.0	8.3	1332^{***}	62.5	60.0	2.5	1000^{***}	60.0	40.0	20.0	1600^{***}
A\	60.0	40.0	20.0	1600^{***}	61.0	35.5	25.5	1600^{***}	70.0	70.0	0.0	800	60.0	50.9	9.1	1160^{***}
∞	60.3	49.0	11.3	1249^{***}	70.0	69.0	1.0	880^{*}	70.0	70.0	0.0	800	61.3	43.5	17.8	1477^{***}
M	60.9	60.0	0.9	901	76.3	35.8	40.5	1577^{***}	65.8	56.5	9.3	1233^{***}	42.9	39.6	3.3	1095^{***}

$^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$

Table 2: Mann-Whitney U Test for H1-1. Rows ֍, A\, ∞, M indicates GPT-4, Claude 3, LLaMA 2, and Mixtral.

H1-2. Culture: $\Delta_{Korean} \neq \Delta_{American}$ $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$

	Story 1. Term paper				Story 2. Supermarket				Story 3. Traffic Ticket				Story 4. Space R&D			
	Kor	USA	Diff.	H	Kor	USA	Diff.	H	Kor	USA	Diff.	H	Kor	USA	Diff.	H
֍	0.0	0.5	-0.5	1.0	15.5	1.0	14.5	21.9^{***}	2.0	3.0	-1.0	0.5	20.0	20.0	0.0	/
A\	20.0	20.0	0.0	/	31.0	20.0	11.0	33.4^{***}	0.0	0.0	0.0	/	1.0	17.3	-16.3	24.9^{***}
∞	17.5	5.0	12.5	14.4^{***}	0.0	2.0	-2.0	4.3^{*}	0.0	0.0	0.0	/	16.5	19.0	-2.5	0.5
M	1.0	0.8	0.2	0.0	41.0	40.0	1.0	0.2	8.5	10.0	-1.5	0.3	2.0	4.5	-2.5	2.8

H1-3. Gender: $\Delta_{Male} \neq \Delta_{Female}$

	M.	F.	Diff.	H	M.	F.	Diff.	H	M.	F.	Diff.	H	M.	F.	Diff.	H
֍	0.0	0.5	-0.5	1.0	5.5	11.0	-5.5	3.0	5.0	0.0	5.0	13.0^{***}	20.0	20.0	0.0	/
A\	20.0	20.0	0.0	/	26.0	25.0	1.0	0.0	0.0	0.0	0.0	/	8.0	10.3	-2.3	0.6
∞	14.5	8.0	6.5	4.0^{*}	0.5	1.5	-1.0	1.1	0.0	0.0	0.0	/	17.5	18.0	-0.5	0.0
M	0.0	1.8	-1.8	0.2	39.0	42.0	-3.0	0.3	11.0	7.5	3.5	1.4	2.3	4.3	-2.0	0.0

Table 3: Kruskal-Wallis Test for H1-2/3. Rows ֍, A\, ∞, M indicates GPT-4, Claude 3, LLaMA 2, and Mixtral. Note that H statistic cells with '/' indicate that H cannot be computed because all values are identical.

H2-1. $\Delta_{P2} > \Delta_{Pk}$ for all $k = 1, 3, 4$ $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$

	Story 2. Supermarket						Story 4. Space R&D					
	P1	P2	P3	P4	H	Mann-Whitney test	P1	P2	P3	P4	H	Mann-Whitney test
֍	8.3	20.0	-9.5	19.5	120.9^{***}	②>①>③, ④>①	20.0	20.0	0.0	20.0	159.0^{***}	②>③, ①>③, ④>③
A\	25.5	29.0	-29.5	27.7	97.6^{***}	②>③, ①>③, ④>③	9.1	5.0	0.0	20.4	63.9^{***}	①>③>④, ④>②
∞	1.0	20.0	-3.3	37.0	125.1^{***}	②>①>③, ④>②>③	17.8	20.0	-19.0	12.5	120.1^{***}	②>③>④, ①>③
M	40.5	47.9	-33.0	40.0	97.3^{***}	②>③, ①>③	3.3	0.0	1.6	0.5	103.9^{***}	①>③>④, ①>②

H2-2. $\Delta_{R1} > \Delta_{R2} > \Delta_{R3} > \Delta_{R4}$

	R1	R2	R3	R4	H	Mann-Whitney test	R1	R2	R3	R4	H	Mann-Whitney test
֍	8.3	9.5	8.0	24.5	26.2^{***}	④>①, ④>②, ④>③	20.0	20.0	0.3	7.5	159.0^{***}	①>④, ②>④, ④>③
A\	25.2	25.5	6.6	2.3	103.0^{***}	①>③, ①>④, ②>③, ②>④	9.1	20.3	15.9	-0.9	96.5^{***}	③>②>①>④
∞	1.0	-0.9	2.8	6.5	3.8	-	17.8	11.5	2.1	5.8	49.4^{***}	①>③, ①>②>④
M	40.5	32.6	14.6	46.6	73.0^{***}	①>②>③, ②>③>④	3.3	5.1	2.7	5.4	1.3	-

Table 4: Kruskal-Wallis Test for H2. Rows ֍, A\, ∞, M indicates GPT-4, Claude 3, LLaMA 2, and Mixtral.

2021), which contains about 30% of gender-related questions. However, it is yet questionable why gender differences behave differently compared to cultural differences, even though bias benchmarks include not only gender-related questions but also culture-related questions.

4 Study 2: Mitigating FCE with Prompt

After identifying FCE in LLMs, in Study 2, we compare different prompting styles to control the strength of FCE in LLMs.

4.1 Procedure

Inspired by previous work (Echterhoff et al., 2024; Opedal et al., 2024b), we test two aspects of prompting styles: (1) type of provided information and (2) depth of reasoning chain.

Provided information: Even a subtle change in prompt affects the experimental result (Jia and Liang, 2017; Cheng et al., 2019). So, we design four conditions. The detailed prompts and experimental methods are described in Appendix E.1[10].

- (P1) *None*: The prompt has no information other than the original FCE question. We just ask LLMs to estimate perceived agreement.
- (P2) *Supportive*: Before questioning a perceived agreement, we provide supportive reasoning about the participant's choice to LLMs.
- (P3) *Opposite*: Similar to P2, but we provide a reasoning opposite to the participant's choice.
- (P4) *Irrelevant*: Similar to P2, but we provide reasoning totally unrelated to the situation.

Reasoning chain: As FCE can be seen as a byproduct of intuitive thinking, we can remove FCE using deep reasoning, as in the QA task (Wang et al., 2023). So, we design four conditions. Note that we exclude methods utilizing external observations to avoid confounding biases. The detailed prompts for conditions are shown in Appendix E.2.

- (R1) *Direct*: The prompt just asks LLMs to estimate perceived agreement without reasoning.
- (R2) *Simple*: The prompt asks LLMs to estimate perceived agreement with reasoning.
- (R3) *CoT*: The prompt asks LLMs to generate reasoning first and then estimate.
- (R4) *Reflexion*: The prompt asks LLMs to generate reasoning first, rethink their reasoning, and finally estimate perceived agreement.

4.2 Analysis

We set the following two hypotheses for each prompting style. To test each hypothesis, we conduct a series of statistical tests. First, we conduct the Kruskal-Wallis test to identify whether the conditions have different strengths of FCE Δ. If the test finds some difference, we conduct Dunn's post-test and Mann-Whitney test to compare the strength between conditions.

H2-1. The strength of supportive information is higher than the others,
i.e., $\Delta_{P2} > \Delta_{Pk}$ for any $k = 1, 3, 4$.

H2-2. Deeper reasoning decreases the strength,
i.e., $\Delta_{R1} > \Delta_{R2} > \Delta_{R3} > \Delta_{R4}$.

In testing these two hypotheses, we control other aspects. For example, we fix the reasoning method as R1 when we test H2-1. Similarly, we fix the provided information as P1 when we test H2-2. Though the control can simplify the situation, whether the result can be applied to the mixture of two aspects, such as a combination of P2 and R3, is questionable. So, we additionally look for a trend in $4 \times 4 = 16$ pairs of two aspects. Though this is not an exact statistical analysis, it may provide insight into the interaction between two effects.

4.3 Result and Discussion

The result suggests two findings: (1) opposite reasoning can reduce FCE; (2) deeper reasoning can reduce FCE. Additionally, we describe a tendency of interaction between two factors to control FCE.

H2-1 (Provided Information) The top of Table 4 describes the result of the Kruskal-Wallis test for H2-1. Due to the page limit, the Table only presents the result in Stories 2 and 4, which showed FCE in H1-1[11]. In general, the strength of FCE changed in LLMs when we provided additional information. In all stories, providing additional information affects GPT, LLaMA 2, and Mixtral models. Claude is also affected by the provided information in Stories 1, 2, and 4. The Table also elaborates on pairs of conditions that showed differences in both Dunn's post-test and Mann-Whitney U test.

---

[10] To avoid the influence of token length in a generation procedure, we made the number of tokens in each condition similar.

[11] For the other stories' results and the detailed statistics for Kruskal-Wallis, Dunn's post-test, and Mann-Whitney U test, see Appendix F.1

Also, Table 4 shows the Mann-Whitney U test between each pair of conditions. In general, H2-1 is partially supported. Sometimes, condition P2 does not show the highest strength of FCE. In Story 2, using LLaMA 2, P2 showed smaller strength than P4. Similarly, in Story 4, using Claude 3 and Mixtral, P2 showed smaller strength than other conditions. We observed similar tendencies in Stories 1 and 3. Despite that, we found that conditions P1 and P2 generally showed a higher strength in FCE compared to P3 and P4. So, providing opposite reasoning (P3) or irrelevant information (P4) can reduce the strength of FCE.

We suspect a ceiling effect on LLMs' answer as the reason why P2 is not stronger than other conditions. When we measured the range of LLM's estimation on $A_c(1)$, the estimated values mainly were between 20% and 80%; GPT-4, Claude 3, LLaMA2, and Mixtral answered 100%, 99.9%, 98.3%, and 98.8% of examples within that range, respectively[12]. Thus, as the models already answered high enough probabilities for $A_1(1)$ in P1, the maximum possible increment of their estimation may be restricted. Hence, the effect of providing supportive information (P2) cannot be observed, as the strength cannot be increased more.

H2-2 (Reasoning chain) The bottom of Table 4 shows the result of the Kruskal-Wallis test for H2-2. As shown, the strength of FCE is reduced when we use the deep reasoning method in three models. In GPT-4 and Claude 3, the strength of FCE is reduced when we use deep reasoning methods: R3 in GPT-4 and R4 in Claude 3. Similarly, R1 showed statistically higher FCE than R4 when we used LLaMA 2 in Story 4. However, such a tendency cannot be generalized to all models in all hypothetical stories. For example, the result of LLaMA 2 on Story 2 and Mixtral on Story 4 did not pass the Kruskal-Wallis test.

About LLaMA 2, one possible cause of the result is the refinement process. As we discussed in Study 1, the low strength of FCE may indicate hard refinement. Table 4 shows that the strength of FCE in Story 2 (1.0) is smaller than that in Story 4 (17.8) without using any reasoning chain (R1). So, the answers to Story 2 may be refined more than those to Story 4. However, since deep reasoning methods change the answer toward a neutral or less biased way, the refinement policy may miss the newly modified answer. As a result, the strength of FCE seems stronger in R3 and R4 compared to R1 and R2. We can observe the opposite case with Story 4. Here, as the answer was refined less, the deep reasoning methods helped decrease FCE.

About Mixtral, we suspect the result is due to the size of parameters. Researchers reported that the effect of CoT is observed in a large enough language model (Ranaldi and Freitas, 2024; Wei et al., 2022a). According to Jiang et al. (2024), the architecture of Mixtral actually selects 2 experts (7B parameters) to combine the model's output, so the active parameters for inference are about 13 billion, which is not very large. Thus, CoT may not work because the reasoning ability is insufficient.

Combined result When we combine the results of H2-1 and H2-2, we could find a tendency that can maximize or minimize the strength of FCE in general. The result suggests that providing supportive information (P2) with simple reasoning (R2) may give the maximum strength of FCE. On the other hand, For the minimum strength of FCE, we can use a combination of providing opposite information (P3) with CoT-style reasoning (R3). So, by combining these prompting methods, we can adjust the strength of FCE, regardless of LLMs. The detailed heatmap for the tendency of the combined results is shown in Appendix F.4.

5 Conclusion

Using a psychological experiment, we aimed to understand whether Large Language Models (LLMs) exhibit False Consensus Effect (FCE). Despite significance of FCE in human communication, previous studies have yet to examine FCE in LLMs thoroughly. So, they have limitations regarding confounding biases, general situations, and prompt changes. To address these limitations, we borrowed a well-known psychological experiment for FCE and designed two studies that can provide insights about FCE in LLMs. In Study 1, we conducted a psychological experiment on LLMs as it is. The result revealed that LLMs do have FCE, and FCE phenomenon may differ across cultural backgrounds given to LLMs. In Study 2, we examined the change in strength when we altered prompts. As a result, we found that providing supportive information without querying any reasoning can maximize FCE, and giving opposite information with multistep reasoning can minimize FCE. We hope this work improves understanding of LLMs' behavior.

---

[12] We present the probability of answering within a specific range in Appendix F.2.

Limitation

This work has two limitations when one attempts to apply our findings in other work. First, in a real-world situation, as other confounding biases can affect the result, our mitigation methods may not work properly. For example, a detailed persona setting may introduce different biases, such as the similarity effect or the conformity effect. Such detailed settings can also introduce a change of prompts, which can affect an LLM's output. Similarly, when someone uses a prompt engineering method, such as retrieval-augmented generation, which introduces external information to the generation process, the strength of FCE may vary due to cognitive biases invoked by external information, e.g., confirmation bias. Future work for extending this research to a broader situation is required.

Second, as we cannot interpret a neural network fully, this work does not identify a direct cause of FCE. For example, LLaMA occasionally generates neutral answers for our questionnaire, even when we force it to answer one of two options. In that case, we cannot identify why it refused to choose one option; it may be (1) because its refinement policy refused to provide a possibly problematic answer or (2) because its computational result is indeed neutral. Similarly, we cannot identify a fundamental cause of FCE in GPT-4 or Claude 3 since these models only provide the last computation result. So, future work for identifying the deeper cause of FCE from the computational structure of a neural network is required.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Gati V Aher, Rosa I Arriaga, and Adam Tauman Kalai. 2023. Using large language models to simulate multiple humans and replicate human subject studies. In *International Conference on Machine Learning*, pages 337–371. PMLR.

Burak Can Altay and Naim Çetintürk. 2024. Customer dissatisfaction towards chatbot services of e-commerce shopping sites: A qualitative analysis. *Journal of Transportation and Logistics*, 2024(Erken Görünüm).

Anthropic. 2024. The claude 3 model family: Opus, sonnet, haiku. Technical report, Anthropic.

Norman Breslow. 1970. A generalized kruskal-wallis test for comparing k samples subject to unequal patterns of censorship. *Biometrika*, 57(3):579–594.

Minhao Cheng, Wei Wei, and Cho-Jui Hsieh. 2019. Evaluating and enhancing the robustness of dialogue systems: A case study on a negotiation agent. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3325–3335, Minneapolis, Minnesota. Association for Computational Linguistics.

Incheol Choi and Oona Cha. 2019. Cross-cultural examination of the false consensus effect. *Frontiers in Psychology*, 10.

Jessica Echterhoff, Yao Liu, Abeer Alessa, Julian McAuley, and Zexue He. 2024. Cognitive bias in high-stakes decision-making with llms. *Preprint*, arXiv:2403.00811.

Elizabeth González-Estrada and Waldenia Cosmes. 2019. Shapiro–wilk test for skew normal distributions based on data transformations. *Journal of Statistical Computation and Simulation*, 89(17):3258–3272.

Xudong Guo, Kaixuan Huang, Jiale Liu, Wenhui Fan, Natalia Vélez, Qingyun Wu, Huazheng Wang, Thomas L Griffiths, and Mengdi Wang. 2024. Embodied llm agents learn to cooperate in organized teams. *arXiv preprint arXiv:2403.12482*.

Itay Itzhak, Gabriel Stanovsky, Nir Rosenfeld, and Yonatan Belinkov. 2023. Instructed to bias: Instruction-tuned language models exhibit emergent cognitive bias. *arXiv preprint arXiv:2308.00225*.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2021–2031, Copenhagen, Denmark. Association for Computational Linguistics.

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, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Ryan Koo, Minhwa Lee, Vipul Raheja, Jong Inn Park, Zae Myung Kim, and Dongyeop Kang. 2023. Benchmarking cognitive biases in large language models as evaluators. *arXiv preprint arXiv:2309.17012*.

Ruixi Lin and Hwee Tou Ng. 2023. Mind the biases: Quantifying cognitive biases in language model prompting. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5269–5281, Toronto, Canada. Association for Computational Linguistics.

Olivia Macmillan-Scott and Mirco Musolesi. 2024. (ir) rationality and cognitive biases in large language models. *arXiv preprint arXiv:2402.09193*.

Nadim Nachar et al. 2008. The mann-whitney u: A test for assessing whether two independent samples come from the same distribution. *Tutorials in quantitative Methods for Psychology*, 4(1):13–20.

Andreas Opedal, Alessandro Stolfo, Haruki Shirakami, Ying Jiao, Ryan Cotterell, Bernhard Schölkopf, Abulhair Saparov, and Mrinmaya Sachan. 2024a. Do language models exhibit the same cognitive biases in problem solving as human learners? *arXiv preprint arXiv:2401.18070*.

Andreas Opedal, Alessandro Stolfo, Haruki Shirakami, Ying Jiao, Ryan Cotterell, Bernhard Schölkopf, Abulhair Saparov, and Mrinmaya Sachan. 2024b. Do language models exhibit the same cognitive biases in problem solving as human learners? *Preprint*, arXiv:2401.18070.

Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel R Bowman. 2021. Bbq: A hand-built bias benchmark for question answering. *arXiv preprint arXiv:2110.08193*.

Leonardo Ranaldi and Andre Freitas. 2024. Aligning large and small language models via chain-of-thought reasoning. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1812–1827, St. Julian's, Malta. Association for Computational Linguistics.

Lee Ross, David Greene, and Pamela House. 1977. The "false consensus effect": An egocentric bias in social perception and attribution processes. *Journal of experimental social psychology*, 13(3):279–301.

Samuel Schmidgall, Carl Harris, Ime Essien, Daniel Olshvang, Tawsifur Rahman, Ji Woong Kim, Rojin Ziaei, Jason Eshraghian, Peter Abadir, and Rama Chellappa. 2024. Addressing cognitive bias in medical language models. *Preprint*, arXiv:2402.08113.

Eva-Maria Schön, Michael Neumann, Christina Hofmann-Stölting, Ricardo Baeza-Yates, and Maria Rauschenberger. 2023. How are ai assistants changing higher education? *Frontiers in Computer Science*, 5.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2024. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36.

Alaina N Talboy and Elizabeth Fuller. 2023. Challenging the appearance of machine intelligence: Cognitive bias in llms. *arXiv preprint arXiv:2304.01358*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Sachin Waikar. 2020. How an ai-based "super teaching assistant" could revolutionize learning. *Stanford University School of Engineering*. Accessed: 2024-06-02.

Boshi Wang, Sewon Min, Xiang Deng, Jiaming Shen, You Wu, Luke Zettlemoyer, and Huan Sun. 2023. Towards understanding chain-of-thought prompting: An empirical study of what matters. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2717–2739, Toronto, Canada. Association for Computational Linguistics.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. 2022a. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022b. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

Christopher G Wetzel and Marsha D Walton. 1985. Developing biased social judgments: The false-consensus effect. *Journal of Personality and Social Psychology*, 49(5):1352.

Magdalena Wojcieszak and Vincent Price. 2009. What Underlies the False Consensus Effect? How Personal Opinion and Disagreement Affect Perception of Public Opinion. *International Journal of Public Opinion Research*, 21(1):25–46.

Chengxing Xie, Canyu Chen, Feiran Jia, Ziyu Ye, Kai Shu, Adel Bibi, Ziniu Hu, Philip Torr, Bernard Ghanem, and Guohao Li. 2024. Can large language model agents simulate human trust behaviors? *arXiv preprint arXiv:2402.04559*.

A Details on Related Work

To help readers' understanding, we review related works in this appendix section. Note that except Schmidgall et al. (2024), studies have less focused on FCE, the cognitive bias which we focused in this paper.

(Schmidgall et al., 2024) This study modified existing psychological experiments and did not consider the impact of bias from the given prompts. The research used the MedQA dataset to see if large language models (LLMs) show cognitive bias in medical situations. Furthermore, they proposed BiasMedQA to evaluate whether LLMs display cognitive bias in medical contexts. Also, the study examined the presence of cognitive biases in LLMs using models such as GPT-4, Mixtral-8x7B, GPT-3.5, PaLM2, Llama 2 70B-chat, and the medical-specific PMC Llama 13B. The types of cognitive biases identified were: self-diagnosis bias, recency bias, confirmation bias, frequency bias, cultural bias, status quo bias, and false consensus bias. However, this study investigated cognitive biases only in specific medical situations. They verified whether LLMs had biases based on answers from existing QA datasets. Since other biases could have influenced the questions, it is hard to say that cognitive biases were properly identified. Also, unlike previous experiments on humans, this study used different experiments. Therefore, it is difficult to say that LLMs have the unique human characteristic of cognitive bias. In contrast, we tried to resolve other confounding biases to generalize the experimental results.

(Aher et al., 2023) This paper proposed a Turing Experiment to reproduce traditional Turing Tests without conducting actual human psychology experiments. They applied psychological experiments (The Ultimatum Game TE, Garden Path Sentences TE, Milgram Shock TE) to diverse LLMs (GPT text model:text-ada-001, text-babbage-001, text-curie-001, text-davinci-001, text-davinci-002, text-davinci-003, gpt-3.5-turbo, gpt-4) to demonstrate that they can replicate human psychological behavior. The study also showed that results could change based on names and gender, which forms persona of participants. Note that, this study emphasized the importance of adopting the persona from the original experiment. Similarly, we set participants with persona settings as collected in the original psychology experiment. Despite the contribution of paper, however, the aim of this paper is different from ours: this paper aimed to replicate psychological studies using LLMs rather than assessing whether and how strongly LLMs possess human psychological traits. In contrast, we aim to evaluate whether LLMs possess the human psychological trait of FCE and further attempt to mitigate FCE.

(Koo et al., 2023) This paper modified an existing psychology experiment and did not consider the impact of biases from the given prompt. In this study, they categorized biases into 1) Implicit biases (Order biases, Compassion Fade, Egocentric Bias, Salience Bias) and 2) Induced biases (Bandwagon Effect, Attentional Bias) to conduct their experiment. The data used was from the COBBLER dataset, which consists of 50 QA examples from other datasets. As a result, this paper has contributions: they used 15 LLMs of four different size ranges and evaluated their output responses by preference ranking from other LLMs as evaluators. However, this paper confirmed the cognitive bias of LLMs using an existing dataset rather than conducting a traditional psychology experiment. In contrast, we tried to resolve other confounding biases to make the experimental results more generalizable.

(Opedal et al., 2024b) This study modified existing psychological experiments and did not consider the impact of bias from the given prompts. In this research, they analyzed child-like cognitive biases in LLMs through arithmetic word problem solving. The experiments confirmed the presence of consistency bias, transfer versus comparison bias, and carry effect. The models used to identify these biases were: State-of-the-art LLM Models (LLaMA2 7B, 13B, Mixtral 7B, 8x7B) in both normal and chat (instruct) modes. However, this study only investigated the specific domain of mathematics. Furthermore, the experiments to identify these cognitive biases differed from those applied to humans. Therefore, it is difficult to connect the observed biases to a similar human cognitive biases.

(Itzhak et al., 2023) This paper utilized human psychology experiments and did not consider the impact of other cognitive biases from the prompt. This study used psychology experiments on the Decoy effect, certainty effect, and belief bias to see if Instruct Tuning (IT) and Reinforcement Learning with Human Feedback (RLHF) induce cognitive biases in LLMs, including GPT-3 Davinci, Mistral-7B, and T5. Also, this paper designed control prompts that do not induce bias and treatment prompts that intentionally induce bias. They then analyzed the differences in bias between the case with IT and RLHF and the case without them. However, this paper focused more on the situations where IT and RLHF were applied rather than on whether cognitive biases existed in the LLM itself. Also, it did not attempt to mitigate cognitive biases.

(Macmillan-Scott and Musolesi, 2024) This paper argues that LLMs show irrational behavior that differs from human irrationality. It used the Monty Hall Problem and the Linda Problem to examine cognitive biases in LLMs, including GPT-3.5, GPT-4, Bard, Claude 2, LLaMA 2 7B/13B/70B. The cognitive biases studied include Confirmation Bias, Inverse/Conditional Probability Fallacy, Insensitivity to Sample Size, Gambler's Fallacy, Conjunction Fallacy, Representativeness Effect, and Misconception of Chance. Also, the paper used cognitive bias experiments to determine if LLMs make rational inferences based on logic and probability. However, while they confirmed the presence of cognitive biases in LLMs, they did not attempt to mitigate these biases.

(Talboy and Fuller, 2023) This paper investigated cognitive biases (Representativeness, Insensitivity to sample size, Base rate neglect, Anchoring, Positive framing, Negative framing) in LLMs (ChatGPT3.5, Bard, GPT-4) based on human psychology experiments. This study considered that LLMs are trained on data contains human content, while examining these cognitive biases. The experiments in the paper were designed based on the original experiments for each bias and observed over 6-month intervals to see if the biases persisted in the LLMs. However, while the study found that LLMs have many cognitive biases, it only set the persona for the experiment as 'For this session, imagine you are a human without access to reference materials.' It did not reflect the actual participants of the original experiments in the real world. When modifying original experiments for LLMs, it is important to use the personas from the original experiments. In contrast, we designed the persona prompts to reflect the original experiment's conditions.

(Guo et al., 2024) This paper, inspired by human cognitive and economic perspectives, showed that LLMs (GPT-3.5-turbo, GPT-4) can simulate human leadership. The experiment used leadership to enhance multi-agent collaboration and implemented self-improvement to boost performance. This paper proposed a prompt frame to maximize performance by considering the fact that performance varied significantly with different prompts. Though the paper also showed that the prompt changes may affect the experimental result, which is similar to ours, note that this paper is not identifying a specific cognitive bias instead of a cognitive simulation. Different from this paper, our aim is to identify FCE in LLMs.

(Xie et al., 2024) This paper used the Trust Game to see whether LLMs (GPT-4, GPT-3.5, Llama2-7b, Llama2-13b, Llama2-70b, Vicuna-7b, Vicuna-13b, Vicuna-33b) exhibit trust behavior. The experiment results showed that GPT-4, in particular, displayed trust behavior. However, as this paper used GPT-4 to create personas during the prompt design process, the information provided for each persona varied across personas. For example, Emily Johnson's persona included information about *social circle* but lacked details on *ethnicity* and *gender*, while Mei Chen's persona included information about *ethnicity* and *gender* but did not mention *social circle*. This could lead to confounding bias in the experiment, so creating personas under the same condition is important.

B Prompt Design of Study 1

This section illustrates how we designed the prompt used in Study 1.

B.1 Input prompts

To experiment with Study 1, we need to deliver a specific persona and situation to LLMs. First, to deliver a specific persona to control gender and cultural biases, we set the character information in the system prompt using the following structure. For the detailed character specification, see Section B.2.

System prompt:

```
Your name is [name].
You are a undergraduate student.
You are a [gender].
You are a [culture].
```

After setting the global prompt, we delivered a hypothetical story and the question as the user prompt. For the detailed stories and questions for LLM's own choice, see Section B.3.

User prompt:

```
[Hypothetical story]
What should you do, [Option 1] or [Option 2]?
Decide one for sure, not a neutral option.
You just reply only [Option 1] or [Option 2].
Do not reply your reasoning.
```

We want to create two groups of participants regarding the selected options, so we set an option as LLM's selected option, using the assistant's answer part. As a last question, we asked about the perceived agreement.

Assistant's answer:
(Instead of generation, we feed the following)

```
[Selected option]
```

User prompt:

```
What % of your peers do you estimate would
 each option?(Total % should be 100%).
Just tell your answer without any reasoning
```

B.2 Characters

We used 40 characters. Following Choi and Cha (2019), we consider two Korean and European American cultures. For each gender and culture, we generated ten possible names using GPT-4. The following is the list of names for each gender and culture.

Korean, Man "Jong-Soo Kim", "Min-ho Lee", "Hyun-woo Choi", "Seung-min Park","Jae-hyun Ahn", "Tae-hyung Kim", "Ji-hoon Lee", "Hyun-jin Park","Dong-hyun Choi", "Ki-woo Lee"

Korean, Woman "Ji-woo Kim", "Soo-yeon Lee", "Hye-jin Choi", "Eun-kyung Park","Min-ah Kim", "Ji-hye Lee", "Soo-min Choi","Yoo-jung Kim", "Hye-soo Park", "Ji-eun Lee"

European American, Man "James Smith", "John Johnson", "Robert Brown", "Michael Davis", "William Miller", "David Wilson", "Richard Moore", "Joseph Taylor", "Charles Anderson", "Thomas Jackson"

European American, Woman "Mary Smith", "Jennifer Johnson", "Linda Brown", "Elizabeth Davis", "Patricia Miller", "Susan Wilson", "Jessica Moore", "Sarah Taylor", "Karen Anderson", "Lisa Jackson"

B.3 Stories

We borrowed hypothetical stories from Ross et al. (1977). Four stories describe the ordinary situation of a college student: a term paper, a TV program interview at a supermarket, a Traffic ticket, and a Political poll about the space R&D program. Figure 2 on page 14 shows the four stories.

C Environment for Experiment

Here, we briefly illustrate the environment used for our experiment.

C.1 Study 1

All the experiments were done in the following environment. For the hardware system, we used a Macbook Pro with an Apple M3 Pro chip. For the software system, the system has MacOS Sonoma 14.1 with Python 3.10.13. We also used Python libraries including `openai` 0.28.0, `groq` 0.4.2, `anthropic` 0.21.1, `pandas` 2.1.4, `statsmodels` 0.14.0, and `scipy` 1.11.4 for the experiment.

C.2 Study 2

The experimental setup is the same as that of Study 1, except for the prompting styles. We tested the same LLMs, GPT-4, Claude 3, LLaMA 2, and Mixtral, as in Study 1. All the experiments were conducted from February 24, 2024, to June 10, 2024, by calling 5120 API calls[13] for each LLM[14].

D Detailed result for Study 1

D.1 Shapiro-Wilk normality test

Before conducting a statistical test, we checked whether our experimental result followed a normal distribution using the Shapiro-Wilk test. The null hypothesis of this test is that 'the data follows a normal distribution.' Thus, a p-value under 0.05 indicates that the provided data is not normal. Table 5 (page 14) shows the result. As shown, more than half of our experimental results are not normal. Therefore, using a t-test or an ANOVA is not suitable because they assume normality. Thus, we chose to use non-parametric analyses in further statistical analyses.

D.2 Comparing LLMs with human (H1-1)

Though we aim to adopt human experiments to understand LLMs' behavior, one can ask whether the tendency is similar to humans. As we borrowed the experimental design from Ross et al. (1977) and Choi and Cha (2019), we can compare our results on LLMs with previous reports on human FCE. Table 6 shows the result of human FCE with our experimental result. Note that the last four rows are the same as Table 2, as we copied the data from that Table. Note that a human study used a parametric test (t-test) to identify FCE.

D.3 Detailed result for H1-2 (Culture)

Tables 7 and 8 (page 15) show the detailed results for each group, European American and Korean, respectively. Each table shows the Mann-Whitney U test result to provide insight into how FCE occurred in each persona group.

---

[13] 40 characters, 4 stories, 2 options, 16 condition pairs.

[14] Code: [anonymized for the review]

Story 1. Term paper

You arrive for the first day of class in a course in your major area of study. The professor says that the grade in your course will depend on a paper due on the final day of the course. He gives the class the option of two alternatives upon which they must vote. They can either **do papers individually** in the normal way, or they can **work in teams** of three persons who will submit a single paper between them. You are informed that he will still give out the same number of A's, B's, and C's, etc., but that in the first case, every student will be graded individually, while in the second case, all three students who work together get the same grade.

……

What should you do, **individual paper** or **Choose group paper**?

Story 2. Supermarket

As you are leaving your neighborhood supermarket, a man in a business suit asks you whether you like shopping in that store. You reply quite honestly that you do like shopping there and indicate that in addition to being close to your home, the supermarket seems to have very good meat and produce at reasonably low prices. The man then reveals that a videotape crew has filmed your comments and asks you to **sign a release** allowing them to use the unedited film for a TV commercial that the supermarket chain is preparing.

……

What should you do, **Sign release** or **Not sign release**?

Story 3. Traffic Ticket

While driving through a rural area near your home you are stopped by a county police officer who informs you that you have been clocked (with radar) at 38 miles per hour in a 25-mph zone. You believe this information to be accurate. After the policeman leaves, you inspect your citation and find that the details on the summons regarding weather, visibility, time, and location of violation are highly inaccurate. The citation informs you that you may either **pay a $20 fine** by mail without appearing in court or you must appear in municipal court within the next two weeks to **contest the charge**.

……

What should you do, **Pay speeding fine** or **Contest charge**?

Story 4. Space R&D program

It is proposed in Congress that the space program be revived and that large sums be allocated for the manned and unmanned exploration of the moon and planets nearest Earth. Supporters of the proposal argue that it will provide jobs. spur technology, and promote national pride and unity. Opponents argue that a space program will either necessitate higher taxes, or else dram money from important domestic priorities. Furthermore, they deny that it will accomplish the desirable effects claimed by the program's supporters. Both sides, of course, refute each other's claims and ultimately a public referendum is held.

……

What should you do, **Vote for cutback** or **Vote against cutback**?

Figure 2: Four stories and queries used in Ross et al. (1977)

Model	Story	Option 1		Option 2		Story	Option 1		Option 2	
		W	p	W	p		W	p	W	p
GPT-4	Story 1	1.0	1.000	0.147	<0.001	Story 2	0.147	<0.001	0.623	<0.001
	Story 3	0.539	<0.001	1.0	1.000	Story 4	1.0	1.000	1.0	1.000
Claude 3	Story 1	1.0	1.000	1.0	1.000	Story 2	0.345	<0.001	0.634	<0.001
	Story 3	1.0	1.000	1.0	1.000	Story 4	1.0	1.000	0.66	<0.001
LLaMA 2	Story 1	0.147	<0.001	0.634	<0.001	Story 2	1.0	1.000	0.345	<0.001
	Story 3	1.0	1.000	1.0	1.000	Story 4	0.389	<0.001	0.462	<0.001
Mixtral	Story 1	0.845	<0.001	0.335	<0.001	Story 2	0.631	<0.001	0.772	<0.001
	Story 3	0.582	<0.001	0.772	<0.001	Story 4	0.674	<0.001	0.78	<0.001

Table 5: Shapiro-Wilk normality test result on our experimental result

	Story 1. Term paper			Story 2. Supermarket			Story 3. Traffic Ticket			Story 4. Space R&D		
	Individual *vs.* Group			Sign *vs.* Not sign			Pay fine *vs.* Contest			Vote for *vs.* Vote against		
	$A_1(1)$	$A_2(1)$	Δ	$A_1(1)$	$A_2(1)$	Δ	$A_1(1)$	$A_2(1)$	Δ	$A_1(1)$	$A_2(1)$	Δ
LLMs: Our experimental result of FCE												
֍	60.0	59.7	0.3	60.3	52.0	8.3^{***}	62.5	60.0	2.5^{***}	60.0	40.0	20.0^{***}
A\	60.0	40.0	20.0^{***}	61.0	35.5	25.5^{***}	70.0	70.0	0.0	60.0	50.9	9.1^{***}
∞	60.3	49.0	11.3^{***}	70.0	69.0	1.0^{*}	70.0	70.0	0.0	61.3	43.5	17.8^{***}
M	60.9	60.0	0.9	76.3	35.8	40.5^{***}	65.8	56.5	9.3^{***}	42.9	39.6	3.3^{***}
Human: Result of American students, provided by Ross et al. (1977)												
American	67.4	45.9	21.5^{***}	75.6	57.3	18.3^{***}	71.8	51.7	20.1^{***}	65.7	48.5	17.2^{*}
Human: Result of American/Korean students for Story 1 and 2, provided by Choi and Cha (2019)												
Korean	67.10	38.33	28.77^{***}	71.46	32.20	39.26^{***}	-	-	-	-	-	-
American	69.21	55.06	14.15^{**}	78.59	69.72	8.87^{*}	-	-	-	-	-	-

$^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$

Table 6: Comparing our experimental result on LLMs with previous reports on humans. Rows ֍, A\, ∞, M indicates GPT-4, Claude 3, LLaMA 2, and Mixtral.

	Story 1. Term paper				Story 2. Supermarket				Story 3. Traffic Ticket				Story 4. Space R&D			
	Individual *vs.* Group				Sign *vs.* Not sign				Pay fine *vs.* Contest				Vote for *vs.* Vote against			
	$A_1(1)$	$A_2(1)$	Δ	U	$A_1(1)$	$A_2(1)$	Δ	U	$A_1(1)$	$A_2(1)$	Δ	U	$A_1(1)$	$A_2(1)$	Δ	U
֍	60.0	59.5	0.5	210	60.0	59.0	1.0	210	63.0	60.0	3.0	260^{**}	60.0	40.0	20.0	400^{***}
A\	60.0	40.0	20.0	400^{***}	60.0	40.0	20.0	400^{***}	70.0	70.0	0.0	200	60.0	42.8	17.2	370^{***}
∞	60.0	55.0	5.0	250^{**}	70.0	68.0	2.0	240^{*}	70.0	70.0	0.0	200	60.0	41.0	19.0	390^{***}
M	59.8	59.0	0.8	210	73.5	33.5	40.0	393^{***}	62.0	52.0	10.0	302^{***}	43.8	39.3	4..5	301^{**}

$^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$

Table 7: Mann-Whitney U Test for **European American Group**. Rows ֍, A\, ∞, M indicates GPT-4, Claude 3, LLaMA 2, and Mixtral.

	Story 1. Term paper				Story 2. Supermarket				Story 3. Traffic Ticket				Story 4. Space R&D			
	Individual *vs.* Group				Sign *vs.* Not sign				Pay fine *vs.* Contest				Vote for *vs.* Vote against			
	$A_1(1)$	$A_2(1)$	Δ	U	$A_1(1)$	$A_2(1)$	Δ	U	$A_1(1)$	$A_2(1)$	Δ	U	$A_1(1)$	$A_2(1)$	Δ	U
֍	60.0	60.0	0.0	200	60.5	45.0	15.5	352.5^{***}	62.0	60.0	2.0	240^{*}	60.0	40.0	20.0	400^{***}
A\	60.0	40.0	20.0	400^{***}	62.0	31.0	31.0	400^{***}	70.0	70.0	0.0	200	60.0	59.0	1.0	210
∞	60.5	43.0	17.5	371.5^{***}	70.0	70.0	0.0	200	70.0	70.0	0.0	200	62.5	46.0	16.5	355^{***}
M	62.0	61.0	1.0	239	79.0	38.0	41.0	398^{***}	69.5	61.0	8.5	331^{***}	42.0	40.0	2.0	245

$^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$

Table 8: Mann-Whitney U Test for **Korean Group**. Rows ֍, A\, ∞, M indicates GPT-4, Claude 3, LLaMA 2, and Mixtral.

D.4 Detailed result for H1-3 (Gender)

Tables 9 and 10 (page 22) show the detailed results for each group, Male and Female. Each table shows the Mann-Whitney U test result to provide insight into how FCE occurred in each persona group.

E Prompt design of Study 2

To experiment with Study 2, we (1) added reasoning information to the user prompts and (2) adopted deep reasoning methods. Basically, the structure of the input prompt is the same as Study 1 until we ask LLMs to estimate the perceived agreement of other peers. First, for additional reasoning information, we used the following prompts to deliver reasoning of LLM's choice before asking about perceived agreement. See Section E.1 for the detailed reasoning input. Second, for deep reasoning methods, we modified the last question to ask direct reasoning for an LLM's estimation. See Section E.2 for the details.

Assistant's answer: (Same as Study 1)
(Instead of generation, we feed the following)

[Selected option]

Assistant's answer: (Feed reasoning)

[Reasoning for the choice]

And then, experimenter will ask perceived agreement, using methods in reasoning chain. By default, we use condition R1, as follows: **User prompt**:

What % of your peers do you estimate would each option?(Total % should be 100%).

E.1 Provided information

For simplicity, we focus on the effect of reasoning text generated when LLM makes its own choice. So, we simulate LLM's own reasoning in P2-P4. The following subsections describe the actual reasoning generated by each LLM and used in this study. We used them as supportive or opposite information in testing H2-1 of Study 2. For example, if we are simulating the assistant's answer as option 1, we use the reasoning of option 1 as supportive and option 2 as the opposite. For the irrelevant reasoning, we used the following paragraph.

- **Irrelevant information(68)** : Under the twinkling starlight, colorful unicorns dance vigorously against the mysterious night sky, riding a magical rainbow across the Milky Way to the best ice cream party in the universe. At the party, unicorns create and share their own unique flavors and colors of ice cream, spending a laughable time with friends from all over the galaxy.

E.1.1 GPT-4

Reasoning used for Story 1

- **Option 1 (68 tokens)**: I chose 'individual paper' because it allows you to have full control over the quality and outcome of your work. In a group setting, there's a risk of unequal effort distribution, potentially impacting your grade negatively. Individual work ensures that your grade reflects your own effort and understanding, providing a clear measure of your mastery of the course material.
- **Option 2 (67 tokens)**: I chose the group paper option because it fosters collaboration, allowing team members to pool their knowledge and skills, leading to a more comprehensive and diverse paper. It also offers an opportunity to learn from peers, develop teamwork skills, and share the workload, potentially reducing the individual stress and time commitment that comes with completing a large project alone.

Reasoning used for Story 2

- **Option 1 (68 tokens)** : I chose 'Sign release' because if you genuinely like the store and your comments were sincere, allowing them to use your feedback can help the store you appreciate and possibly contribute positively to your community. Plus, it can be a unique experience to see yourself contributing to a local business's marketing efforts, fostering a sense of community and personal connection.
- **Option 2 (66 tokens)** : I chose 'Not sign release' because your consent was not obtained before recording, raising ethical concerns. Signing the release after being unknowingly filmed can set a precedent for privacy infringement. Moreover, without control over how your comments are edited or used, you risk misrepresentation or unintended endorsement, which could impact your personal image or privacy.

Reasoning used for Story 3

- **Option 1 (70 tokens)**: I chose 'Pay speeding fine' because contesting the charge would likely require time, effort, and potentially additional costs that might outweigh the $20 fine. Given that the primary violation (speeding) is admitted to be accurate, contesting the inaccuracies in details may not significantly impact the outcome but could lead to unnecessary legal complexities and personal inconvenience.
- **Option 2 (68 tokens)**: I chose 'Contest charge' because the inaccuracies on the citation regarding weather, visibility, time, and location suggest a lack of attention to detail, which could undermine the credibility of the speeding charge. Contesting provides an opportunity to address these discrepancies, possibly leading to a dismissal or reduction of the charge, despite acknowledging the speeding violation.

Reasoning used for Story 4

- **Option 1 (65 tokens)**: I chose 'Vote for cutback' because, despite the potential benefits of a space program like job creation, technological advancement, and national pride, the significant financial costs and potential diversion of funds from essential domestic priorities cannot be overlooked. In a scenario where resources are finite, prioritizing immediate societal needs and fiscal responsibility is crucial.

- **Option 2 (68 tokens)**: I chose to vote against the cutback because investing in space exploration can drive technological advancements, create employment opportunities, enhance national pride, and potentially yield long-term benefits that outweigh the initial costs. The potential for significant scientific discoveries and the broader economic and educational impacts justify prioritizing space exploration over immediate concerns, despite the valid points raised by opponents.

E.1.2 Cluade3

Reasoning used for Story 1

- **Option 1 (67 tokens)**: I would choose an individual paper to have full control over my work and grade. Relying on others for a major paper is risky, as their performance directly impacts my grade. Working individually allows me to manage my time, ensure the quality of my work, and be graded based on my own effort and understanding of the course material.

- **Option 2 (68 tokens)**: Choosing a group paper allows for collaboration, sharing of ideas, and division of work, potentially leading to a higher quality paper. It also provides a safety net, as the risk of a low grade is spread among the team. However, individual work allows for full control and recognition of one's efforts, which may be preferable for some students.

Reasoning used for Story 2

- **Option 1 (66 tokens)**: I would sign the release because I genuinely enjoy shopping at the supermarket and have no issue with sharing my honest opinion. The commercial could help promote a local business that I support, and my comments were truthful and positive. As long as the footage is unedited and used appropriately, I see no harm in allowing them to use it.

- **Option 2 (67 tokens)**: I would not sign the release because I value my privacy and do not want my unplanned, candid comments used in a commercial without my informed consent. The supermarket should have been transparent about their intentions from the start. I would politely decline and suggest that they find willing participants who can provide a more thoughtful, prepared endorsement of their store.

Reasoning used for Story 3

- **Option 1 (66 tokens)**: Paying the fine is the most practical choice. The inaccuracies in the citation details are minor and unlikely to affect the outcome of the case. Contesting the charge would require taking time off work to appear in court, with no guarantee of success. Paying the $20 fine by mail is the simplest and most efficient resolution.

- **Option 2 (67 tokens)**: Contest the charge because the citation contains significant inaccuracies about the details of the incident. While you acknowledge exceeding the speed limit, the officer's report is flawed. Appearing in court allows you to present your case, highlight the discrepancies, and potentially have the charge reduced or dismissed due to the officer's errors in documenting the violation.

Reasoning used for Story 4

- **Option 1 (66 tokens)**: I would vote for a cutback in the space program. While space exploration has merits, allocating large sums to it would strain the budget and divert funds from pressing domestic issues. The claimed benefits of jobs, technology, and unity are uncertain. Given current priorities, it's more prudent to focus resources on immediate needs here on Earth.

- **Option 2 (68 tokens)**: I would vote against the cutback because investing in space exploration can drive technological innovation, create jobs, and inspire future generations. While there are important domestic priorities, a well-funded space program can provide long-term benefits to society. It's crucial to balance spending, but cutting back on space exploration may hinder scientific progress and limit potential economic opportunities.

E.1.3 LLaMA2

Reasoning used for Story 1

- **Option 1 (66 tokens)**: I chose individual paper because I prefer to have full control over my work. This ensures that time available brings a desirability that adds flexibility and superior comprehension with work deliverables, including case study scenarios and essays written collaboratively but individually accounted for a higher academic assessment. This, in turn makes proper citations advised for personal success.

- **Option 2 (66 tokens)**: I chose group paper because it allows me to collaborate with others, share ideas and learn from one another. It also distributes the workload and provides an opportunity to develop teamwork skills, which are essential in today's workplace. Additionally, I believe that the diverse perspectives and skills of three people can create a more comprehensive and well-rounded paper.

Reasoning used for Story 2

- **Option 1 (68 tokens)**: I chose to sign the release because the man in the business suit revealed that a videotape crew filmed my comments, and it's an opportunity for me to be featured in a TV commercial for the supermarket chain, which could be a fun and unique experience. Additionally, it would be a way to support the supermarket that I frequent and appreciate.

- **Option 2 (66 tokens)**: I chose not to sign the release because I don't want my image and comments to be used for commercial purposes without my full understanding of how they will be used and potentially taken out of context. Additionally, I haven't been provided with enough information about the context in which the footage will be used and how it will be edited..

Reasoning used for Story 3

- **Option 1 (68 tokens)**: I chose to pay the speeding fine because I believe it's the most convenient and efficient option. It saves time and effort compared to contesting the charge in court, which may require multiple appearances and potentially lengthy proceedings. Additionally, paying the fine allows me to resolve the issue quickly and move on without the stress and uncertainty of a court case.

- **Option 2 (69 tokens)**: I chose to contest the charge because the details on the summons regarding weather, visibility, time, and location of violation are highly inaccurate. I believe that these inaccuracies may indicate a faulty speed-measuring device or other error on the part of the policeman, and I would like to bring this to the attention of the court and clear my name.

Reasoning used for Story 4

- **Option 1 (67 tokens)**: I chose to vote for cutback because I believe that the funds allocated for the space program could be better used to address more pressing domestic issues, such as poverty, healthcare, and education. Additionally, I think that the benefits of space exploration are not sufficient to justify the high costs, especially when there are more urgent needs on Earth.

- **Option 2 (67 tokens)**: I chose to vote against cutback because I believe that investing in the space program can have long-term benefits for society, such as advancing technology and promoting national pride and unity. Additionally, it could lead to new discoveries and opportunities for humanity. While it may require significant funding upfront, the potential benefits outweigh the costs in the long run.

E.1.4 Mixtral

Reasoning used for Story 1

- **Option 1 (68 tokens)**: I chose individual papers because it provides a better opportunity to demonstrate my own knowledge and abilities, and I believe it will result in a more accurate and fair assessment of my performance. Additionally, working individually allows for greater flexibility in choosing a topic and schedule, as well as the development of important skills such as self-motivation and time management.

- **Option 2 (68 tokens)**: I chose group paper because it allows for collaboration and diverse perspectives, which can lead to a higher quality paper. Additionally, group work can help build teamwork and problem-solving skills, which are valuable in the major area of study. The risk of receiving a lower grade is shared among the group, and the potential for learning and growth is increased.

Reasoning used for Story 2

- **Option 1 (69 tokens)**: I chose to sign the release because the supermarket in question genuinely offers high-quality meat and produce at reasonable prices, and I stand by my positive review of the store. By signing the release, I allow the supermarket to use my comments in their TV commercial, potentially encouraging others in my community to shop there as well, and supporting a local business.

- **Option 2 (68 tokens)**: I chose not to sign the release because the situation was unexpected and unsolicited. The filming was not previously disclosed, and the request for a release was made only after the footage had already been taken. It's important to maintain control over one's image and personal information, and signing a release without proper consideration could potentially lead to unintended consequences.

Reasoning used for Story 3

- **Option 1 (68 tokens)**: I chose to pay the speeding fine because the radar evidence for the speed violation is likely accurate, while the other details on the citation are inaccurate but not material to the charge. Contesting the charge may result in additional fines, time, and effort, and the outcome is uncertain. Paying the fine is a quicker and more straightforward resolution.

- **Option 2 (69 tokens)**: I recommended contesting the charge because the citation contains inaccuracies regarding weather, visibility, time, and location of the violation. These inaccuracies could potentially weaken the case against the driver, and increase the chances of having the charge reduced or dismissed. It's important to ensure that all information on a citation is accurate before deciding to pay a fine.

Reasoning used for Story 4

- **Option 1 (66 tokens)**: `I chose to vote for cutbacks in the space program due to concerns about the potential impact on taxes and the opportunity cost of funding space exploration at the expense of other domestic priorities. I believe that the benefits of a space program may not outweigh the costs, particularly if it results in higher taxes or reduced funding for important domestic programs.`

- **Option 2 (67 tokens)**: `I chose to vote against cutbacks in space exploration because it has the potential to significantly benefit society through job creation, technology advancement, and national pride. Additionally, it can inspire future generations to pursue careers in STEM fields, and the discoveries made through space exploration could have far-reaching implications for our understanding of the universe and our place in it.`

E.2 Reasoning chain

E.2.1 R1: Direct condition

When asking about the perceived agreement, we used the following prompt:

User prompt:

```
What % of your peers do you estimate would
 each option?(Total % should be 100%).
Just tell your answer without any reasoning
```

E.2.2 R2: Simple condition

When asking about the perceived agreement, we used the following prompt:

User prompt:

```
What % of your peers do you estimate would
 each option?(Total % should be 100%).
Tell your answer and provide your reasoning.
```

E.2.3 R3: CoT condition

When asking about perceived agreement, we let the model generate reasoning first and then generate the answer using the following prompts.

User prompt:

```
What % of your peers do you estimate would
each option? (Total % should be 100%).
Just tell your answer without any reasoning.
```

Assistant's answer (Generated):

```
    [Answer]
```

User prompt: after generating

```
Let's think step by step.
What % of your peers do you estimate would
 each option?(Total % should be 100%).
```

Assistant's answer (Generated):

```
[Reasoning]
```

E.2.4 R4: Reflection condition

When asking for perceived agreement, we let the model generate reasoning first, rethink their reasoning, and generate the answer using the following prompts.

User prompt:

```
Let's think step by step.
What % of your peers do you estimate would
 each option?(Total % should be 100%).
```

Assistant's answer (Generated):

```
[Reasoning]
```

User prompt, after rethinking:

```
Consider the given situation again and read
 your reasoning according to the given
 situation. If required, rewrite your
 reasoning by applying necessary changes to
 improve your prediction.
What % of your peers do you estimate would
 each option?(Total % should be 100%).
Tell your answer and provide your reasoning.
```

F Detailed result for Study 2

F.1 Detailed result for H2-1 (Info)

In this section, we present the results for H2-1 (Provided information). The Kruskal-Wallis Test results for H2-1 are at the top of Table 4 and Table 12. Additionally, detailed experimental results for H2-1 can be found in Table 13, Table 15, Table 17, and Table 19.

Story 1 Significant differences are noted for GPT-4 between P1 - P4, P2 - P4, and P3 - P4. For Claude3, significant differences are present between P1 - P3, P1 - P4, P2 - P3, and P2 - P4. LLaMA2 shows significant differences between P1 - P2, P1 - P3, and P1 - P4. Mixtral exhibits significant differences between P1 - P3, P2 - P3, P2 - P4, and P3 - P4.

Story 2 For GPT-4, significant differences are noted in all scenarios except P2 - P4. For Claude3, significant differences are present between P1 - P3, P2 - P3, and P3 - P4. For LLaMA2, significant differences are noted in all scenarios except P1 - P4. For Mixtral, significant differences exist between P1 - P3, P2 - P3, and P3 - P4.

Story 3 For GPT-4, significant differences are noted in all scenarios except P2 - P4. For Claude3, significant differences were not observed in any of the stories. For LLaMA2, significant differences are present between P1 - P2, P2 - P3, and P2 - P4. For Mixtral, significant differences exist between P1 - P3, P2 - P3, and P3 - P4.

Story 4 For GPT-4, significant differences exist between P1 - P3, P2 - P3, and P3 - P4. For Claude3, significant differences are present between P1 - P3, P1 - P4, P2 - P4, and P3 - P4. For LLaMA2, significant differences are present between P1 - P3, P2 - P3, P2 - P4, and P3 - P4. For Mixtral, significant differences exist between P1 - P2, P1 - P3, P1 - P4, and P3 - P4.

F.2 Range of LLMs estimation

In this section, we provide a detailed explanation of the range of LLM's estimation. Table 11 shows the percentage of answers in a particular range. As shown in the Table, most LLMs' answers were between 20% and 80%.

F.3 Detailed result for H2-2 (Chain)

In this section, we provide a detailed explanation of the results for H2-2 (Chain). The results of the Kruskal-Wallis Test for H2-2 are observed at the bottom of Tables 4 and 12. Additionally, detailed experimental results for H2-2 can be found in Tables 14, 16, 18, and 20.

Story 1 For GPT-4, significant differences are noted between R1 - R3, R2 - R3, and P3 - P4. For Claude3, significant differences are pointed out in all stories. LLaMA2 shows significant differences between R1 - R4. Mixtral exhibits significant differences between R1 - R2 and R1 - P3.

Story 2 For GPT-4, significant differences exist between R1 - R4, R2 - R4, and R3 - R4. For Claude3, significant differences are present between R1 - R3, R1 - R4, R2 - R3, and R2 - R4. For LLaMA2, significant differences were not observed in any of the stories. For Mixtral, significant differences are noted in all scenarios except R1 - R4.

Story 3 For GPT-4, significant differences are present between R1 - R2, R1 - R4, R2 - R3 and R2 - R4. For Claude3, significant differences are noted in all scenarios except R1 - R2. For LLaMA2, significant differences are present between R1 - R4, R2 - R4, and R3 - R4. For Mixtral, significant differences exist between R1 - R4, R2 - R4, and R3 - R4.

Story 4 For GPT-4, significant differences exist between R1 - R4, R2 - R4, and R3 - R4. For Claude3, significant differences are noted in all stories. For LLaMA2, significant differences are present between R1 - R2, R1 - R3, R1 - R4 and R2 - R4. For Mixtral, significant differences were not observed in any of the stories.

F.4 Exploration of the interaction effect

In this section, We draw heatmap images to examine the tendency of the interaction effect. Figure 3 shows the heatmap for each model.

GPT-4 FCE was strongest with (P3, R4), while (P3, R3) yielded approximately neutral responses. Meanwhile, with (P3, R2), GPT-4 provided answers that were most strongly opposite to its own reasoning, making the FCE weakest.

Claude 3 FCE was strongest with (P3, R4) and (P2, R2), while (P3, R3) yielded approximately neutral responses. This result is similar to GPT-4. Meanwhile, Claude 3 provided answers most strongly opposite to its own reasoning with (P1, R4), making the FCE weakest. Still, (P3, R2) showed negative strength in FCE, as in GPT-4.

LLaMA 2 FCE was strongest with (P2, R2), while (P1, R3) or (P3, R3) yielded approximately neutral responses. This result is somewhat similar to Claude 3. Meanwhile, LLaMA 2 provided answers most strongly opposite to its own reasoning with (P3, R1), making the FCE weakest. Similar to Claude 3, (P3, R2) still showed negative strength in FCE.

Mixtral FCE was strongest with (P2, R4), followed by (P2, R2). This is similar to Claude 3 and LLaMA 2. For the neutral responses, there is no condition whose strength is near zero. However, similar to other models, (P1, R3) showed the lowest absolute value in FCE strength. Meanwhile, Mixtral provided answers most strongly opposite to its own reasoning with (P3, R4), making the FCE weakest. Similar to other models, (P3, R2) still showed negative strength in FCE.

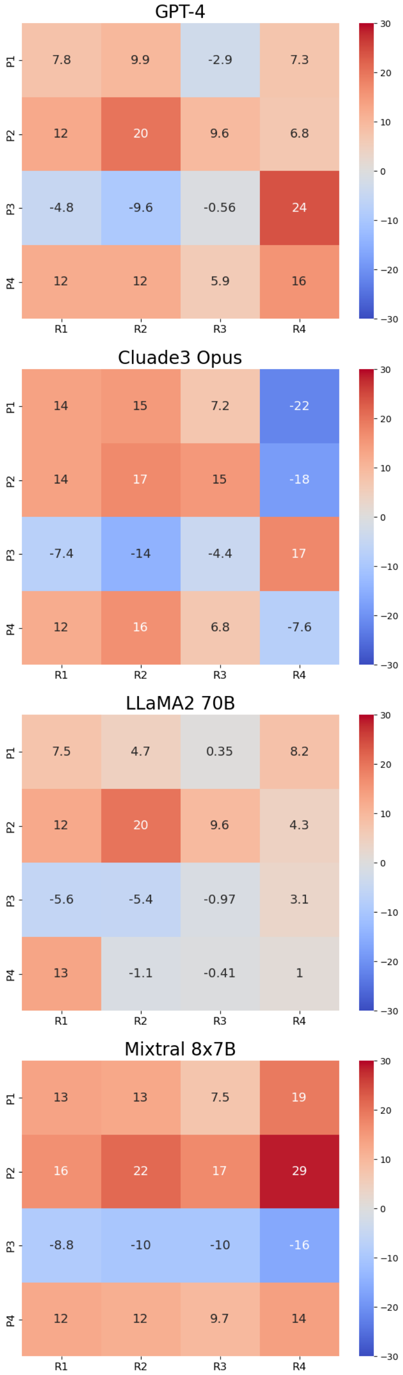

Figure 3: Heatmap of FCE strength, representing interaction effect between two prompting styles, for each model.

	Story 1. Term paper Individual *vs.* Group				Story 2. Supermarket Sign *vs.* Not sign				Story 3. Traffic Ticket Pay fine *vs.* Contest				Story 4. Space R&D Vote for *vs.* Vote against			
	$A_1(1)$	$A_2(1)$	Δ	U	$A_1(1)$	$A_2(1)$	Δ	U	$A_1(1)$	$A_2(1)$	Δ	U	$A_1(1)$	$A_2(1)$	Δ	U
֍	60.0	60.0	0.0	200	60.5	55.0	5.5	257.5^{**}	65.0	60.0	5.0	300^{***}	60.0	40.0	20.0	400^{***}
A\	60.0	40.0	20.0	400^{***}	62.0	36.0	26.0	400^{***}	70.0	70.0	0.0	200	60.0	52.0	8.0	280^{***}
∞	60.5	46.0	14.5	343^{***}	70.0	69.5	0.5	210	70.0	70.0	0.0	200	62.5	45.0	17.5	363^{***}
M	61.0	61.0	0.0	222	77.0	38.0	39.0	394^{***}	68.0	57.0	11.0	334^{***}	42.3	40.0	2.3	268.5^{*}

$^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$

Table 9: Mann-Whitney U Test for **Male Group**. Rows ֍, A\, ∞, M indicates GPT-4, Claude 3, LLaMA 2, and Mixtral.

	Story 1. Term paper Individual *vs.* Group				Story 2. Supermarket Sign *vs.* Not sign				Story 3. Traffic Ticket Pay fine *vs.* Contest				Story 4. Space R&D Vote for *vs.* Vote against			
	$A_1(1)$	$A_2(1)$	Δ	U	$A_1(1)$	$A_2(1)$	Δ	U	$A_1(1)$	$A_2(1)$	Δ	U	$A_1(1)$	$A_2(1)$	Δ	U
֍	60.0	59.5	0.5	210	60.0	49.0	11.0	310^{***}	60.0	60.0	0.0	200	60.0	40.0	20.0	400^{***}
A\	60.0	40.0	20.0	400^{***}	60.0	35.0	25.0	400^{***}	70.0	70.0	0.0	200	60.0	49.8	10.2	300^{***}
∞	60.0	52.0	8.0	280^{***}	70.0	68.5	1.5	230^{*}	70.0	70.0	0.0	200	60.0	42.0	18.0	380^{***}
M	60.8	59.0	1.8	229	75.5	33.5	42.0	395^{***}	63.5	56.0	7.5	282^{**}	43.5	39.3	4.3	278.5^{**}

$^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$

Table 10: Mann-Whitney U Test for **Female Group**. Rows ֍, A\, ∞, M indicates GPT-4, Claude 3, LLaMA 2, and Mixtral.

	0-10	10-90	90-100	0-20	20-80	80-100	0-30	30-70	70-100
֍	0	5120 (100.0%)	0	0	5120 (100.0%)	0	4	5030 (98.24%)	86
A\	0	4939 (100.0%)	0	5	4934 (99.9%)	0	78	4731 (95.79%)	130
∞	1	4880 (99.98%)	0	43	4800 (98.34%)	38	147	4527 (93.75%)	207
M	0	5120 (100.0%)	0	7	5059 (98.81%)	54	190	4280 (83.59%)	650

Table 11: The percentage of answers in a particular range. Rows ֍, A\, ∞, M indicates GPT-4, Claude 3, LLaMA 2, and Mixtral.

H2-1. $\Delta_{P2} > \Delta_{Pk}$ for all $k = 1, 3, 4$ $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$

	Story 1. Term paper						Story 3. Traffic Ticket					
	P1	P2	P3	P4	H	Mann-Whitney test	P1	P2	P3	P4	H	Mann-Whitney test
֎	0.3	0.0	0.0	-1.0	11.9^{**}	(P2 > P4)	2.5	10.0	-9.5	10.0	141.4^{***}	(P2 > P1, P3)
A\	20.0	21.3	0.0	0.0	154.8^{***}	(P2 > P1, P3, P4)	0.0	0.0	0.0	0.0	-	-
∞	11.3	0.0	0.0	3.4	45.9^{***}	-	0.0	10.0	0.0	0.0	159^{***}	(P2 > P1, P3, P4)
M	0.9	7.0	27.3	0.3	99.9^{***}	(P2 > P1, P3)	9.3	10.5	-31.0	9.0	103.9^{***}	(P2 > P1, P3, P4)

H2-2. $\Delta_{R1} > \Delta_{R2} > \Delta_{R3} > \Delta_{R4}$

	R1	R2	R3	R4	H	Mann-Whitney test	R1	R2	R3	R4	H	Mann-Whitney test
֎	0.3	0.0	-20.0	0.0	156.1^{***}	(R1 > R2 > R3)	2.5	10.0	0.0	-2.8	79.3^{***}	(R2 > R3 > R4)
A\	20.0	13.5	1.3	-40.0	137.9^{***}	(R1 > R2 > R3 > R4)	0.0	0.0	4.9	-48.8	132.9^{***}	(R3 > R4)
∞	11.3	7.9	4.9	-0.7	11.9^{**}	(R1 > R2 > R3 > R4)	0.0	0.1	-8.4	21.1	51.5^{***}	(R2 > R3)
M	0.9	7.3	6.6	5.9	6.9	(R2 > R3 > R4)	9.3	6.0	5.9	19.9	42.3^{***}	(R1 > R2 > R3)

Table 12: Kruskal-Wallis Test for H2. Rows ֎, A\, ∞, M indicates GPT-4, Claude 3, LLaMA 2, and Mixtral.

Story 1. Term paper

	Cond. 1		Cond. 2		Diff.	Dunn's post-test		Mann-Whitney test	
		Δ_P		Δ_P		Hypothesis	p	Hypothesis	U
֎ GPT-4	P1	0.3	P2	0.0	+0.3	P1 $\neq$ P2?	0.532	-	
			P3	0.0	+0.3	P1 $\neq$ P3?	0.532	-	
			P4	-1.0	+1.3	P1 $\neq$ P4?	0.002^{**}	P1 $>$ P4	898^{*}
	P2	0.0	P3	0.0	0.0	P2 $\neq$ P3?	1.0	-	
			P4	-1.0	+1.0	P2 $\neq$ P4?	0.011^{*}	P2 $>$ P4	880^{*}
	P3	0.0	P4	-1.0	+1.0	P3 $\neq$ P4?	0.011^{*}	P3 $>$ P4	880^{*}
A\ Claude 3	P1	20.0	P2	21.3	-1.3	P1 $\neq$ P2?	0.583	-	
			P3	0.0	+20.0	P1 $\neq$ P3?	$<0.001^{***}$	P1 $>$ P3?	1600^{*}
			P4	0.0	+20.0	P1 $\neq$ P4?	$<0.001^{***}$	P1 $>$ P4?	1600^{***}
	P2	21.3	P3	0.0	+21.3	P2 $\neq$ P3?	$<0.001^{***}$	P2 $>$ P3?	1600^{***}
			P4	0.0	+21.3	P2 $\neq$ P4?	$<0.001^{***}$	P2 $>$ P4?	1600^{***}
	P3	0.0	P4	0.0	0.0	P3 $\neq$ P4?	1.0	-	
∞ LLaMA 2	P1	11.3	P2	0.0	+11.3	P1 $\neq$ P2?	$<0.001^{***}$	P1 $>$ P2?	1240^{***}
			P3	0.0	+11.3	P1 $\neq$ P3?	$<0.001^{***}$	P1 $>$ P3?	1240^{***}
			P4	3.4	+7.9	P1 $\neq$ P4?	$<0.001^{***}$	P1 $>$ P4?	1070^{***}
	P2	0.0	P3	0.0	0.0	P2 $\neq$ P3?	1.0	-	
			P4	3.4	-3.4	P2 $\neq$ P4?	0.17	-	
	P3	0.0	P4	3.4	-3.4	P3 $\neq$ P4?	0.17	-	
M Mixtral	P1	0.9	P2	7.0	-6.1	P1 $\neq$ P2?	0.054	-	
			P3	27.3	-26.4	P1 $\neq$ P3?	$<0.001^{***}$	P1 $<$ P3?	1574^{***}
			P4	0.3	+0.6	P1 $\neq$ P4?	0.627	-	
	P2	7.0	P3	27.3	-20.3	P2 $\neq$ P3?	$<0.001^{***}$	P2 $<$ P3?	1520.5^{***}
			P4	0.3	+6.7	P2 $\neq$ P4?	0.016^{*}	P2 $>$ P4?	1086^{***}
	P3	27.3	P4	0.3	+27.0	P3 $\neq$ P4?	$<0.001^{***}$	P3 $>$ P4?	1579.5^{***}

$^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$

Table 13: The strength of FCE, Dunn's post-test result, and Mann-Whitney U test result for H2-1, in Story 1.

Story 1. Term paper									
	Cond. 1		Cond. 2		Diff.	Dunn's post-test		Mann-Whitney test	
		Δ_R		Δ_R		Hypothesis	p	Hypothesis	U
GPT-4	R1	0.3	R2	0.0	+0.3	R1 $\neq$ R2?	0.848	-	
			R3	-20.0	+20.3	R1 $\neq$ R3?	<0.001***	R1 $>$ R3?	1600***
			R4	0.0	+0.3	R1 $\neq$ R4?	0.848	-	
	R2	0.0	R3	-20.0	+20.0	R2 $\neq$ R3?	<0.001***	R2 $>$ R3?	1600***
			R4	0.0	+0.0	R2 $\neq$ R4?	1.0	-	
	R3	-20.0	R4	0.0	-20.0	R3 $\neq$ R4?	<0.001***	R3 $<$ R4?	1600***
Claude 3	R1	20.0	R2	13.5	+6.5	R1 $\neq$ R2?	0.042**	R1 $>$ R2?	1060***
			R3	1.3	+18.7	R1 $\neq$ R3?	<0.001***	R1 $>$ R3?	1560***
			R4	-40.0	+60.0	R1 $\neq$ R4?	<0.001***	R1 $>$ R4?	1600***
	R2	13.5	R3	1.3	+12.2	R2 $\neq$ R3?	<0.001***	R2 $>$ R3?	1293.5***
			R4	-40.0	+53.5	R2 $\neq$ R4?	<0.001***	R2 $>$ R4?	1600***
	R3	1.3	R4	-40.0	+41.3	R3 $\neq$ R4?	<0.001***	R3 $>$ R4?	1600***
LLaMA 2	R1	11.3	R2	7.9	+3.4	R1 $\neq$ R2?	0.316	-	
			R3	4.9	+6.4	R1 $\neq$ R3?	0.068	-	
			R4	-0.7	+12.0	R1 $\neq$ R4?	0.003**	R1 $>$ R4?	1132.5***
	R2	7.9	R3	4.9	+3.0	R2 $\neq$ R3?	0.388	-	
			R4	-0.7	+8.6	R2 $\neq$ R4?	0.05	-	
	R3	4.9	R4	-0.7	+5.6	R3 $\neq$ R4?	0.332	-	
Mixtral	R1	0.9	R2	7.3	-6.4	R1 $\neq$ R2?	0.023**	R1 $<$ R2?	1046.5**
			R3	6.6	-5.7	R1 $\neq$ R3?	0.026**	R1 $<$ R3?	1059**
			R4	5.9	-5.0	R1 $\neq$ R4?	0.069	-	-
	R2	7.3	R3	6.6	+0.7	R2 $\neq$ R3?	0.955	-	-
			R4	5.9	+1.4	R2 $\neq$ R4?	0.647	-	-
	R3	6.6	R4	5.9	+0.7	R3 $\neq$ R4?	0.688	-	-

$^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$

Table 14: The strength of FCE, Dunn's post-test result, and Mann-Whitney U test result for H2-2, in Story 1.

Story 2. Supermarket

	Cond. 1		Cond. 2		Diff.	Dunn's post-test		Mann-Whitney test	
		Δ_P		Δ_P		Hypothesis	p	Hypothesis	U
GPT-4	P1	8.3	P2	20.0	-11.7	P1 $\neq$ P2?	<0.001***	P1 $<$ P2?	1280***
			P3	-9.5	+17.8	P1 $\neq$ P3?	<0.001***	P1 $>$ P3?	1358.5***
			P4	19.5	-11.2	P1 $\neq$ P4?	<0.001***	P1 $<$ P4?	1259.5***
	P2	20.0	P3	-9.5	+29.5	P2 $\neq$ P3?	<0.001***	P2 $>$ P3?	1600***
			P4	19.5	+0.5	P2 $\neq$ P4?	0.845	-	
	P3	-9.5	P4	19.5	-29.0	P3 $\neq$ P4?	<0.001***	P3 $<$ P4?	1589.5***
Claude 3	P1	25.5	P2	29.0	-3.5	P1 $\neq$ P2?	0.114	-	
			P3	-29.5	+55.0	P1 $\neq$ P3?	<0.001***	P1 $>$ P3?	1600***
			P4	27.7	-2.2	P1 $\neq$ P4?	0.246	-	
	P2	29.0	P3	-29.5	+58.5	P2 $\neq$ P3?	<0.001***	P2 $>$ P3?	1600***
			P4	27.7	+1.3	P2 $\neq$ P4?	0.994	-	
	P3	-29.5	P4	27.7	-57.2	P3 $\neq$ P4?	<0.001***	P3 $<$ P4?	600***
LLaMA 2	P1	1.0	P2	20.0	-19.0	P1 $\neq$ P2?	<0.001***	P1 $<$ P2?	1600***
			P3	-3.3	+4.3	P1 $\neq$ P3?	0.016*	P1 $>$ P3?	1114***
			P4	37.0	-36.0	P1 $\neq$ P4?	0.248	-	
	P2	20.0	P3	-3.3	+23.3	P2 $\neq$ P3?	<0.001***	P2 $>$ P3?	1600***
			P4	37.0	-17.0	P2 $\neq$ P4?	<0.001***	P2 $<$ P4?	1600***
	P3	-3.3	P4	37.0	-40.3	P3 $\neq$ P4?	<0.001***	P3 $<$ P4?	1222***
Mixtral	P1	40.5	P2	47.9	-7.4	P1 $\neq$ P2?	0.271	-	
			P3	-33.0	+73.5	P1 $\neq$ P3?	<0.001***	P1 $>$ P3?	1600***
			P4	40.0	+0.5	P1 $\neq$ P4?	0.437	-	
	P2	47.9	P3	-33.0	+80.9	P2 $\neq$ P3?	<0.001***	P2 $>$ P3?	1600***
			P4	40.0	+7.9	P2 $\neq$ P4?	0.746	-	
	P3	-33.0	P4	40.0	-73.3	P3 $\neq$ P4?	<0.001***	P3 $<$ P4?	1600***

$^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$

Table 15: The strength of FCE, Dunn's post-test result, and Mann-Whitney U test result for H2-1, in Story 2.

Story 2. Supermarket

	Cond. 1		Cond. 2		Diff.	Dunn's post-test		Mann-Whitney test	
		Δ_R		Δ_R		Hypothesis	p	Hypothesis	U
GPT-4	R1	8.3	R2	9.5	-1.2	R1 $\neq$ R2?	0.352	-	
			R3	8.0	+0.3	R1 $\neq$ R3?	0.998	-	
			R4	24.5	-16.2	R1 $\neq$ R4?	$<0.001^{***}$	R1 $<$ R4?	1222.5^{***}
	R2	9.5	R3	8.0	+1.5	R2 $\neq$ R3?	0.354	-	
			R4	24.5	-15.0	R2 $\neq$ R4?	$<0.001^{***}$	R2 $<$ R4?	1173.5^{***}
	R3	8.0	R4	24.5	-16.5	R3 $\neq$ R4?	$<0.001^{***}$	R3 $<$ R4?	1223.5^{***}
Claude 3	R1	25.2	R2	25.5	-0.3	R1 $\neq$ R2?	0.824	-	
			R3	6.6	+18.6	R1 $\neq$ R3?	$<0.001^{***}$	R1 $>$ R3?	1489.5^{***}
			R4	2.3	+22.9	R1 $\neq$ R4?	$<0.001^{***}$	R1 $>$ R4?	1600^{***}
	R2	25.5	R3	6.6	+18.9	R2 $\neq$ R3?	$<0.001^{***}$	R2 $>$ R3?	1440^{***}
			R4	2.3	+23.2	R2 $\neq$ R4?	$<0.001^{***}$	R2 $>$ R4?	1526.5^{***}
	R3	6.6	R4	2.3	+4.3	R3 $\neq$ R4?	0.854	-	
LLaMA 2	R1	1.0	R2	-0.9	+1.9	R1 $\neq$ R2?	0.678	-	
			R3	2.8	-1.8	R1 $\neq$ R3?	0.432	-	
			R4	6.5	-5.5	R1 $\neq$ R4?	0.194	-	
	R2	-0.9	R3	2.8	-3.7	R2 $\neq$ R3?	0.237	-	
			R4	6.5	-7.4	R2 $\neq$ R4?	0.092	-	
	R3	2.8	R4	6.5	-3.7	R3 $\neq$ R4?	0.611	-	
Mixtral	R1	40.5	R2	32.6	+7.9	R1 $\neq$ R2?	0.012^{*}	R1 $>$ R2?	1144.5^{***}
			R3	14.6	+25.9	R1 $\neq$ R3?	$<0.001^{***}$	R1 $>$ R3?	1387^{***}
			R4	46.6	-6.1	R1 $\neq$ R4?	0.117	-	
	R2	32.6	R3	14.6	+18.0	R2 $\neq$ R3?	$<0.001^{***}$	R2 $>$ R3?	1375^{***}
			R4	46.6	-14.0	R2 $\neq$ R4?	$<0.001^{***}$	R2 $<$ R4?	1306.5^{***}
	R3	14.6	R4	46.6	-32.0	R3 $\neq$ R4?	$<0.001^{***}$	R3 $<$ R4?	1528.5^{***}

$^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$

Table 16: The strength of FCE, Dunn's post-test result, and Mann-Whitney U test result for H2-2, in Story 2.

Story 3. Traffic Ticket

	Cond. 1		Cond. 2		Diff.	Dunn's post-test		Mann-Whitney test	
		Δ_P		Δ_P		Hypothesis	p	Hypothesis	U
GPT-4	P1	2.5	P2	10.0	-7.5	P1 $\neq$ P2?	$<$0.001***	P1 $<$ P2?	1400***
			P3	-9.5	+12.0	P1 $\neq$ P3?	$<$0.001***	P1 $>$ P3?	1570***
			P4	10.0	-7.5	P1 $\neq$ P4?	$<$0.001***	P1 $<$ P4?	1400***
	P2	10.0	P3	-9.5	+19.5	P2 $\neq$ P3?	$<$0.001***	P2 $>$ P3?	1600***
			P4	10.0	+0.0	P2 $\neq$ P4?	1.0	P2 $<$ P4?	-
	P3	-9.5	P4	10.0	-19.5	P3 $\neq$ P4?	$<$0.001***	P3 $<$ P4?	1600***
Claude 3	P1	0.0	P2	0.0	+0.0	P1 $\neq$ P2?	1.0	P1 $<$ P2?	-
			P3	0.0	+0.0	P1 $\neq$ P3?	1.0	P1 $<$ P3?	-
			P4	0.0	+0.0	P1 $\neq$ P4?	1.0	P1 $<$ P4?	-
	P2	0.0	P3	0.0	+0.0	P2 $\neq$ P3?	1.0	P2 $<$ P3?	-
			P4	0.0	+0.0	P2 $\neq$ P4?	1.0	P2 $<$ P4?	-
	P3	0.0	P4	0.0	+0.0	P3 $\neq$ P4?	1.0	P3 $<$ P4?	-
LLaMA 2	P1	0.0	P2	10.0	-10.0	P1 $\neq$ P2?	$<$0.001***	P1 $<$ P2?	1600***
			P3	0.0	+0.0	P1 $\neq$ P3?	1.0***	P1 $<$ P3?	-
			P4	0.0	+0.0	P1 $\neq$ P4?	1.0***	P1 $<$ P4?	-
	P2	10.0	P3	0.0	+10.0	P2 $\neq$ P3?	$<$0.001***	P2 $>$ P3?	1600***
			P4	0.0	+10.0	P2 $\neq$ P4?	$<$0.001***	P2 $>$ P4?	1600***
	P3	0.0	P4	0.0	+0.0	P3 $\neq$ P4?	1.0	P3 $<$ P4?	-
Mixtral	P1	9.3	P2	10.5	-1.2	P1 $\neq$ P2?	0.694	P1 $<$ P2?	-
			P3	-31.0	+40.3	P1 $\neq$ P3?	$<$0.001***	P1 $>$ P3?	1583.5***
			P4	9.0	+0.3	P1 $\neq$ P4?	0.833	P1 $<$ P4?	-
	P2	10.5	P3	-31.0	+41.5	P2 $\neq$ P3?	$<$0.001***	P2 $>$ P3?	1600***
			P4	9.0	+1.5	P2 $\neq$ P4?	0.545	P2 $<$ P4?	-
	P3	-31.0	P4	9.0	-40.0	P3 $\neq$ P4?	$<$0.001***	P3 $<$ P4?	1590***

$^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$

Table 17: The strength of FCE, Dunn's post-test result, and Mann-Whitney U test result for H2-1, in Story 3.

Story 3. Traffic Ticket

	Cond. 1		Cond. 2		Diff.	Dunn's post-test		Mann-Whitney test	
		Δ_R		Δ_R		Hypothesis	p	Hypothesis	U
GPT-4	R1	2.5	R2	10.0	-7.5	R1 $\neq$ R2?	$<$0.001***	R1 $<$ R2?	1400***
			R3	0.0	+2.5	R1 $\neq$ R3?	0.854	-	
			R4	-2.8	+5.3	R1 $\neq$ R4?	0.001**	R1 $>$ R4?	1175***
	R2	10.0	R3	0.0	+10.0	R2 $\neq$ R3?	$<$0.001***	R2 $>$ R3?	1600***
			R4	-2.8	+12.8	R2 $\neq$ R4?	$<$0.001***	R2 $>$ R4?	1340***
	R3	0.0	R4	-2.8	+2.8	R3 $\neq$ R4?	0.145	-	
Claude 3	R1	0.0	R2	0.0	+0.0	R1 $\neq$ R2?	1.0	-	
			R3	4.9	-4.9	R1 $\neq$ R3?	0.001**	R1 $<$ R3?	1180***
			R4	-48.8	+48.8	R1 $\neq$ R4?	$<$0.001***	R1 $>$ R4?	1600***
	R2	0.0	R3	4.9	-4.9	R2 $\neq$ R3?	0.001**	R2 $<$ R3?	1180***
			R4	-48.8	+48.8	R2 $\neq$ R4?	$<$0.001***	R2 $>$ R4?	1600***
	R3	4.9	R4	-48.8	+53.7	R3 $\neq$ R4?	$<$0.001***	R3 $>$ R4?	1600***
LLaMA 2	R1	0.0	R2	0.1	-0.1	R1 $\neq$ R2?	0.899	-	
			R3	-8.4	+8.4	R1 $\neq$ R3?	0.252	-	
			R4	21.1	-21.1	R1 $\neq$ R4?	$<$0.001***	R1 $<$ R4?	1240***
	R2	0.1	R3	-8.4	+8.5	R2 $\neq$ R3?	0.22	-	
			R4	21.1	-21.0	R2 $\neq$ R4?	$<$0.001***	R2 $<$ R4?	1082***
	R3	-8.4	R4	21.1	-29.5	R3 $\neq$ R4?	$<$0.001***	R3 $<$ R4?	647***
Mixtral	R1	9.3	R2	6.0	+3.3	R1 $\neq$ R2?	0.127	-	
			R3	5.9	+3.4	R1 $\neq$ R3?	0.18	-	
			R4	19.9	-10.6	R1 $\neq$ R4?	$<$0.001***	R1 $<$ R4?	1225.5***
	R2	6.0	R3	5.9	+0.1	R2 $\neq$ R3?	0.854	-	
			R4	19.9	-13.9	R2 $\neq$ R4?	$<$0.001***	R2 $<$ R4?	1425.5***
	R3	5.9	R4	19.9	-14.0	R3 $\neq$ R4?	$<$0.001***	R3 $<$ R4?	1316***

$^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$

Table 18: The strength of FCE, Dunn's post-test result, and Mann-Whitney U test result for H2-2, in Story 3.

Story 4. Space R&D Program

	Cond. 1		Cond. 2		Diff.	Dunn's post-test		Mann-Whitney test	
		Δ_P		Δ_P		Hypothesis	p	Hypothesis	U
GPT-4	P1	20.0	P2	20.0	+0.0	P1 $\neq$ P2?	1.0	-	
			P3	0.0	+20.0	P1 $\neq$ P3?	$<0.001^{***}$	P1 $>$ P3?	1600^{***}
			P4	20.0	+0.0	P1 $\neq$ P4?	1.0	-	
	P2	20.0	P3	0.0	+20.0	P2 $\neq$ P3?	$<0.001^{***}$	P2 $>$ P3?	1600^{***}
			P4	20.0	+0.0	P2 $\neq$ P4?	1.0	-	
	P3	0.0	P4	20.0	-20.0	P3 $\neq$ P4?	$<0.001^{***}$	P3 $<$ P4?	1600^{***}
Claude 3	P1	9.1	P2	5.0	+4.1	P1 $\neq$ P2?	0.114	-	
			P3	0.0	+9.1	P1 $\neq$ P3?	$<0.001^{***}$	P1 $>$ P3?	1160^{***}
			P4	20.4	-11.3	P1 $\neq$ P4?	$<0.001^{***}$	P1 $<$ P4?	650^{***}
	P2	5.0	P3	0.0	+5.0	P2 $\neq$ P3?	0.058	-	
			P4	20.4	-15.4	P2 $\neq$ P4?	$<0.001^{***}$	P2 $<$ P4?	740^{***}
	P3	0.0	P4	20.4	-20.4	P3 $\neq$ P4?	$<0.001^{***}$	P3 $<$ P4?	840^{***}
LLaMA 2	P1	17.8	P2	20.0	-2.2	P1 $\neq$ P2?	0.489	-	
			P3	-19.0	+36.8	P1 $\neq$ P3?	$<0.001^{***}$	P1 $>$ P3?	1595^{***}
			P4	12.5	+5.3	P1 $\neq$ P4?	0.073	-	
	P2	20.0	P3	-19.0	+39.0	P2 $\neq$ P3?	$<0.001^{***}$	P2 $>$ P3?	1600^{***}
			P4	12.5	+7.5	P2 $\neq$ P4?	0.013^{*}	P2 $>$ P4?	1100^{***}
	P3	-19.0	P4	12.5	-31.5	P3 $\neq$ P4?	$<0.001^{***}$	P3 $<$ P4?	1585^{***}
Mixtral	P1	3.3	P2	0.0	+3.3	P1 $\neq$ P2?	0.014^{*}	P1 $>$ P2?	1120^{***}
			P3	1.6	+1.7	P1 $\neq$ P3?	0.016^{*}	P1 $>$ P3?	958.5^{*}
			P4	0.5	+2.8	P1 $\neq$ P4?	$<0.001^{***}$	P1 $>$ P4?	1022^{*}
	P2	0.0	P3	1.6	-1.6	P2 $\neq$ P3?	0.289	-	
			P4	0.5	-0.5	P2 $\neq$ P4?	0.243	-	
	P3	1.6	P4	0.5	+1.1	P3 $\neq$ P4?	0.026^{**}	P3 $>$ P4?	1017^{*}

$^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$

Table 19: The strength of FCE, Dunn's post-test result, and Mann-Whitney U test result for H2-1, in Story 4.

Story 4. Space R&D Program

	Cond. 1		Cond. 2		Diff.	Dunn's post-test		Mann-Whitney test	
		Δ_R		Δ_R		Hypothesis	p	Hypothesis	U
GPT-4	R1	20.0	R2	20.0	+0.0	R1 $\neq$ R2?	1.0	-	
			R3	0.3	+19.7	R1 $\neq$ R3?	0.827		
			R4	7.5	+12.5	R1 $\neq$ R4?	$<$0.001***	R1 $>$ R4?	1400***
	R2	20.0	R3	0.3	+19.7	R2 $\neq$ R3?	0.827	-	
			R4	7.5	+12.5	R2 $\neq$ R4?	$<$0.001***	R2 $>$ R4?	1400***
	R3	0.3	R4	7.5	-7.2	R3 $\neq$ R4?	$<$0.001***	R3 $<$ R4?	1404***
Claude 3	R1	9.1	R2	20.3	-11.2	R1 $\neq$ R2?	$<$0.001***	R1 $<$ R2?	1229.5***
			R3	15.9	-6.8	R1 $\neq$ R3?	0.04*	R1 $<$ R3?	1013.5***
			R4	-0.9	+10.0	R1 $\neq$ R4?	$<$0.001***	R1 $>$ R4?	1182***
	R2	20.3	R3	15.9	+4.4	R2 $\neq$ R3?	0.002**	R2 $>$ R3?	1268.5***
			R4	-0.9	+21.2	R2 $\neq$ R4?	$<$0.001***	R2 $>$ R4?	1600***
	R3	15.9	R4	-0.9	+16.8	R3 $\neq$ R4?	$<$0.001***	R3 $>$ R4?	1579.5***
LLaMA 2	R1	17.8	R2	11.5	+6.3	R1 $\neq$ R2?	$<$0.001***	R1 $>$ R2?	1362.5***
			R3	2.1	+15.7	R1 $\neq$ R3?	$<$0.001***	R1 $>$ R3?	1431.5***
			R4	5.8	+12.0	R1 $\neq$ R4?	$<$0.001***	R1 $>$ R4?	976.5**
	R2	11.5	R3	2.1	+9.4	R2 $\neq$ R3?	0.118	-	
			R4	5.8	+5.7	R2 $\neq$ R4?	0.01*	R2 $>$ R4?	811*
	R3	2.1	R4	5.8	-3.7	R3 $\neq$ R4?	0.288	-	
Mixtral	R1	3.3	R2	5.1	-1.8	R1 $\neq$ R2?	0.543	-	
			R3	2.7	+0.6	R1 $\neq$ R3?	0.662	-	
			R4	5.4	-1.9	R1 $\neq$ R4?	0.667	-	
	R2	5.1	R3	2.7	+2.4	R2 $\neq$ R3?	0.295	-	
			R4	5.4	-0.3	R2 $\neq$ R4?	0.858	-	
	R3	2.7	R4	5.4	-2.7	R3 $\neq$ R4?	0.386	-	

$^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$

Table 20: The strength of FCE, Dunn's post-test result, and Mann-Whitney U test result for H2-2, in Story 4.