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

Anonymous ACL submission

Abstract

False Consensus Effect(FCE) is a cognitive bias in which a person considers his/her own behavioral choices as relatively common choices in a given situation while viewing choices as uncommon in society. FCE acts as an obstacle to communication, yet this has not been scrutinized meticulously in prior studies. Our research aims to determine whether the FCE, a cognitive bias inherent in humans, is also exhibited by Large Language Models(LLMs). To achieve this, we emulate conditions as close as possible to human experiments and conduct experiments under rigorous controls to minimize the influence of other cognitive biases. Through these experiments, we have been able to confirm the manifestation of the FCE in LLMs. Moreover, within an environment unimpeded by the influence of other cognitive biases, we introduce a methodology that applies 16 different variables to either maximize the expression of the FCE, yield a neutral choice outcome, or produce results that are the antithesis of the FCE.

1 Introduction

Recently, Large Language Model(LLM) is widely adopted for communication-based tasks. As such task involves communication between human and LLMs, cognitive bias on LLMs may affect usability or experience of LLMs. So, researchers have studied the underlying reason and the mitigation method for cognitive biases (Echterhoff et al., 2024; Itzhak et al., 2023; Lin and Ng, 2023). Among such biases, false consensus effect (FCE) is important because sometimes that bias hinders smooth communication(Wojcieszak and Price, 2009; Wetzel and Walton, 1985). However, researchers have less focused on investigating the effect of FCE on LLMs. Thus, this paper aims to investigate how FCE appear in LLMs and how we can mitigate FCE using prompts.

False consensus effect(FCE) is a cognitive bias in which a person considers his/her own behavioral choices as relatively common choices in a given situation while viewing choices as uncommon in society(Ross et al., 1977). Previous studies on investigating cognitive bias of LLMs have three issues: (1) lack of daily situation, (2) lack of consideration on alternative biases, and (3) lack of investigation on prompting methods.

First, necessary to investigate FCE on daily situations. Previous research indicates that when FCE is expressed during conversation, the other person can negatively view their perception of me.(Wojcieszak and Price, 2009) However, most studies have investigated other cognitive biases except FCE.(Echterhoff et al., 2024; Macmillan-Scott and Musolesi, 2024) To our knowledge, the only FCE investigation study was conducted in medical rather than routine situations. Hence, we must study whether FCE is expressed even in non-specific situations(Schmidgall et al., 2024).

Second, it is necessary to separate FCE from other cognitive biases. As shown in psychological experiments, several cognitive bases are often expressed simultaneously if other cognitive bases are not controlled during the experiment. In previous studies, it was necessary to block factors that would cause other cognitive biases, but this was not attempted in previous studies. In addition, statistical analysis based on the analysis method of psychological experiments is required. Therefore, experiments and analysis in a controlled environment based on psychological experiments are needed.

Third, necessary to understand the impact of FCE depending on the LLM prompting method. Previous research indicates that LM's performance can significantly vary based on the information provided and the prompt engineering methods employed.(Jia and Liang, 2017; Cheng et al., 2019) Therefore, examining the variations in FCE by altering prompt methodology without being influenced by other cognitive biases is necessary.

Therefore, We propose two studies. In the Study 1, we try to examine FCE of LLM. We investigate LLM's FCE in the context of controlling other Cognitive vias that may occur during the experiment. Additionally, our study examines changes in the patterns of FCE due to the effects of gender bias and cultural bias. In Study 2, we try to mitigate FCE with Prompt. Our study applied two variables to the prompt design. We categorized the variables into two groups: (i) the relevance of the provided information, and (ii) the depth of the reasoning process. Additionally, by examining the interaction between these two variables, we aim to identify methodologies that can either mitigate or intensify the characteristics of the FCE.

As a result, our study makes the following contributions.

- As far as we know, this is the first paper that adopts psychological experiment to confirm LLM's FCE. Using that experiment, we show that it is possible to adopt psychological experiments directly to check LLM's cognitive bias.
- Also, we demonstrate that one can control alternative biases even in an experiment which is similar to a daily situation.
- We compare various prompting styles, in order to figure out the most appropriate style that can mitigate or intensify the strength of FCE.

2 Related Work

Our goal is to investigate the False Consensus Effect (FCE) on Large Language Models (LLMs) deeply. FCE is a type of cognitive bias that views behavioral choices based on one's beliefs, judgments, and attitudes as relatively general and appropriate for existing situations (Ross et al., 1977; Choi and Cha, 2019). As FCE is a sort of cognitive bias, in this section, we summarize existing literature examining such cognitive biases in LLMs into two categories: (1) investigated biases and the way of testing them and (2) prompting methods when testing the biases.

2.1 Examining bias of LLMs

To investigate biases of LLMs, researchers have recently focused on measuring cognitive biases in LLMs inspired by psychological experiments (Echterhoff et al., 2024; Schmidgall et al., 2024; Leng, 2024; Wang et al., 2023b; Itzhak et al., 2023; Kaneko et al., 2024). Some researchers modified existing tasks to perform psychological experiments on LLMs. For instance, Schmidgall et al. (2024) attempted to detect cognitive biases such as Self-diagnosis bias, Recency bias, Confirmation bias, and False consensus bias during interactions with LLMs in a medical QA task.

In previous studies, there was a study on the cognitive bias of LLM that applied psychological experimental methodology as it was.Echterhoff et al. (2024); Itzhak et al. (2023); Leng (2024) Echterhoff et al. (2024) designed a separate prompt to check the cognitive bias of LLM, referring to experimental methodologies of cognitive bias such as anchoring effect and framing. In addition, Itzhak et al. (2023) questioned LLM by applying psychological experimental methodology to confirm decoy-effect and certify-effect. However, despite applying traditional experimental methodologies, this research has limitations due to the failure to thoroughly control for external biases such as character settings for utilizing LLMs, the influence of different stories, and the effects of temperature settings on LLMs.

In summary, prior research needs to thoroughly control the alternative bias or the external environment in the method of measuring cognitive bias. In particular, it is difficult to fully interpret the experimental results due to the effect of the alternative bias when using a psychologically unverified experiment in situations where other bias is prone to confusion, such as FCE. Therefore, it is necessary to confirm the FCE effect through controlled experiments.

2.2 Mitigating biases of LMs

To mitigate biases of LMs, previous studies have less focused on controlling alternative biases. In this subsection, we first review (a) previous attempts at mitigating cognitive biases. And, as mitigating biases can be seen as prompt engineering methods improving the quality of reasoning, we summarize (b) previous attempts of designing prompts for reasoning.

In previous research, efforts to mitigate cognitive biases involved inserting additional sentences or further employing existing models. For example, Echterhoff et al. (2024) uses the additional sentence[1] to mitigate cognitive biases. Additionally,

---

[1] The sentence is "Be mindful of not being biased by cognitive bias." (Echterhoff et al., 2024)

Itzhak et al. (2023) attempted to mitigate cognitive bias using an existing model (Holtzman et al., 2021). However, these studies have less considered the effect of prompt engineering techniques on cognitive biases. As studies reported that context given by a prompt can affect the performance of a specific reasoning task, such changes on a prompt can change the strength of a cognitive bias.

Thus, we also review various prompting methods which can be used to improve the reasoning process (Wei et al., 2022; Chen et al., 2023; Yao et al., 2023, 2022; Shinn et al., 2024). We classify prompting methods into two parts in terms: (1) The influence of the information provided in the prompt, and (2) the influence of how the prompt asks reasoning process.

First, the influence of information provided in the prompt was usually tested by checking whether reasoning was interfered by injecting irrelevant or confusing information into the reasoning process (Jia and Liang, 2017; Cheng et al., 2019; Wu et al., 2024; Iyyer et al., 2018). For example, Jia and Liang (2017) confirmed that injecting adversarial examples which contains opposite information may decrease performance on a reading comprehension task. Similarly, Cheng et al. (2019) showed that inserting a sentence which is grammatically correct but semantically strange changes the performance of a reasoning model.

Second, the influence of how the prompt asks reasoning process was usually tested by dividing the process into several steps to enhance logical thinking. For example, Wei et al. (2022) showed that LLMs performance increased when they asked LLMs to generate reasoning before generating answer. Moreover, Shinn et al. (2024) showed that the performance can be improved more when they asked LLMs to use self-reflective reasoning methods. Thus, as the answering process of FCE can be seen as a reasoning process, it is necessary to check whether these prompting styles can affect the strength of FCE.

3 Study 1: Examining FCE of LLM

To investigate whether the False Consensus Effect (FCE) emerges in LLMs, the Study 1 conducts an experiment inspired by psychological experiments (Ross et al., 1977; Choi and Cha, 2019). In this section, we illustrate the design of experiment done in Ross et al. (1977), the result of our experiment, and a discussion about the result.

> 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.

Figure 1: An example story used in the psychological experiment for false consensus effect (Ross et al., 1977). Bold-faced texts represent two options given in this story.

3.1 Procedure

To confirm whether LLMs have FCE through psychology-alike experiments, our experimental design 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, procedure, and analysis method for our experiment. Each paragraph begins with how psychologists conducted their experiment, to help the readers understanding.

Participants: Originally, psychological studies usually recruited college students. There is no screening procedure to select a specific kind of student in order to identify the general phenomenon of human students. It is worth noticing that psychologists reported that cultural or gender biases may affect FCE(Choi and Cha, 2019).

In Study 1, we make LLMs pretend as a college student 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 different characters for each of the two cultures and each gender[2]. For the detailed information about 40 characters that we used, see Appendix A.

Procedure: Originally, psychologists give participants several hypothetical situations which can be occur in their daily life. Figure 1 shows a sample

---

[2] For the culture, we selected European American and Korean, inspired by (Choi and Cha, 2019). And for the gender, we used the words ‘man’ and ‘woman.’

	GPT-4		Claude 3		LLaMA 2	
Story 1	40	0	40	0	40	0
Story 2	40	0	40	0	0	40
Story 3	0	40	0	40	20	20
Story 4	9	31	31	9	0	40

Table 1: Skewness of LLMs answer. A cell shows the number of choosing option 1 and 2 in each story, respectively. For detailed models, We use GPT-4 and Claude 3 Opus, LLaMA 2 70B Model

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 situation; the choice may vary across people because there is no social agreement about them.

In Study 1, we follow the exact same procedure except asking 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 options since the analysis procedure requires comparing two groups; LLMs who choose one option and LLMs who choose the other. Instead, we directly feed each option as if LLMs have answered that option, since we found that LLMs stick on a specific choice, as shown in Table 1. Detailed prompt structure and four hypothetical stories are illustrated in Appendix A.

Also, note that we do not modify the experimental structure to control the other cognitive biases. Psychologists carefully designed their experiment to control other alternative biases like confirmation bias, in-group bias, or accumulation effect. As these biases arise due to externally given or preconceived opinions, we excluded such information in designing prompts in this study. For example, in the experiment, we do not give any information related to external social consensus about the given story. Also, we do not ask or provide reasoning for the answer in this Study 1, as the original experiment does not ask participants' reasoning to avoid deep thinking about the social consensus. Such effect of reasoning will be discussed in Study 2.

Analysis: Using the response, psychological studies have shown that when estimating perceived agreement on a specific option, participants who choose that option provide higher estimates on that option compared to the other people. Mathematically, let $A_c(p)$ be the average perceived agreement for option p by other peers, provided by those who selected option c as their own choice. And, let $A_1(1) - A_2(1)$, which denotes how one overestimates his/her choice compared to the others. Then, FCE can be proven by checking alternative hypothesis $A_1(1) > A_2(1)$ using statistical methods, such as two-sample t-test.

In Study 1, we follow the same hypothesis of the original experiments. The alternative hypotheses are stated below. The first hypothesis H1-1 is an analogy of hypothesis stated in Ross et al. (1977). Also, the other hypotheses are inspired by Choi and Cha (2019). To test these hypotheses, We conduct Mann-Whitney U-test as the data distribution does not follow the normal distribution. Note that, following Choi and Cha (2019), we regard $|A_1(1) - A_2(1)|$ in a group G (e.g., 'Korean') as the strength of FCE on G. Throughout this paper, we briefly write the strength of FCE on G as Δ_G.

H1-1. LLMs have FCE; i.e., $A_1(1) > A_2(1)$.

H1-2. Cultural bias affects the strength of FCE in LLMs; i.e., $|\Delta_{Korean}| \neq |\Delta_{American}|$.

H1-3. Gender bias affects the strength of FCE in LLMs; i.e., $|\Delta_{male}| \neq |\Delta_{female}|$.

3.2 Tested LLMs

For the experiment, we use three LLMs: GPT-4 (Achiam et al., 2023), Claude 3 Opus (Anthropic, 2024), and LLaMA 2 70B (Touvron et al., 2023). We selected these models because they have shown outstanding performance on question answering task, and the largest model in the family has been published publicly through API or model parameters. Here, we briefly summarize refinement policy of each LLM since FCE experiment forces LLMs to answer about unknown social norms. According to Achiam et al. (2023); Anthropic (2024), GPT-4 and Claude 3 have a procedure that refuses answers to questions concerning personal information or questions sensitive politically, religiously, or culturally. On the other hand, LLaMA 2 was trained to avoid making dangerous making dangerous or unethical utterances and to retain strong neutrality when making choices. We think that this difference may affect the result of Study 1.

All the experiments were done by calling APIs. We used official API provided by GPT-4 and

Claude 3. For LLaMA 2, we used free API provided by Groq[3]. All the experiments were conducted from February 24, 2024 to April 12, 2024, by calling 320 API calls[4] for each LLM. Also, to reduce the effect of stochastic generation procedure and focus on LLMs' deterministic computation of generating tokens, we set the temperature value as zero for all API calls[5].

After the generation procedure, two of the authors manually extracted probability regarding perceived agreement from LLMs' answers. As the probability is clearly stated in each generated answer, the extraction procedure is straight-forward. Based on the extracted values, we conducted statistical analysis to verify the hypotheses.

3.3 Environment for the experiment

All the experiments were done in the following environment. For the hardware system, we used a Macbook Pro with 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.

3.4 Result and Discussion

The results on examining FCE of LLMs are as follows. (1) LLMs do have FCE in general. (2) FCE exists regardless of the character that we provided: culture and gender.

H1-1 (FCE) Table 2 shows the result of statistical analysis for H1-1. In general, FCE was observed in all of three models. GPT-4 has FCE in three of four stories ($p < 0.05$) except Story 1. Claude 3 and LLaMA 2 have FCE in three of four stories ($p < 0.05$) except Story 3. As the average value of $A_1(1)$ is smaller than that of $A_2(1)$, we conclude that there is a general tendency of FCE in Story 1, though the difference is marginal ($U = 820$, $p = 0.16$) in GPT-4. So, the only story that the LLMs do not demonstrate FCE is Story 3.

We suspect that refinement policy of LLMs affect demonstration of FCE, when the options are related to any ethical/legal issues. Regarding the Story 3, the story provides an option of contesting traffic ticket because of the incorrect information even though the driver drove 38 mph in 25-mph zone actually. Thus, it is possible that the refinement policy of LLM interprets 'contesting' option as refusing legal judgment, which seems illegal or unethical. Such interpretation may affect the model to adjust its answer to 'paying fine,' when we ask LLMs to make decision of other peers' thought.

H1-2 (Culture) Table 3 shows the difference of $A_1(1) - A_2(2)$, whose absolute value (Δ) indicates the strength of FCE[6]. The result shows that FCE was observed on characters with different cultural backgrounds. The statistical test result is not much different across cultures on each story.

Additionally, the result indicates that LLMs sometimes give culturally biased result when we invoke FCE. The average change on strength $|\Delta_{Korean} - \Delta_{American}|$ for each model reveal such cultural bias; the large value indicates strong bias. In GPT-4, the average change on strength is 4.0, because the strength values are 0.5, 14.5, 1.0, and 0.0 on each story. Also, in Claude 3, the average change is 4.8; strength values are 15.0, 2.0, 0.0, and 2.5. Meanwhile, compared to the other two models, LLaMA 2 showed a larger value of 7.5; strength values are 2.5, 11.0, 0.0 and 16.3.

The result suggests that the LLMs may have cultural biases in their parameters. The result showed that the strength of FCE changes when we just change the cultural background of a character. Since the hypothetical stories ask culturally insensitive questions, the refinement process may not be able to catch subtle changes on the answer due to changes on cultural backgrounds.

H1-3 (Gender) Table 3 also shows the strength of FCE across different gender groups. The result shows that FCE was observed on characters with different cultural backgrounds. The statistical test result is not much different across cultures on each story.

Additionally, the result suggests that different LLMs reveal gender biases different situations. Similar to H1-2, we compute the average change on strength $|\Delta_{male} - \Delta_{female}|$ for each model. In GPT-4, the average change on strength is 2.8, and the change is larger than 5.0 in two stories: Story 2 and 3. In Claude 3, the average change on strength is 3.5, and the change is larger than 5.0 in Story 4. And, in LLaMA 2, the average change on strength is 0.8, and there is no story whose change

---

[3] http://groq.com

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

[5] Code: [anonymized for the review]

[6] Due to the page limit, we do not describe the entire statistical results including $A_1(1)$ or standard errors, here. For the detailed results, please refer to the Appendix B.

		GPT-4		Claude 3 Opus		LLaMA 2 70B	
Story	Option c	$A_c(1)$ ($\pm$SE)	U	$A_c(1)$ ($\pm$SE)	U	$A_c(1)$ ($\pm$SE)	U
Story 1 (Paper)	1. Individual	60.00 ($\pm$0.0)	820	60.00 ($\pm$0.0)	1600 ***	60.25 ($\pm$0.3)	1249 ***
	2. Group	59.75 ($\pm$0.3)		40.00 ($\pm$0.0)		49.00 ($\pm$1.6)	
Story 2 (Supermarket)	1. Sign	60.25 ($\pm$0.3)	1132 ***	61.00 ($\pm$0.5)	1600 ***	70.00 ($\pm$0.0)	880 *
	2. Not sign	52.00 ($\pm$1.6)		35.50 ($\pm$0.8)		69.00 ($\pm$0.5)	
Story 3 (Traffic Ticket)	1. Pay fine	62.50 ($\pm$0.7)	1000 ***	70.00 ($\pm$0.0)	800	70.00 ($\pm$0.0)	800
	2. Contest	60.00 ($\pm$0.0)		70.00 ($\pm$0.0)		70.00 ($\pm$0.0)	
Story 4 (Space Program)	1. Vote for	60.00 ($\pm$0.0)	1600 ***	60.00 ($\pm$0.0)	1160 ***	61.25 ($\pm$0.5)	1477.5 ***
	2. Vote against	40.00 ($\pm$0.0)		50.87 ($\pm$1.6)		43.50 ($\pm$1.2)	

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

Table 2: Result of H1-1, “LLMs have FCE." SE and U stands for standard error and U statistic, respectively.

on strength is larger than 5.0. Thus, we conclude that GPT-4 reveals gender bias on daily situations (Story 2 and 3) but Claude 3 and LLaMA 2 do not. Especially, LLaMA 2 successfully controlled gender bias in all hypothetical stories.

4 Study 2: Mitigating FCE with Prompt

We found False Consensus Effect (FCE) exists in LLMs in Study 1. To control the strength of FCE in LLMs, Study 2 focuses on an experiment that compares different prompting styles. This section illustrates the prompting styles that we compare, the result of such comparison, and findings from the result.

4.1 Procedure and Hypotheses

In this study, we compare how the strength of FCE changes when using different prompting styles. As the question-answering procedure of FCE experiment can be seen as a reasoning task, we refer to prompting methods for reasoning when designing the second experiment. Specifically, we focus on two parts in a prompt: (1) provided information and (2) requested type of reasoning chain.

Provided information (P1-P4). To see whether LLM is affected by its own reasoning, we set four levels for provided information. (P1) *No information*: The prompt does not have any information other than the original FCE question. (P2) *Supportive information*: The prompt has a supportive reasoning about the option about the LLM’s choice. (P3) *Opposite information*: The prompt has an opposite reasoning about the LLM’s choice. (P4) *Irrelevant information*: The prompt has a reason-

Story 1	GPT-4	Claude 3	LLaMA 2
Korean	0.0	+20.0***	+17.5***
American	+0.5	+5.0**	+20.0***
Male	0.0	+4.5***	+20.0***
Female	+0.5	+8.0***	+20.0***
Story 2	GPT-4	Claude 3	LLaMA 2
Korean	+15.5***	0.0	+31.0***
American	+1.0	+2.0*	+20.0***
Male	+5.5**	+0.5	+26.0***
Female	+11.0***	+1.5*	+25.0***
Story 3	GPT-4	Claude 3	LLaMA 2
Korean	+2.0*	0.0	0.0
American	+3.0**	0.0	0.0
Male	+5.0***	0.0	0.0
Female	0.0	0.0	0.0
Story 4	GPT-4	Claude 3	LLaMA 2
Korean	+20.0***	+16.5***	+1.0
American	+20.0***	+19.0***	+17.3***
Male	+20.0***	+17.5***	+8.0***
Female	+20.0***	+8.0***	+10.3***

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

Table 3: Difference $A_1(1) - A_2(1)$ for H1-2 and H1-3.

ing irrelevant to the given situation. We set these four levels because of two reasons. First, LLMs often asked to provide their reasoning to the users. Second, studies reported that information provided in prompt can change LM's output(Gonen et al., 2023). For simplicity, we focus on the effect of reasoning text generated when LLM make its own choice. So, we simulate LLM's own reasoning in P2-P4. The detailed methods for simulating each LLM's reasoning and the detailed prompts for each level are described in Appendix C.

Reasoning chain (R1-R4). To see whether LLM is affected by reasoning methods on estimating $A_c(1)$, we test four types of reasoning chain. (R1) *Direct Answer*: The prompt just ask LLM to generate $A_c(1)$ without reasoning. (R2) *Simple Reasoning*: The prompt ask LLM to generate $A_c(1)$ with a reasoning. (R3) *CoT-style*: The prompt asks LLM to generate reasoning first, and $A_c(1)$ for the second. (R4) *Reflexion-style*: The prompt asks LLM to generate reasoning first, re-think about its reasoning, and finally generate $A_c(1)$. We set these four levels because of two reasons. First, as FCE can be seen as a byproduct of intuitive thinking, we suspect methods for deep or complicated reasoning may reduce the strength of FCE. Second, studies confirmed that reasoning methods invoke deep thinking can improve the accuracy of problem solving(Wang et al., 2023a). We excluded reasoning methods that utilizes external observations to avoid introducing alternative biases. The detailed prompts for each level are described in Appendix **??**.

Analysis: We follow the same analysis method in Study 1. The hypotheses are stated below. To test whether FCE is observed, we conduct Mann-Whitney U-test. And to reveal the trend in strength of FCE, we compare Δ_G for each prompting style.

H2-1. Providing supportive information has higher FCE than the others; i.e., $|\Delta_{P2}| > |\Delta_{Pk}|$ for any $k = 1, 3, 4$.

H2-2. As reasoning method requires deeper thinking, the strength of FCE decreases; i.e., $|\Delta_{R1}| > |\Delta_{R2}| > |\Delta_{R3}| > |\Delta_{R4}|$.

Additionally, we try to investigate the interaction effect between the provided information and reasoning chain. This is because studies reported that adding additional information within reasoning chain can improve the performance (Baek et al., 2023; Brown et al., 2020). As our aim is to control the strength of FCE, we examine interaction effect by finding pair of two factors when Δ_G is zero, the maximum, or the minimum among 16 pairs.

GPT-4	R1	R2	R3	R4
P1	$+20.0^{***}$	$+20.0^{***}$	$+0.3^{***}$	$+7.5^{***}$
P2	$+20.0^{***}$	$+20.0^{***}$	$+23.5^{***}$	0.0
P3	0.0	-20.0	-6.8	$+16.5^{***}$
P4	$+20.0^{***}$	$+20.0^{***}$	$+15.5^{***}$	$+37.9^{***}$
Claude 3	R1	R2	R3	R4
P1	$+9.1^{***}$	$+20.3^{***}$	$+15.9^{***}$	-0.9
P2	$+5.0^{***}$	$+20.0^{***}$	$+14.1^{***}$	-2.5
P3	0.0	-20.0	-6.6	$+1.8^{*}$
P4	$+20.4^{***}$	$+20.5^{***}$	$+12.4^{***}$	-9.9
LLaMA 2	R1	R2	R3	R4
P1	$+17.8^{***}$	+11.51	+2.1	+5.8
P2	$+20.0^{***}$	$+20.0^{***}$	$+23.5^{***}$	0.0
P3	-19.0	-2.15	+2.1	-6.7
P4	$+12.5^{***}$	-2.0	$+6.2^{**}$	+2.5

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

Table 4: Result of H2-1 and H2-2 on **Story 4**. Number in each cell represents $A_1(1) - A_2(1)$.

4.2 Environment for the experiment

The experimental setup is the same as that of Study1, except the prompting styles. We tested the same LLMs, GPT-4, Claude 3, and LLaMA 2, as in the Study 1. All the experiments were conducted from February 24, 2024 to April 12, 2024, by calling 5120 API calls[7] for each LLM[8].

4.3 Result and Discussion

The results on mitigating FCE with Prompt are as follows. (1) (2) (3).

H2-1 (Provided Information) Table 4 describes the result of statistical analysis for H2-1. As the Story 4 showed the highest FCE among the four stories, we show the result on Story 4 as a representative result and illustrate other stories in Appendix D. In general, the strength of FCE is lower in P3 (Opposite Information) compared to the other conditions. When we control the reasoning type to R1 (no reasoning), GPT-4 has strength value zero on P3 but 20.0 on other information prompts. Similarly, in Claude 3, P3 method showed strength

---

[7] 40 characters, 4 stories, 2 options, 16 condition pairs.

[8] Code: [anonymized for the review]

value zero but the other cases have strength value between 5.0 to 20.4. In LLaMA 2, Δ_{P2} is -19.0, which is lower than zero, when we used P3 method. So, the H2-1 is partially supported for the case between Δ_{P2} and Δ_{P3}.

We suspect the partial rejection of H2-1 is due to a ceiling effect on LLMs probability estimation. When we measured the range of LLM's estimation on $A_c(1)$, the estimated values are mostly in between 30% and 70%. On GPT-4, Claude 3 and LLaMA 2, 98.24%, 95.79%, and 92.75% of the estimated values are in that range. Thus, as the models already answered high enough probabilities for $A_1(1)$ in P1, there is not much room to increase the strength of FCE. Thus, it is not easy to observe the effect of providing supportive information (P2). On the other hand, we can observe the effect of providing opposite information in Table 4, because the estimated value can be decreased by a large amount.

H2-2 (Reasoning chain) Table 4 also describes the result of statistical analysis for H2-2. In general, the strength of FCE becomes lower as we change our reasoning chain method from R1 (no reasoning) to R3 (CoT-style). When we control the information type to P1 (no information), R3 has the lowest strength of FCE (0.3) among four conditions in GPT-4. Similarly, in LLaMA 2, R3 has the lowest strength of FCE (2.1) among the four conditions. But, it is worth noticing that R4 (Reflexion-style), the strength of FCE becomes higher than R3. For example, in GPT-4, the value changes from 0.3 to 7.5.

We suspect that the thinking procedure in P2, P3, and P4 makes LLMs weigh more on the given information during the decision process. In P2 and P3, as there is not much information provided, LLMs cannot decide an option easily. So, the strength of FCE generally becomes lower in P2 and P3. Also, in P4, as there is some information provided by LLMs themselves in the previous thought, LLMs try to decide their action based on that information. Thus, the strength of FCE becomes higher in P4, as they lean more on their opinion.

Combined result When we combine the results of H2-1 and H2-2, we could find conditions that can maximize or minimize the strength of FCE in general. To maximize the strength of FCE, the result suggests that providing supportive information (P2) with simple reasoning approach may give the maximum strength of FCE in general. On the other hand, to minimize the strength of FCE in general, the result suggests that providing opposite information (P3) with CoT-style prompt may give the minimum strength of FCE, which is near zero. So, by combining these prompting methods, we can adjust the strength of FCE, regardless of LLMs.

5 Conclusion

To confirm whether LLMs also exhibit FCE, a cognitive bias inherent in humans, we conducted experiments under conditions that closely simulate those experienced by humans. Subsequently, we implemented extensive controls to minimize the impact of other cognitive biases. Additionally, to regulate the manifestation of FCE, we modified both the provided information and the depth of reasoning. we confirmed the interaction between these changes and established a total of 16 variables to control the expression of FCE in LLMs. Through this process, we categorized our results into five major findings. We applied psychological experiments conducted on humans under conditions very similar to humans and confirmed the results. We minimized the effects of other Cognitive biases as much as possible and then purely confirmed the bias of FCE. We identified differences in the expression of FCE according to culture and Gender. We presented a methodology that can maximize the expression of FCE or lead to choices that are contrary to one's preferences by adding a neutral situation.

Limitation

Among the LLM-generated answers, there are cases in which the answer was rejected. This phenomenon is especially prevalent in LLaMA, where, due to the structural characteristics of LLMs, sometimes inappropriate answers are filtered out. Therefore, in situations where the LLM has refused to answer, it is impossible to determine whether it has chosen neutrality or an extreme option. Additionally, when provided with irrelevant sentences in LLaMA, errors occurred[9]. Error values were excluded from the analysis.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman,

---

[9] Sure! Here are my estimates: 55% choose individual paper 95% choose group paper

Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

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

Jinheon Baek, Alham Fikri Aji, and Amir Saffari. 2023. Knowledge-augmented language model prompting for zero-shot knowledge graph question answering. *arXiv preprint arXiv:2306.04136*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Wenhu Chen, Xueguang Ma, Xinyi Wang, and William W. Cohen. 2023. Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks. *Transactions on Machine Learning Research*.

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.

Hila Gonen, Srini Iyer, Terra Blevins, Noah Smith, and Luke Zettlemoyer. 2023. Demystifying prompts in language models via perplexity estimation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10136–10148, Singapore. Association for Computational Linguistics.

Ari Holtzman, Peter West, Vered Shwartz, Yejin Choi, and Luke Zettlemoyer. 2021. Surface form competition: Why the highest probability answer isn't always right. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7038–7051, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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*.

Mohit Iyyer, John Wieting, Kevin Gimpel, and Luke Zettlemoyer. 2018. Adversarial example generation with syntactically controlled paraphrase networks. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1875–1885, New Orleans, Louisiana. Association for Computational Linguistics.

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.

Masahiro Kaneko, Danushka Bollegala, Naoaki Okazaki, and Timothy Baldwin. 2024. Evaluating gender bias in large language models via chain-of-thought prompting. *arXiv preprint arXiv:2401.15585*.

Yan Leng. 2024. Can llms mimic human-like mental accounting and behavioral biases? *Available at SSRN 4705130*.

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*.

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.

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.

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*.

Boshi Wang, Sewon Min, Xiang Deng, Jiaming Shen, You Wu, Luke Zettlemoyer, and Huan Sun. 2023a. 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.

Yiwei Wang, Yujun Cai, Muhao Chen, Yuxuan Liang, and Bryan Hooi. 2023b. Primacy effect of ChatGPT. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 108–115, Singapore. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. 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.

Siye Wu, Jian Xie, Jiangjie Chen, Tinghui Zhu, Kai Zhang, and Yanghua Xiao. 2024. How easily do irrelevant inputs skew the responses of large language models? *Preprint*, arXiv:2404.03302.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. In *Advances in Neural Information Processing Systems*, volume 36, pages 11809–11822. Curran Associates, Inc.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2022. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*.

A Prompt of Study 1

A.1 Stories from Ross et al. (1977)

Story 1

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, Choose **individual paper** or **Choose group paper**?

Story 2

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

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

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**?

A.2 Full prompt for the experiment

System Prompt

Name "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 Americans Man": ["James Smith", "John Johnson", "Robert Brown", "Michael Davis", "William Miller", "David Wilson", "Richard Moore", "Joseph Taylor", "Charles Anderson", "Thomas Jackson"]

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

B Detailed result from Study 1

B.1 Effect of cultural bias in FCE

Tables 5 and 6 show the detailed result for H1-2.

B.2 Effect of gender bias in FCE

Tables 7 and 8 show the detailed result for H1-3.

C Prompt of Study 2

C.1 Provided information

The following paragraphs are actual reasoning generated each LLM. We used these paragraphs to give supportive or opposite information in testing H2-1 of Study 2.

C.1.1 GPT

Reasoning information 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 information 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

		GPT-4		Claude 3 Opus		LLaMA 2 70B	
Story	Option c	$A_c(1)$ (±SE)	U	$A_c(1)$ (±SE)	U	$A_c(1)$ (±SE)	U
Story 1	1. Individual	60.00 (±0.0)	210	60.00 (±0.0)	400 ***	60.00 (±0.0)	250 ***
(Paper)	2. Group	59.50 (±0.0)		40.00 (±0.0)		55.00 (±1.6)	
Story 2	1. Sign	60.00 (±0.0)	210	60.00 (±0.0)	400 ***	70.00 (±0.0)	240 ***
(Supermarket)	2. Not sign	59.00 (±0.0)		40.00 (±0.0)		68.00 (±0.0)	
Story 3	1. Pay fine	63.00 (±0.0)	260 ***	70.00 (±0.0)	200	70.00 (±0.0)	200
(Traffic Ticket)	2. Contest	60.00 (±0.0)		70.00 (±0.0)		70.00 (±0.0)	
Story 4	1. Vote for	60.00 (±0.0)	400 ***	60.00 (±0.0)	370 ***	61.25 (±0.0)	390 ***
(Space Program)	2. Vote against	40.00 (±0.0)		42.75 (±0.0)		41.00 (±0.0)	

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

Table 5: Result of **H1-2 European Americans group**, “LLMs have FCE ." SE and U stands for standard error and U statistic, respectively.

		GPT-4		Claude 3 Opus		LLaMA 2 70B	
Story	Option c	$A_c(1)$ (±SE)	U	$A_c(1)$ (±SE)	U	$A_c(1)$ (±SE)	U
Story 1	1. Individual	60.00 (±0.0)	200	60.00 (±0.0)	400 ***	60.50 (±0.0)	371.5 ***
(Paper)	2. Group	60.00 (±0.0)		40.00 (±0.0)		43.00 (±1.6)	
Story 2	1. Sign	60.50 (±0.0)	352.5 ***	62.00 (±0.0)	400 ***	70.00 (±0.0)	200 *
(Supermarket)	2. Not sign	45.00 (±0.0)		31.00 (±0.0)		70.00 (±0.0)	
Story 3	1. Pay fine	62.00 (±0.0)	240 *	70.00 (±0.0)	200	70.00 (±0.0)	200
(Traffic Ticket)	2. Contest	60.00 (±0.0)		70.00 (±0.0)		70.00 (±0.0)	
Story 4	1. Vote for	60.00 (±0.0)	400 ***	60.00 (±0.0)	210 ***	62.50 (±0.0)	355
(Space Program)	2. Vote against	40.00 (±0.0)		59.00 (±0.0)		46.00 (±0.0)	

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

Table 6: Result of **H1-2 Korean group**, “LLMs have FCE." SE and U stands for standard error and U statistic, respectively.

		GPT-4		Claude 3 Opus		LLaMA 2 70B	
Story	Option c	$A_c(1)$ (±SE)	U	$A_c(1)$ (±SE)	U	$A_c(1)$ (±SE)	U
Story 1	1. Individual	60.00 (±0.0)	200	60.00 (±0.0)	400 ***	60.5 (±0.0)	343 ***
(Paper)	2. Group	60.00 (±0.0)		40.00 (±0.0)		40.00 (±0.1)	
Story 2	1. Sign	60.50 (±0.0)	257.5	62.00 (±0.0)	400 ***	70.00 (±0.0)	210 *
(Supermarket)	2. Not sign	66.00 (±0.0)		36.00 (±0.0)		69.50 (±0.0)	
Story 3	1. Pay fine	65.00 (±0.0)	300 ***	70.00 (±0.0)	200	70.00 (±0.0)	200
(Traffic Ticket)	2. Contest	60.00 (±0.0)		70.00 (±0.0)		70.00 (±0.0)	
Story 4	1. Vote for	60.00 (±0.0)	400 ***	60.00 (±0.0)	280 ***	62.50 (±0.5)	362.5 ***
(Space Program)	2. Vote against	40.00 (±0.0)		52.00 (±0.0)		45.00 (±0.0)	

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

Table 7: Result of **H1-3 Man Group**, “LLMs have FCE." SE and U stands for standard error and U statistic, respectively.

		GPT-4		Claude 3 Opus		LLaMA 2 70B	
Story	Option c	$A_c(1)$ ($\pm$SE)	U	$A_c(1)$ ($\pm$SE)	U	$A_c(1)$ ($\pm$SE)	U
Story 1 (Paper)	1. Individual 2. Group	60.00 ($\pm$0.0) 59.50 ($\pm$0.0)	210	60.00 ($\pm$0.0) 40.00 ($\pm$0.0)	400 ***	60.00 ($\pm$0.0) 52.00 ($\pm$0.0)	280 ***
Story 2 (Supermarket)	1. Sign 2. Not sign	60.00 ($\pm$0.0) 49.00 ($\pm$0.0)	210 ***	60.00 ($\pm$0.0) 35.00 ($\pm$0.0)	400 ***	70.00 ($\pm$0.0) 68.50 ($\pm$0.5)	230 *
Story 3 (Traffic Ticket)	1. Pay fine 2. Contest	60.00 ($\pm$0.0) 60.00 ($\pm$0.0)	200	70.00 ($\pm$0.0) 70.00 ($\pm$0.0)	200	70.00 ($\pm$0.0) 70.00 ($\pm$0.0)	200
Story 4 (Space Program)	1. Vote for 2. Vote against	60.00 ($\pm$0.0) 40.00 ($\pm$0.0)	400 ***	60.00 ($\pm$0.0) 49.75 ($\pm$0.0)	300 ***	60.00 ($\pm$0.0) 42.00 ($\pm$0.0)	380 ***

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

Table 8: Result of **H1-3 Woman Group**, “LLMs have FCE." SE and U stands for standard error and U statistic, respectively.

comments are edited or used, you risk misrepresentation or unintended endorsement, which could impact your personal image or privacy."

Reasoning information 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 information 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."

C.1.2 Cluade

Reasoning information 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 information 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 information 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 information 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."

C.1.3 LLaMA

C.2 Reasoning information 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."

C.3 Reasoning information 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.."

C.4 Reasoning information 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."

C.5 Reasoning information 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

GPT-4	R1	R2	R3	R4
P1	+0.3	0.0	-20.0	1.3*
P2	0.0	+20.0***	-5.3	-5.5
P3	0.0	0.0	+2.75	+24.5***
P4	+12.5***	-1.0	+1.0	+15.5***
Claude 3	R1	R2	R3	R4
P1	+20.0***	+13.5***	+1.25*	-40.0
P2	+21.25***	+9.5***	+20.0***	-3.0
P3	0.0	-6.5	+1.9*	-4.0
P4	0	+7.8***	+0.0	-39.0
LLaMA 2	R1	R2	R3	R4
P1	+11.25***	+7.92***	+4.9*	-0.7
P2	+0.0	+20.0***	+23.5***	0.0
P3	-19.0	-2.15	+2.1	-6.7
P4	+12.5***	-2.0	+6.2**	+2.5

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

Table 9: Result of H2-1 and H2-2 on **Story 1**. Number in each cell represents $A_1(1) - A_2(1)$.

`potential benefits outweigh the costs in the long run."`

- **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.

D Detailed result from Study 2

Tables 9, 10 and 12 show the detailed result for H2 on Story 1, 2, and 3.

GPT-4	R1	R2	R3	R4
P1	+8.25***	+9.5***	+8.0***	+24.45***
P2	+20.0***	+29.75***	+15.0***	+13.75***
P3	-9.5	-11.0	+2.5	+34.5***
P4	+19.5***	+6.25***	+7.0***	+17.5***
Claude 3	R1	R2	R3	R4
P1	+25.5***	+25.5***	+6.6***	+2.3***
P2	+29.0***	+30.0***	+21.5***	-33.8
P3	-29.6	-30.0	-9.4	+33.0***
P4	+27.7***	+36.7***	+12.5***	-1.8
LLaMA 2	R1	R2	R3	R4
P1	+1.0*	-0.9	+2.8	+6.5***
P2	+20.0***	+29.8***	+15.0***	13.8***
P3	-3.25	-12.6	+7.0***	-16.6
P4	+37.0***	-8.8	-3.0	-12.8

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

Table 10: Result of H2-1 and H2-2 on **Story 2**. Number in each cell represents $A_1(1) - A_2(1)$.

GPT-4	R1	R2	R3	R4
P1	+2.25***	+10.0***	0.0	-2.8
P2	+10.0***	+10.0***	+5.0***	+19.0***
P3	-9.5	-7.25	-0.8	+20.5***
P4	+10.0***	+5.8***	0.0	-7.5
Claude 3	R1	R2	R3	R4
P1	+0.0	+0.0	+4.9***	-48.8
P2	+0.0	+9.0***	+5.3***	-34.8
P3	+0.0	-1.3	-3.5	+39.0***
P4	+0,0	+0.0	+2.5***	+20.2***
LLaMA 2	R1	R2	R3	R4
P1	+0.0	+0.1	-8.4	+21.1***
P2	+10.0***	+10.0***	+5.0***	+9.0***
P3	+0.0	-10.3	+5.1**	+11.9***
P4	+0.0	+4.4**	-8.0	+13.1***

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

Table 12: Result of H2-1 and H2-2 on **Story 3**. Number in each cell represents $A_1(1) - A_2(1)$.