Cognitive Biases in Artificial Intelligence: Susceptibility of a Large Language Model to Framing Effect and Confirmation Bias*

Li Hao[1], *Wang You*[1, 2], *Yang Xueling*[1, 2]

([1]Department of Psychology, School of Public Health, Southern Medical University (Guangdong Provincial Key Laboratory of Tropical Disease Research), Guangzhou, 510515)([2]Department of Psychiatry, Zhujiang Hospital, Southern Medical University, Guangzhou, 510280)

Abstract The rapid advancement of Artificial Intelligence (AI) and Large Language Models (LLMs) has led to their increasing integration into various domains, from text generation and translation to question-answering. However, a critical question remains: do these sophisticated models, much like humans, exhibit susceptibility to cognitive biases? Understanding the presence and nature of such biases in AI is paramount for assessing their reliability, enhancing their performance, and predicting their societal impact. This research specifically investigates the susceptibility of Google's Gemini 1.5 Pro and DeepSeek, two prominent LLMs, to framing effects and confirmation bias. The study meticulously designed a series of experimental trials, systematically manipulating information proportions and presentation orders to evaluate these biases.

In the framing effect experiment, a genetic testing decision-making scenario was constructed. The proportion of positive and negative information (e.g., 20%, 50%, or 80% positive) and their presentation order were varied. The models' inclination towards undergoing genetic testing was recorded. For the confirmation bias experiment, two reports—one positive and one negative—about "RoboTaxi" autonomous vehicles were provided. The proportion of erroneous information within these reports (10%, 30%, and 50%) and their presentation order were systematically altered, and the models' support for each report was assessed.

The findings demonstrate that both Gemini 1.5 Pro and DeepSeek are susceptible to framing effects. In the genetic testing scenario, their decision-making was primarily influenced by the proportion of positive and negative information presented. When the proportion of positive information was higher, both models showed a greater inclination to recommend or proceed with genetic testing. Conversely, a higher proportion of negative information led to greater caution or a tendency not to recommend the testing. Importantly, the order in which this information was presented did not significantly influence their decisions in the framing effect scenarios.

Regarding confirmation bias, the two models exhibited distinct behaviors. Gemini 1.5 Pro did not show an overall preference for either positive or negative reports. However, its judgments were significantly influenced by the order of information presentation, demonstrating a "recency effect," meaning it tended to support the report presented later. The proportion of erroneous information within the reports had no significant impact on Gemini 1.5 Pro's decisions. In contrast, DeepSeek exhibited an overall confirmation bias, showing a clear preference for positive reports. Similar to Gemini 1.5 Pro, DeepSeek's decisions were also significantly affected by the order of information presentation, while the proportion of misinformation had no significant effect.

These results reveal human-like cognitive vulnerabilities in advanced LLMs, highlighting critical challenges to their reliability and objectivity in decision-making processes. Gemini 1.5 Pro's sensitivity to presentation order and DeepSeek's general preference for positive information, coupled with its sensitivity to order, underscore the need for careful evaluation of potential cognitive biases during the development and application of AI. The study suggests that effective measures are necessary to mitigate these biases and prevent potential negative societal impacts. Future research should include a broader range of models for comparative analysis and explore more complex interactive scenarios to further understand and address these phenomena. The findings contribute significantly to understanding the limitations and capabilities of current AI systems, guiding their responsible development, and anticipating their potential societal implications.

Key words artificial intelligence, large language models, cognitive bias, confirmation bias, framing effect

* This work was supported by the Guangdong Provincial Undergraduate Teaching Quality and Teaching Reform Project (Yue Jiao Gao Han [2024] No. 30) and the Ministry of Education of China, Humanities and Social Sciences Research Project (22YJCZHI82).

** Corresponding Author：Yang Xueling, E-mail: yhtyxl2006@126.com

DOI:10.16719/j.cnki.1671-6981.20250411

Introduction

Imagine a news app subtly curating your reality, feeding you articles that echo your existing beliefs while silencing dissenting voices. This isn' t science fiction, but a potential consequence of confirmation bias, a well-documented flaw in human cognition. As artificial intelligence (AI) becomes increasingly integrated into our lives, a crucial question emerges: do AI systems, designed to mimic and even surpass human intelligence, also inherit our cognitive biases? And if so, what are the implications for a future increasingly reliant on AI decision-making?

The rise of AI has profoundly impacted human life, transforming how we live, work, and learn. Simultaneously, AI has revolutionized research methods, techniques, and paradigms within psychology, facilitating the exploration of human cognitive patterns and driving advancements in the field（刘冬予等，2024）. Concurrently, AI has revolutionized psychological research, providing new tools and methodologies for understanding the human mind (Botvinick, 2022). AI now augments human capabilities in diverse applications, from healthcare (Gandhi et al., 2023) to finance (Lee & Yoon, 2021; Mukhamediev et al., 2022), raising questions about the nature of intelligence itself. Yet, despite these advancements, a critical gap remains in our understanding of AI' s cognitive abilities, particularly its susceptibility to the same biases that shape human judgment. Cognitive biases, systematic deviations from rational judgment arising from heuristics and limitations in information processing (Tversky & Kahneman, 1974), can distort our perception of reality, leading to suboptimal decisions (Sergio et al., 2023). If these biases also manifest in AI systems, the implications for their reliability and safety are profound.

This research investigates the susceptibility of Gemini 1.5 Pro and DeepSeek to confirmation bias and framing effects. Confirmation bias, the tendency to selectively favor information that confirms pre-existing beliefs, can create echo chambers (Gu et al., 2024) and polarized discourse (张瀚予等 , 2024; Piao et al., 2023). In AI systems like Large Language Models (LLMs), critical for information retrieval and generation, such bias may stem from skewed training data or algorithmic prioritization, potentially leading to skewed predictions, reinforcing societal biases (Michel & Peters, 2020), and compromising information fairness and diversity. In this study, we operationalize confirmation bias by presenting LLMs with positive and negative reports on a controversial event (e.g., autonomous vehicles) and examining if altering report order or embedding varying misinformation levels systematically influences the model's stance or report preference. While not directly measuring ‘prior beliefs,' this approach assesses if sensitivity to information presentation and quality reveals processing patterns analogous to human confirmation bias.

Framing effects, on the other hand, demonstrate how the presentation of information influences judgment and decision-making. In AI, framing effects can manifest through variations in data preprocessing, training data selection, or the framing of decision-making problems (Binz & Schulz, 2023; Tversky & Kahneman, 1981). Given the pervasive and profound impact of framing effects on human decision-making, investigating whether LLMs exhibit similar sensitivities is crucial for evaluating their objectivity and reliability in tasks such as information summarization and recommendation generation. In this study, we operationalize the framing effect as follows: when an LLM is presented with equivalent information about the same decision problem (e.g., genetic testing), we examine whether merely altering the proportion or presentation order of positive versus negative information leads to systematic changes in its attitude or decision tendency regarding that problem.

While human cognitive biases have been extensively studied, research on these biases in AI remains limited. Existing work primarily focuses on the application and effectiveness of AI in specific domains (e.g., healthcare, autonomous driving, finance) (Gandhi et al., 2023; Lee &

Yoon, 2021; Mukhamediev et al., 2022). While research on AI decision-making (Binz & Schulz, 2023), emotions （侯焊超等 , 2024）, and empathy （赵立等 , 2024） has broadened our understanding of AI capabilities, a significant gap persists in understanding AI' s vulnerability to cognitive biases.

However, investigating the cognitive characteristics of rapidly evolving AI, such as Large Language Models (LLMs), presents unique challenges, particularly given that their iteration speed far outpaces traditional research cycles. Despite this, exploring the behavioral patterns of advanced models at specific points in their development—for instance, their susceptibility to cognitive biases—remains crucial. Such research offers significant theoretical and practical value for understanding the capability boundaries of current AI systems, guiding their responsible development and deployment, and anticipating their potential societal impacts.

This study addresses this gap by investigating Gemini' s responses to experimentally manipulated framing effects and confirmation bias scenarios. Specifically, we examine Gemini' s susceptibility to framing effects manipulated through the proportion and order of positive and negative information, and to confirmation bias manipulated through the proportion of misinformation and the order of information presentation. We hypothesize that: (1) Gemini will exhibit similar confirmation bias to humans, manifesting as decision bias; (2) Gemini will be susceptible to framing effects, with the presentation of information influencing its decisions; and (3) compared to humans, Gemini may exhibit distinct characteristics in its susceptibility to these biases, potentially due to its reliance on training data. By exploring these cognitive biases in AI, this research aims to illuminate the challenges and opportunities for human-AI collaboration and inform the development of safer, more reliable AI systems.

1 Methods

1.1 Subjects

This study focuses on cutting-edge multi-modal AI models, specifically examining Google' s Gemini 1.5 Pro and DeepSeek. Gemini 1.5 Pro, recently introduced by Google, possesses robust language comprehension and generation capabilities, enabling it to process information in diverse formats, including text, code, images, and videos. Alongside Gemini, this research also utilizes DeepSeek, another advanced AI model known for its strong language understanding, reasoning, and coding abilities. Certain DeepSeek models also offer multimodal functionalities, allowing for the processing of various data types. This experiment leverages both Gemini 1.5 Pro and DeepSeek models to conduct its analysis.

1.2 Instruments

1.2.1 Scenarios Design

This study utilized a series of scenarios to investigate AI model' s responses to framing effects and confirmation bias. To demonstrate the potential influence of each bias, multiple scenarios were designed, with each scenario encompassing two or more versions that manipulated key aspects of information presentation. Specifically, two types of scenarios were developed: framing effect scenarios and confirmation bias scenarios. The goal was to assess AI model' s sensitivity to these biases by analyzing its responses across different scenario versions. Based on sample sizes commonly used in human cognitive bias and decision-making research (Berthet et al., 2024; Zhao et al., 2024), we collected 300 responses from AI models for each version of both scenario types.

1.2.2 Framing Effect Scenarios

These scenarios utilized gene testing as the context. Positive and negative information points were carefully curated to ensure comparability in quantity, tone, and similar text lengths. Scenarios systematically manipulated both the proportion of positive to negative information (20% positive/80% negative; 50% positive/50% negative; 80% positive/20% negative), and their presentation order (positive information first then negative, or vice versa). These manipulations resulted in six unique combinations, with 50 responses collected for each, totaling 300 responses. Gemini and DeepSeek were then asked about its inclination to undergo gene testing, based on the presented information, using a structured prompt: "introduction - positive/negative information - negative/positive information - concluding statement." The concluding question was standardized, for example: "Based on all the information provided above, would you be inclined to undergo gene testing? Please explain your main considerations." Crucially, each trial was conducted in a new, independent interaction session, ensuring no carry-over effects or contextual learning from previous trials. This rigorous control over materials and systematic manipulation of variables aimed to ensure measurement reliability and the consistency and comparability of stimuli across conditions, thereby enhancing the likelihood of replicable results.

This scenario' s validity lies in its operationalization of the framing effect construct. It tests whether systematic variations in the "frame"—the proportion and order of positive versus negative arguments regarding gene testing—elicit corresponding shifts in Gemini' s expressed inclination, a principle aligned with classic framing research (Tversky & Kahneman, 1981). Consistent variation in Gemini' s decisions corresponding to these framing manipulations would support the scenario' s validity. Although the cognitive mechanisms in LLMs may differ from humans, this quantifiable and controllable method allows for investigation into framing-induced response biases. Gene testing offered a suitable context with inherent positive and negative aspects, though the ecological validity of such lab-based textual stimuli for AI should be considered.

1.2.3 Confirmation Bias Scenarios

These scenarios employed autonomous vehicles, specifically 'RoboTaxi' as the context. Supporting and opposing reports, based on standardized authentic news materials (controlled for comparable length and narrative style), were presented. Key manipulations involved embedding varying levels of misinformation (10%, 30%, 50%—exaggerated descriptions, inaccurate data) within one of the reports and altering the presentation order (supporting report first, or vice versa). This resulted in twelve unique combinations of proportion and order, with 25 responses collected for each combination, totaling 300 responses. Gemini and DeepSeek' s attitude towards RoboTaxi was recorded after it processed the reports presented via a structured prompt: "positive/negative new - negative/positive news - concluding statement." The concluding question was, for example: "Please carefully read the following two news reports and, based on the information provided, tell me which report you are more inclined to support?" As with the framing effect scenarios, each trial for the confirmation bias experiment was conducted in a new, independent interaction session to prevent contextual influences from prior interactions.

The validity of this scenario design lies in its operationalization of confirmation bias, defined here as the model' s systematic preference for selecting a specific news article over alternatives in a forced-choice task, despite containing higher misinformation. This bias is quantified by a statistically significant higher selection rate of the preferred article across multiple trials. The experiment aimed to determine if LLMs exhibited tendencies analogous to human confirmation bias, such as showing preferential support for initially presented reports or for reports containing less misinformation. While the mechanisms of "belief" formation in LLMs differ from those in humans, this experimental design provides a controllable method to observe systematic tendencies when processing conflicting or imperfect information. It is acknowledged that the ecological validity of such laboratory-based textual measures for AI may warrant consideration, yet this approach offers a crucial methodology for initial explorations of AI's susceptibility to cognitive phenomena like confirmation bias.

1.3 Data Analysis

The study found that the large language models, Gemini 1.5 Pro and DeepSeek, explicitly articulated its preferences in response to the experimental scenarios. For instance, in the framing effect experiment, Gemini and Deepseek' s responses typically centered on its inclination, with statements such as, "I am inclined to undergo gene testing," "I am inclined to consider this cautiously," or "I am inclined not to undergo gene testing." Consequently, these responses were coded on a 3-point scale, where 1 indicated being "inclined not to perform," 2 signified "cautious consideration," and 3 represented being "inclined to perform."

Similarly, in the confirmation bias experiment, Gemini and DeepSeek' s responses directly stated its leaning, for example, "I am more supportive of the first report" or "I am more supportive of the second report." Therefore, for the confirmation bias experiment, responses were coded dichotomously: 1 was assigned if the model supported the positive report, and 2 if it supported the negative report.

Data from AI model' s responses were then analyzed using SPSS. Descriptive statistics summarized response patterns across scenario versions. Independent samples t-tests compared Gemini' s responses between different presentation orders within each scenario type (framing effect and confirmation bias). To examine the impact of information proportion (framing effect) and misinformation proportion (confirmation bias), ANOVAs were conducted. The significance level was set at $p < .05$.

2 Results of Gemini 1.5 Pro

2.1 Experiment 1: Framing Effects

2.1.1 Impact of Information Proportion

Table 1 presents Gemini' s response patterns across different framing effect scenarios, manipulating both the proportion and order of positive and negative information regarding gene testing. Table 2 displays the results of a one-way ANOVA, comparing Gemini' s overall decision-making tendency (measured on a scale where 1 = "inclined not to perform", 2 = "cautious consideration", and 3 = "inclined to perform") across the different proportions of positive and negative information. The ANOVA revealed a significant main effect of information proportion ($F = 90.666, p < .001$).

As seen in Table 1, when presented with a higher proportion of positive information (80%), Gemini predominantly expresses an inclination towards "performing" gene testing (96%). Conversely, when faced with a higher proportion of negative information (80%), Gemini' s responses shift towards "not performing" gene testing (32%), with a substantial portion also expressing "cautious consideration" (55%).

These findings suggest that Gemini' s decision-

Table 1 Response Patterns of Gemini in Different Framing Effect Scenarios (%)

	Positive then
	Negative
		Negative then
		Positive
			Positive(20%)
			Negative(80%)
				Positive(50%)
				Negative(50%)
					Positive(80%)
					Negative(20%)

Inclined to perform	52	49.3	13	43	96
Inclined not to perform	10.7	15.3	32	7	0
Cautious consideration	37.3	35.3	55	50	4

Table 2: Comparison of Gemini' s Responses Across Different Information Proportions（*M* ± *SD*）

Item	Mean	*F*	*p*
Positive (20%) & Negative (80%)	2.42±.7132	90.666	<.001
Positive (50%) & Negative (50%)	2.07±.9667		
Positive (80%) & Negative (20%)	1.08±.3939		
Total	1.86±.923		

Table 3 Comparison of Gemini' s Responses Across Different Presentation Orders（*M* ±*SD*）

Item	Mean	*t*	*p*
Positive then Negative	1.85±.937	-.062	.950
Negative then Positive	1.86±.912		

making in the context of gene testing is susceptible to framing effects, with the proportion of positive and negative information significantly influencing its inclination to perform the test.

2.1.2 Comparison of Different Order Presentations in Framing Effect Scenarios

Table 3 presents the results of an independent samples t-test comparing Gemini' s overall decision-making tendency (as measured on the same scale used in Table 2) across the two different presentation orders of positive and negative information. The results indicate that the order of information presentation did not have a significant effect on Gemini' s decision-making tendency (t = -.062, p > .05). This suggests that, while Gemin' s responses are sensitive to the proportion of positive and negative information (as shown in Table 2), the order in which this information is presented does not appear to significantly influence its overall inclination to perform or not perform gene testing.

2.2 Experiment 2: Confirmation Bias

2.2.1 Comparison of Different Order Presentations

Table 4 presents Gemini' s response patterns in different confirmation bias scenarios concerning autonomous vehicles, while Table 5 demonstrates its responses across various presentation orders (measured on a scale where 1 = supported the positive report, and 2 = supported the negative report). Notably, the study found that in 300 responses, Gemini supported positive reports 150 times and negative reports 150 times. This balanced outcome suggests that Gemini may not possess an inherent confirmation bias towards consistently favoring either positive or negative information overall.

However, presentation order significantly affects Gemini' s decisions (p < .001, from Table 5). Specifically, Gemini exhibits a recency effect, meaning it tends to support the report presented later. For instance, if a positive report was followed by a negative report, Gemini was more likely to support the later (negative) report; conversely, if a negative report was followed by a positive one, it was more inclined to support the later (positive) report.

Therefore, while Gemini may not exhibit an

Table 4 Response Patterns of Gemini in Different Confirmation Bias Scenarios (%)

Misinformation Error (%)	Positive Report First						Negative Report First						Total (300)	*p*
	Positive Report			Negative Report			Positive Report			Negative Report				
	10	30	50	10	30	50	10	30	50	10	30	50		
Support Positive Report	7	9	12	6	6	8	15	16	19	14	19	19	150	1.000
Support Negative Report	18	16	13	19	19	17	10	9	6	11	6	6	150	

Table 5: Comparison of Gemini' s Responses Across Different Presentation Orders（*M*±*SD*）

Item	Mean	*t*	*p*	Cohen's *d*
Positive Report First	1.68±.468	6.661	<.001	.46804
Negative Report First	1.32± .468			

Table 6 Simple Effects Analysis of Gemini's Attitude Tendency Across Different Presentation Orders and Information Proportions（*M* ± *SD*）

Misinformation proportion	Presentation order		Type III Sum of Squares	Degrees of Freedom	*F*	*p*
	Positive Report First	Negative Report First				
Positive report (10% errors)	1.72±.46	1.40±.50	1.480	5	1.349	.244
Positive report (30% errors)	1.64±.49	1.36±.49				
Positive report (50% errors)	1.52±.51	1.24±.44				
Negative report (10% errors)	1.76±.44	1.44±.51				
Negative report (30% errors)	1.76±.44	1.24±.44				
Negative report (50% errors)	1.68±.48	1.24±.44				
Type III Sum of Squares	9.720		.600			
Degrees of Freedom	1			5		
F	44.294				.547	
P	<.001					.741

overall confirmation bias in terms of its preference for positive versus negative content, its decision-making is significantly influenced by the sequence of information, demonstrating a strong recency effect ($p < .001$).

2.2.2 Impact of Information Proportion

To further examine the impact of different presentation orders and misinformation proportions on Gemini' s attitude tendency in confirmation bias scenarios, a two-way ANOVA was conducted (shown in Table 6). The ANOVA revealed a significant main effect of presentation order on Gemini' s attitude ($F(1, 594) = 44.294$, $p < .001$), suggesting that Gemini' s attitude towards RoboTaxi was influenced by the order in which the positive and negative reports were presented. The main effect of misinformation proportion was not significant ($F(5, 594) = 1.349$, $p > .05$), indicating that the overall level of misinformation in the reports did not have a strong impact on Gemini's attitude.

Importantly, the interaction effect between presentation order and misinformation proportion was not significant ($F(5, 594) = .547$, $p > .05$). This suggests that the influence of misinformation proportion on Gemini' s attitude tendency is not dependent on the order in which the reports are presented. In other words, the effect of misinformation (or lack thereof) appears to be relatively consistent regardless of whether the positive or negative report is presented first.

Simple effects analyses were conducted to examine the effect of misinformation proportion separately for each presentation order. These analyses confirmed that the proportion of errors in either the positive or negative report did not significantly influence Gemini's attitude towards RoboTaxi, regardless of whether the positive or negative report was presented first (*ps* > .05).

3 Results of DeepSeek

3.1 Experiment 1: Framing Effects

3.1.1 Impact of Information Proportion

Table 7 presents DeepSeek' s response patterns across different framing effect scenarios, which involved manipulating both the proportion and the order of positive and negative information concerning gene testing. Table 8 displays results from a one-way ANOVA that compared DeepSeek' s overall decision-making tendency—measured on a scale where 1 = "inclined not to perform," 2 = "cautious consideration," and 3 = "inclined to perform"—across these varying proportions of positive and negative information. The ANOVA revealed a

Table 7 Response Patterns of DeepSeek in Different Framing Effect Scenarios (%)

	Positive First	Negative First	Positive(20%)
			Negative(80%)
				Positive(50%)
				Negative(50%)
					Positive(80%)
					Negative(20%)

Inclined to perform	57.3	54	13	57	97
Inclined not to perform	26	30	38	43	3
Cautious consideration	16.7	16	49	0	0

Table 8 Comparison of DeepSeek' s Responses Across Different Information Proportions ($\bar{X}$ ±*SD*)

Item	Mean	*F*	*p*
Positive (20%) & Negative (80%)	2.36±.70	180.870	<.001
Positive (50%) & Negative (50%)	1.43±.50		
Positive (80%) & Negative (20%)	1.03±.17		
Total	1.61±.75		

significant main effect for the proportion of information ($F = 180.870, p < .001$).

As detailed in Table 7, when presented with a predominantly positive information frame (80% positive), DeepSeek overwhelmingly favored "performing" gene testing (97%). Conversely, when faced with a predominantly negative information frame (80% negative), DeepSeek' s responses shifted: 49% indicated an inclination towards "not performing" gene testing, while a substantial 38% expressed "cautious consideration."

These findings suggest that DeepSeek' s decision-making in the context of gene testing is susceptible to framing effects. Specifically, the proportion of positive versus negative information significantly influences its inclination to recommend or proceed with the test.

3.1.2 Comparison of Different Order Presentations in Framing Effect Scenarios

Table 9 provides the results of an independent samples t-test that examined DeepSeek' s overall decision-making tendency when positive and negative information were presented in two different orders. The analysis reveals that the presentation order did not have a statistically significant effect on DeepSeek' s decision-making tendency (t = - .3306, $p > .05$). This suggests that while DeepSeek' s responses are sensitive to the

Table 9 Comparison of DeepSeek' s Responses Across Different Presentation Orders ($\bar{X}$ ±*SD*)

Item	Mean	*t*	*p*	Cohen's *d*
Positive First	1.59±.76	-.3306	.677	.75426
Negative First	1.62±.75			

Table 10 Response Patterns of DeepSeek in Different Confirmation Bias Scenarios (%)

Misinformation Error (%)	Positive Report First						Negative Report First						Total (%)	*p*
	Positive Report			Negative Report			Positive Report			Negative Report				
	10	30	50	10	30	50	10	30	50	10	30	50		
Support Positive Report	18	19	15	13	21	15	21	22	22	22	17	20	75%	<.001
Support Negative Report	7	6	10	12	4	10	4	3	3	3	8	5	25%	

proportion of positive versus negative information (as detailed in Table 8), the sequence in which this information is presented does not appear to substantially alter its overall inclination regarding gene testing. This finding is consistent with the results observed in Gemini' s experiment.

3.2 Experiment 2: Confirmation Bias

3.2.1 Comparison of Different Order Presentations

Table 10 presents DeepSeek' s response patterns in various confirmation bias scenarios concerning autonomous vehicles. Additionally, a binomial test was conducted (details in Table 10), which showed that DeepSeek supported positive reports in 75% of all its responses ($p < .001$). This finding indicates that DeepSeek exhibits confirmation bias and is distinctly more inclined to support positive reports.

The results concerning presentation order (Table 11, $p < .05$) revealed that DeepSeek' s frequency of supporting negative reports was actually higher when positive reports were presented first, compared to when negative reports were presented first.

Therefore, DeepSeek demonstrates confirmation bias, evidenced by its overall tendency to favor positive reports (75% support, $p < .001$, as detailed in Table 10). Furthermore, its decision-making is significantly susceptible to the presentation order of information (p=.002, based on Table 11), with the order influencing outcomes in the specific, nuanced manner described above.

3.2.2 Impact of Information Proportion

To further investigate how different presentation orders and proportions of misinformation affect DeepSeek' s attitude in confirmation bias scenarios, a two-way ANOVA was performed (results detailed in Table 12).

The analysis revealed a significant main effect for presentation order on DeepSeek' s attitude (F(1, 594) =

Table 11 Comparison of DeepSeek' s Responses Across Different Presentation Orders（$M \pm SD$）

Item	Mean	t	p	Cohen’s d
Positive Report First	1.32±.471	3.105	.002	.4276
Negative Report First	1.17±.380			

Table 12 Simple Effects Analysis of DeepSeek' s Attitude Tendency Across Different Presentation Orders and Information Proportions（$M \pm SD$）

Misinformation proportion	Presentation order		Type III Sum of Squares	Degrees of Freedom	F	p
	Positive Report First	Negative Report First				
Positive report (10% errors)	1.28±.46	1.16±.37	0.550	5	.610	.692
Positive report (30% errors)	1.24±.44	1.12±.33				
Positive report (50% errors)	1.40±.50	1.12±.33				
Negative report (10% errors)	1.48±.51	1.12±.33				
Negative report (30% errors)	1.16±.37	1.32±.48				
Negative report (50% errors)	1.4±.50	1.17±.38				
Type III Sum of Squares	1.763		2.017			
Degrees of Freedom	1			5		
F	9.781				2.237	
P	.002					.051

9.781, p < .05). This indicates that DeepSeek' s stance towards RoboTaxi was indeed influenced by the sequence in which positive and negative reports were presented. In contrast, the main effect of misinformation proportion was not significant (F(5, 594) = .610, p > .05), suggesting that the overall level of misinformation in the reports did not substantially impact DeepSeek' s attitude.

Notably, the interaction effect between presentation order and misinformation proportion did not reach statistical significance (F(5, 594) = 2.237, p > .05). This suggests that the influence of misinformation proportion on DeepSeek' s attitude did not significantly differ based on the order in which the reports were presented.

4 Discussion

4.1 LLMs' s Susceptibility to Cognitive Biases

4.1.1 LLM' s Susceptibility to Framing Effect

Framing effects lead people to make different choices based on how a problem is described, even when the underlying options are identical. Variations in presentation can sway decisions. This study found that, in the context of gene testing, Gemini and DeepSeek are both more likely to recommend testing when presented with predominantly positive information. Conversely, they both exhibit greater caution when presented with more negative information (50% and 80%), opting against testing more frequently when the negative information proportion is higher (80%). Interestingly, presentation order did not influence Gemini and DeepSeek' s decisions in these framing effect scenarios. These results demonstrate that Gemini and Deepseek both exhibit framing effects, with their decisions correlating positively with the proportion of positive information, regardless of presentation order.

4.1.2 LLMs' s Susceptibility to Confirmation Biases

Confirmation bias, a common human cognition bias, leads people to favor information confirming their existing beliefs. They often dismiss or ignore contradictory information (Moravec et al., 2019).

In the experiment, the ANOVA results of Gemini showed that the proportion of misinformation did not have a statistically significant main effect on Gemini' s attitude tendency. Although descriptive statistics suggest a slight decrease in Gemini' s support for positive reports as the error rate in those reports increased (e.g., mean decreased from 1.72 at 10% error to 1.52 at 50% error, see Table 6), these changes did not reach statistical significance. Therefore, we cannot conclude that the proportion of misinformation directly influenced Gemini's decisions. Future research with larger sample sizes or more sensitive experimental designs could further investigate this issue.

This phenomenon suggests that even without explicit pre-existing "beliefs," the model may exhibit behaviors similar to confirmation bias when processing sequential information, i.e., giving greater weight to specific (e.g., later presented) information, which is similar to the mechanism of the recency effect. Future research could further explore how to distinguish between recency effects and confirmation bias based on complex internal representations in large language models.

However, the binomial test in DeepSeek' s confirmation bias experiment indicated (see Table 10) that DeepSeek exhibited confirmation bias during the experiment. Through ANOVA analysis, we found that DeepSeek' s decision-making tendency was related to the presentation order of the reports, which is consistent with the experimental findings for Gemini.

Furthermore, Gemini and DeepSeek both demonstrate some ability to identify misinformation and maintain skepticism. It can present both sides of an argument, and even after forming an opinion, express reservations, suggesting the need for "further observation and understanding of the real situation."

4.2 Potential Reasons for LLMs' s Susceptibility to Cognitive Biases

4.2.1 The Potential Influences of Questionnaire Information

First, this study found that AI models may can learn from past conversations, potentially leading to responses consistent with prior exchanges and thus, converging answers.

Second, AI models may weigh different benefits and risks differently within the context of framing effects. Consequently, the presence of certain high-weighted factors can influence its decisions. For example, regarding the risks of genetic testing, they might prioritize"privacy breaches"over "errors in testing results," leading to greater caution when presented with privacy breach concerns. Future research could explore this by presenting large language models with individual risks or benefits as framing elements and observing their decision tendencies in various framing contexts.

Furthermore, in framing effect scenarios, AI models are likely already possessed information about gene testing. Therefore, their analysis and reasoning may not rely solely on the information provided by researchers but also integrate its existing knowledge and experience. Future research could investigate this by using knowledge probing techniques, designing questions to assess the model' s understanding of specific information and comparing its responses to its internal knowledge base.

Additionally, the length of the presented information may also influence AI model' s responses. In the framing effect scenarios, the gene testing information was considerably shorter than the news reports in the confirmation bias scenarios. This length discrepancy could contribute to the significant impact of presentation order observed in the confirmation bias scenarios. Future research could control for information length across different scenarios to examine whether it affects AI model' s interpretation and susceptibility to framing effects and confirmation bias.

4.2.2 Attention Mechanisms and Confirmation Bias: A Potential Explanation for Order Effects

This study found that in confirmation bias scenarios, Gemini and DeepSeek were more sensitive to the order of information presented than to the proportion of errors within that information. This heightened sensitivity to order may stem from the training mechanisms of large language models, particularly the Transformer model' s attention mechanism.

The Transformer model' s mechanism (赵立等 , 2024; DeRose et al., 2021; Yeh et al., 2024) allows the model to differentially weight words in sequential data based on their importance. This enables the model to prioritize key words during processing and reference them more heavily during output generation, improving not only data processing capabilities but also stability and training efficiency. However, this same mechanism could also increase the model's susceptibility to anchoring effects from prior information.

Anchoring effect (Liu et al., 2021; Turner & Schley, 2016) is a cognitive bias where initial information (the "anchor"), even if inaccurate or incomplete, disproportionately influences subsequent judgments and decisions. Specifically, if a model encounters errors early in a sequence, the attention mechanism may overemphasize these errors, biasing its understanding of subsequent information. This could explain Gemini's observed sensitivity to presentation order in this study.

Conversely, variations in error rates appeared to have less impact on AI model' s decisions. This resilience to errors likely stems from the model' s extensive training data, which allows it to filter noise and make reasonably accurate judgments even when presented with flawed information.

However, these findings are specific to Gemini. Further research is needed to determine whether other large language models exhibit similar behavior. Moreover, the limited sample size and scenario design in this study necessitate further validation with broader scenarios and larger samples.

Future research should investigate the relationship between the Transformer model' s attention mechanism and confirmation bias more deeply. This could involve analyzing how different attention mechanism designs influence susceptibility to confirmation bias and developing training methods to mitigate this sensitivity.

4.3 Ethical and Societal Implications of Cognitive Biases in LLMs

4.3.1 Amplification of Biases and Unfair Decision-Making

Framing effects and confirmation bias within LLMs can amplify existing biases in their training data, potentially leading to algorithmic discrimination and exacerbating social inequalities (Bengio et al., 2024; Buslón et al., 2023; O' Connor, 2024). For example, if training data contains stereotypes about certain groups, the model may inadvertently perpetuate these stereotypes in generated text, resulting in algorithmic discrimination (Kostick-Quenet et al., 2022).

This phenomenon is well documented. Obermeyer et al. (2019) found racial bias in algorithms used to predict patients complex health need in patients. Similarly, racial bias has been identified in several clinical algorithms, including pulmonary function tests that incorporate "racial correction," leading to the underdiagnosis of lung disease prevalence and severity in Black patients (Moffett et al., 2023). Such biases perpetuate social prejudices and disproportionately harm marginalized communities (O' Connor, 2024).

In decision-making tasks like hiring, loan applications, and treatment planning, cognitive biases in models can lead to unfair outcomes. For example, a model influenced by racial biases in its training data unfairly rejects loan applications from minority groups, even when applicants meet the required criteria.

4.3.2 Personalized Recommendations and the Filter Bubble Effect

Confirmation bias in large language models can reinforce users' existing viewpoints by preferentially recommending aligning information and filtering out dissenting perspectives, thereby exacerbating the filter bubble effect (Gu et al., 2024). This narrowed information access hinders exposure to diverse viewpoints and impedes the development of independent thinking and judgment. For example, a user frequently browsing websites espousing a particular political view might receive further recommendations for similar content, while alternative perspectives are neglected. This can trap users in a filter bubble, limiting their understanding and fostering one-sided perspectives.

This phenomenon can intensify social fragmentation and polarization (张瀚予等, 2024; Gu et al., 2024; Piao et al., 2023), hindering social consensus and stability. For instance, on social media platforms, if AI recommendation algorithms primarily promote information confirming users' existing beliefs, individuals with different viewpoints may become isolated within their respective information bubbles, preventing meaningful communication and exchange and exacerbating social division and conflict.

4.3.3 Group Polarization

The filter bubble effect can exacerbate divergent viewpoints between different groups, potentially pushing them towards extremism and hindering societal consensus and stability. Research indicates that exclusive exposure to confirming information can radicalize perspectives and foster hostility towards opposing viewpoints (张瀚予等, 2024; Gu et al., 2024; Piao et al., 2023). Driven by personal preferences, filter bubbles contributes to societal polarization.

In AI recommendation systems, algorithms solely recommend information aligning with users' viewpoints can amplify this group polarization, intensifying social fragmentation and conflict. For example, during political elections, if AI systems only recommend news and information supporting a specific party, this can bolster support for that party while diminishing support for others, potentially influencing the election outcome.

4.3.4 The Spread of Misinformation

Confirmation bias in large language models can increase their susceptibility to misinformation. Models may readily accept information aligning with their pre-existing knowledge, even if false or misleading, contributing to the rapid spread of misinformation online and misleading users (Cui et al., 2024; Zhu et al., 2018). If a model has learned erroneous information, it may retain this belief even when presented with accurate information, subsequently disseminating the misinformation and potentially exacerbating social chaos and negatively impacting people's decisions and actions.

4.4.4 Considerations and Future Directions for Research

When interpreting the results of this study, the dynamic and rapidly evolving nature of LLMs must be fully acknowledged. LLMs are updated and iterated at an unprecedented pace. This implies that findings obtained from a specific model version, such as Gemini 1.5 Pro and DeepSeek used in this research, may have a relatively limited 'longevity' as future iterations might exhibit different patterns of cognitive biases or may have even mitigated certain known biases through targeted training. However, this does not diminish the value of such research. Systematically investigating cognitive biases in advanced models at their current stage of development can reveal prevalent cognitive patterns and potential risks associated with this technological phase. This provides empirical evidence for model developers to continuously optimize algorithms and enhance model robustness, while also offering crucial insights for the formulation of AI ethics guidelines and the selection of responsible AI application scenarios.

Regarding the study subject, Gemini 1.5 Pro and DeepSeek were selected as the test model primarily because, at the time of research, it was an accessible and widely recognized leading LLMs, indicative of the general performance of then state-of-the-art (SOTA) models on cognitive tasks. However, we must emphasize that the extent to which Gemini 1.5 Pro and DeepSeek's performance can be generalized to all LLMs, let alone to the broader spectrum of artificial intelligence systems, requires careful consideration. Different LLMs vary in architecture, training data, parameter scale, and training strategies, all of which could lead to variations in their cognitive characteristics and bias manifestations. Therefore, the conclusions of this study should be viewed as an initial exploration of cognitive bias phenomena in a specific advanced model, rather than as universally applicable statements about all AI. Future research should strive for broader multi-model comparisons, incorporating models from different technological paths and origins, to hopefully derive more generalizable conclusions.

Furthermore, this study adapted established research paradigms from human cognitive psychology to investigate cognitive biases in AI. The rationale for this interdisciplinary approach lies in the fact that one of the design goals of LLMs is to simulate and understand human language and cognition; thus, theoretical frameworks of human cognitive biases provide valuable perspectives and tools for examining potential flaws in AI decision-making. However, directly applying paradigms derived from human research to AI also has inherent limitations. AI' s information processing mechanisms, learning methods (based on training with vast amounts of data), and lack of subjective experience and biological underpinnings are fundamentally different from human cognition. This implies that the manifestation, underlying mechanisms, and even the very existence of certain biases observed in humans might differ in AI. Therefore, future research should not only cautiously apply existing paradigms but also actively explore and develop new research methodologies and theoretical frameworks specifically tailored to AI' s cognitive characteristics to more accurately understand and evaluate its complex cognitive behaviors.

5 Conclusion

The findings indicate that both Gemini and DeepSeek are susceptible to framing effects, which appear to be linked to the proportion of positive and negative misinformation they process. However, the models differ in their susceptibility to confirmation bias. Gemini does not exhibit confirmation bias; nonetheless, the order in which positive and negative reports are presented does influence its decision-making. In contrast, DeepSeek is affected by confirmation bias, and this bias is also influenced by the presentation order of positive and negative reports.

References

侯悍超，倪士光，林书亚，王蒲生. (2024). 当AI学习共情：心理学视角下共情计算的主题、场景与优化. *心理科学进展*, *32*(5), 845–858.

刘冬予，骆方，屠焯然，饶思敬，沈阳. (2024). 人工智能技术赋能心理学发展的现状与挑战. *北京师范大学学报(自然科学版)*, *60*(1), 30–37.

张瀚予，丁怡宁，郭思琪. (2024). 信息茧房效应下用户群体极化形成机理研究. *图书与情报*, *3*, 132–144.

赵立，赵宏坚，高智伟，王黎明，刘越，罗渝，廖勇. (2024). 认知传感网中基于Transformer网络的MAC协议识别方法. *电讯技术*, *64*(10), 1–8.

Bengio, Y., Hinton, G., Yao, A., Song, D., Abbeel, P., Darrell, T., Harari, Y. N., Zhang, Y. Q., Xue, L., Shalev-Shwartz, S., Hadfield, G., Clune, J., Maharaj, T., Hutter, F., Baydin, A. G., McIlraith, S., Gao, Q., Acharya, A., Krueger, D., & Dragan, A. (2024). Managing extreme AI risks amid rapid progress. *Science, 384*(6698), 842–845.

Berthet, V., Teovanovic, P., & de Gardelle, V. (2024). A common factor underlying individual differences in confirmation bias. *Scientific Reports, 14*(1), PP 27795.

Binz, M., & Schulz, E. (2023). Using cognitive psychology to understand GPT-3. *Proceedings of the National Academy of Sciences, 120*(6), e2218523120.

Botvinick, M. M. (2022). Realizing the promise of AI: A new calling for cognitive science. *Trends in Cognitive Sciences, 26*(12), 1013–1014.

Buslón, N., Cortés, A., Catuara-Solarz, S., Cirillo, D., & Rementería, M. J. (2023). Raising awareness of sex and gender bias in artificial intelligence and health. *Frontiers in Global Women' s Health, 4,* 970312.

Cui, W., Wang, D., & Han, N. (2024). Survey on fake information generation, dissemination and detection. *Chinese Journal of Electronics, 33*(3), 573–583.

DeRose, J. F., Wang, J., & Berger, M. (2021). Attention flows: Analyzing and comparing attention mechanisms in language models. *IEEE Transactions on Visualization and Computer Graphics, 27*(2), 1160–1170.

Gandhi, T. K., Classen, D. C., Sinsky, C. A., Rhew, D. C., Garde, N. V., Roberts, A., & Federico, F. (2023). How can artificial intelligence decrease cognitive and work burden for front line practitioners? *JAMIA Open, 6*(3), ooad079.

Gu, M., Zhao, T. F., Yang, L., Wu, X. K., & Chen, W. N. (2024). Modeling information cocoons in networked populations: Insights from backgrounds and preferences. *IEEE Transactions on Computational Social Systems, 11*(3), 4497–4510.

Kostick-Quenet, K., Cohen, I. G., Gerke, S., Lo, B., Antaki, J., Movahedi, F., Njah, H., Schoen, L., Estep, J. E., & Blumenthal-Barby, J. S. (2022). Mitigating racial bias in machine learning. *Journal of Law, Medicine and Ethics, 50*(1), 92–100.

Lee, D., & Yoon, S. N. (2021). Application of artificial intelligence-based technologies in the healthcare industry: Opportunities and challenges. *International Journal of Environmental Research and Public Health, 18*(1), 271. mdpi.

Liu, M., Zeng, J., & Gao, Z. (2021). The interval anchoring effect. *Experimental Psychology, 68*(6), 295–304.

Michel, M., & Peters, M. A. K. (2020). Confirmation bias without rhyme or reason. *Synthese, 199*(1–2), 2757–2772.

Moffett, A. T., Bowerman, C., Stanojevic, S., Eneanya, N. D., Halpern, S. D., & Weissman, G. E. (2023). Global, race-neutral reference equations and pulmonary function test interpretation. *JAMA Network Open, 6*(6), e2316174.

Moravec, P. L., Minas, R. K., & Dennis, A. R. (2019). Fake news on social media: People believe what they want to believe when it makes no sense at all. *MIS Quarterly, 43*(4), 1343.

Mukhamediev, R. L., Popova, Y., Kuchin, Y., Zaitseva, E., Kalimoldayev, A., Symagulov, A., Levashenko, V., Abdoldina, F., Gopejenko, V., Yakunin, K., Muhamedijeva, E., & Yelis, M. (2022). Review of artificial intelligence and machine learning technologies: Classification, restrictions, opportunities and challenges. *Mathematics, 10*(15), 2552.

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science, 366*(6464), 447–453.

O' Connor, M. I. (2024). Equity360: Gender, race, and ethnicity—The power of AI to improve or worsen health disparities. *Clinical Orthopaedics and Related Research, 482*(4), 591–594.

Piao, J., Liu, J., Zhang, F., Su, J., & Li, Y. (2023). Human-AI adaptive dynamics drives the emergence of information cocoons. *Nature Machine Intelligence, 5*(11), 1214–1224.

Sergio, D. S., Rashmi, G., & Dario, M. (2023). Editorial: Highlights in psychology: Cognitive bias. *Frontiers in Psychology, 14,* 1242809.

Turner, B. M., & Schley, D. R. (2016). The anchor integration model: A descriptive model of anchoring effects. *Cognitive Psychology, 90,* 1–47.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*(4157), 1124–1131.

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science, 211*(4481), 453–458.

Yeh, C., Chen, Y., Wu, A., Chen, C., Vi é gas, F., & Wattenberg, M. (2024). AttentionViz: A global view of transformer attention. *IEEE Transactions on visualization and computer graphics, 30*(1), 262–272.

Zhao, Y., Huang, Z., Seligman, M., & Peng, K. (2024). Risk and prosocial behavioural cues elicit human-like response patterns from AI chatbots. *Scientific Reports, 14*(1), 7095.

Zhu, H., Wu, H., Cao, J., Fu, G., & Li, H. (2018). Information dissemination model for social media with constant updates. *Physica A-Statistical Mechanics and Its Applications, 502,* 469–482.

人工智能的认知偏差：大语言模型对框架效应与确认偏误的易感性*

李　好[1]　王　优[1,2]　杨雪岭[**1,2]

（[1]南方医科大学公共卫生学院心理学系，广州，510515）

（[2]南方医科大学珠江医院精神心理科，广州，510280）

摘　要　随着人工智能（AI）技术和大语言模型（LLMs）的飞速发展，其在文本生成、翻译、问答等诸多领域展现出惊人的能力。LLMs不仅提升了人类的能力，也在心理学领域革新了研究方法、技术和范式，促进了对人类认知模式的探索，推动了该领域的进步。然而，一个亟待解决的关键问题浮出水面：这些旨在模仿甚至超越人类智能的模型，是否也像人类一样容易受到认知偏差的影响？理解人工智能模型是否存在类似的认知偏差，对于评估其可靠性、改进其性能以及预测其潜在的社会影响至关重要。认知偏差是源于启发式和信息处理局限性而导致系统性偏离理性判断的现象，它们可以扭曲我们对现实的感知，从而导致次优决策。如果这些偏差也出现在AI系统中，将对其可靠性和安全性产生深远影响。

研究旨在深入探究谷歌Gemini 1.5 Pro和DeepSeek这两款大语言模型对框架效应和确认偏误的易感性。框架效应是指人们对相同信息的不同表述方式做出不同反应的现象，而确认偏误则考察模型在处理信息时是否存在系统性偏好。对于LLMs等AI系统而言，这种偏见可能源于训练数据的偏差或算法的优先排序，从而可能导致预测结果的偏差，加剧社会偏见，并损害信息公平性和多样性。为评估模型是否存在这两类认知偏差，研究系统地操控了信息的比例和呈现顺序，揭示LLMs是否继承了人类的认知脆弱性，并探讨其潜在的伦理和社会影响。

在框架效应实验中，研究构建了基因检测的决策场景。实验通过控制积极信息和消极信息的比例（例如20%积极/80%消极；50%积极/50%消极；80%积极/20%消极），并变换信息呈现的先后顺序，来记录模型对是否进行基因检测的倾向。这种设计旨在评估信息呈现方式（即“框架”）的系统性变化是否会引发Gemini表达倾向的相应转变。每次试验都在一个新的、独立的交互会话中进行，以确保没有来自先前试验的结转效应或上下文学习 。

在确认偏误实验中，研究提供了关于“萝卜快跑”自动驾驶汽车的积极性和消极性两篇报道。实验系统改变了报道中错误信息的比例（10%，30%和50%），并同样测试了不同信息呈现顺序下模型对报道的支持倾向。本研究通过强制选择任务中模型的系统性偏好对确认偏误进行操作性定义，通过统计学显著的更高选择率来量化这种偏误。与框架效应场景一样，确认偏误实验的每次试验都在一个新的、独立的交互会话中进行，以防止先前交互的上下文影响。

研究结果表明，Gemini 1.5 Pro和DeepSeek均表现出对框架效应的易感性。具体而言，在基因检测场景下，两者的决策态度主要受到所呈现积极信息和消极信息比例的影响。当积极信息占比较高时，模型更倾向于选择进行基因检测。反之，当消极信息占比较高时，则更倾向于不进行或持谨慎态度。而信息呈现的先后顺序对框架效应的实验结果没有显著影响。这些结果表明，Gemini和DeepSeek都表现出框架效应，其决策与积极信息的比例呈正相关。

在确认偏误的实验中，Gemini 1.5 Pro并未表现出对积极或消极报道的整体偏好。在300个响应中，Gemini支持积极报道150次，支持消极报道150次，这种平衡的结果表明Gemini可能不具备始终偏爱积极或消极信息的内在确认偏误。然而，其判断更多地受到信息呈现顺序的显著影响，表现出“近因效应”，即更倾向于支持后呈现的报道。因此，尽管Gemini可能没有表现出对其偏好积极或消极内容的整体确认偏误，但其决策受到信息顺序的显著影响，而错误信息的比例对其影响不显著。简单效应分析也证实，无论积极或消极报道先呈现，错误信息的比例变化对Gemini对RoboTaxi的态度均无显著影响（所有ps>.05）。

DeepSeek在确认偏误实验中则表现出对正面报道的整体偏好，其支持正面报道的比例显著更高。二项检验显示，DeepSeek在其所有响应中，有75%支持积极报道 ($p<.001$)。这一发现表明DeepSeek表现出确认偏误，并且明显更倾向于支持积极报道。尽管如此，DeepSeek的决策同样也受到了信息呈现顺序的显著影响。即当先呈现积极报道后呈现消极报道时，DeepSeek支持消极报道的频率会更高，错误信息比例则无显著影响 ($F(5,594)=.610, p>.05$)。这表明报道中错误信息的总体水平并未实质性影响DeepSeek的态度。此外，Gemini和DeepSeek都表现出一定识别错误信息并保持怀疑的能力。它们可以呈现论证的两面性，甚至在形成意见后，表达保留意见，暗示需要“进一步观察和理解实际情况” 。

这些发现揭示了先进大语言模型中存在类似人类的认知脆弱性，对人工智能在决策过程中的可靠性和客观性提出了严峻挑战。这也提示在开发和应用人工智能时，需更加审慎地评估其潜在的认知偏误，并采取有效措施以避免可能带来的负面社会影响。本研究作为对特定前沿模型（Gemini 1.5 Pro, DeepSeek）在特定发展阶段的探索，其结论在推广至所有LLMs及模型的未来版本时需持谨慎态度。同时，将人类认知范式应用于机制根本不同的AI，虽具启发性，也凸显了开发针对AI认知特性的新研究框架的必要性。

关键词　人工智能　大语言模型　认知偏差　确认偏误　框架效应

* 本研究得到广东省本科高校教学质量与教学改革工程建设项目（粤教高函〔2024〕30号）和教育部人文社会科学研究项目（22YJCZH182）的赞助

** 通讯作者：杨雪岭，E-mail: yhtyxl2006@126.com