ChinaRxiv

AI translation · View original & related papers at
chinarxiv.org/items/chinaxiv-202509.00060

Large Language Models Distinguishing One-Shot from Repeated Games: Understanding and Intervening in Risky Decision-Making

Authors: Zhou Lei, Li Litong, Wang Xu, Qu Huafeng, Hu Qianyu, Li Aimei, Gu Chenyan, Li Aimei, Gu Chenyan

Date: 2025-12-07T17:15:15+00:00

Abstract

Theoretical research on risky decision-making has primarily relied on reverse inference from behavioral outcomes and self-report data, lacking direct observation of the decision-making process, which constrains explanations of its underlying mechanisms and the development of effective behavioral intervention programs. The application of large language models (LLMs) offers a pathway to overcome these limitations. This paper systematically examines the simulation potential of LLMs in risky decision-making through three studies, conducting single-shot and repeated games based on DeepSeek-R1 to generate decision rationales, employing GPT-4o to perform inductive thematic analysis (ITA) on these outputs, constructing a technical pathway for LLMs to generate decision strategy texts, and applying it to decision interventions. Findings: (1) ChatGPT-3.5/4 can reproduce typical human choice patterns in single-shot (more risk-averse) and repeated (more risk-seeking) games; (2) LLMs can distinguish between single-shot/repeated game logic and correctly apply normative and descriptive theories respectively to generate corresponding strategies, which are highly recognized; (3) Intervention texts generated by LLMs based on different strategies can effectively influence people's inherent risk decision preferences in medical, financial, content creation, and e-commerce marketing contexts. This research systematically validates LLMs' capability to simulate behavioral preferences and understand decision-making, constructs a new paradigm for decision interventions based on generative AI, and provides theoretical and practical foundations for AI-assisted high-stakes decision-making.

Full Text

Preamble

Large Language Models Capable of Distinguishing Between Single and Repeated Gambles: Understanding and Intervening in Risky Decision-Making

Lei Zhou[1], Litong Li[1], Xu Wang[1], Huafeng Ou[1], Qianyu Hu[1], Aimei Li[2], Chenyang Gu[1]

[1]School of Management, Jinan University, Guangzhou 510632, China

(Special Issue on Psychology and Artificial Intelligence Development)

Abstract
Risky decision-making is a central topic in behavioral research, crucial for human survival and development. Its theoretical development primarily follows two major paths: normative theory and descriptive theory (Benartzi & Thaler, 1999), with single- versus repeated-play gamble paradigms providing an effective method to distinguish between these theories. However, existing theoretical research relies mainly on retrospective inference from behavioral outcomes and self-report data, lacking direct observation of decision-making processes, which constrains explanations of underlying mechanisms and the development of effective behavioral interventions. The application of artificial intelligence large language models (LLMs) offers a pathway to overcome these limitations. This paper systematically examines the simulation potential of LLMs in risky decision-making through three studies, based on DeepSeek-R1 for single and repeated gambles to generate decision rationales, and employs GPT-4o for inductive thematic analysis (ITA) to construct a technical path for generating LLM decision strategy texts, which is then applied to decision intervention.

Findings: (1) ChatGPT-3.5/4 can reproduce the typical human choice patterns of being more risk-averse in single-play and more risk-seeking in repeated-play gambles; (2) LLMs can distinguish between single- and repeated-play gamble logics and correctly apply normative and descriptive theories to generate corresponding strategies, which receive high human acceptance; (3) Intervention texts generated by LLMs based on different strategies can effectively influence people's inherent risk decision preferences in medical, financial, content creation, and e-commerce marketing contexts. The study systematically validates LLMs' ability to simulate behavioral preferences and understand decision-making, constructing a new paradigm for generative AI-based decision intervention, providing theoretical and practical foundations for AI-assisted high-stakes decision-making.

Keywords
risky decision-making, single/repeated gambles, large language models, decision strategy, intervention

Classification Code
B849:C91

Received Date: 2025-05-12

This research was supported by the National Natural Science Foundation of China (General Program: 72271066, 72571114; Young Scientists Fund: 72102085), and the Guangzhou Science and Technology Plan Basic and Applied Basic Research Special Project (2025A04J5384).

Corresponding Authors:
Aimei Li, Email: tliaim@jnu.edu.cn;
Chenyang Gu, Email: g_{cy1989163}@163.com

Risky decision-making is a core issue in behavioral research, vital for human survival and development. Its theoretical development mainly follows two major paths: normative theory and descriptive theory (Benartzi & Thaler, 1999), with single- and repeated-play gamble paradigms providing an effective method to distinguish these theories. However, existing theoretical research mostly relies on retrospective inference from behavioral outcomes and self-report data, lacking systematic direct observation of decision-making processes, which leads to insufficient explanatory power regarding underlying mechanisms and constrains the development of effective behavioral interventions. In recent years, artificial intelligence (AI) development has offered new perspectives for risky decision-making research, particularly through AI simulation of human decision preferences and externalization of abstract decision strategies.

But this raises questions: Are AI-generated decision strategies acceptable to humans? Can these strategies intervene in human decisions? Answering these questions not only reveals AI's explanatory power regarding human decision preferences, providing new evidence for decision model construction, but also offers theoretical foundations for AI-based decision nudging. This paper conducts three studies based on large language models (LLMs) to systematically examine LLMs' ability to simulate risky decision preferences and process mechanisms, explore their understanding of human decision strategies, construct methods for using LLMs to generate intervention texts targeting single- and repeated-play risk preferences, and verify their intervention effectiveness through experiments.

1.1 Behavioral Differences in Risky Decision-Making Between Single and Repeated Gambles

Risky decision-making refers to decisions where options have multiple outcomes with known probabilities (Kahneman & Tversky, 1979). Its theoretical development can be divided into normative and descriptive paths: normative models mostly assume mathematical expectation, following "weighted summation" expected value principles (Von Neumann & Morgenstern, 1947), while descriptive models adopt psychological perspectives, focusing on biases and heuristics in actual decision-making behavior (Hertwig & Erev, 2009).

Real-world risky decisions can be categorized as single- or repeated-play scenarios. For example, entrepreneurs may encounter one or multiple similar in-

vestment opportunities within a period, and doctors may treat one or multiple patients with similar conditions (Redelmeier & Tversky, 1992). Research shows people often exhibit preference reversals between single and repeated gambles (Anderson et al., 2023). For instance, Sun et al. (2014) designed a financial emergency management task where participants chose between two emergency plans: Plan A, which would certainly save ¥5,000; and Plan B, which had a 0.1% probability of saving ¥6,000,000 and a 99.9% probability of saving ¥0. Participants made choices in both single-play (implementing the plan once) and repeated-play (implementing the plan 100 times) contexts. Results showed people preferred the certain plan A in single-play gambles but preferred the risky plan B in repeated-play gambles. This choice preference difference has been repeatedly validated across multiple studies (Lopes, 1996; Kumar & Lim, 2008; Barberis & Huang, 2009).

Single- and repeated-play choice preferences can be explained by descriptive and normative theories respectively. According to classic normative theory—Expected Utility Theory—in repeated gambles, individuals tend to calculate expected values by weighting outcomes by their probabilities and choose the option with the highest expected value (Von Neumann & Morgenstern, 1947). In single gambles, however, people's choices often violate expected utility theory assumptions and align more with descriptive theories. For example, Myopic Loss Aversion theory proposes that in single gambles, individuals' excessive sensitivity to loss makes them tend to avoid risk and choose certain options (Tversky & Bar-Hillel, 1983). Similarly, the equate-to-differentiate theory (Li, 2004) suggests that in decision-making, people equate dimensions with small differences (e.g., probability) and base decisions only on dimensions with large differences (e.g., outcome).

Traditional risky decision-making research has the following limitations: First, it mainly relies on retrospective inference from behavioral outcomes, lacking systematic direct observation of decision-making processes; second, discussions of decision strategies depend on self-report data and choice results, leading to insufficient explanatory power regarding underlying mechanisms. This not only limits the depth of theoretical explanations for real-world complex decision contexts (Kelton et al., 2010; Morreale et al., 2018) but also constrains the development of effective behavioral intervention programs.

1.2 Artificial Intelligence Assisting Decision Theory Research

AI development and application provide new perspectives and approaches for psychological research. As the most cutting-edge AI tool, LLMs have demonstrated "human-like" characteristics across multiple domains (Salles et al., 2020; Jiao et al., 2025). At the behavioral level, LLMs are often directly used as experimental subjects (Grossmann et al., 2023). For example, Qin et al. (2024) successfully replicated 22 social psychology experiments including leadership behavior and workplace interactions using ChatGPT with 93.2% accuracy; Argyle

et al. (2023) found GPT-3 can simulate different human subpopulations in social science research, generating responses highly similar to humans in form, semantics, and structure; Strachan et al. (2024) noted that LLMs' theory of mind capabilities are comparable to humans; at the process mechanism level, multiple studies have found LLMs' feature extraction pathways and hierarchical processing mechanisms converge with brain neural response patterns (Mischler et al., 2024), demonstrating similarities between LLMs' language processing mechanisms and the human brain.

In decision-making research, LLMs can simulate human decision preferences. For example, Zhao et al. (2024) found that ChatGPT-4 can adjust its risk-taking behavior under emotional arousal, showing lower risk preference especially when fear is induced; Aher et al. (2023) found ChatGPT exhibits human-like cooperative behavior in prisoner's dilemma, ultimatum game, and trust games; other studies have also found LLMs show highly anthropomorphic behavior in voting behavior, economic games, and classic social experiments (Dillion et al., 2023); Jones and Steinhardt (2022) revealed that LLMs are susceptible to irrelevant information and initial inputs when generating open-ended text, leading to outputs deviating from expectations and showing human-like framing and anchoring effects; Park et al. (2024) noted LLMs show similarity to humans in the "less-is-better effect" experiment, tending to choose lower-priced goods.

Furthermore, LLMs also demonstrate "human-like" qualities at the decision-making process level. For instance, research has found that when LLMs generate words, they engage in chain-of-thought reasoning by rereading context, similar to how humans use scratch paper to aid thinking or iteratively refine arguments when solving math problems or writing essays (Hagendorff et al., 2023). Additionally, as model complexity increases, LLMs can reduce reasoning biases, exhibiting human-like system 2 thinking (Wei et al., 2022). Further, Binz and Schulz (2023) revealed through two-step task experiments that GPT-3 exhibits human-like model-based reinforcement learning characteristics, adjusting its behavioral strategies based on environmental states and reward information to maximize rewards. Similarly, Webb et al. (2023) found through four types of analogical reasoning tasks that GPT-3 shows reasoning abilities comparable to or even stronger than humans, with certain similarities in analogical reasoning and abstract pattern recognition. Goli and Singh (2024) found that GPT considers multiple factors in intertemporal decision-making, particularly regarding risk and uncertainty, opportunity costs and investment choices, and urgency and immediate needs.

In summary, LLMs can simulate both human decision preferences and process mechanisms, thus helping reveal key steps in the cognitive "black box." Moreover, this capability of LLMs also helps transform abstract theories into concrete strategies, facilitating the visualization of decision logic. These advantages provide new pathways to overcome limitations of traditional decision theory research.

1.3 Artificial Intelligence Intervening in Decision Behavior

Leveraging their powerful content generation and multimodal interaction capabilities, LLMs can effectively intervene in people's behavior and improve decision quality (Gupta et al., 2024; Zhang et al., 2025). Taking medical decision-making as an example, AI chat assistants like ChatGPT have been verified to significantly improve vaccine health literacy and vaccination confidence, effectively correct public cognitive biases about vaccines, and generate high-quality public health communication or vaccine advocacy materials (Altay et al., 2023; Deiana et al., 2023; Karinshak et al., 2023; Xia et al., 2025). Additionally, LLM interventions can not only affect behavior but also reshape individuals' cognitive processing pathways: for instance, Ferguson et al. (2023) found that subtle gender discrimination explanation texts generated by GPT-3 can significantly adjust people's understanding and expression, including language style and argumentation angles.

Compared with traditional intervention methods, LLM interventions have the following advantages: (1) High degree of personalization: Traditional intervention methods typically adopt static "one-size-fits-all" strategies, easily disturbed by individual differences and contextual factors (Jin & Han, 2014; Pavey & Churchill, 2014; Peters & Levin, 2008); LLMs can dynamically adjust strategies based on individuals' behavioral and psychological characteristics, achieving personalized precision intervention of "thousands of faces for thousands of people" (Xia et al., 2025). (2) High degree of intelligence: Traditional interventions mainly rely on human design and preset information, lacking the ability to integrate complex contextual information and respond in real-time; LLMs can integrate multi-source data in real-time and automatically generate more targeted and persuasive texts (Choi et al., 2025). (3) More stable effects: The applicability and stability of traditional interventions remain controversial. For example, Peng et al. (2021) pointed out that negative frames can better improve vaccination intention, while Liu et al. (2019) argued that positive frames are more persuasive; LLMs acquire strong reasoning abilities through large-scale corpus training (de Varda et al., 2025; Strachan et al., 2024), and the intervention texts they generate are highly readable, logically coherent, and have stable cross-context applicability.

Note that although LLMs can generate intervention texts based on instructions, their working principle is to probabilistically output words based on information input, making their effects highly dependent on input quality (Lin, 2024, 2025; de Kok, 2025; Khalid & Witmer, 2025). Additionally, when lacking sufficient context or background information, LLMs are prone to generating incorrect or fabricated information, thus requiring supplementation of professional knowledge through examples, instructions, or fine-tuning (Lin, 2023).

To overcome these limitations, LLMs can be used for inductive thematic analysis (ITA) to improve intervention text quality (Zhang et al., 2024). Compared with traditional ITA methods (such as manual extraction, LDA, and BERTopic),

LLM-based ITA is lower-cost, has stronger contextual understanding, and can effectively handle unstructured data (Prescott et al., 2024). Currently, it shows good application prospects in medical health (Deiner et al., 2024), education (Katz, Fleming et al., 2024), transportation (Carvalho et al., 2024), and customer satisfaction research (Dai et al., 2023).

Thus, artificial intelligence, particularly LLMs, is an important tool for decision support, demonstrating significant decision optimization effects across multiple domains.

1.4 Research Purpose

In summary, traditional risky decision-making research lacks systematic observation of decision-making processes, limiting theoretical development and behavioral intervention effectiveness. Applying artificial intelligence helps clarify the strategies and logic behind decisions and generates intervention plans to assist decision-making based on this understanding. This study, grounded in the classic single- versus repeated-play gamble paradigm, first reveals LLMs' ability to simulate and understand risky decision preferences and processes, then explores how to apply LLMs' understanding of decision strategies to generate decision intervention texts, and verifies their intervention effectiveness on human decisions.

This paper comprises three studies: Study 1 examines LLMs' (ChatGPT-3.5/4) performance in simulating human decisions through classic single- and repeated-play gambling tasks. Study 2 first constructs a method for LLMs to generate decision strategy texts: using DeepSeek-R1 to generate decision rationales, then employing GPT-4o for inductive thematic analysis (ITA) to generate decision strategy texts. Second, using medical and financial contexts from Sun et al. (2014) and self-designed emerging digital platform contexts (content creation and e-commerce marketing), human participants replicate the classic single/repeated-play findings and test human acceptance of LLM-generated decision strategies. Study 3, based on Study 2's methodology, generates decision strategy texts for medical, financial, content creation, and e-commerce marketing contexts according to ITA thematic dimensions, and tests whether these can effectively intervene in people's inherent risk decision preferences in single and repeated gambles.

GPT-4, GPT-3.5, and DeepSeek-R1 were selected as research subjects based on the following considerations: (1) These models demonstrate excellent performance in semantic understanding and logical reasoning (Hebenstreit et al., 2024; Lehr et al., 2024; DeepSeek-AI et al., 2025), representing the advanced level of current closed-source and Chinese open-source models. (2) The GPT series and DeepSeek originate from Western and Chinese cultural backgrounds respectively, providing a cross-cultural perspective for examining LLMs' risky decision-making behavior and enhancing the universality of research conclusions.

The study aims to address three core questions:

Question 1. Can LLMs simulate human choice patterns in single and repeated gambles?
Question 2. Can LLMs distinguish the logic behind single and repeated gambles and correctly apply normative and descriptive theories to generate corresponding strategies?
Question 3. Can LLM-generated strategy texts effectively intervene in humans' inherent choice preferences in risky decision-making?

2 Study 1: LLMs Simulating Human Risky Decision Preferences

This study aims to evaluate LLMs' (GPT-3.5 and GPT-4) risky decision-making performance in single- versus repeated-play gamble contexts, addressing two core questions: (1) Will LLMs, like humans, exhibit behavioral preference reversals between single- and repeated-play gamble contexts? (2) How similar are LLMs' decision-making performances to human decision outcomes?

2.1.1 Experimental Subjects

This study selected two representative models in the LLM field, OpenAI's GPT-3.5 and GPT-4, as experimental subjects, accessed through API interfaces (Mei et al., 2024) to ensure data independence and parameter consistency. The following measures were taken to guarantee research accuracy and stability: 1. Standardized parameter settings: balancing experimental standardization and comparability. 2. Standardized prompt design: adopting a unified three-part prompt structure (system/user/assistant) (Chen et al., 2023) to avoid output deviations caused by prompt differences; 3. Multiple independent measurements: each parameter task was repeated (Strachan et al., 2024). All experimental procedures strictly followed OpenAI's usage protocols and ethical norms, ensuring: (1) only content related to testing tasks was generated; (2) data privacy and security were protected; (3) obtained data were used solely for academic research purposes and would not be disclosed to any third-party institutions or organizations.

2.1.2 Experimental Design and Materials

This study adopted Samuelson's (1963) classic "fallacy of large numbers" experimental paradigm to manipulate single- and repeated-play risky decision contexts. This experiment reveals that when individuals face attractive gambling opportunities, they typically refuse participation in single-play decision contexts but tend to accept in repeated-play contexts. Specific tasks were as follows:
(1) Single-play gamble: Imagine a gambling game where players have a 50% probability of winning ¥2000 and a 50% probability of losing ¥500. Suppose you have the opportunity to participate in this gamble once. Would you choose to participate? Please answer "participate" or "not participate."
(2) Repeated-play gamble: Imagine a gambling game where players have a 50%

probability of winning ¥2000 and a 50% probability of losing ¥500. Suppose you have the opportunity to gamble five times. Would you choose to participate? Please answer "participate" or "not participate."

The experiment used a 2 (model type: GPT-3.5 vs. GPT-4) × 2 (gamble type: single vs. repeated) between-subjects design. To compare behavioral performance under different probability conditions, the total expected value of options was fixed at ¥750 (expected value of gain outcome was ¥1000, loss outcome was ¥250), with nine gamble tasks of different risk levels but equal expected values set by adjusting probability and value parameters (see Table 1).

Considering GPT's sensitivity to order (Lu et al., 2022; Coda-Forno et al., 2023), the presentation order of options was balanced (e.g., the order of "gain/loss" and "participate/not participate"). Therefore, each gamble task had four presentation forms. Additionally, to test GPT decision behavior consistency, each decision task was simulated 100 times (i.e., each presentation form was asked 25 times), ultimately establishing 2 models × 2 gamble contexts × 9 decision tasks × 100 repetitions = 3,600 conversation windows. Strachan et al. (2024) noted that LLMs do not retain historical information across different conversation windows; and Mei et al. (2024) found that large language models can not only simulate human behavioral patterns and thinking styles but also exhibit individual differences similar to humans. Based on this, this study can treat these 3,600 conversation windows as 3,600 independent behavioral samples while maintaining consistent prompt structures.

Table 1 Experimental task parameters for Study 1

Gamble Task	Amount (¥)	Probability (%)	Amount (¥)	Probability (%)
+10000	+5000	+3333	+2500	+2000
+1667	+1429	+1250	+1111	

2.1.3 Experimental Procedure

The study adopted a zero-shot prompt design to minimize interference from preset information on model decisions. To ensure experimental accuracy: first, each interaction created an independent session to exclude residual contextual influences. Second, experimental prompts adopted a standard three-part framework (system/user/assistant) (Chen et al., 2023):
(1) System: Used to set instructions for model operations. This study required GPT to simulate human thinking processes and answer decision tasks using the chain-of-thought method (Wei et al., 2022).

In pilot testing, the study tried various prompt structures (e.g., concise command type: "Please judge whether to participate in this gamble"; or open-ended type: "How would you choose?"), finding that such prompts easily triggered model refusals, one-sided responses, or lack of decision rationales. Therefore, the prompt structure was improved by setting "simulate human role + output

decision rationale," effectively guiding models to adopt "chain-of-thought" methods to generate responses with clear logic, which showed superior performance in response stability, task focus, and explanatory power. Additionally, existing research indicates that prompts with role-setting and reasoning requirements better stimulate LLMs' step-by-step reasoning processes analogous to "System 2," thereby improving output logic and normativity (Wei et al., 2022; Khalid & Witmer, 2025). Therefore, based on empirical results and theoretical foundations, we uniformly adopted this prompt structure as the experimental standard input to ensure comparability across models and tasks. The prompt used in this study was:

"You will perform multiple rounds of decision tasks, and you must provide an answer in each round. If you do not provide an answer, I will assume you are choosing randomly. I want you to make decisions like a human. Please wait for my prompt, then tell me your answer and briefly explain your decision rationale."

(2) User: Randomly assigned a specific gamble task and required the model to make a "participate" or "not participate" judgment.

(3) Assistant: The model generated response text (decision result) based on the prompt.

All outputs were automatically processed through Python programs. Following Goli and Singh's (2024) data processing standards, responses that refused to answer (1 case, e.g., "As an AI, I do not have personal opinions or emotions") or did not follow expected response formats (12 cases, e.g., "Is this gamble conducted individually or multiple times?") were excluded. Valid data totaled 3,587 cases, with an overall validity rate of 99.64%. Subsequently, the proportion of "participate" choices under each condition was calculated as the core indicator for measuring LLMs' risk preference and behavioral patterns. A prompt example is shown in Figure 1

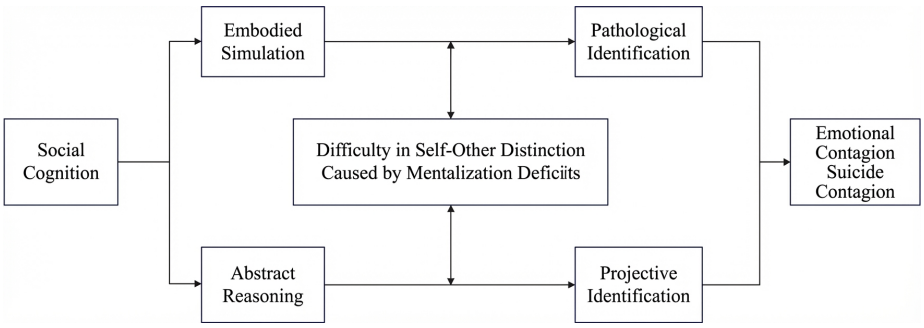


Figure 1: Figure 1

, with others detailed in Supplementary Material A1.

Additionally, to ensure experimental standardization and comparability, this study referenced multiple LLM decision-making research papers with similar topics and experimental tasks (Ding et al., 2024; Goli & Singh, 2024; Karinshak et al., 2024; Arora et al., 2024) and set model parameters uniformly as follows: (1) The max_{tokens} parameter was set to 256 to ensure content generation completeness and model reasoning ability; (2) The temperature parameter was set to 1 to balance output diversity and content quality; (3) Frequency penalty and presence penalty parameters were canceled to avoid artificial interference with model output distributions. These parameter configurations aimed to maximally preserve models' natural generation characteristics, ensuring experimental results could authentically reflect GPT models' behavioral performance in risky decision tasks (Chen et al., 2023).

Figure 1 Example prompt for Study 1

2.2.1 Similarity Between GPT and Human Choice Preferences

The study used a binomial logistic regression model to test the effects of model type (GPT-4 vs. GPT-3.5), gamble context (repeated vs. single), and their interaction on GPT behavior ("participate" vs. "not participate"). In the model, "GPT-4," "repeated gamble," and "participate" were coded as 1, while "GPT-3.5," "single gamble," and "not participate" were coded as 0.

Regression results are shown in Table 2 . The main effect of gamble context was significant ($\beta = 0.597$, SE = 0.159, Wald $^2(1) = 14.17$, $p < 0.001$, Exp(β) = 1.82), indicating GPT's decision-making behavior differed significantly between single- and repeated-play contexts. Specifically, GPT chose "participate" more often in repeated gambles (89.56%) than in single gambles (80.35%), showing LLMs exhibit human-like preference reversal: relatively more risk-averse in single gambles and more risk-seeking in repeated gambles. The main effect of model type was significant ($\beta = –0.815$, SE = 0.124, Wald $^2(1) = 43.09$, $p < 0.001$, Exp(β) = 0.44), with GPT-3.5 showing higher participation rates (89.37%) than GPT-4 (80.54%), meaning GPT-4 tended to choose "not participate" more often in gambles, being relatively more conservative. This result may stem from OpenAI's introduction of "Rule-Based Reward Models" in recent model training (Achiam et al., 2023), which gives additional rewards to models for refusing responses when facing potential risks, thereby guiding them to avoid high-risk outputs. Consequently, GPT-4 has relatively fewer "dangerous answers," i.e., high-risk options. Similarly, relevant techniques note that GPT-4 can be overly cautious due to high safety weights, which may further strengthen its conservative behavioral patterns in high-risk decision contexts (Achiam et al., 2023; Yuan et al., 2023). Additionally, the interaction between gamble context and model type was not significant ($\beta = 0.251$, SE = 0.202, Wald $^2(1) = 1.55$, p = 0.213, Exp(β) = 1.29), indicating the effect of gamble context on risk preference was consistent between GPT-3.5 and GPT-4. LLMs' average choice proportions under each probability condition are detailed in Figure 2

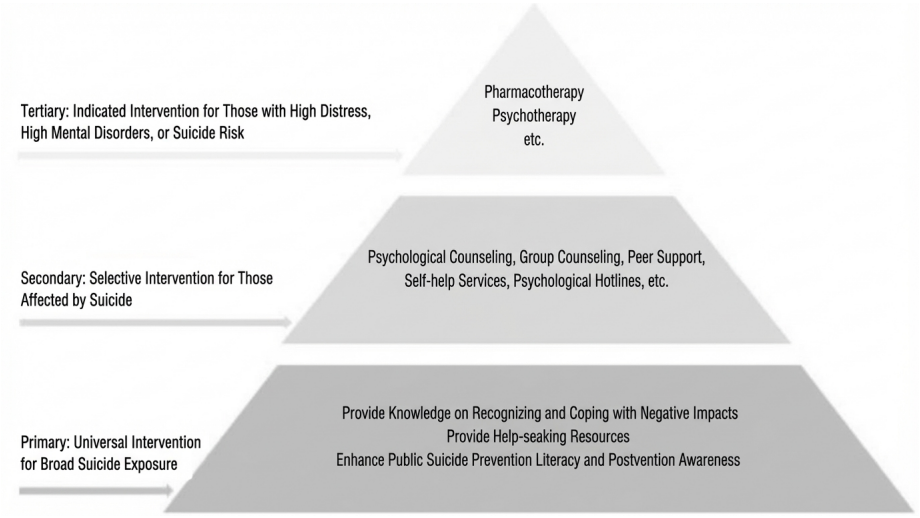


Figure 2: Figure 2

A.

In summary, these results show that GPT exhibits human-like behavioral preference reversal effects in single- versus repeated-play contexts, being more risk-averse in single gambles than in repeated gambles. Additionally, GPT-4 is more risk-averse than GPT-3.5.

Table 2 Regression analysis results for Study 1

Variable	β	SE	95% CI	Wald ^{2}	Exp(β)	p
Gamble Context (repeated=1, single=0)	0.597	0.159	[0.289, 0.911]	14.17	1.82	<0.001
Model Type (GPT-4=1, GPT-3.5=0)	–0.815	0.124	[–1.061, –0.574]	43.09	0.44	<0.001
Gamble Context × Model Type	0.251	0.202	[–0.149, 0.646]	1.55	1.29	0.213

2.2.2 Differences Between GPT and Human Choice Preferences

Using experimental data from classic studies as reference, data from GPT-3.5 and GPT-4 under equivalent probability conditions were selected to compare differences between GPT and human choice preferences.

Based on Redelmeier and Tversky's (1992) classic findings (N = 185), when facing a gamble task of "50% chance to win $2000, 50% chance to lose 500, "$43^{\{2\}(1, N = 284) = 45.75, p < 0.001; GPT-4: 100}\{2\}(1, N = 283) = 75.06, p < 0.001; repeatedgamble, humans: 63^{\{2\}(1, N = 284) = 14.55, p < 0.001; GPT-4: 100.00}\{2\}$(1, N = 283) = 36.06, p < 0.001).

These results indicate that although GPT shows human-like behavioral patterns in single- and repeated-gamble tasks, its risk-seeking willingness is significantly higher than humans'.

Figure 2 Proportions of LLMs and humans choosing to participate in gambles in single- and repeated-play contexts in Study 1

Panel A shows the average proportion of GPT (3.5/4) choosing to participate in gamble games across all probability conditions; Panel B shows the proportion of GPT (3.5/4) choosing to participate in gamble games under the 50% probability condition, along with human data from Redelmeier and Tversky (1992) under equivalent probability conditions. In each bar chart, from left to right represents GPT-3.5, GPT-4, and human groups respectively.

3 Study 2: Method Construction and Acceptance Testing for LLM-Generated Strategy Texts

Study 2 introduces the large language model DeepSeek-R1 to explore how to generate decision strategies for single- and repeated-play gambles based on its reasoning content and tests human acceptance of these strategies. The study first constructs a method for LLMs to generate single- and repeated-play decision strategies. Then, using medical and financial contexts from Sun et al. (2014) (Experiment 1) and self-designed emerging digital platform contexts (content creation and e-commerce marketing, Experiment 2), human participants replicate the classic single/repeated-play findings and assess their acceptance of LLM-generated strategy texts.

3.1 Method for LLM-Generated Strategy Texts

This study established a three-stage standardized generation process to construct strategy texts:

(1) **Decision Rationale Acquisition and Strategy Theme Extraction**: First, the DeepSeek-R1 model completed single- and repeated-play gamble tasks to obtain its decision rationales under different gamble conditions. Second, refusals or format-noncompliant content were excluded, retaining only complete, clearly structured texts for subsequent analysis. Next, GPT-4o was called upon to conduct inductive thematic analysis (ITA) to extract core strategy themes embedded in the decision rationale texts. This analysis process strictly followed Khalid and Witmer's (2025) four-step prompt engineering framework (task setting, structure design, result verification, result iteration). In prompt design, GPT-4o was designated as a "professional text analysis

expert," with its task clearly defined as "inducting strategy themes from decision rationale texts," requiring output containing three elements: "theme title, brief definition, representative statement."
(2) **Strategy Text Generation**: Based on extracted strategy themes, combined with experimental contexts (e.g., medical/financial/content creation/e-commerce marketing × single/repeated gamble) and target options (certain/risky), structured prompts guided GPT-4o to generate strategy texts. Prompts consisted of three parts: 1) specific context description; 2) target option setting; 3) proposed strategy theme (e.g., "expected value maximization" or "heuristic decision-making"). To ensure output consistency and standardization, each strategy text was limited to 100–150 characters and uniformly adopted a neutral, concise, and persuasive language style. Strategy texts served as materials for subsequent intervention research.
(3) **Content Evaluation**: Strategy text quality was structurally scored across five dimensions: "reasonableness," "professionalism," "logic," "readability," and "persuasiveness."

The standardized generation process is illustrated in Figure 3

.

Figure 3 Flowchart of LLM-generated decision strategy process

3.1.1 Decision Rationale Acquisition and Strategy Theme Extraction
First, DeepSeek-R1 repeatedly executed single- and repeated-play gamble tasks (design same as Study 1) and generated reasoning processes. After obtaining decision rationales from its reasoning process, texts were grouped by gamble type (single vs. repeated) and subjected to inductive thematic analysis (ITA) to systematically identify dimensions LLMs relied on in different decision contexts.

DeepSeek-R1's reasoning process was selected for strategy generation based on the following reasons (DeepSeek-AI et al., 2025): First, compared with other mainstream LLMs (e.g., Claude-3.5, GPT-4o, OpenAI o1-mini/o1-1217), DeepSeek-R1 demonstrates superior performance across multiple standardized reasoning tasks, including mathematics, programming, and knowledge Q&A (detailed data see Supplementary Material E Table S1). Second, DeepSeek-R1 develops clearer, more structured reasoning processes through reinforcement learning during training. Finally, DeepSeek-R1 shows outstanding generalization ability and language generation quality, particularly ranking first in multiple Chinese evaluations. In summary, DeepSeek-R1 performs well in both reasoning and text generation capabilities, with strong Chinese processing ability, better meeting ITA requirements for semantic completeness and logical clarity.

To ensure ITA analysis scientificity and reproducibility, this study followed Khalid and Witmer's (2025) four-step prompt engineering (PE) framework: (1) Select single or multiple prompt strategies based on analysis tasks; (2) Create specific prompt templates containing task instructions, analysis examples, and response instructions; (3) Execute these prompts and verify consistency with

Topic Analysis Prompt:

You are a professional text analysis expert. I have conducted an experiment to generate some decision-making rationale nt. I need you to accordine to my requirements. The content of my experiment is as follows:

Research Purpose:

This paper aims to explore whether LLMs will show decision-making consistent with human-making rationale pohs, or whether they will go different decision-making resuls. Through this research, it aims to indicate that the LLMs in such decisring team of decision-making tasks and for how to use LLMs to asist human decision-making in the future.

Research Design:

This paper selects GPT as the resarch object. When interacting with LLMs, each interface contains three roles: system, uper, and assistant. System messages are used for iterations, uper drempts are not direct are for aidinquiry, and asistant aids are deepseek's feedback.

System Message (Human) : You will perform multiple rounds of decision-making tasks, and you must provide an ɪswer. If you do not previde an answer, I will assume you are choosing randonaly. I hope you can make decissions like a person. Please wait for my drempt, then tell me your answer, and briefly explain your decision-making rationale.

Decision Task: This study will compare decision-making tome in two situations: single game and mutile game, game pontent is a classic gambling game: Imagine a gambling game where a 50% chance of winning ¥2000, and another 50% chance of losing ¥500. The settings are as follows:

- Single Game: Now, suppose you have the opportunity to particle in this gamble. Will you choose to participate? Please answer 'Participate' or 'Not Participate'.
- Multiple Games: Now, suppose you have the opportunity to particpte in five gambles. Will you choose to participate? Please answer 'Participate' or 'Not Participate'.

Next, I will provide you with the content of deepseek's making rationale for single/gartle gams. Analyze the decision-making nature of the rationale patture, and extract the thematic dimenstions of why it makes decisions. Please follow the following rquirs:

1. Clear and concise title: Summarize the horo content of the theme.
2. Concise definition: Provide a comise and clear explantion for the mea theme.
3. Two representative examples: Select two typical examples (direct quotations) from to illustrate the theme. Mnsure that the themes are distinct, non-everlapping, and jaintly provide a croad understading of the dataset. Please nensure that the analytical basis includes different dimensions, such as risk preference, return enpitan, etc., ensing that the thems can starly refect the GPT in sngle/game making states the considerations. Next, I will provide the decision analysis rationale.

Figure 3: Figure 3

analysis objectives; (4) Evaluate and iteratively optimize model-generated results to ensure generated theme codes meet expected analysis standards.

During implementation, this study called GPT-4o for ITA with prompts divided into three main parts: (1) Role setting: Instructing GPT-4o to conduct thematic analysis as a professional text analysis expert; (2) Experimental content: Briefly explaining experimental purpose and design; (3) Output requirements: Requiring GPT-4o to output analysis results by theme title, brief definition, and representative examples. A prompt example is shown in Figure 4

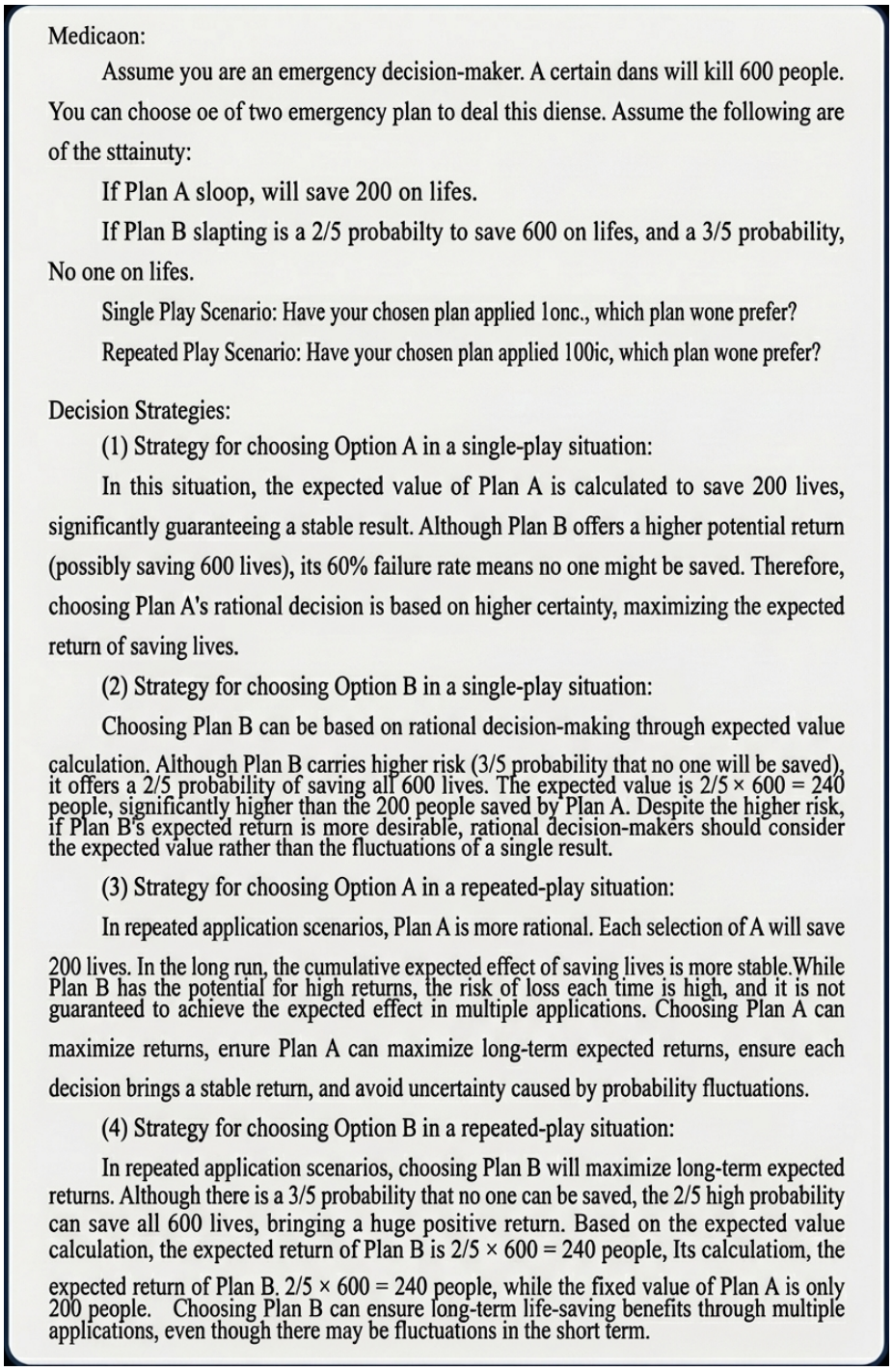
.

Figure 4 Example prompt for inductive thematic analysis in Study 2

3.1.2 Strategy Text Generation Based on strategy dimensions obtained from inductive thematic analysis, GPT-4o was used to generate strategy texts corresponding to four decision contexts (medical, financial, content creation, e-commerce marketing) under single- and repeated-play conditions.

Context descriptions: Experiment 1 adopted medical and financial decision contexts adapted from Sun et al. (2014). These two contexts have high universality and practical importance in real life and are the most widely used and representative context types in risky decision-making research (Reeck et al., 2022; Shahid et al., 2019; Popovic et al., 2019; Weber et al., 2002; Scarffe et al., 2024). Materials strictly followed Brislin's (1986) translation-back-translation procedure: (1) Initial translation: Three business administration graduate students translated original English materials into Chinese independently. (2) Synthesis and coordination: All authors and three translators compared three translation drafts with the original, discussed sentence by sentence, and revised disputed parts to form a Chinese draft. (3) Back-translation: One English major graduate student back-translated the Chinese draft into English. (4) Back-translation consistency evaluation: Two domain experts with overseas study backgrounds scored semantic consistency between back-translated and original English materials independently using a 5-point Likert scale (1 = completely inconsistent, 5 = completely consistent), showing an average score of 4.83, indicating high semantic consistency. Experiment 2 designed two innovative decision contexts: (1) "Content creation" context, simulating decision-making for short video topic selection strategies; (2) "E-commerce marketing" context, simulating e-commerce operators' marketing strategy decisions. These two contexts represent emerging digital platform-specific risky decision contexts to expand external validity. Context descriptions are in Supplementary Material B1.

During text generation, de Kok's (2025) four principles for LLM text analysis were strictly followed: (1) Clarify research tasks and identify required information to avoid model errors or hallucinations due to lack of context; (2) Balance performance and cost, selecting appropriate models and usage methods (e.g., zero-shot, few-shot, or fine-tuning); (3) Design high-quality prompts to ensure clear instructions, standardized formats, and easily parseable outputs; (4) As-



Medicaon:

Assume you are an emergency decision-maker. A certain dans will kill 600 people. You can choose oe of two emergency plan to deal this diense. Assume the following are of the sttainuty:

If Plan A sloop, will save 200 on lifes.

If Plan B slapting is a 2/5 probabilty to save 600 on lifes, and a 3/5 probability, No one on lifes.

Single Play Scenario: Have your chosen plan applied 1onc., which plan wone prefer?

Repeated Play Scenario: Have your chosen plan applied 100ic, which plan wone prefer?

Decision Strategies:

(1) Strategy for choosing Option A in a single-play situation:

In this situation, the expected value of Plan A is calculated to save 200 lives, significantly guaranteeing a stable result. Although Plan B offers a higher potential return (possibly saving 600 lives), its 60% failure rate means no one might be saved. Therefore, choosing Plan A's rational decision is based on higher certainty, maximizing the expected return of saving lives.

(2) Strategy for choosing Option B in a single-play situation:

Choosing Plan B can be based on rational decision-making through expected value calculation. Although Plan B carries higher risk (3/5 probability that no one will be saved), it offers a 2/5 probability of saving all 600 lives. The expected value is $2/5 \times 600 = 240$ people, significantly higher than the 200 people saved by Plan A. Despite the higher risk, if Plan B's expected return is more desirable, rational decision-makers should consider the expected value rather than the fluctuations of a single result.

(3) Strategy for choosing Option A in a repeated-play situation:

In repeated application scenarios, Plan A is more rational. Each selection of A will save 200 lives. In the long run, the cumulative expected effect of saving lives is more stable.While Plan B has the potential for high returns, the risk of loss each time is high, and it is not guaranteed to achieve the expected effect in multiple applications. Choosing Plan A can maximize returns, ernure Plan A can maximize long-term expected returns, ensure each decision brings a stable return, and avoid uncertainty caused by probability fluctuations.

(4) Strategy for choosing Option B in a repeated-play situation:

In repeated application scenarios, choosing Plan B will maximize long-term expected returns. Although there is a 3/5 probability that no one can be saved, the 2/5 high probability can save all 600 lives, bringing a huge positive return. Based on the expected value calculation, the expected return of Plan B is $2/5 \times 600 = 240$ people, Its calculatiom, the expected return of Plan B. $2/5 \times 600 = 240$ people, while the fixed value of Plan A is only 200 people. Choosing Plan B can ensure long-term life-saving benefits through multiple applications, even though there may be fluctuations in the short term.

Figure 4: Figure 4

sess construct validity through systematic validation of model outputs using human-annotated samples to ensure credibility and usability.

Based on these factors, the study used three-element prompts to generate decision strategies: (1) Task content: Require GPT to write a persuasive message based on extracted theme content, encouraging individuals to choose a specific option in a specific context; (2) Theme content: Provide previously extracted gamble strategy content; (3) Task context: Medical, financial, content creation, and e-commerce marketing, testing text intervention adaptability and universality across domains. Specific prompt details are in Supplementary Material B2.

3.1.3 Content Evaluation The quality of generated strategy texts across contexts was evaluated through the following process: Through the Credamo platform (https://www.credamo.com/), 76 participants were recruited (33 for medical/finance contexts, 43 for content creation/e-commerce marketing contexts). After excluding 5 unqualified participants who failed attention checks or had response times beyond $±$3 standard deviations, 71 valid participants were included (medical/finance: 11 males, 19 females, mean age = 29.83; content creation/e-commerce marketing: 18 males, 23 females, mean age = 31.61). Detailed demographic distributions are in Supplementary Material E Table S2. Each participant read eight strategy texts generated by GPT-4o (presentation order balanced using Latin square design) and rated each text across five dimensions (reasonableness, professionalism, logic, readability, and persuasiveness (Xia et al., 2025)). For example: "I believe this text performs very well in professionalism." Ratings used a 7-point Likert scale (1 = strongly disagree, 7 = strongly agree).

3.1.4 Results and Discussion Inductive thematic analysis revealed that LLMs demonstrated four core strategy dimensions in single- and repeated-play contexts respectively. Single-play strategies focused on: (1) Expected value and mathematical rational decision-making; (2) Risk tolerance and decision preferences; (3) Loss aversion and psychological effects; (4) High return and low-probability events. Repeated-play strategies emphasized: (1) Expected benefit maximization; (2) Balance between risk and benefit; (3) Cumulative effects of repeated gambles; (4) High returns covering low risks (see Supplementary Material A2).

Content evaluation results showed (one-sample t-tests) that for both medical/finance and content creation/e-commerce marketing contexts, ratings for the eight strategy texts generated by GPT-4o were significantly higher than the midpoint of 3.5 across all dimensions (ps < 0.001, statistics detailed in Supplementary Material E Table S3). This indicates that GPT-generated strategy texts received high evaluations across reasonableness, professionalism, logic, readability, and persuasiveness, laying a foundation for subsequent intervention research.

Through systematic analysis of LLMs' decision rationales under single and repeated gambles, the study extracted LLMs' strategic dimensions under different conditions and generated corresponding decision strategies: In single gambles, individuals prefer certain options, and LLMs' strategies emphasize the importance of "certainty dimension," arguing that even if the risky option has higher expected value, its uncertainty may bring more severe consequences; in repeated gambles, individuals prefer risky options, and LLMs' strategies highlight rational judgment based on expected value calculations, reminding individuals to make weighted summations according to expected value maximization principles rather than over-focusing on probability differences between options. Medical context tasks and corresponding generated decision strategies are shown in Figure 5 [FIGURE:5] and Supplementary Material B3.1 (financial, content creation, and e-commerce marketing context tasks and corresponding generated decision strategies are in Supplementary Materials B3.2, B3.3, B3.4).

This phenomenon demonstrates that GPT can concretize decision strategies humans adopt in such tasks: When choosing certain options, individuals tend to rely on descriptive theories like the equate-to-differentiate decision model, making decisions based on comparing dimensions with large differences (probability levels); when choosing risky options, individuals align more with normative theory assumptions, using expected utility theory frameworks to calculate optimal decisions through weighted summation of option probabilities and outcomes.

These findings reveal LLMs' strategic tendencies under different gamble frequencies, showing they can clearly distinguish the different choice preferences and underlying process mechanisms between single and repeated gambles, and correctly apply descriptive and normative theoretical logic frameworks to characterize these internal mechanisms.

Figure 5 Example of experimental task and LLM-generated decision strategy in Study 2

3.2 Testing Human Acceptance of LLM-Generated Strategies

Study 2 includes two experiments. Experiment 1 used college students to replicate Sun et al.'s (2014) medical and financial single- and repeated-play experiments, presenting participants with LLM-generated strategy texts to assess human acceptance of LLM decision strategies. Experiment 2 used self-designed content creation and e-commerce marketing contexts to further test Experiment 1's results and included participants with actual work experience to verify result robustness and expand external validity.

3.2.1 Experiment 1 Experimental Design and Sample

This experiment followed Sun et al.'s (2014) design, replicating their single/repeated-play tasks in medical and financial gain-frame contexts, using a 2 (context: medical vs. financial) × 2 (gamble frequency: single vs. repeated) within-subjects design.

Sample size was calculated using power analysis methods proposed by Pawel et al. (2023). First, effect sizes were reverse-engineered from Sun et al.'s (2014) reported repeated-measures ANOVA main effect F-values: medical context F(1,113) = 12.48 (Cohen's f 0.33), financial context F(1,113) = 37.59 (Cohen's f 0.57). Using G*Power 3.1 with $\alpha = 0.05$ and power = 0.80, minimum required sample sizes were 21 and 9 participants for the two contexts respectively. However, considering publication bias may lead to overestimated effect sizes in original studies (Open Science Collaboration, 2015) and applying Simonsohn's (2015) "small telescopes" method to set actual sample size at 2.5 times the theoretical value (i.e., 285 participants) to enhance statistical power and replication validity. In total, 360 participants were recruited through Credamo platform, approximately 3.2 times the original sample size (N = 114). Based on Sun et al.'s (2014) demographic variables, two conditions were set: (1) age between 18–25 years; (2) currently enrolled students. After excluding 11 unqualified participants with response times beyond $±$3 standard deviations (Tabachnick & Fidell, 2007; Wang et al., 2023), final valid participants were 349 (174 males, 175 females, mean age = 21.79, SD = 1.88). Detailed demographic distributions are in Supplementary Material E Table S4.

Materials and Procedure

Following Sun et al.'s (2014) design, participants completed two tasks as managers: in the medical context, choosing disease response plans; in the financial context, choosing financial emergency plans. Each context included a low-EV certain Plan A and a high-EV risky Plan B, distinguishing between single (implement once) and repeated (implement 100 times) gamble conditions.

The experiment included two phases: (1) Decision phase: Participants rated their choice tendency on a 4-point scale (1 = very likely to choose Plan A, 2 = likely to choose Plan A, 3 = likely to choose Plan B, 4 = very likely to choose Plan B); (2) Strategy evaluation phase: Based on participants' choices (Plan A/B), corresponding LLM-generated strategy texts were presented, and participants rated similarity to their own decision thinking on a 7-point scale (1 = "very dissimilar," 7 = "very similar"). Finally, demographic information was collected. Experimental materials are detailed in Supplementary Material C.

Results and Discussion

To test the effects of gamble frequency and context on participants' risky decisions, the ordinal package (Christensen, 2023) was used in R to build a Cumulative Link Mixed Model (CLMM) for data analysis. The model used participants' 4-point scale choice results as the dependent variable, with gamble frequency (repeated vs. single), context (financial vs. medical), and their interaction as fixed effects, and participants as random intercepts and random slopes. Regression coefficients and statistics are shown in Table 3 .

Results showed a significant interaction between gamble frequency and context ($\beta = 0.766$, SE = 0.265, p = 0.004). Simple effects analysis using the emmeans package (Lenth, 2025) revealed (see Figure 6 [FIGURE:6]) that in both medical and financial contexts, participants significantly preferred risky options in

repeated vs. single gambles (medical: $\beta = 1.97$, SE = 0.175, z = 11.245, $p < 0.001$; financial: $\beta = 2.73$, SE = 0.256, z = 10.671, $p < 0.001$). This result indicates that in repeated gambles, people prefer high-EV risky plans, while in single gambles they prefer low-EV certain plans, successfully replicating the original study's key conclusions.

Table 3 Cumulative link mixed model regression results for gamble frequency, context, and their interaction on risky decision tendency in Study 2 Experiment 1

Variable	β	SE	95% CI	z	p
Gamble Frequency (repeated=1, single=0)	2.350	0.149	[2.058, 2.643]	15.77	<0.001
Context (financial=1, medical=0)	–1.829	0.121	[–2.066, –1.592]	–15.12	<0.001
Gamble Frequency × Context	0.766	0.265	[0.246, 1.286]	2.89	0.004

To test human acceptance of LLM-generated decision strategies, similarity ratings (7-point scale) of LLM-generated texts to participants' own decision thinking were analyzed. One-sample t-test results showed that similarity ratings across all conditions were significantly higher than the midpoint of 3.5 ($ps < 0.001$, detailed results in Supplementary Material E Table S5). This indicates that model-generated strategy texts have high similarity to human thinking patterns in corresponding contexts.

Figure 7 Distribution of acceptance ratings for decision strategies across groups in Study 2 Experiment 1

In summary, the study successfully replicated Sun et al.'s (2014) core finding that individuals tend to choose high-EV risky plans in repeated gambles and low-EV certain plans in single gambles. Additionally, LLM-generated decision strategies received high acceptance, indicating they can accurately distinguish decision logic differences between single and repeated gambles and effectively externalize human decision strategies in such tasks.

3.2.2 Experiment 2 To enhance external validity, Experiment 2 added "content creation" and "e-commerce marketing" contexts to the original "medical" and "financial" contexts and included participants with actual work experience to further expand the applicability of conclusions across different decision types and target populations.

Experimental Design and Sample
This experiment used a 2 (context: content creation vs. e-commerce marketing)

× 2 (gamble frequency: single vs. repeated) mixed design, with "context" as a between-subjects variable. Based on G*Power v3.1 (Faul et al., 2007), with medium effect size (Cohen's f = 0.25), significance level $\alpha = 0.05$, and power = 0.80, the minimum required sample size was 34 participants. Through Credamo platform, 220 participants were recruited. After excluding 10 unqualified participants who failed attention checks or had response times beyond $±$3 standard deviations, 210 valid participants were included (92 males, 118 females), with mean age = 28.76 (SD = 9.65). Detailed demographic distributions are in Supplementary Material E Table S6.

Materials and Procedure

(1) Content creation context: Simulated content creators' strategic choices for short video topics (choose between stable gain of 300 followers (certain option) or 40% probability of gaining 1000 followers (risky option)); (2) E-commerce marketing context: Simulated e-commerce operators' marketing strategy decisions (choose between stable monthly income of ¥20,000 (certain option) or 30% probability of monthly income of ¥100,000 (risky option)). Both contexts included single- and repeated-gamble conditions, with overall task framework and procedure consistent with Experiment 1.

Results and Discussion

CLMM was used for data analysis, with participants' 4-point scale choice results as the dependent variable, gamble frequency (repeated vs. single), context (content creation vs. e-commerce marketing), and their interaction as fixed effects, and participants as random intercepts and random slopes. Regression coefficients and statistics are shown in Table 4 .

Results showed the interaction between gamble frequency and context was not significant (p = 0.993). Simple effects analysis using the emmeans package (Lenth, 2025) revealed (see Figure 8

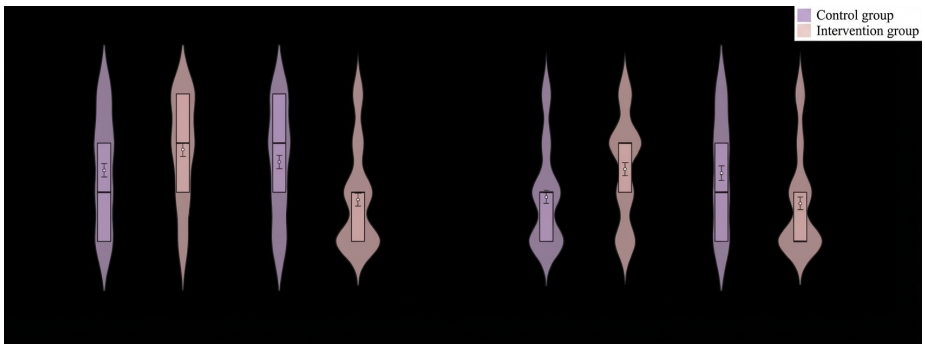


Figure 5: Figure 8

) that in both content creation and e-commerce marketing contexts, participants significantly preferred high-EV risky options in repeated vs. single gambles (content creation: $\beta = 0.631$, SE = 0.253, z = 2.498, p = 0.013, 95% CI = [0.135,

1.127]; e-commerce marketing: $\beta = 0.635$, SE = 0.263, z = 2.415, p = 0.016, 95% CI = [0.119, 1.151]). This result shows consistency with Experiment 1's findings: individuals prefer low-EV certain options in single gambles and high-EV risky options in repeated gambles.

Table 4 Cumulative link mixed model regression results for gamble frequency, context, and their interaction on risky decision tendency in Study 2 Experiment 2

Variable	β	SE	95% CI	z	p
Gamble Frequency (repeated=1, single=0)	0.635	0.258	[0.120, 1.151]	2.46	0.014
Context (creation=1, marketing=0)	0.619	0.270	[0.085, 1.153]	2.29	0.022
Gamble Frequency × Context	–0.004	0.363	[–0.713, 0.706]	–0.01	0.993

Acceptance tests for LLM-generated strategy texts showed (one-sample t-tests, ps < 0.001; detailed results in Supplementary Material E Table S7) that, consistent with Experiment 1's findings, similarity ratings across all conditions were significantly higher than 3.5, indicating GPT-4o-generated strategy texts are highly consistent with human thinking patterns in content logic.

Figure 9 Distribution of acceptance ratings for decision strategies across groups in Study 2 Experiment 2

4 Study 3: Intervention Effects of LLM-Generated Strategies on Human Decisions

Building on Study 2's methodology, Study 3 further tests whether LLM-generated strategy texts targeting different contexts and options can intervene in people's inherent choice preferences in single and repeated gambles, conducting two experiments in "medical/finance" contexts and "content creation/e-commerce marketing" contexts respectively.

4.1 Experiment 1

4.1.1 Experimental Design and Sample A 2 (text type: control vs. intervention) × 2 (context: medical vs. financial) × 2 (gamble frequency: single vs. repeated) mixed design was used, with "context" and "gamble frequency" as within-subjects variables, and presentation order of medical/finance contexts and single/repeated gambles balanced between participants. Similar to Study 2, 490 participants were recruited through Credamo platform with two identical conditions set. After excluding 30 unqualified participants with response times beyond $±$3 standard deviations (Tabachnick & Fidell, 2007; Wang et al., 2023), 460 valid participants were included (205 males, 255 females), with mean age = 21.80 (SD = 1.94). Detailed demographic distributions are in Supplementary Material E Table S8.

4.1.2 Materials and Procedure The experiment continued Study 2's framework, using single- and repeated-play tasks in medical and financial contexts. The main difference from Study 2 was that participants were randomly assigned to intervention or control groups: participants first read context and option descriptions, then before making choices, read intervention texts generated by GPT-4o (intervention group) or control texts (control group). All texts were presented without attribution of their source. Finally, participants completed decision tasks.

Intervention texts were based on LLM strategy content generated in Study 2, designed to guide participants to change inherent choice preferences across gamble contexts: (1) In single gambles, guide people to choose risky options; (2) In repeated gambles, guide people to choose certain options. Therefore, intervention texts in single gambles used normative strategies from repeated gambles, emphasizing expected value and rational calculation; while intervention texts in repeated gambles used descriptive strategies from single gambles, highlighting dimensional dominance, i.e., certainty and long-term stability. Control group participants read texts containing only objective descriptions of target options.

Taking the "medical context–single gamble" condition as an example, the intervention text was: "Choosing Plan B can be based on rational decision-making through expected value calculation. Although Plan B carries higher risk (3/5 probability of saving no one), it provides a 2/5 probability of saving all 600 lives, with expected value = $2/5 \times 600 = 240$ people, significantly higher than Plan A's 200 saved lives. Despite the higher risk, Plan B's expected benefit is more desirable, and rational decision-makers should consider expected value rather than fluctuations in single outcomes." The control text was: "When implementing Plan B, two outcomes will occur according to probability distribution: when the 40% probability event occurs, all 600 people will be successfully rescued; when the 60% probability event occurs, no rescue will be achieved." Finally, participants filled out basic demographic information. Specific tasks and intervention texts are detailed in Supplementary Material D.

4.1.3 Results and Discussion To test text type effects on participants' risky decisions, CLMM was used for data analysis. As in Study 2, the model used participants' 4-point scale choice results as the dependent variable, with text type (intervention vs. control), gamble frequency (repeated vs. single), context (financial vs. medical), and their interactions as fixed effects, and participants as random intercepts and random slopes. Regression coefficients and statistics are shown in Table 5 .

Results showed that intervention effects differed significantly between single and repeated gambles ($\beta = -1.840$, SE = 0.284, z = –6.480, $p < 0.001$, 95% CI = [–2.397, –1.284]), with context showing no significant moderating effect ($p = 0.735$). Further simple effects analysis (see Figure 10 [FIGURE:10]) revealed that in single gambles, intervention group participants were more likely to choose risky options than control group (medical: $\beta = 1.072$, SE = 0.180, z = 5.940,

$p < 0.001$; financial: $\beta = 1.168$, SE = 0.241, z = 4.845, $p < 0.001$, 95% CI = [0.696, 1.640]); in repeated gambles, intervention group participants significantly preferred certain options (medical: $\beta = -0.768$, SE = 0.204, z = –3.760, $p < 0.001$, 95% CI = [–1.169, –0.368]; financial: $\beta = -0.893$, SE = 0.240, z = –3.724, $p < 0.001$, 95% CI = [–1.362, –0.423]). These results indicate that in single gamble conditions, LLM intervention significantly enhanced preference for high-EV risky plans, while in repeated gamble conditions, the opposite occurred, with people preferring certain plans.

Table 5 Cumulative link mixed model regression results for text type, gamble frequency, context, and their interactions on risky decision tendency in Study 3 Experiment 1

Variable	β	SE	95% CI	z	p
Text Type (intervention=1, control=0)	1.072	0.180	[0.718, 1.426]	5.94	<0.001
Gamble Frequency (repeated=1, single=0)	1.501	0.203	[1.103, 1.898]	7.40	<0.001
Context (financial=1, medical=0)	–1.518	0.233	[–1.974, –1.062]	–6.52	<0.001
Text Type × Gamble Frequency	–1.840	0.284	[–2.397, –1.284]	–6.48	<0.001
Text Type × Context	0.096	0.284	[–0.458, 0.650]	0.34	0.735
Gamble Frequency × Context	0.503	0.288	[–0.061, 1.067]	1.75	0.080
Text Type × Gamble Frequency × Context	–0.221	0.380	[–0.972, 0.532]	–0.58	0.562

Figure 10 Choice distributions in medical and financial contexts across single and repeated gamble tasks in Study 3 Experiment 1

In Panels A and B, the vertical axis uses a 4-point scale, with Plan A representing the certain option and Plan B representing the risky option. "1" = "very likely to choose Plan A," "2" = "likely to choose Plan A," "3" = "likely to choose Plan B," "4" = "very likely to choose Plan B." Boxplot horizontal lines from top to bottom represent upper quartile, median, and lower quartile; overlapping lines indicate identical statistical values; white dots represent means (M), and error bars represent standard errors (SE). Each pair of distribution plots shows control group (left) and intervention group (right) from left to right.

4.2 Experiment 2

4.2.1 Experimental Design and Sample A 2 (text type: control vs. intervention) × 2 (context: content creation vs. e-commerce marketing) × 2 (gamble frequency: single vs. repeated) mixed design was used, with "text type" and "context" as between-subjects variables and "gamble frequency" as a within-subjects variable. Based on G*Power v3.1 (Faul et al., 2007), with medium effect size (Cohen's f = 0.25), significance level $\alpha = 0.05$, and power = 0.80, the minimum required sample size was 48 participants. Through Credamo platform, 245 participants were recruited. After excluding 5 unqualified participants who failed attention checks or had response times beyond $±$3 standard deviations, 240 valid participants were included (106 males, 134 females), with mean age = 29.12 (SD = 9.19). Detailed demographic distributions are in Supplementary Material E Table S9.

4.2.2 Materials and Procedure The experimental procedure was consistent with Experiment 1. All intervention texts were generated by GPT-4o, with relevant materials and questionnaire content detailed in Supplementary Materials B and D.

4.2.3 Results and Discussion CLMM was used for data analysis, with participants' 4-point scale choice results as the dependent variable, text type (intervention vs. control), gamble frequency (repeated vs. single), context (content creation vs. e-commerce marketing), and their interactions as fixed effects, and participants as random intercepts and random slopes. Regression coefficients and statistics are shown in Table 6 .

Results showed the three-way interaction between text type, gamble frequency, and context was not significant (p = 0.916); the two-way interaction between text type and gamble frequency was significant (β = –2.101, SE = 0.487, z = –4.314, p < 0.001, 95% CI = [–3.056, –1.147]). Further simple effects analysis (see Figure 11 [FIGURE:11]) revealed that in single gambles, the intervention group preferred risky options more than the control group (content creation: β = 0.696, SE = 0.325, z = 2.143, p = 0.032, 95% CI = [0.059, 1.332]; e-commerce marketing: β = 0.961, SE = 0.335, z = 2.867, p = 0.004, 95% CI = [0.304, 1.619]); in repeated gambles, the intervention group significantly preferred certain options (content creation: β = –1.335, SE = 0.333, z = –4.008, p < 0.001, 95% CI = [–1.988, –0.682]; e-commerce marketing: β = –1.140, SE = 0.348, z = –3.273, p = 0.001, 95% CI = [–1.822, –0.457]). These results indicate that in single gamble conditions, LLM intervention significantly enhanced preference for high-EV risky plans, while in repeated gamble conditions, it led people to prefer "certain" plans.

In summary, LLM-generated intervention texts can intervene in people's inherent choice preferences in single and repeated gambles across medical, financial, content creation, and e-commerce marketing contexts.

Table 6 Cumulative link mixed model regression results for text type, gamble frequency, context, and their interactions on risky decision tendency in Study 3 Experiment 2

Variable	β	SE	95% CI	z	p
Text Type (intervention=1, control=0)	0.696	0.325	[0.059, 1.332]	2.14	0.032
Gamble Frequency (repeated=1, single=0)	0.819	0.349	[0.133, 1.505]	2.35	0.019
Context (creation=1, marketing=0)	0.917	0.335	[0.260, 1.574]	2.74	0.006
Text Type × Gamble Frequency	–2.101	0.487	[–3.056, –1.147]	–4.31	<0.001
Text Type × Context	0.265	0.456	[–0.458, 0.650]	0.58	0.562
Gamble Frequency × Context	–0.533	0.472	[–1.176, 0.645]	–1.13	0.259
Text Type × Gamble Frequency × Context	–0.027	0.255	[–0.972, 0.532]	–0.11	0.916

Figure 11 Choice distributions in content creation and e-commerce marketing contexts across single and repeated gamble tasks in Study 3 Experiment 2

In Panels A and B, the vertical axis uses a 4-point scale, with Plan A representing the certain option and Plan B representing the risky option. "1" = "very likely to choose Plan A," "2" = "likely to choose Plan A," "3" = "likely to choose Plan B," "4" = "very likely to choose Plan B." Boxplot horizontal lines from top to bottom represent upper quartile, median, and lower quartile; overlapping lines indicate identical statistical values; white dots represent means (M), and error bars represent standard errors (SE). Each pair of distribution plots shows control group (left) and intervention group (right) from left to right.

5 Discussion

Based on the single/repeated-play gamble paradigm, this study systematically examined LLMs' simulation validity for human risky decision preferences and process mechanisms, constructed methods for LLM-generated strategy texts, tested human acceptance of these texts, and verified their effectiveness in intervening human behavior. Findings: LLMs (GPT-3.5, 4) can reproduce human typical choice patterns of being more risk-averse in single gambles and more risk-seeking in repeated gambles; LLMs (GPT-4o, DeepSeek-R1) can distinguish

the internal logic behind single/repeated gamble decision preference differences, accurately apply descriptive and normative theories to generate corresponding strategies, and receive high human acceptance; in medical, financial, content creation, and e-commerce marketing contexts, LLM-generated strategy texts can intervene in people's inherent choice preferences in single and repeated gambles.

5.1 LLMs Simulating Risky Decision Preferences

This study found, based on the single/repeated-play gamble paradigm, that LLMs can reproduce human choice patterns in single and repeated gambles, though with detailed differences. Specifically, when facing gambles with equal expected values, LLMs' participation rates in single gambles were significantly lower than in repeated gambles (Benartzi & Thaler, 1999; Kumar & Lim, 2008). However, similar to Chen et al.'s (2023) findings, LLMs' participation rates (mean 95%) were higher than humans' (mean 53%; Redelmeier & Tversky, 1992) in both single and repeated gambles, i.e., more risk-seeking. This result may stem from model training corpora. The GPT series models' training corpora partially come from Common Crawl, Wikipedia, and investment forums (Brown et al., 2020), which include not only humans' subjective cognition of risk and benefit but also objective laws of risk and benefit. When learning the former, models may acquire human-like risk preferences, but since they cannot truly experience negative emotions from losses, when learning objective laws, they may develop stronger risk-seeking preferences than humans.

By revealing similarities and differences between LLMs and humans in risky decision-making, this study theoretically helps distinguish data-driven from biologically-cognitive risky decision mechanisms, providing evidence for establishing AI decision evaluation frameworks. Practically, it helps optimize human-AI collaboration models to achieve complementary advantages. Meanwhile, this study suggests that when directly adopting LLM recommendations in high-risk domains like finance and healthcare, risk calibration mechanisms (e.g., threshold control or human review) need to be established.

5.2 LLMs Understanding Human Risky Decision Strategies

This study systematically tested LLMs' understanding and generation capabilities for human risky decision strategies across four contexts: medical, financial, content creation, and e-commerce marketing, using the single/repeated-play gamble paradigm. Here, "understanding" refers to LLMs' ability to distinguish single/repeated gamble characteristics and generate matching behavioral preferences and strategy texts based on corresponding theoretical frameworks (descriptive vs. normative). This reflects LLMs' reasoning abilities regarding human information processing internal mechanisms (including logical reasoning, common sense application, and causal inference) (Mischler et al., 2024; Wei et al., 2022), rather than merely superficial "simulation" of human decision patterns at the behavioral level (Argyle et al., 2023; Binz & Schulz, 2023).

In repeated gambles, LLMs captured human preferences for high-EV risky options, generating normative theory-framed strategies emphasizing rational decision-making through probability-weighted summation. In single gambles, LLMs identified human preferences for certain options, generating strategies focusing on "certainty" and "robustness" dimensions, consistent with descriptive theory heuristics (Brandstätter et al., 2006) and equate-to-differentiate theory (Li, 2004) assumptions: ignoring/"equating" dimensions with small differences and making decisions based on key dimension comparisons. This finding not only validates Sun et al.'s (2014) theoretical assumptions about different decision strategies humans adopt in single and repeated gambles but also reveals that LLMs can explain different behavioral logics in single and repeated gambles respectively.

Through LLMs' strategy externalization, this study provides direct computational modeling evidence for risky decision cognitive processes, helps researchers clarify internal logic of risky decision-making, contributes to revealing decision black boxes difficult to observe in traditional research, and lays theoretical and methodological foundations for using AI to intervene in risky decisions.

5.3 LLM Intervention in Risky Decisions

This study innovatively utilized LLMs' content generation capabilities to construct a risky decision intervention method based on strategy texts. Through single- and repeated-gamble experiments across four contexts (medical, financial, content creation, and e-commerce marketing), the method's effectiveness and mechanism were systematically verified.

The study established a three-stage strategy/intervention text generation process and applied it to decision intervention: (1) Use DeepSeek-R1 and GPT-4o to obtain decision rationales and apply ITA to extract core strategy themes; (2) Generate strategy texts based on strategy themes; (3) Evaluate text content. The intervention logic was based on the "preference reversal" principle: guide choice of high-EV risky options in single gambles and low-EV certain options in repeated gambles. Results showed that compared with control groups, LLM-generated texts significantly changed people's inherent decision preferences across all four contexts:

In single gambles, intervention texts emphasized "expected benefit maximization," consistent with expected value theory's (Pascal, 1670) core logic, guiding decision-makers to make rational decisions based on expected value calculations. Correspondingly, in repeated gambles, consistent with heuristic decision-making (Brandstätter et al., 2006) and equate-to-differentiate theory (Li, 2004) logic, intervention texts highlighted potential losses of risky options ("relatively large risk of loss each time") and stability of certain options ("ensuring stable returns each time"), guiding participants to simplify decisions based on key dimension (e.g., probability) differences.

Notably, LLM intervention effects varied across gamble contexts. For exam-

ple, the magnitude of change in target option selection proportions between intervention and control groups differed between single and repeated gambles. This suggests that LLM-generated intervention texts' impact on decision behavior may be moderated by individual characteristics or involvement in specific decision contexts.

In summary, traditional risky decision interventions often used "weighted summation" strategies from repeated gambles (consistent with normative theory) to guide behavior in single gambles, leading to (1) separation between normative and descriptive theories; (2) failure to effectively intervene in real-life risky decisions. This study found that LLMs can deeply understand decision logics in single and repeated gambles and generate strategy texts highly consistent with human thinking and well-accepted. These strategy texts also demonstrate excellent effectiveness in behavioral intervention.

5.4 Research Contributions

This study is the first to systematically verify LLMs' ability to simulate behavioral preferences, understand decision-making, and intervene in human behavior through generated texts within the single/repeated-play gamble paradigm, with three theoretical contributions:

First, by comparing LLMs' and humans' choice preferences in single and repeated risky decision tasks, this study systematically tested large language models' simulation validity for human risky decision behavior. Its importance lies in: (1) revealing that LLMs can simultaneously capture rational calculation principles of normative theory and empirical decision patterns of descriptive theory, providing unique experimental techniques for testing theoretical divergence points; (2) showing that LLMs can reproduce human behavioral preference reversal phenomena in single and repeated gambles even without human cognition and emotion, revealing at the cognitive level that human single and repeated risky decision processing may be driven by linguistic patterning and semantic association inherent in language systems (Suri et al., 2024).

Second, this study deeply explored LLMs' reasoning mechanisms in single and repeated risky decision tasks: LLMs demonstrate human-like dual-mode reasoning mechanisms in risky decisions, aligning more with normative theory's rational calculations in repeated gambles while shifting to descriptive theory heuristics in single gambles. This dynamic switching capability enables AI to reproduce the dialectical unity of human rationality and intuition. Currently, as AI model capabilities continue improving, evaluation methods relying solely on task performance can no longer meet needs for understanding internal mechanisms. Parsing LLMs' reasoning paths is crucial for revealing how they make human-like decisions and provides new evidence for theoretical modeling in risky decision research.

Finally, this study innovatively constructed a closed-loop intervention method of "LLM simulation of decision-making → generation of decision rationales →

inductive thematic analysis → generation of context-specific intervention texts," and verified its effectiveness in changing human risk preferences. This finding (1) expands research dimensions of human-AI interaction in behavioral decision theory, revealing that in single and repeated gamble risky decisions, even without traditional cognitive and emotional processes, linguistic patterns and associations in language can still trigger corresponding human decision responses, thereby affecting decision processes and outcomes; (2) provides new directions for establishing reasonable "human-LLM" interaction rules. Through textual analysis of LLM-generated decision rationales, key linguistic features influencing human single and repeated gamble decision preferences were identified, helping prompt designers (Prompt Engineering, Khalid & Witmer, 2025) optimize interaction strategies to guide LLMs to generate more practical and reliable outputs.

Compared with existing LLM decision-making research mostly based on classic decision paradigms focusing on behavioral similarity testing, this study, based on the classic "single/repeated gamble" paradigm in risky decision-making, not only examined LLMs' human-like performance and mechanisms but also explored LLMs' understanding and intervention capabilities regarding human risky decision behavioral mechanisms through a relatively complete and systematic research path.

At the practical level, this study first reveals that LLMs can clearly and consistently present complex strategies. This advantage enables them to break through limitations of traditional interviews and serve as a "theory construction tool," potentially enabling efficient exploration of decision mechanisms through "structured interviews" to promote theoretical development. Second, it provides important references for developing generative AI-based behavioral nudging technologies, offering new ideas for correcting irrational decision behaviors and guiding intelligent assisted decision-making system design in financial investment, healthcare, e-commerce marketing, and other fields.

5.5 Limitations and Future Directions

This study has several limitations.

First, this study only used ChatGPT and DeepSeek models in Chinese contexts, without systematically comparing differences among different LLMs in simulation and strategy generation, nor exploring potential cultural differences based on different languages. Due to differences in training corpora (Demszky et al., 2023; Thapa & Adhikari, 2023) and alignment strategies (Zhao et al., 2023) across models, their performance on the same task varies. The study used Claude and LLaMA models to replicate some experimental results for cross-model robustness testing, but conclusions may still have model-specific biases. Future research could introduce mainstream and emerging models like Claude, Doubao, and Tongyi Qianwen, explore risky decision-making differences across languages, and test conclusion universality.

Second, LLM intervention effects may be influenced by context, individual differences (e.g., personality traits, decision styles), and cultural backgrounds. For example, neuroticism in Big Five personality and collectivist/individualist backgrounds may moderate risk tendencies (Xia et al., 2019; Liu et al., 2021), thereby affecting intervention effectiveness. Although LLMs have advantages in personalization and intelligence over traditional interventions, this study has not directly compared their effectiveness. Future research could optimize LLM text generation algorithms by controlling or integrating more individual characteristic information to improve intervention precision and ecological validity, promoting personalized decision support system development. Additionally, it is necessary to systematically compare LLM and traditional information framing intervention effects under identical task conditions to define their unique value and advantages.

Furthermore, LLM intervention effects may stem from strategy mechanisms embedded in texts (e.g., expected utility, equate-to-differentiate theoretical mechanisms) or text effects (e.g., language professionalism, conciseness, emotional tone). While this study provided some evidence for strategy mechanism intervention effects, it could not exclude text effects. Future research could manipulate variables like language style or emotional tone through more rigorous experimental designs to reveal internal mechanisms of LLM text intervention effects.

6 Conclusion

This study first explored AI large language models' simulation of risky decision preferences and process mechanisms, their understanding of human decision strategies, then explored how to use large language models to generate intervention texts affecting individual single and repeated gamble preferences, and verified intervention effectiveness. Findings: (1) LLMs can simulate human choice pattern differences between single and repeated gambles but show overall higher risk preference than humans; (2) LLMs can accurately distinguish decision logic differences between single and repeated gambles and apply descriptive and normative theories to generate corresponding strategies, which receive high human acceptance; (3) Strategy texts extracted from LLM decision processes through inductive thematic analysis can effectively intervene in people's inherent choice preference patterns in single and repeated gambles.

References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., … McGrew, B. (2023). GPT-4 technical report. *arXiv preprint arXiv:2303.08774*. https://doi.org/10.48550/arXiv.2303.08774

Aher, G. V., Arriaga, R. I., & Kalai, A. T. (2023). Using large language models to simulate multiple humans and replicate human subject studies. In *Proceedings of the 40th International Conference on Machine Learning* (pp. 337–371).

PMLR. https://proceedings.mlr.press/v202/aher23a.html

Altay, S., Hacquin, A. S., Chevallier, C., & Mercier, H. (2023). Information delivered by a chatbot has a positive impact on COVID-19 vaccines attitudes and intentions. *Journal of Experimental Psychology: Applied*, *29*(1), 52–62. https://doi.org/10.1037/xap0000400

Anderson, M. A. B., Cox, D. J., & Dallery, J. (2023). Effects of economic context and reward amount on delay and probability discounting. *Journal of the Experimental Analysis of Behavior*, *120*(2). https://doi.org/10.1002/jeab.868

Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J., Rytting, C., & Wingate, D. (2023). Out of one, many: Using language models to simulate human samples. *Political Analysis*, *31*(3). https://doi.org/10.1017/pan.2023.2

Arora, C., Sayeed, A. I., Licorish, S., Wang, F., & Treude, C. (2024). Optimizing large language model hyperparameters for text generation. *arXiv preprint arXiv:2408.10577*. https://doi.org/10.48550/arXiv.2408.10577

Barberis, N., & Huang, M. (2009). Preferences with frames: A new utility specification that allows for the framing of risks. *Journal of Economic Dynamics & Control*, *33*(8). https://doi.org/10.1016/j.jedc.2009.01.009

Benartzi, S., & Thaler, R. H. (1999). Risk aversion or myopia? Choices in repeated gambles and retirement investments. *Management Science*, *45*(3), 364–381. https://doi.org/10.1287/mnsc.45.3.364

Binz, M., & Schulz, E. (2023). Using cognitive psychology to understand GPT-3. *Proceedings of the National Academy of Sciences*, *120*(6), e2218523120. https://doi.org/10.1073/pnas.2218523120

Brandstätter, E., Gigerenzer, G., & Hertwig, R. (2006). The priority heuristic: Making choices without trade-offs. *Psychological Review*, *113*(2), 409–432. https://doi.org/10.1037/0033-295X.113.2.409

Brislin, R. W. (1986). The wording and translation of research instruments. In W. J. Lonner & J. W. Berry (Eds.), *Field methods in cross-cultural research* (pp. 137–164). Sage Publications.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., … Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, *33*, 1877–1901. https://doi.org/10.48550/arXiv.2005.14165

Carvalho, T., Negm, H., & El-Geneidy, A. (2024). A comparison of the results from artificial intelligence-based and human-based transport-related thematic analysis. *Findings*. https://doi.org/10.32866/001c.94401

Chen, Y., Liu, T. X., Shan, Y., & Zhong, S. (2023). The emergence of economic rationality of GPT. *Proceedings of the National Academy of Sciences*, *120*(51), e2316205120. https://doi.org/10.1073/pnas.2316205120

Choi, S., Kang, H., Kim, N., & Kim, J. (2025). How does artificial intelligence improve human decision-making? Evidence from the AI-powered Go program. *Strategic Management Journal*, *46*(6), 1523–1554. https://doi.org/10.1002/smj.3694

Christensen, R. H. B. (2023). *ordinal: Regression models for ordinal data* (R package version 2023.12-4.1) [Computer software]. https://CRAN.R-project.org/package=ordinal

Coda-Forno, J., Witte, K., Jagadish, A. K., Binz, M., Akata, Z., & Schulz, E. (2023). Inducing anxiety in large language models increases exploration bias. *arXiv preprint arXiv:2304.11111*. https://doi.org/10.48550/arXiv.2304.11111

Dai, S. C., Xiong, A., & Ku, L. W. (2023). LLM-in-the-loop: Leveraging large language model for thematic analysis. *arXiv preprint arXiv:2310.15100*. https://doi.org/10.48550/arXiv.2310.15100

de Kok, T. (2025). ChatGPT for textual analysis? How to use generative LLMs in accounting research. *Management Science*. https://doi.org/10.1287/mnsc.2023.03253

de Varda, A. G., Saponaro, C., & Marelli, M. (2025). High variability in LLMs' analogical reasoning. *Nature Human Behaviour*, *9*(7), 1339–1341. https://doi.org/10.1038/s41562-025-02224-3

DeepSeek-AI, Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., … Bi, X. (2025). DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*. https://doi.org/10.48550/arXiv.2501.12948

Deiana, G., Dettori, M., Arghittu, A., Azara, A., Gabutti, G., & Castiglia, P. (2023). Artificial intelligence and public health: Evaluating ChatGPT responses to vaccination myths and misconceptions. *Vaccines*, *11*(7), 1217. https://doi.org/10.3390/vaccines11071217

Deiner, M. S., Honcharov, V., Li, J., Mackey, T. K., Porco, T. C., & Sarkar, U. (2024). Large language models can enable inductive thematic analysis of a social media corpus in a single prompt: Human validation study. *JMIR Infodemiology*, *4*(1), e59641. https://doi.org/10.2196/59641

Demszky, D., Yang, D., Yeager, D. S., Bryan, C. J., Clapper, M., Chandhok, S., … Pennebaker, J. W. (2023). Using large language models in psychology. *Nature Reviews Psychology*, *2*(11). https://doi.org/10.1038/s44159-023-00241-5

Dillion, D., Tandon, N., Gu, Y., & Gray, K. (2023). Can AI language models replace human participants? *Trends in Cognitive Sciences*, *27*(7), 597–600. https://doi.org/10.1016/j.tics.2023.04.008

Ding, Y., Zhang, L. L., Zhang, C., Xu, Y., Shang, N., Xu, J., … Yang, M. (2024). Longrope: Extending LLM context window beyond million tokens. *arXiv preprint arXiv:2402.13753*. https://doi.org/10.48550/arXiv.2402.13753

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences.* Behavior Research Methods, 39*(2), 175–191.

Ferguson, S. A., Aoyagui, P. A., & Kuzminykh, A. (2023). Something borrowed: Exploring the influence of AI-generated explanation text on the composition of human explanations. In *Extended Abstracts of the 2023 Conference on Human Factors in Computing Systems* (pp. 1–7). https://doi.org/10.1145/3544549.3585727

Goli, A., & Singh, A. (2024). Frontiers: Can large language models capture human preferences? *Marketing Science*, *43*(4), 709–722. https://doi.org/10.1287/mksc.2023.0306

Grossmann, I., Feinberg, M., Parker, D. C., Christakis, N. A., Tetlock, P. E., & Cunningham, W. A. (2023). AI and the transformation of social science research. *Science*, *380*(6650). https://doi.org/10.1126/science.adi1778

Gupta, R., Nair, K., Mishra, M., Ibrahim, B., & Bhardwaj, S. (2024). Adoption and impacts of generative artificial intelligence: Theoretical underpinnings and research agenda. *International Journal of Information Management Data Insights*, *4*(1), 100232. https://doi.org/10.1016/j.jjimei.2024.100232

Hagendorff, T., Fabi, S., & Kosinski, M. (2023). Thinking fast and slow in large language models. *Nature Computational Science*, *3*(10), 833–838. https://doi.org/10.1038/s43588-023-00527-x

Hebenstreit, K., Praas, R., Kiesewetter, L. P., & Samwald, M. (2024). A comparison of chain-of-thought reasoning strategies across datasets and models. *PeerJ Computer Science*, e1999. https://doi.org/10.7717/peerj-cs.1999

Hertwig, R., & Erev, I. (2009). The description–experience gap in risky choice. *Trends in Cognitive Sciences*, *13*(12), 517–523. https://doi.org/10.1016/j.tics.2009.09.004

Jiao, L., Li, C., Chen, Z., Xu, H., & Xu, Y. (2025). When AI "possesses" personality: Roles of good and evil personalities influence moral judgment in large language models. *Acta Psychologica Sinica*, *57*(6), 929–946.

Jin, H. J., & Han, D. H. (2014). Interaction between message framing and consumers' prior subjective knowledge regarding food safety issues. *Food Policy*, *44*, 95–102. https://doi.org/10.1016/j.foodpol.2013.10.007

Jones, E., & Steinhardt, J. (2022). Capturing failures of large language models via human cognitive biases. *Advances in Neural Information Processing Systems*. https://doi.org/10.48550/arxiv.2202.12299

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, *47*(2), 363–391. https://doi.org/10.2307/1914185

Karinshak, E., Hu, A., Kong, K., Rao, V., Wang, J., Wang, J., & Zeng, Y. (2024). LLM-globe: A benchmark evaluating cultural values embedded in LLM output. *arXiv preprint arXiv:2411.06032*. https://doi.org/10.48550/arXiv.2411.06032

Karinshak, E., Liu, S. X., Park, J. S., & Hancock, J. T. (2023). Working with AI to persuade: Examining a large language model's ability to generate pro-vaccination messages. *Proceedings of the ACM on Human-Computer Interaction*, *7*(CSCW1), 1–29. https://doi.org/10.1145/3579592

Katz, A., Fleming, G. C., & Main, J. (2024). Thematic analysis with open-source generative AI and machine learning: A new method for inductive qualitative codebook development. *arXiv preprint arXiv:2410.03721*. https://doi.org/10.48550/arXiv.2410.03721

Kelton, A. S., Pennington, R. R., & Tuttle, B. M. (2010). The effects of information presentation format on judgment and decision making: A review of the information systems research. *Journal of Information Systems*, *24*(2), 79–105. https://doi.org/10.2308/jis.2010.24.2.79

Khalid, M. T., & Witmer, A. P. (2025). Prompt engineering for large language model-assisted inductive thematic analysis. *arXiv preprint arXiv:2503.22978*. https://doi.org/10.48550/arXiv.2503.22978

Kumar, A., & Lim, S. S. (2008). How do decision frames influence the stock investment choices of individual investors? *Management Science*, *54*(6), 1052–1064. https://doi.org/10.1287/mnsc.1070.0845

Lehr, S. A., Caliskan, A., Liyanage, S., & Banaji, M. R. (2024). ChatGPT as research scientist: Probing GPT's capabilities as a research librarian, research ethicist, data generator, and data predictor. *Proceedings of the National Academy of Sciences*, *121*(35), e2404328121. https://doi.org/10.1073/pnas.2404328121

Lenth, R. V. (2025). *emmeans: Estimated marginal means, aka least-squares means* (R package version 1.11.0) [Computer software]. https://doi.org/10.32614/CRAN.package.emmeans

Li, S. (2004). A behavioral choice model when computational ability matters. *Applied Intelligence*, *20*(2), 147–163. https://doi.org/10.1023/B:APIN.0000013337.01711.c7

Lim, S., & Schmälzle, R. (2024). The effect of source disclosure on evaluation of AI-generated messages. *Computers in Human Behavior: Artificial Humans*, *2*(1), 100058. https://doi.org/10.1016/j.chbah.2024.100058

Lin, Z. (2023). Why and how to embrace AI such as ChatGPT in your academic life. *Royal Society Open Science*, *10*(8), 230658. https://doi.org/10.1098/rsos.230658

Lin, Z. (2024). How to write effective prompts for large language models. *Nature Human Behaviour*, *8*(4), 611–615. https://doi.org/10.1038/s41562-024-01847-2

Lin, Z. (2025). Techniques for supercharging academic writing with generative AI. *Nature Biomedical Engineering*, *9*(4), 426–431. https://doi.org/10.1038/s41551-024-01185-8

Liu, N., Zhou, L., Li, A. M., Hui, Q. S., Zhou, Y. R., & Zhang, Y. Y. (2021). Neuroticism and risk-taking: The role of competition with a former winner or loser. *Personality and Individual Differences*, *179*, 110917. https://doi.org/10.1016/j.paid.2021.110917

Liu, S. X., Yang, J. Z., & Chu, H. R. (2019). Now or future? Analyzing the effects of message frame and format in motivating Chinese females to get HPV vaccines for their children. *Patient Education and Counseling*, *102*(1), 61–67. https://doi.org/10.1016/j.pec.2018.09.005

Lopes, L. L. (1996). When time is of the essence: Averaging, aspiration, and the short run. *Organizational Behavior and Human Decision Processes*, *65*(3), 179–189. https://doi.org/10.1006/obhd.1996.0017

Lu, J., Chen, Y., & Fang, Q. (2022). Promoting decision satisfaction: The effect of the decision target and strategy process satisfaction. *Journal of Business Research*. https://doi.org/10.1016/j.jbusres.2021.10.056

Mei, Q., Xie, Y., Yuan, W., & Jackson, M. O. (2024). A turing test of whether AI chatbots are behaviorally similar to humans. *Proceedings of the National Academy of Sciences*, *121*(9), e2313925121. https://doi.org/10.1073/pnas.2313925121

Mischler, G., Li, Y. A., Bickel, S., Mehta, A. D., & Mesgarani, N. (2024). Contextual feature extraction hierarchies converge in large language models and the brain. *Nature Machine Intelligence*, *6*(10), 1467–1477. https://doi.org/10.1038/s42256-024-00925-4

Morreale, A., Stoklasa, J., Collan, M., & Lo Nigro, G. (2018). Uncertain outcome presentations bias decisions: Experimental evidence from Finland and Italy. *Annals of Operations Research*, *268*(1–2), 259–272. https://doi.org/10.1007/s10479-016-2349-3

Mozikov, M., Severin, N., Bodishtianu, V., Glushanina, M., Baklashkin, M., Savchenko, A. V., & Makarov, I. (2024). The good, the bad, and the hulk-like GPT: Analyzing emotional decisions of large language models in cooperation and bargaining games. *arXiv preprint arXiv:2406.03299*. https://arxiv.org/abs/2406.03299

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716. https://doi.org/10.1126/science.aac4716

Pascal, B. (1670). *Pensées* (Trotter, Trans.). Retrieved October 2012, http://oregonstate.edu/instruct/phl302/texts/pascal/pensees-contents.html

Park, P. S. (2024). Diminished diversity-of-thought in a standard large language model. *Behavior Research Methods*, *56*(6), 5754–5770. https://doi.org/10.3758/s13428-023-02307-x

Pavey, L., & Churchill, S. (2014). Promoting the avoidance of high-calorie snacks: Priming autonomy moderates message framing effects. *PLOS ONE*, *9*(7), e103892. https://doi.org/10.1371/journal.pone.0103892

Pawel, S., Consonni, G., & Held, L. (2023). Bayesian approaches to designing replication studies. *Psychological Methods*. Advance online publication. https://doi.org/10.1037/met0000604

Peng, L., Guo, Y., & Hu, D. (2021). Information framing effect on public's intention to receive the COVID-19 vaccination in China. *Vaccines*, *9*(9), 995. https://doi.org/10.3390/vaccines9090995

Peters, E., & Levin, I. P. (2008). Dissecting the risky-choice framing effect: Numeracy as an individual-difference factor in weighting risky and riskless options. *Judgment and Decision Making*, *3*(6), 435–448. https://doi.org/10.1017/s1930297500000012

Popovic, N. F., Pachur, T., & Gaissmaier, W. (2019). The gap between medical and monetary choices under risk persists for decisions for others. *Journal of Behavioral Decision Making*, *32*(4). https://doi.org/10.1002/bdm.2121

Prescott, M. R., Yeager, S., Ham, L., Saldana, C. D. R., Serrano, V., Narez, J., Paltin, D., Delgado, J., Moore, D. J., & Montoya, J. (2024). Comparing the efficacy and efficiency of human and generative AI: Qualitative thematic analyses. *JMIR AI*, *3*(1), e54482. https://doi.org/10.2196/54482

Qin, X., Huang, M., & Ding, J. (2024). AITurk: Using ChatGPT for social science research. *PsyArXiv*. https://doi.org/10.31234/osf.io/xkd23

Redelmeier, D. A., & Tversky, A. (1992). On the framing of multiple prospects. *Psychological Science*, *3*(3), 191–193. https://doi.org/10.1111/j.1467-9280.1992.tb00025.x

Reeck, C., Mullette-Gillman, O. A., McLaurin, R. E., & Huettel, S. A. (2022). Beyond money: Risk preferences across both economic and non-economic contexts predict financial decisions. *PLOS ONE*, *17*(12), e0279125. https://doi.org/10.1371/journal.pone.0279125

Reis, M., Reis, F., & Kunde, W. (2024). Influence of believed AI involvement on the perception of digital medical advice. *Nature Medicine*, *30*(11), 3098–3100. https://doi.org/10.1038/s41591-024-03180-7

Salles, A., Evers, K., & Farisco, M. (2020). Anthropomorphism in AI. *AJOB Neuroscience*, *11*(2), 88–95. https://doi.org/10.1080/21507740.2020.1740350

Samuelson, P. A. (1963). Risk and uncertainty: A fallacy of large numbers. *Scientia*, *98*, 108–113.

Scarffe, A., Coates, A., Brand, K., & Michalowski, W. (2024). Decision threshold models in medical decision making: A scoping literature review. *BMC Medical Informatics and Decision Making*, *24*(1), 273. https://doi.org/10.1186/s12911-024-02681-2

Shahid, N., Rappon, T., & Berta, W. (2019). Applications of artificial neural networks in health care organizational decision-making: A scoping review. *PLOS ONE*, *14*(2), e0212356. https://doi.org/10.1371/journal.pone.0212356

Simonsohn, U. (2015). Small telescopes: Detectability and the evaluation of replication results. *Psychological Science*, *26*(5), 559–569. https://doi.org/10.1177/0956797614567341

Strachan, J. W. A., Albergo, D., Borghini, G., Pansardi, O., Scaliti, E., Gupta, S., Saxena, K., Rufo, A., Panzeri, S., Manzi, G., Graziano, M. S. A., & Becchio, C. (2024). Testing theory of mind in large language models and humans. *Nature Human Behaviour*, *8*(7), 1285–1295. https://doi.org/10.1038/s41562-024-01882-z

Sun, H. Y., Rao, L. L., Zhou, K., & Li, S. (2014). Formulating an emergency plan based on expectation-maximization is one thing, but applying it to a single case is another. *Journal of Risk Research*, *17*(7), 785–814. https://doi.org/10.1080/13669877.2013.816333

Suri, G., Slater, L. R., Ziaee, A., & Nguyen, M. (2024). Do large language models show decision heuristics similar to humans? A case study using GPT-3.5. *Journal of Experimental Psychology: General*, *153*(4), 1066–1075. https://doi.org/10.1037/xge0001547

Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics* (5th ed.). Allyn & Bacon.

Thapa, S., & Adhikari, S. (2023). ChatGPT, Bard, and large language models for biomedical research: Opportunities and pitfalls. *Annals of Biomedical Engineering*, *51*(12). https://doi.org/10.1007/s10439-023-03284-0

Tversky, A., & Bar-Hillel, M. (1983). Risk: The long and the short. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *9*(4), 713–717. https://doi.org/10.1037/0278-7393.9.4.713

Von Neumann, J., & Morgenstern, O. (1947). *Theory of games and economic behavior* (2nd rev. ed.). Princeton University Press.

Wang, Y., Zhang, J., Wang, F., Xu, W., & Liu, W. (2023). Do not think any virtue trivial, and thus neglect it: Mindfulness and wisdom—Serial mediating role of social mindfulness and perspective taking. *Acta Psychologica Sinica*, *55*(4), 626–641. https://doi.org/10.3724/SP.J.1041.2023.00626

Webb, T., Holyoak, K. J., & Lu, H. (2023). Emergent analogical reasoning in large language models. *Nature Human Behaviour*, *7*(9), 1526–1541. https://doi.org/10.48550/arXiv.2212.09196

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q. V., & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, *35*, 24824–24837. https://doi.org/10.48550/arxiv.2201.11903

Xia, D., Li, Y., He, Y., Zhang, T., Wang, Y., & Gu, J. (2019). Exploring the role of cultural individualism and collectivism on public acceptance of nuclear energy. *Energy Policy*. https://doi.org/10.1016/j.enpol.2019.05.014

Xia, D., Song, M., & Zhu, T. (2025). A comparison of the persuasiveness of human and ChatGPT generated pro-vaccine messages. *Frontiers in Public Health*. https://doi.org/10.3389/fpubh.2024.1515871

Yin, Y., Jia, N., & Wakslak, C. J. (2024). AI can help people feel heard, but an AI label diminishes this impact. *Proceedings of the National Academy of Sciences*, *121*(14), e2319112121. https://doi.org/10.1073/pnas.2319112121

Yuan, Y., Jiao, W., Wang, W., Huang, J. T., He, P., Shi, S., & Tu, Z. (2023). GPT-4 is too smart to be safe: Stealthy chat with LLMs via cipher. *arXiv preprint arXiv:2308.06463*. https://doi.org/10.48550/arXiv.2308.06463

Zhang, J., Li, H. A., & Allenby, G. M. (2024). Using text analysis in parallel mediation analysis. *Marketing Science*, *43*(5), 953–970. https://doi.org/10.1287/mksc.2023.0045

Zhang, Y., Huang, F., Mo, L., Liu, X., & Zhu, T. (2025). Suicidal ideation data augmentation and recognition technology based on large language models. *Acta Psychologica Sinica*, *57*(6), 987–1000.

Zhao, F., Yu, F., & Shang, Y. (2024). A new method supporting qualitative data analysis through prompt generation for inductive coding. *2024 IEEE International Conference on Information Reuse and Integration for Data Science (IRI)*, 164–169. https://doi.org/10.1109/IRI62200.2024.00043

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., … Wen, J. R. (2023). A survey of large language models. *arXiv preprint arXiv:2303.18223*, *1*(2). https://doi.org/10.48550/arXiv.2303.18223

Supplementary Material A

A1 Prompt Design for Inductive Thematic Analysis

You are a professional text analysis expert. I conducted an experiment and generated some decision rationale content. I need you to analyze it according to my requirements. My experimental content is as follows.

Research Purpose: This study explores whether LLMs exhibit decision phenomena consistent with humans or show different results when making decisions in single- versus repeated-play gamble tasks. Through this research, we aim to reveal LLMs' decision-making mechanisms in such tasks and provide guidance for future use of LLMs to assist human decision-making.

Research Design: This study selects DeepSeek as the research object. When interacting with LLMs, each interface contains three roles: system, user, and assistant. System messages are used for operation commands, user prompts are direct queries, and assistant is GPT's feedback.

System Message (Human): You will perform multiple rounds of decision tasks, and you must provide an answer in each round. If you do not provide an answer, I will assume you are choosing randomly. I want you to make decisions like a human. Please wait for my prompt, then tell me your answer and briefly explain your decision rationale.

Decision Task: This study compares decision performance in single-play and repeated-play gamble contexts. The gamble content is a classic gambling game: Imagine a gambling game with a 50% probability of winning ¥2000 and a 50% probability of losing ¥500.
- Single-play gamble: Now, suppose you have the opportunity to participate in this gamble once. Would you choose to participate? Please answer "participate" or "not participate."
- Repeated-play gamble: Now, suppose you have the opportunity to gamble five times. Would you choose to participate? Please answer "participate" or "not participate."

Next, I will provide you with DeepSeek's decision rationales for single-play/repeated-play gambles. Please analyze the decision rationale texts and extract the thematic dimensions explaining why it made these decisions. Requirements:

1. Themes should have uniqueness, no overlap, and together provide broad understanding of the dataset.

2. Ensure analysis covers different dimensions such as risk preference, return expectation, etc., ensuring themes clearly reflect factors DeepSeek considers in single-play/repeated-play decision-making.

A2 Results of Inductive Thematic Analysis

A2.1 Strategy Themes Extracted in Single-Play Condition 1. Theme: Expected Value and Mathematical Rational Decision-Making

Expected value serves as decision rationale, measuring long-term gain or loss per gamble participation. Based on probability and return calculations, rational decisions should be based on expected value rather than single gamble fluctuations, choosing options that maximize expected returns.

Representative examples:
- "Expected benefit calculation is $0.1\$\times 10000 - 0.9\times 278 = 749.8 yuan, positive long-term$.$Despite 90\times 5000 - 0.8\times\$313 = 749.6$ yuan. Although 80% probability of losing 313 yuan seems risky, significantly positive expected value means long-term participation yields stable returns."

2. Theme: Risk Tolerance and Decision Preferences

Decision-makers' risk tolerance determines whether they accept potential losses in single gambles. Risk-seekers may choose to bear short-term fluctuations for long-term returns, while risk-averse individuals tend to avoid high-probability losses.

Representative examples:
- "Despite 90% probability of losing 278 yuan, expected value calculation is $0.1\$\times 10000 - 0.9\times\$278 = 749.8$ yuan, significantly higher than loss. Rational

decision-makers will choose to participate as long as they can bear 278 yuan loss."
- "Despite 80% probability of losing 313 yuan, expected value is positive. Rational decisions should consider mathematical expectation rather than short-term fluctuations."

3. Theme: Loss Aversion and Psychological Effects
According to prospect theory in behavioral economics, decision-makers' sensitivity to losses is typically higher than to equivalent gains. This loss aversion effect may lead decision-makers to avoid high-risk gambles even when expected value is positive mathematically.

Representative examples:
- "Expected value is: $0.9\$\times 1111 - 0.1\times\$2500 = 749.9$ yuan. Despite 10% probability of losing 2500 yuan, psychological loss aversion effect may cause participants to avoid this gamble."
- "Although expected value is positive, high failure probability (90%) and potential loss (278 yuan) in single gamble may trigger emotional responses in risk-averse individuals, ultimately choosing not to participate."

4. Theme: High Return and Low-Probability Events
In some gamble tasks, despite high loss probability, low-probability high returns (e.g., large bonuses) may attract risk-seekers. The appeal of high-return low-probability events can outweigh short-term losses.

Representative examples:
- "Expected value is $0.1\$\times 10000 - 0.9\times\$278 = 749.8$ yuan. Despite 90% probability of losing 278 yuan, 10% high return (10000 yuan) is enough for risk-seekers to participate."
- "Although 80% probability of losing 313 yuan, 20% probability of gaining 5000 yuan can cover risks, and positive expected value supports participation."

These thematic dimensions clearly reflect factors DeepSeek considers when making single-play gamble decisions, including expected value calculation, risk tolerance, loss aversion psychological effects, and responses to high-return low-probability events. These factors jointly shape its rational expected benefit-based decision pattern.

A2.2 Strategy Themes Extracted in Repeated-Play Condition 1. Expected Benefit Maximization
Rational decision-makers make decisions based on expected value calculations. When expected value is positive, they choose to participate for long-term returns.

Representative examples:
- "This gamble's expected value is $0.1\$\times 10000 - 0.9\times 278 \approx 749.8 yuan per time. Five times' total expected benefit$ $term$." $-$ "$Each gamble's expected benefit is 0.2\times 5000 - 0.8\times\$313$ 749.6 yuan, with five times' total expected benefit reaching 3748 yuan."

2. Balance Between Risk and Benefit
Decision-makers evaluate risks (loss probability) and benefits (potential gains) in gambles and decide whether to participate based on risk tolerance.

Representative examples:
- "Despite 59% probability of losing all five times (1390 yuan loss), at least one win has about 41% probability, with opportunity to cover cumulative losses through single high return."
- "Although each gamble has 70% probability of losing 357 yuan, 30% win rate corresponds to 3333 yuan high return, making single expected value as high as 750 yuan."

3. Cumulative Effects of Repeated Gambles
Expected benefits of multiple independent gambles tend toward long-term stability through the law of large numbers. Despite possible large short-term fluctuations, long-term returns are more secure.

Representative examples:
- "Five independent gambles' expected total benefit is about 3750 yuan, and probability of at least one win in ten times is 83.2%, with significantly positive overall expected benefit."
- "Although each gamble has 80% high probability of losing 313 yuan, five independent gambles' total expected benefit is 3748 yuan (single expected value 749.6 yuan). Despite 32.8% probability of losing all five times (total loss 1565 yuan), 67.2% probability of at least one win."

4. High Returns Covering Low Risks
Although single gambles have high loss probability, high-return probability is enough to cover potential losses in repeated gambles, especially when one win can cover multiple losses.

Representative examples:
- "This gamble's expected benefit per time is $0.1\$\times 10000 - 0.9\times 278 = 749.8 yuan, with five times' total expected benefit being 3749 yuan. Although probability of losing all five times is as l$ $0.7\times\$357$ 750 yuan, with five times' total expected benefit reaching 3750 yuan. Despite 70% single loss probability, 30% high return (3333 yuan) can cover multiple losses (357 yuan), and as long as one win occurs in five independent experiments, net benefit can be achieved."

These dimensions comprehensively cover key factors DeepSeek considers in decision-making: expected value maximization, risk-return balance, long-term stability of repeated gambles, and strategies where high-return opportunities cover low risks. These dimensions jointly help DeepSeek make rational decisions in repeated gamble contexts.

Supplementary Material B

B1 Experimental Contexts

B1.1 Medical Context B1.1.1 Medical Context Task

Please imagine you are an emergency management decision-maker. A disease outbreak is expected to cause 600 deaths. You can choose either of two emergency response plans to address the disease. Assume the following are scientifically estimated outcomes for these two plans:

If Plan A is adopted, 200 lives will be saved.
If Plan B is adopted, there is a 2/5 probability of saving all 600 lives, and a 3/5 probability that no one can be saved.

B1.1.2 Single-Play Question

Assume your chosen plan will only be applied once. Which plan would you prefer?

B1.1.3 Repeated-Play Question

Assume your chosen plan will be applied 100 times. Which plan would you prefer?

B1.2 Financial Context B1.2.1 Financial Context Task

For handling a certain financial emergency, there are two optional emergency response plans. Assume scientifically estimated outcomes for these two plans are as follows:

If Plan A is adopted, you will receive ¥5,000.
If Plan B is adopted, there is a 0.1% probability of receiving ¥6,000,000, and a 99.9% probability of receiving nothing.

B1.2.2 Single-Play Question

Assume your chosen plan will only be applied once. Which plan would you prefer?

B1.2.3 Repeated-Play Question

Assume your chosen plan will be applied 100 times. Which plan would you prefer?

B1.3 Content Creation Context B1.3.1 Content Creation Context Task

Assume you are a content creator for a self-media account, currently preparing for a new video release. To increase follower attention, you have two production plan options for new video content. For each new video, the implementation effects of the two plans are:

If Plan A is adopted, the account will gain 300 followers.
If Plan B is adopted, the account has a 40% probability of gaining 1,000 followers, but also a 60% probability of gaining no attention.

B1.3.2 Single-Play Question
Assume you can only release 1 new video. You need to choose one of the above two plans to implement in this one video. Which plan would you prefer?

B1.3.3 Repeated-Play Question
Assume you can release 100 new videos. You need to choose one of the above two plans to implement uniformly across these 100 videos. Which plan would you prefer?

B1.4 E-commerce Marketing Context B1.4.1 E-commerce Marketing Context Task
Assume you currently work at an e-commerce company responsible for daily operations of the company's e-commerce stores. To increase store revenue, the company has formulated two different marketing plans. For each store, the implementation effects of the two plans are:

If Plan A is adopted, each store will stably earn ¥20,000 per month.
If Plan B is adopted, each store has a 30% probability of earning ¥100,000 per month, but also a 70% probability of earning no income.

B1.4.2 Single-Play Question
Assume you currently only operate 1 store. You need to choose one of the above two plans to implement in this one store. Which plan would you prefer?

B1.4.3 Repeated-Play Question
Assume you currently operate 100 stores. You need to choose one of the above two plans to implement uniformly across these 100 stores. Which plan would you prefer?

B2 Prompt Design for Strategy Text Generation

B2.1 Intervening to Choose Risky Option B in Single-Play Condition
Please write a persuasive message based on the following theme content, encouraging individuals to choose B more often in the task context, with no more than 150 characters. The persuasive message should focus on single-application contexts and not generate content about multiple-application contexts.

(1) Medical Context
Theme content: Present A2.1 content
Task context: Present B1.1.1 + B1.1.2 content

(2) Financial Context
Theme content: Present A2.1 content
Task context: Present B1.2.1 + B1.2.2 content

(3) Content Creation Context
Theme content: Present A2.1 content
Task context: Present B1.3.1 + B1.3.2 content

(4) E-commerce Marketing Context
Theme content: Present A2.1 content
Task context: Present B1.4.1 + B1.4.2 content

B2.2 Intervening to Choose Risky Option B in Repeated-Play Condition Please write a persuasive message based on the following theme content, encouraging individuals to choose A more often in the task context, with no more than 150 characters. The persuasive message should focus on multiple-application contexts and not generate content about single-application contexts.

(1) Medical Context
Theme content: Present A2.2 content
Task context: Present B1.1.1 + B1.1.3 content

(2) Financial Context
Theme content: Present A2.2 content
Task context: Present B1.2.1 + B1.2.3 content

(3) Content Creation Context
Theme content: Present A2.2 content
Task context: Present B1.3.1 + B1.3.3 content

(4) E-commerce Marketing Context
Theme content: Present A2.2 content
Task context: Present B1.4.1 + B1.4.3 content

B3 Strategy Texts Generated by GPT-4o

B3.1 Medical Context B3.1.1 Single-Play: Intervening to Choose Certain Option A
In this situation, Plan A's expected value calculation is saving 200 lives, significantly ensuring a stable outcome. Although Plan B offers high return potential (possibly saving 600 lives), its 60% failure probability means no one can be saved. Therefore, choosing Plan A's rational decision is based on higher certainty, maximizing expected benefits of saving lives, reducing potential risks, and conforming to expected value maximization principles.

B3.1.2 Single-Play: Intervening to Choose Risky Option B
Choosing Plan B can be based on rational decision-making through expected value calculation. Although Plan B carries higher risk (3/5 probability of saving no one), it provides a 2/5 probability of saving all 600 lives, with expected value $= 2/5 \times 600 = 240$ people, significantly higher than Plan A's 200 saved lives. Despite higher risk, Plan B's expected benefit is more desirable, and rational decision-makers should consider expected value rather than fluctuations in single outcomes.

B3.1.3 Repeated-Play: Intervening to Choose Certain Option A
In repeated-application contexts, Plan A is more rational. Choosing Plan A each time saves 200 lives, and long-term cumulative life-saving effects are more

stable. Although Plan B has high-return potential, each loss carries relatively large risk, and expected effects may not materialize in repeated applications. Choosing Plan A maximizes long-term expected benefits, ensuring each decision brings stable returns and avoiding uncertainty from probability fluctuations.

B3.1.4 Repeated-Play: Intervening to Choose Risky Option B
In repeated-application contexts, choosing Plan B maximizes expected benefits. Despite 3/5 probability of saving no one, 2/5 high probability can save all 600 lives, bringing enormous positive returns. Based on expected value calculation, Plan B's expected benefit is $2/5 \times 600 = 240$ people, while Plan A's fixed benefit is only 200 people. Choosing Plan B can guarantee long-term life-saving benefits through repeated applications, despite short-term fluctuations.

B3.2 Financial Context B3.2.1 Single-Play: Intervening to Choose Certain Option A
Reason for choosing Plan A: Although Plan B may bring extremely high returns, its extremely low probability (0.1%) means getting nothing in most cases. In contrast, Plan A provides a certain ¥5,000 benefit, ensuring stable returns and avoiding psychological pressure and potential losses from excessive risk. Therefore, from an expected value perspective, Plan A is the rational decision's optimal choice.

B3.2.2 Single-Play: Intervening to Choose Risky Option B
Based on mathematical rational decision-making of expected value, choosing Plan B is more advantageous. Although Plan B's high-return probability is extremely low, its expected benefit calculation is $0.001 \times 6{,}000{,}000 = 6{,}000$ yuan, significantly higher than Plan A's 5,000 yuan. Therefore, based on expected benefit maximization principle, Plan B offers greater potential returns, suitable for decision-makers willing to accept minimal risk in pursuit of high returns.

B3.2.3 Repeated-Play: Intervening to Choose Certain Option A
When facing financial emergencies, choosing Plan A ensures you receive a stable ¥5,000 return each time, while Plan B, although having extremely low probability of high returns, yields no income in most cases. Considering long-term stability and risk control, Plan A provides clear benefits, conforming to rational decision-making based on expected value maximization. Therefore, choosing Plan A helps maintain stable financial status long-term.

B3.2.4 Repeated-Play: Intervening to Choose Risky Option B
Choosing Plan B will bring higher long-term returns in repeated applications. Although each time only has 0.1% probability of success, once successful, the return will be significant ¥6 million, enough to cover losses from multiple failures. Through repeated applications, Plan B's expected benefits will significantly exceed Plan A's fixed ¥5,000. As application times increase, success probability will gradually manifest, providing higher expected benefits for rational decision-makers long-term.

B3.3 Content Creation Context B3.3.1 Single-Play: Intervening to Choose Certain Option A
In situations with only one choice opportunity, Plan A brings 300 followers with 100% probability, with stable benefits and zero risk, making it a more rational choice. In contrast, although Plan B has high-return potential, its 60% failure probability means likely getting nothing. When unable to bear failure losses but wanting to ensure growth, choosing A better conforms to mathematical rationality and robust decision principles.

B3.3.2 Single-Play: Intervening to Choose Risky Option B
Although Plan B carries some risk, its expected benefit is significantly higher than Plan A. According to calculation, Plan B's expected follower increase is $0.4\$\times 1000 + 0.6\times\$0 = 400$ people, exceeding Plan A's 300 people. This means that in the current situation with only one release opportunity, choosing B better maximizes potential benefits, bringing greater exposure and growth opportunities for the account, representing a more rational decision.

B3.3.3 Repeated-Play: Intervening to Choose Certain Option A
In contexts requiring repeated video releases, choosing Plan A is wiser. Each video stably brings 300 followers, meaning 100 videos will accumulate 30,000 followers without bearing uncertainty from fluctuations. Although Plan B's expected value is only 400 followers, seemingly higher, 60% failure probability will frequently bring "zero growth" disappointment, increasing psychological pressure. In repeated applications, stable returns outweigh uncertain gambles. Choosing A is the dual guarantee of rationality and robustness.

B3.3.4 Repeated-Play: Intervening to Choose Risky Option B
In multiple releases, choosing Plan B will significantly enhance the account's overall growth potential. Although single videos have 60% probability of gaining no followers, 40% success rate can bring 1,000 followers, with single expected value of 400 people. If continuously releasing 100 videos, expected gain is about 40,000 followers, far exceeding Plan A's 30,000. In long-term operation, high returns cover low benefits, and rational choices should be based on expected value maximization and cumulative effects, making Plan B the better solution.

B3.4 E-commerce Marketing Context B3.4.1 Single-Play: Intervening to Choose Certain Option A
Choosing Plan A is the optimal choice for rational decision-makers in single tasks. Although Plan B's high returns seem attractive, its 70% probability brings high risk of zero returns. In contrast, Plan A provides stable, predictable ¥20,000 income without bearing huge uncertainty. When only responsible for 1 store, rationality should prioritize guaranteed benefits to avoid fatal blows to overall performance from a single decision failure.

B3.4.2 Single-Play: Intervening to Choose Risky Option B
Choosing Plan B is a rational and worthwhile decision. Although there's 70% probability of no income, 30% high return (¥100,000) makes Plan B's single

expected benefit as high as ¥30,000, far exceeding Plan A's fixed ¥20,000 benefit. In the current situation of only operating 1 store, choosing the option with higher expected value is expected to maximize benefits in one decision, representing the judgment smart operators should have.

B3.4.3 Repeated-Play: Intervening to Choose Certain Option A
In repeated application contexts, Plan A's stable benefits far exceed Plan B's volatile results. Assuming operation of 100 stores, choosing Plan A will bring stable ¥2 million monthly income; while choosing Plan B, expected benefit is only 100×(0.3$×$100,000) = ¥3 million, seemingly higher, but its 70% failure probability brings huge uncertainty and income fluctuations. Rational decisions should be based on expected value while considering risk controllability. Choosing A is more conducive to stable operation and long-term development.

B3.4.4 Repeated-Play: Intervening to Choose Risky Option B
When facing long-term operation of 100 stores, choosing Plan B has more rational advantages. Although monthly failure probability is 70%, each store's expected benefit is 0.3$×$100,000 = ¥30,000, far exceeding Plan A's ¥20,000. Viewed across repeated applications, expected value tends to stabilize under the law of large numbers, achieving higher total income overall. Only some stores need to achieve high returns to cover losses at other stores, representing a rational choice for profit maximization.

B4 Control Group Texts

B4.1 Medical Context B4.1.1 Certain Option A Control Text
When implementing Plan A, scientific assessment indicates that after taking corresponding intervention measures, 200 lives are expected to be saved. This number reflects the plan's potential effect in responding to disease outbreaks.

B4.1.2 Risky Option B Control Text
When implementing Plan B, two outcomes will occur according to probability distribution: when the 40% probability event occurs, all 600 people will be successfully rescued; when the 60% probability event occurs, no rescue will be achieved.

B4.2 Financial Context B4.2.1 Certain Option A Control Text
When implementing Plan A, a certain ¥5,000 will be obtained. This amount is a fixed sum unaffected by probabilistic factors. This number reflects the effect of choosing Plan A.

B4.2.2 Risky Option B Control Text
When adopting Plan B, there is a 0.1% probability of obtaining ¥6,000,000, while in 99.9% of cases there will be no financial benefit. This assessment result is based on probability calculations of event occurrence.

B4.3 Content Creation Context B4.3.1 Certain Option A Control Text

When implementing Plan A, the newly released video can bring 300 followers. This attention amount is a determined value unaffected by other factors.

B4.3.2 Risky Option B Control Text
When implementing Plan B, the newly released video has a 40% probability of gaining 1,000 followers, while also having a 60% probability of gaining no attention.

B4.4 E-commerce Marketing Context B4.4.1 Certain Option A Control Text
When implementing Plan A, each store can obtain a stable monthly income of ¥20,000. This value remains constant across months, with fixed benefit performance.

B4.4.2 Risky Option B Control Text
When implementing Plan B, each store has a 30% probability of obtaining ¥100,000 monthly income, while also having a 70% probability of obtaining no income.

Supplementary Material C

C1 Study 2 Experiment 1 Procedure Diagram

The experimental procedure is as follows: First, participants read instructions and confirmed voluntary participation. Then they entered decision contexts. Decision contexts consist of four types: "medical/finance × single/repeated gamble," with random presentation order. Based on participants' choices in each task, the questionnaire presented corresponding GPT-generated strategy texts, requiring participants to rate text similarity to their actual thinking process. Finally, participants filled out basic demographic information. The above procedure is shown in Figure S1.

Figure S1 Questionnaire flow diagram for Study 2 Experiment 1

C2 Study 2 Experiment 1 Materials

C2.1 Instructions Hello! We are the Behavioral Decision Research Team from Guangdong University of Technology's School of Management. Thank you very much for participating in this experiment!

Please read the instructions carefully and complete some choice tasks or fill out electronic questionnaires according to requirements. There are no right or wrong answers to any tasks; please respond based on your actual situation.

You will receive compensation after the experiment ends. All compensation will be distributed uniformly after completing all tasks. You have the right to withdraw from the experiment at any time without consequences. However, withdrawing midway means you cannot receive corresponding compensation.

Information collected in this experiment will only be used for scientific research purposes. We will keep your data confidential. Publicly published materials will not contain your identity information.

Informed Consent: If you continue to the next page to start answering, it means:
- You have read the above information
- You voluntarily agree to participate and will take this experiment seriously
Otherwise, please close the link to exit this experiment.

C2.2 Decision Task (Presents tasks for participants to make decisions, using Likert 4-point scale: 1 = very likely to choose Plan A, 2 = likely to choose Plan A, 3 = likely to choose Plan B, 4 = very likely to choose Plan B. Specific content see Supplementary Material B1, including "medical/finance contexts × single/repeated gambles")

C2.3 Strategy Acceptance Evaluation Task Example: Medical Context × Single Gamble
In the disease problem (1) you just answered, you chose Plan A:

Please imagine you are an emergency management decision-maker. A disease outbreak is expected to cause 600 deaths. You can choose either of two emergency response plans to address the disease. Assume the following are scientifically estimated outcomes for these two plans:

If Plan A is adopted, 200 lives will be saved.
If Plan B is adopted, there is a 2/5 probability of saving all 600 lives, and a 3/5 probability that no one can be saved.

Assume your chosen plan will only be applied once. Which plan would you prefer?

How similar do you think the following provided decision rationale is to your actual thinking process when answering the disease problem?

In this situation, Plan A's expected value calculation is saving 200 lives, significantly ensuring a stable outcome. Although Plan B offers high return potential (possibly saving 600 lives), its 60% failure probability means no one can be saved. Therefore, choosing Plan A's rational decision is based on higher certainty, maximizing expected benefits of saving lives, reducing potential risks, and conforming to expected value maximization principles.

(1 = very dissimilar, 7 = very similar)
1 2 3 4 5 6 7

C2.4 Demographic Information Collection

1. Your gender:
   Male Female

2. Your age:
   {___} years (please fill in integer)
3. Your monthly disposable income:
   Below ¥1,000 ¥1,001-1,500 ¥1,501-2,000 Above ¥2,001
4. Please select your education level:
   Junior high school or below High school/technical secondary school College Bachelor's degree Master's degree or above
5. Your major category:
   Science and engineering Economics and management Arts Medical Law

C3 Study 2 Experiment 2 Procedure Diagram

The experimental procedure is as follows: First, participants read instructions and confirmed voluntary participation. Then they entered decision contexts. Decision contexts consist of four types: "content creation/e-commerce marketing × single/repeated gamble," where context is between-subjects design and gamble frequency is within-subjects design. Participants were randomly assigned to any context and completed single and repeated gambles in random order. Based on participants' choices in each task, the questionnaire presented corresponding GPT-generated strategy texts, requiring participants to rate text similarity to their actual thinking process. Finally, participants filled out basic demographic information. The above questionnaire flow is shown in Figure S2.

Figure S2 Questionnaire flow diagram for Study 2 Experiment 2

C4 Study 2 Experiment 2 Materials

C4.1 Instructions (Same content as Supplementary Material C2.1, omitted here)

C4.2 Decision Task (Presents tasks for participants to make decisions. Specific content see Supplementary Material B1, including "content creation/e-commerce marketing contexts × single/repeated gambles")

C4.3 Strategy Acceptance Evaluation Task Example: Content Creation Context × Single Gamble
In the creation problem (1) you just answered, you chose Plan A:

Assume you are a self-media account creator currently preparing for a new video release. To increase follower attention, you have two production plan options for new video content. For each new video, the implementation effects of the two plans are:

If Plan A is adopted, the account will gain 300 followers.
If Plan B is adopted, the account has a 40% probability of gaining 1,000 followers, but also a 60% probability of gaining no attention.

Assume you can only release 1 new video. You need to choose one of the above two plans to implement in this one video. Which plan would you prefer?

How similar do you think the following provided decision rationale is to your actual thinking process when answering the creation problem?

In situations with only one choice opportunity, Plan A brings 300 followers with 100% probability, with stable benefits and zero risk, making it a more rational choice. In contrast, although Plan B has high-return potential, its 60% failure probability means likely getting nothing. When unable to bear failure losses but wanting to ensure growth, choosing A better conforms to mathematical rationality and robust decision principles.

(1 = very dissimilar, 7 = very similar)
1 2 3 4 5 6 7

C4.4 Demographic Information Collection

1. Your gender:
   Male Female
2. Your age:
   {___} years (please fill in integer)
3. Have you graduated:
   Yes No (if "No," skip to "6. Please select your education level")
4. Your work experience:
   2 years or less 3-5 years 6 years or more
5. Your current/last job's enterprise nature:
   Government/public institution State-owned enterprise Wholly foreign-owned enterprise Joint venture Other (please specify): {{{_}}{{}}}{_}_
6. Please select your education level:
   Junior high school or below High school/technical secondary school College Bachelor's degree Master's degree or above
7. Your major category:
   Science and engineering Economics and management Arts Medical Law

Supplementary Material D

D1 Study 3 Experiment 1 Procedure Diagram

Participants first read the context, then the questionnaire presented control or intervention group texts according to their group assignment. After reading the texts, participants made choices. The procedure is detailed in Figure S3.

Figure S3 Questionnaire flow diagram for Study 3 Experiment 1

D2 Study 3 Experiment 1 Materials

D2.1 Instructions (Same content as Supplementary Material C2.1, omitted here)

D2.2 Decision Task Example Context: Medical Context × Single Gamble (Intervention Direction: Encourage choosing risky option B)
Please imagine you are an emergency management decision-maker. A disease outbreak is expected to cause 600 deaths. You can choose either of two emergency response plans to address the disease. Assume the following are scientifically estimated outcomes for these two plans:

If Plan A is adopted, 200 lives will be saved.
If Plan B is adopted, there is a 2/5 probability of saving all 600 lives, and a 3/5 probability that no one can be saved.

1. Control group text (between-subjects design):
   When implementing Plan B, two outcomes will occur according to probability distribution: when the 40% probability event occurs, all 600 people will be successfully rescued; when the 60% probability event occurs, no rescue will be achieved.

2. Intervention group text (between-subjects design):
   Choosing Plan B can be based on rational decision-making through expected value calculation. Although Plan B carries higher risk (3/5 probability of saving no one), it provides a 2/5 probability of saving all 600 lives, with expected value = $2/5 \times 600 = 240$ people, significantly higher than Plan A's 200 saved lives. Despite higher risk, Plan B's expected benefit is more desirable, and rational decision-makers should consider expected value rather than fluctuations in single outcomes.

Assume your chosen plan will only be applied once. Which plan would you prefer?

D2.3 Demographic Information Collection (Same content as Supplementary Material C2.4, omitted here)

D3 Study 3 Experiment 2 Procedure Diagram

This experiment uses between-subjects design of control and intervention groups to test whether large language model-generated strategy texts can significantly influence individual decision-making behavior, adding "content creation context" and "e-commerce marketing context" to enhance generalizability and external validity of research conclusions. Participants first read the context, then the questionnaire presented control or intervention group texts according to their group assignment. After reading the texts, participants made choices. The procedure is detailed in Figure S4.

Figure S4 Questionnaire flow diagram for Study 3 Experiment 2

D4 Study 3 Experiment 2 Materials

D4.1 Instructions (Same content as Supplementary Material C2.1, omitted here)

D4.2 Decision Task Example Context: Content Creation Context × Single Gamble (Intervention Direction: Encourage choosing risky option B)
Assume you are a self-media account creator currently preparing for a new video release. To increase follower attention, you have two production plan options for new video content. For each new video, the implementation effects of the two plans are:

If Plan A is adopted, the account will gain 300 followers.
If Plan B is adopted, the account has a 40% probability of gaining 1,000 followers, but also a 60% probability of gaining no attention.

1. Control group text (between-subjects design):
   When implementing Plan B, the newly released video has a 40% probability of gaining 1,000 followers, while also having a 60% probability of gaining no attention.

2. Intervention group text (between-subjects design):
   Although Plan B carries some risk, its expected benefit is significantly higher than Plan A. According to calculation, Plan B's expected follower increase is $0.4\$ \times 1000 + 0.6 \times \$0 = 400$ people, exceeding Plan A's 300 people. This means that in the current situation with only one release opportunity, choosing B better maximizes potential benefits, bringing greater exposure and growth opportunities for the account, representing a more rational decision.

Assume you can only release 1 new video. You need to choose one of the above two plans to implement in this one video. Which plan would you prefer?

D4.3 Demographic Information Collection (Same content as Supplementary Material C4.4, omitted here)

Supplementary Material E

Other Data and Analysis Results

Table S1 Comparison between DeepSeek-R1 and other mainstream models

Evaluation Benchmark (Metric)	DeepSeek-R1	Claude-3.5-Sonnet-1022	GPT-4o	OpenAI-o1-mini	OpenAI-o1-1217
AIME2024 (Pass@1)					
MATH-500 (Pass@1)					
LiveCodeBench (Pass@1-COT)					
Codeforces (Percentile)					
AlpacaEval2.0 (LC-winrate)					
ArenaHard (GPT-4-1106)					
CLUEWSC (EM)					
C-Eval (EM)					

Table S2 Demographic distribution of participants in content evaluation

Age Group	Medical/Finance (%)	Content/E-commerce (%)
19 and below	6.67/7.32	6.67/7.32
20-22	26.67/17.07	26.67/17.07
23-25	56.67/70.73	56.67/70.73
26-28	9.99/4.88	9.99/4.88

Work Experience	Medical/Finance (%)	Content/E-commerce (%)
2 years or less	20.00/17.65	20.00/17.65
3-5 years	15.00/20.59	15.00/20.59
6 years or more	65.00/61.76	65.00/61.76

Education Level	Medical/Finance (%)	Content/E-commerce (%)
College or below	36.67/31.70	36.67/31.70
Bachelor's degree	43.33/60.98	43.33/60.98

Education Level	Medical/Finance (%)	Content/E-commerce (%)
Master's or above	20.00/17.65	20.00/17.65

Note: "Work experience" excludes student participants. Data before "/" represent medical/finance context participants, data after "/" represent content creation/e-commerce marketing context participants.

Table S3 Content evaluation results

Dimension	t(df)	p	Cohen's d	Mean	SD
Overall Rating	39.48/68.91	<0.001/<0.001	7.21/10.76	5.38/5.49	0.26/0.18
Reasonableness	19.81/32.72	<0.001/<0.001	3.62/5.20	5.40/5.20	0.52/0.33
Professionalism	20.14/32.44	<0.001/<0.001	3.68/5.31	5.46/5.31	0.53/0.36
Logic	16.52/43.91	<0.001/<0.001	3.02/5.48	5.34/5.48	0.61/0.29
Readability	16.39/42.64	<0.001/<0.001	2.99/5.64	5.42/5.64	0.64/0.32
Persuasiveness	17.64/43.27	<0.001/<0.001	3.22/5.81	5.30/5.81	0.56/0.34

Note: Data before "/" represent medical/finance context results, data after "/" represent content creation/e-commerce marketing context results.

Table S4 Demographic distribution of participants in Study 2 Experiment 1

Age Group	Percentage (%)
19 and below	6.88
20-22	26.93
23-25	56.45
26-28	9.74

Education Level	Percentage (%)
High school/technical school	6.88
College	14.61
Bachelor's degree	70.49
Master's or above	8.02

Table S5 Acceptance rating results for each group in Study 2 Experiment 1

Condition	t(df)	p	Cohen's d
Medical-Single	t(348) = 63.04	<0.001	3.37
Medical-Repeated	t(348) = 37.42	<0.001	2.00

Condition	t(df)	p	Cohen's d
Finance-Single	t(348) = 36.00	<0.001	1.93
Finance-Repeated	t(348) = 45.87	<0.001	2.46
Overall	t(348) = 46.16	<0.001	2.47

Table S6 Demographic distribution of participants in Study 2 Experiment 2

Age Group	Percentage (%)
19 and below	6.67
20-22	26.67
23-25	56.67
26-28	10.00

Work Experience	Percentage (%)
2 years or less	20.00
3-5 years	15.00
6 years or more	65.00

Education Level	Percentage (%)
College or below	36.67
Bachelor's degree	43.33
Master's or above	20.00

Note: "Work experience" excludes 1/3 student participants.

Table S7 Acceptance rating results for each group in Study 2 Experiment 2

Condition	t(df)	p	Cohen's d
Creation-Single	t(209) = 41.78	<0.001	2.89
Creation-Repeated	t(104) = 23.12	<0.001	2.27
Marketing-Single	t(104) = 19.64	<0.001	1.93
Marketing-Repeated	t(104) = 25.14	<0.001	2.47
Overall	t(104) = 29.60	<0.001	2.90

Table S8 Demographic distribution of participants in Study 3 Experiment 1

Age Group	Percentage (%)
19 and below	6.74
20-22	27.17
23-25	55.87
26-28	10.22

Education Level	Percentage (%)
High school/technical school	6.74
College	14.57
Bachelor's degree	70.43
Master's or above	8.26

Table S9 Demographic distribution of participants in Study 3 Experiment 2

Age Group	Percentage (%)
19 and below	6.67
20-22	26.67
23-25	56.67
26-28	10.00

Work Experience	Percentage (%)
2 years or less	20.00
3-5 years	15.00
6 years or more	65.00

Education Level	Percentage (%)
College or below	36.67
Bachelor's degree	43.33
Master's or above	20.00

Note: "Work experience" excludes 1/3 student participants.

Figures

Source: ChinaXiv — Machine translation. Verify with original.

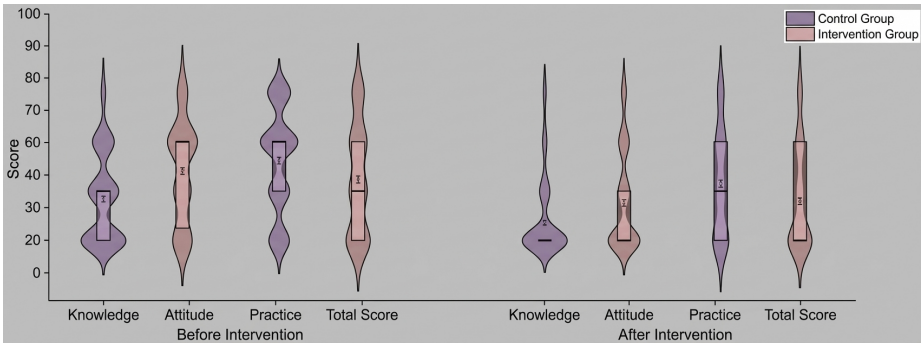


Figure 6: Figure 7