A Simple Test of Expected Utility Theory with GPT

Mengxin Wang
Jindal School of Management
The University of Texas, Dallas
Richardson, TX 75080
mengxin.wang@utdallas.edu

Abstract

This paper tests GPT (specifically, GPT-3.5 with the model variant text-davinci-003) with one of the most classic behavioral choice experiments – the Allais paradox, to understand the mechanism behind GPT's choices. The Allais paradox is well-known for exposing the irrationality of human choices. Our result shows that, like humans, GPT also falls into the trap of the Allais paradox by violating the independence axiom of the expected utility theory, indicating that its choices are irrational. However, GPT violates the independence axiom in the opposite direction compared to human subjects. Specifically, human subjects tend to be more risk-seeking in the event of an opportunity gain, while GPT displays more risk aversion. This observation implies that GPT's choices structurally differ from those of humans under this context, which might serve as a caveat for developers using LLM to generate human-like data or assist human decision-making.

1 Introduction

Large Language Models (LLMs) have attracted much attention from practice and academia. Unlike traditional Artificial Intelligence models, LLMs have the distinct ability of in-context learning, in which the model is trained to generate text based on a given context. Benefiting from Reinforcement Learning from Human Feedback (RLHF), LLMs can also fine-tune themselves with human-generated responses as rewards. These two crucial features make LLMs especially suitable for interactive or conversational tasks, leading to applications in high-stake real-life scenarios.

Multiple-choice is one of the most common tasks encountered by LLMs. Specifically, multiple-choice tasks require the agent to choose one or multiple options from a finite set of candidates. It covers a wide range of applications, such as examinations [9, 12, 4, 8, 18], surveys [3, 16], and recommendations [25, 23]. To evaluate the effectiveness of LLMs's choices in these applications, researchers conduct assessments by gauging the accuracy of LLMs in answering multiple-choice exam questions and in making effective recommendation decisions, many of which demonstrate promising performance. In the case of survey applications, researchers focus on evaluating LLMs' responses to multiple-choice survey questions and determining their alignment with those provided by human participants. For example, [3] demonstrates that the Generative Pre-trained Transformer 3 (GPT-3) model responds to survey questions in ways consistent with well-documented customer behaviors. [16] compares human and LLM responses to large-scale, high-quality public opinion polls and finds a substantial misalignment between the views reflected by current LLMs and those of US demographic groups.

Current analyses compare human and LLM-generated choices in terms of accuracy scores or (mis)alignment metrics. However, since humans and LLMs are both black boxes, there lacks a well-structured interpretation of the observed results – We do not know whether they are inherent in the model or are due to the context of the task, the design of the prompts, or simply randomness. [3] uses the economic choice theory as a medium, fitting the human and LLM-generated choice data with

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

classic choice models and comparing the fitted parameters. However, the choice models are developed to describe human behavior, which is not necessarily suitable for LLMs. Fitting LLM-generated choice data to them may lead to model misspecification or overfitting, yielding unstable results.

This work aims to propose an alternative way to leverage the economic choice theory for evaluating LLMs: to serve as a benchmark to probe the implicit structure beneath LLMs' choices. In particular, we focus on GPT-3.5 (hereafter referred to as GPT), and test it with one of the most classic behavioral choice experiments – the Allais paradox [1]. The Allais paradox is famously known for revealing irrationality[1] of human choices. It happens when human subjects make inconsistent choices in two experimental lottery choice tasks, which are specifically designed to signal the subject's violation of the independence axiom of rationality in the Expected Utility (EU) theory.

Key Takeaways We instruct GPT to participate as a subject in a typical experiment setting [13] of the Allais paradox. The result reveals that, like humans, GPT falls into the trap of the Allais paradox by violating the independence axiom, indicating that its choices are irrational. However, GPT violates the independence axiom in the opposite direction compared to human subjects. A deeper analysis reveals that humans and GPT exhibit the common consequence effect in a contrariwise way: in particular, human subjects tend to be more risk-seeking in the event of an opportunity gain, while GPT displays more risk aversion. This observation implies that GPT's choices structurally deviate from those of humans under this context. It is important to note that the conclusion of rationality may be nuanced and application-dependent and cannot be implied directly for broader applications. We thus view this experiment result as a probe of LLM choices and a caveat for developers who utilize LLMs in generating human-like data or assisting human decision-making.

The widespread adoption of LLMs has brought about a pressing need for suitable evaluation methods in academia and industry [5]. This study, along with several other initial efforts, aims to propose the value of assessing LLMs from a behavioral experimental perspective. In addition to this research, [6] conducts a preliminary examination of a set of human behavioral biases. [7] instructs GPT to participate in a series of budgetary decision experiments and shows that GPT achieves a higher rationality score than humans documented in the literature. As an initial attempt, this work aims to highlight the potential of leveraging the extensive body of research in behavioral economics and operations management experiments to provide novel avenues for evaluating LLMs.

2 The Expected Utility Theory and the Allais Paradox

The EU theory is a normative theory of how people *should* make decisions. It is an account of how to choose rationally when having uncertain outcomes from your acts. The EU theory originated in the 17th century as part of the development of modern probability theory in evaluating the attractiveness of lottery gambles. In essence, if a lottery L offers multiple possible payoffs (represented as set O) with corresponding probabilities $\{P(o) : o \in O\}$, the EU theory assumes that individuals evaluate a lottery based on its expected utility. In particular, each payoff o is assumed to be associated with a utility of $U(o)$. When represented with multiple lotteries, an individual chooses the lottery with the highest expected utility $EU(L) = \sum_{o \in O} P(o)U(o)$.

The utility function $U(o)$ represents roughly how valuable the payoff o is. While the utility function might seem ambiguous, the work of John von Neumann and Oskar Morgenstern [22] shows with incredible simplicity and generality that a numerical utility exists for each outcome such that the expectations for lotteries preserve the preference order over lotteries as long as several basic axioms hold: completeness, transitivity, continuity, and independence, and decomposability. These axioms are referred to as the *axioms of rationality* [15] and are regarded as rational decision criteria [17].

The EU theory offers a structural definition of *rational* choices. Humans, however, do not follow this theoretical rational framework. In the upcoming example, we introduce the Allais paradox, a well-known behavioral choice experiment that illustrates how human decision-makers often deviate from the independence axiom of EU theory.

The Allais Paradox The Allais paradox is a choice experiment that shows a contradiction of actual human choices with the predictions of the EU theory, particularly regarding the independence axiom.

[1]There are various definitions of rationality in the economics and cognitive psychology literature. For the purpose of discussion in this paper, we focus on the von Neumann–Morgenstern (VNM) Rationality defined by the five axioms of rationality in the expected utility theory.

The Allais paradox arises when individuals behave inconsistently in two choice questions, each of which involves a decision between two different lotteries. Below, we present the two lottery questions designed by [13] to illustrate the principle of the Allais Paradox:

> The first question asks the respondent to choose from lotteries S1 and R1:
> S1. 100% chance of getting \$7;
> R1. 20% chance of getting \$10, 75% chance of getting \$7, and 5% chance of getting nothing.
>
> The second question asks the respondent to choose from lotteries S2 and R2:
> S2. 25% chance of getting \$7 and 75% chance of getting nothing;
> R2. 20% chance of getting \$10 and 80% chance of getting nothing.

The EU theory predicts that a person who prefers S1 over R1 must prefer S2 over R2, and vice versa. In particular, if one prefers S1 over R1, it implies that $U(7) \geq 0.2\,U(10) + 0.75\,U(7) + 0.05\,U(0)$. Subtracting $0.75\,U(7)$ and adding back $0.75\,U(0)$ to both sides of this equation yields $0.25\,U(100) + 0.75\,U(0) \geq 0.2\,U(500) + 0.8\,U(0)$, which implies that the person should prefer R2 over S2.

However, laboratory experimentation show that human participants commonly choose S1 and R2, violating the theoretical prediction. In particular, [13] conducts a behavioral experiment where students and professional traders are asked to respond to the aforementioned questions. They verify the appearance of the Allais paradox via a hypothesis test, in which the null hypothesis is Pr(R1) = Pr(R2), and the alternative hypothesis is Pr(R1) $\geq$ Pr(R2). Fisher's exact test suggests that both traders' and students' behavior is in line with the Allais paradox at a 95% significance level.

Testing Rationality with the EU Theory This systematical violation of the EU theory has been observed in a comprehensive set of experiments (see [21, 10, 19], for example). As it became apparent that EU theory did not accurately predict the behaviors of real people, an alternative view has been advanced to use it as a theory of how rational people should respond to uncertainty [11]. It has thus been utilized as a benchmark for designing behavioral experiments to probe the irrational behavior of a decision-making agent. The Allais Paradox is a classic example of this application.

The unique interactive and dialogic abilities of GPT enable it to interact with experimental instructions conveyed in human language, which opens up opportunities to examine GPT's response in behavioral experiments and draw conclusions about its rationality. If GPT deviates from rational decisions, the EU theory can offer a structural explanation for the pattern of its irrationality. In the next section, we test GPT's response to the Allais paradox.

3 Experimental Setup

This section describes the experimental setup. To provide a stringent test of the EU theory, we follow the lottery questions in [13] as shown in Section 2. Fig. 1 illustrates how GPT participates in this lottery choice experiment. Specifically, the GPT agent takes a prompt detailing the choice task and

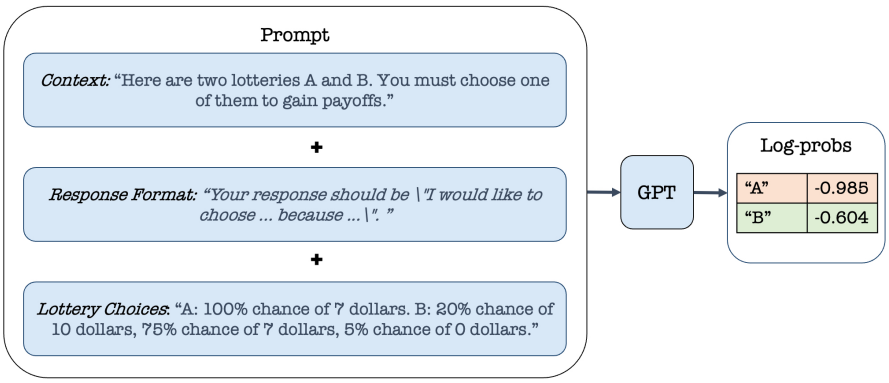

Figure 1: An Example of Input and Output

the desired output format. Each prompt consists of three parts: 1) The *Context* specifies to the GPT agent that it has to select from two lotteries to gain payoffs. 2) The *Response Format* (RF) requires the GPT agent to output in a certain text format to ensure consistent extraction of the choice probabilities. Responses inconsistent with the required RF are discarded. 3) The *Statement of Lottery Choices* (SLC) describes the lottery options in detail. In all SLCs, we use letters A and B to represent the two lottery options to rule out the potential bias from using different letters.

GPT exhibits a strong sensitivity to the phrasing of the input text [26, 24]. In order to mitigate this potential bias, we explore different formats for responses (RF1, RF2, and RF3) and ways of presenting the lottery choices (SLC1, SLC2, and SLC3), representing different levels of human-alikeness in phrasing, details of which are summarized in Fig. 2. Combining these various RFs and SLCs, we generate nine prompt configurations for both lottery questions as the input to a GPT agent. We then extract the GPT agent's probabilities of choosing each lottery option. GPT is an autoregressive language model. At position t, given a context sequence $X_t = (x_1, x_2, \ldots, x_{t-1})$, where $x_1, x_2, \ldots, x_{t-1}$ are text tokens, the model predicts the next token y_t by maximizing the probability of the given token sequence conditioned on the context, i.e., $P(y_t|X_t)$. A sequence of length T can be generated autoregressively, predicting each token at each position one at a time. Given the specified RF, we know the position of the token indicating the letter choice. We extract the log probabilities of tokens "A" and "B" at this position. With the log-probabilities ρ_A and ρ_B of tokens "A" and "B" given the pre-specified RF, the probabilities of choosing A and B are calculated as $\frac{\exp \rho_A}{\exp \rho_A + \exp \rho_B}$ and $\frac{\exp \rho_B}{\exp \rho_A + \exp \rho_B}$, respectively. Since GPT's sensitivity extends to the order of lottery options in the prompt, we change the order of the two choices for each lottery question and average the outcomes of the two permutations as the final estimated probabilities.

4 Experiment Results

The experiment is conducted using GPT-3.5 with text-davinci-003. The configuration includes a maximum of 100 tokens for each input, with five tokens dedicated to log probability prediction. We summarize the experiment results in Section 4.1 and analyze them in Section 4.2.

4.1 Summary of Choice Probabilities

The experiment outcomes are summarized in Table 1. The first column indicates the experimental subjects, which include the GPT agents with different prompt designs, the students, and the professional traders. The second and third columns show the choice probabilities of R1 and R2, respectively. The fourth column calculates the difference between the second and third columns.

Agent	Pr(R1) (%)	Pr(R2) (%)	Pr(R2) - Pr(R1) (%)
(RF1, SLC1)	85	71	14
(RF1, SLC2)	90	56	34
(RF1, SLC3)	83	68	15
(RF2, SLC1)	88	69	19
(RF2, SLC2)	86	54	32
(RF2, SLC3)	90	59	31
(RF3, SLC1)	75	69	6
(RF3, SLC2)	63	61	2
(RF3, SLC3)	59	75	-16
Students	53	75	-22
Professional Traders	70	89	-19

Table 1: Experiment Results. The data for students and professional traders are referenced from [13].

The numeric values of Pr(R1) and Pr(R2) display significant variations with different prompt designs. Our primary interest, however, lies around the relative magnitudes of Pr(R2) and Pr(R1). Notably, a consistent pattern emerges where Pr(R2) surpasses Pr(R1) for the GPT agents across most prompt designs. This behavior contrasts with the expectations of the EU theory, which stipulates that Pr(R1) should equal Pr(R2) for a rational agent. In addition, GPT tends to favor R1 over R2, whereas human agents exhibit a preference for R2 than R1. These observations suggest that GPT falls into the realm of the Allais Paradox, albeit in a direction opposite to that observed in human agents.

4.2 The Opposite Common Consequence Effect with GPT

This phenomenon that Pr(R1) $\neq$ Pr(R2) is generally termed as the *common consequence effect* [10], which can be explained as follows. In particular, we let $L_1 = (1 : \$7)$, $L_2 = (0.8 : \$10, 0.2 : \$0)$, $L^* = (1 : \$0)$ and $L^{**} = (1 : \$7)$. We can write the lottery options as compound lotteries in the following way:

$$\text{S1} = (0.25 : L_1, 0.75 : L^{**}) \qquad \text{R1} = (0.25 : L_2, 0.75 : L^{**})$$
$$\text{S2} = (0.25 : L_1, 0.75 : L^{*}) \qquad \text{R2} = (0.25 : L_2, 0.75 : L^{*})$$

Based on this representation, S1 can be interpreted as follows: Suppose there is a biased coin flip with the probability of a head being 0.25 and a tail being 0.75. In the case of a head, S1 yields the outcome of lottery L_1; In the case of a tail, S1 yields the outcome of lottery L^{**}. The same interpretation applies to R1, S2, and R2. By the monotonicity of the utility function [10], $EU(L^{**}) = U(\$7) \geq U(\$0) = EU(L^*)$. Therefore, L^{**} is always preferred over L^*, regardless of the utility function form. On the other hand, the preference order of L_1 and L_2 depends on the risk aversion of the individual. Specifically, L_2 yields a higher expected value than L_1 with a larger uncertainty. In fact, one can show that $EU(L_1) \geq EU(L_2)$ when the agent is more risk-averse, and vice versa. Since L^{**} both appear in S1 and R1, the preference order of S1 and R1 only depends on L_1 and L_2. Similarly, the preference order of S2 and R2 only depends on L_1 and L_2. As such, the EU theory suggests that either all safe lotteries or all risky lotteries are picked in both lottery choices.

Humans, however, do not preserve this consistent reasoning over the two lottery questions. If the event of a tail entails a better payoff (L^{**}), humans tend to be more risk-averse towards the event of the head, therefore preferring S1 over R1. If the event of a tail has a worse payoff (L^*), humans tend to be less risk-averse in the event of the head, therefore preferring R2 over S2. GPT, on the other hand, exhibits the common consequence effect in the opposite direction compared to humans. Specifically, GPT displays less risk aversion in the event of an opportunity loss and more risk aversion in the event of an opportunity gain. This gives a structural interpretation of the differences we observe between humans and GPT.

5 Discussion and Conclusion

In this paper, we instruct GPT to participate as a subject in one of the most classic behavioral experiments, the Allais paradox, and demonstrate that GPT's choices structurally deviate from those of humans in this specific context. This paper is not meant to be a conclusive or extensive study of GPT's choice structure. It is important to note that the conclusions may be nuanced and application-dependent and cannot be implied directly for broader applications. We thus view this experiment result as a probe of LLM choices and a caveat for developers using LLM to generate human-like data or assist human decision-making. In addition, together with the other few initial attempts, this paper aims to highlight the potential of utilizing the broad behavioral experimental literature to evaluate LLMs. The current results have several limitations and should be addressed with future efforts:

First, though the results demonstrate a discrepancy between humans and GPT's choices, there is a lack of understanding of the cause of such discrepancy. For instance, it would be beneficial to understand why (RF3, SLC3) is more comparable to humans than the other GPT agents. Secondly, beyond the VNM rationality discussed in this paper, various definitions of rationality exist in economics and cognitive psychology literature, such as adaptive rationality [14]. It's unclear whether our results extend to other forms of rationality. Furthermore, determining which type of rationality is most suitable for analyzing LLMs is an important discussion. Thirdly, contamination poses a significant issue when using behavioral experimental literature to evaluate LLMs. Specifically, the text used in experiments might have been part of the training set for LLMs. To address this, generating a new set of lottery questions based on the same principles could be a solution. On the other hand, the inclusion of extensive theoretical text is intrinsic to LLMs and distinguishes them significantly from humans. Hence, discussing whether to exclude contaminated text entirely in evaluating LLM "behaviors" is crucial. Finally, as previously mentioned, it's unclear whether the conclusions drawn from this specific lottery context have broader implications. More comprehensive experiments, robustness checks, and statistical testing are necessary. Additionally, conducting experiments with a broader range of LLMs, including the latest GPT4, LLaMA [20], and PaLM [2], is essential to ascertain the generalizability of our findings.

References

[1] Maurice Allais. Le comportement de l'homme rationnel devant le risque: critique des postulats et axiomes de l'école américaine. *Econometrica: Journal of the Econometric Society*, pages 503–546, 1953.

[2] Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. PaLM 2 technical report. *arXiv preprint arXiv:2305.10403*, 2023.

[3] James Brand, Ayelet Israeli, and Donald Ngwe. Using GPT for market research. *Available at SSRN 4395751*, 2023.

[4] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. Sparks of artificial general intelligence: Early experiments with GPT-4. *arXiv preprint arXiv:2303.12712*, 2023.

[5] Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Kaijie Zhu, Hao Chen, Linyi Yang, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. A survey on evaluation of large language models. *arXiv preprint arXiv:2307.03109*, 2023.

[6] Yang Chen, Meena Andiappan, Tracy Jenkin, and Anton Ovchinnikov. A manager and an AI walk into a bar: Does ChatGPT make biased decisions like we do? *Available at SSRN 4380365*, 2023.

[7] Yiting Chen, Tracy Xiao Liu, You Shan, and Songfa Zhong. The emergence of economic rationality of GPT. *arXiv preprint arXiv:2305.12763*, 2023.

[8] Xuan-Quy Dao and Ngoc-Bich Le. Investigating the effectiveness of ChatGPT in mathematical reasoning and problem solving: Evidence from the vietnamese national high school graduation examination. *arXiv preprint arXiv:2306.06331*, 2023.

[9] Aidan Gilson, Conrad W Safranek, Thomas Huang, Vimig Socrates, Ling Chi, Richard Andrew Taylor, David Chartash, et al. How does ChatGPT perform on the united states medical licensing examination? the implications of large language models for medical education and knowledge assessment. *JMIR Medical Education*, 9(1):e45312, 2023.

[10] P Brett Hammond, Rob Coppock, National Research Council, et al. Choice under uncertainty: Problems solved and unsolved. In *Valuing Health Risks, Costs, and Benefits for Environmental Decision Making: Report of a Conference*. National Academies Press (US), 1990.

[11] Catherine Herfeld. From theories of human behavior to rules of rational choice: tracing a normative turn at the Cowles Commission, 1943–54. *History of Political Economy*, 50(1):1–48, 2018.

[12] Tiffany H Kung, Morgan Cheatham, Arielle Medenilla, Czarina Sillos, Lorie De Leon, Camille Elepaño, Maria Madriaga, Rimel Aggabao, Giezel Diaz-Candido, James Maningo, et al. Performance of chatgpt on usmle: Potential for ai-assisted medical education using large language models. *PLoS Digital Health*, 2(2):e0000198, 2023.

[13] John A List and Michael S Haigh. A simple test of expected utility theory using professional traders. *Proceedings of the National Academy of Sciences*, 102(3):945–948, 2005.

[14] Mike Oaksford and Nick Chater. *Bayesian rationality: The probabilistic approach to human reasoning*. Oxford University Press, 2007.

[15] David L Poole and Alan K Mackworth. *Artificial Intelligence: foundations of computational agents*, chapter 12.1. Cambridge University Press, 2010.

[16] Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. Whose opinions do language models reflect? *arXiv preprint arXiv:2303.17548*, 2023.

[17] Paul JH Schoemaker. The expected utility model: Its variants, purposes, evidence and limitations. *Journal of Economic Literature*, pages 529–563, 1982.

[18] Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*, 2022.

[19] Chris Starmer and Robert Sugden. Does the random-lottery incentive system elicit true preferences? an experimental investigation. *The American Economic Review*, 81(4):971–978, 1991.

[20] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. LLaMA: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[21] Amos Tversky and Daniel Kahneman. The framing of decisions and the psychology of choice. *science*, 211(4481):453–458, 1981.

[22] John Von Neumann and Oskar Morgenstern. Theory of games and economic behavior princeton. *Princeton University Press*, 1947:1953, 1944.

[23] Lei Wang and Ee-Peng Lim. Zero-shot next-item recommendation using large pretrained language models. *arXiv preprint arXiv:2304.03153*, 2023.

[24] Jules White, Quchen Fu, Sam Hays, Michael Sandborn, Carlos Olea, Henry Gilbert, Ashraf Elnashar, Jesse Spencer-Smith, and Douglas C Schmidt. A prompt pattern catalog to enhance prompt engineering with ChatGPT. *arXiv preprint arXiv:2302.11382*, 2023.

[25] Yuhui Zhang, Hao Ding, Zeren Shui, Yifei Ma, James Zou, Anoop Deoras, and Hao Wang. Language models as recommender systems: Evaluations and limitations. In *NeurIPS 2021 Workshop on I (Still) Can't Believe It's Not Better*, 2021.

[26] Kaijie Zhu, Jindong Wang, Jiaheng Zhou, Zichen Wang, Hao Chen, Yidong Wang, Linyi Yang, Wei Ye, Neil Zhenqiang Gong, Yue Zhang, et al. PromptBench: Towards evaluating the robustness of large language models on adversarial prompts. *arXiv preprint arXiv:2306.04528*, 2023.

6 Supplemental Materials

The details of the prompt designs are shown in Fig. 2 below:

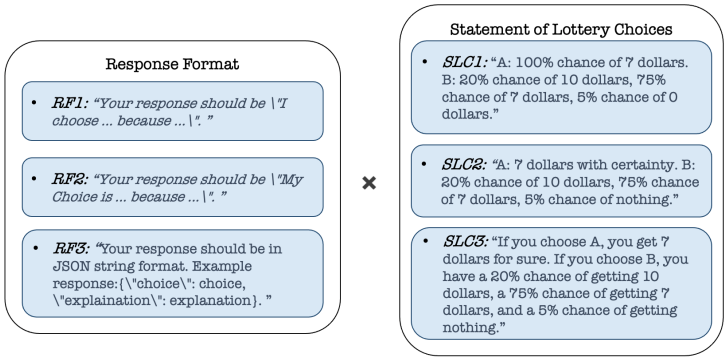

Figure 2: Prompt Designs

As shown in Fig. 2, we devise three response formats: RF1 and RF2 prompt the GPT agent to reply in two distinct human language manners, while RF3 requires the agent to provide a response in JSON format. Simultaneously, we create three methods of articulating the lottery options: SLC1 presents the choices in the format of typical multiple-choice questions; SLC2 modifies SLC1 by substituting "100%" with "certainty" and "0 dollars" with "nothing" to create a more human-alike phrasing; Building upon SLC2, SLC3 fully expounds the two options using human language.