When Experimental Economics Meets Large Language Models: Evidence-based Tactics[*]

Shu Wang   Zijun Yao   Shuhuai Zhang   Jianuo Gai

Tracy Xiao Liu   Songfa Zhong

October 17, 2025

Abstract

Advancements in large language models (LLMs) have sparked a growing interest in measuring and understanding their behavior through experimental economics. However, there is still a lack of established guidelines for designing economic experiments for LLMs. Inspired by principles from experimental economics with insights from LLM research in artificial intelligence, we outline key considerations in the experimental design and implementation stage, and perform two sets of experiments to assess the impact of these considerations on LLMs' responses. Based on our findings, we discuss seven practical tactics for conducting experiments with LLMs. Our study enhances the design, replicability, and generalizability of LLM experiments, and broadens the scope of experimental economics in the digital age.

Keywords: large language models (LLMs), experimental economics, decision-making, rationality, preference

1 Introduction

The rise of large language models (LLMs), particularly since the launch of ChatGPT in November 2022, has sparked significant interest in understanding their capacities. Research has shown that LLMs exhibit impressive abilities not only in language generation, but also in sophisticated reasoning and problem-solving. These capabilities translate into a wide range of applications.[1] This growing prominence has led to a surge in studies that examine how LLMs perform in economic experiments using decision-making tasks—referred to as "LLM experiments". These studies measure the preferences, rationality, heuristics, and biases of LLMs, and assess whether they can pass a form of the Turing test that evaluates behavioral similarity to humans (Chen et al., 2023; Filippas et al., 2024; Leng and Yuan, 2023; Mei et al., 2024; Meng, 2024). As LLMs continue to be integrated into various aspects of daily life and reshape our interactions with technology, these studies are becoming increasingly crucial and popular. However, despite this rapid progress, little has been done to examine the underlying rules of conducting economic experiments in the context of LLMs.

This study investigates the rules of LLM experiments. Evaluating LLM experiments requires a different perspective compared to traditional natural language processing (NLP) benchmarks. In typical NLP benchmarks—such as language translation, mathematical reasoning, and code generation—there exist correct answers. The objective of prompt engineering in these scenarios is to improve the accuracy of LLMs' outputs. In contrast, the main objective of LLM experiments is to measure the behavior of LLMs and to assess whether they exhibit human-like behavior, a domain where the outputs do not have definitive right or wrong answers.[2] Consequently, while LLMs demonstrate robustness to minor prompt variations in standard benchmark tasks, the influence of prompt formulation on decision-making tasks remains largely unexplored (Wei et al., 2022; Razavi et al., 2025). This issue is crucial for ensuring the robustness and generalizability of the conclusions from these experiments.

To address this issue, we synthesize principles from experimental economics with insights derived from LLM research in artificial intelligence. First, it is natural to evaluate the results of LLM experiments using established principles in experimental economics. For example, experimental economists emphasize the clarity and neutrality of instructions to ensure that experiments align closely with theoretical models, as well as the importance of monetary incentives and the incentive

---

[1]These include deconstructing textual disclosures (Kim and Nikolaev, 2024; de Kok, 2025; Siano, 2025; Niu et al., 2025), simulating experts' forecasts (Hansen et al., 2024), augmenting professional workflows (Noy and Zhang, 2023; Brynjolfsson et al., 2025) and enhancing the creative process (Chen and Chan, 2024; Zhou and Lee, 2024).

[2]For studies of typical NLP benchmarks, see Cobbe et al. (2021); Chen et al. (2021); Papineni et al. (2002). For studies of LLM behavior, see Leng and Yuan (2023); Goli and Singh (2024); Leng (2024); Leng et al. (2024); Liu et al. (2025); Deng et al. (2025); Fedyk et al. (2025); Sun et al. (2025); Zhu et al. (2025).

compatibility of mechanisms (Smith, 1982; Davis and Holt, 1993; Friedman and Sunder, 1994). Second, the unique characteristics of LLM experiments introduce new questions for experimental economics. For example, when addressing multiple decisions, researchers can either consolidate them into a single prompt (single-turn dialogue) or separate them into distinct prompts while incorporating prior interactions as memory (multi-turn dialogue). The lack of standardization regarding the choice between single-turn and multi-turn dialogues in existing LLM experiments further complicates the evaluation process. Therefore, it is essential to systematically identify and consider the characteristics, parameters, and procedures at each stage of LLM experiments that can influence the results. Although universal rules may not be applicable, some common practices can provide valuable guidelines for future research.

In this paper, we design two sets of experiments to examine the effect of some key considerations in experimental protocols on LLMs' responses from different perspectives. In Study 1, we examine the economic rationality of LLMs (Chen et al., 2023; Kim et al., 2024), in which LLMs make a series of budgetary decisions (Andreoni and Miller, 2002; Choi et al., 2007; Chen et al., 2023; Ellis et al., 2024). In Study 2, we examine the economic preferences of LLMs (Mei et al., 2024; Xie et al., 2024), in which LLMs make decisions in a series of behavioral games, including the dictator game, ultimatum game, public goods game and bomb risk game. We conduct the experiments using four popular LLMs, including GPT-4o-2024-11-20 (hereafter referred to as GPT), DeepSeek-V3-0324 (DeepSeek), Llama-3.1-8B-Instruct (Llama) and Qwen2.5-7B-Instruct (Qwen).

We have three main observations. First, assigned personas, such as occupation, significantly affect preferences but not rationality. Second, the single-turn dialogue approach significantly decreases the rationality of Llama and Qwen, while that of GPT and DeepSeek is not affected by the dialogue type. Third, limiting responses to multiple-choice options instead of allowing open-ended responses reduces the rationality of Llama and Qwen. It also significantly changes the outputs of all four LLMs in approximately half of the behavioral games. Building on these observations, we discuss several actionable tactics and the significance of design, replicability, and generalizability in LLM experiments.

Our paper adds to the emerging literature at the intersection of LLMs and experimental economics, with a particular focus on advancing methodological approaches. Previous research mainly explores how LLMs can be effectively used as supplementary tools at various stages of human experiments (Charness et al., 2025), and how to use generative AI in experimental research (Chang et al., 2024). In addition to these studies, we shift the perspective by treating LLMs as experimental objects and develop a framework for standardizing experimental protocols in LLM experiments. Moreover, this framework is grounded in two studies, which further enhances its validity.

Our study also contributes to improving the replicability and generalizability of LLM experi-

ments. Several studies have assessed the replicability of experimental studies with human subjects (Camerer et al., 2016, 2018; Holzmeister et al., 2025). The findings demonstrate that experiments in economics exhibit a high rate of replicability, which may be in part attributed to the rigorous methodological standards and established norms developed over decades within experimental economics (Croson, 2005; Camerer et al., 2016; Fréchette et al., 2022; Niederle, 2025). As LLM experiments are still in their early stages, similarly standardized protocols have yet to be formalized. Our study addresses this gap by proposing protocols that aim to increase the methodological rigor. In addition, aligning these protocols with those used in human experiments can improve the generalizability of LLM-based findings to human experiments, offering new insights into human behavior.

The paper is organized as follows. Section 2 provides an overview of LLMs. Section 3 summarizes the principles of experimental economics and reviews existing LLM experiments. Section 4 and 5 report the results of Study 1 and Study 2, respectively. Section 6 proposes seven tactics for conducting LLM experiments based on our experimental results and existing literature. Section 7 concludes.

2 LLMs: An Overview

In this section, we provide an overview of LLMs, including its history, categories of LLMs, key parameters, and tactics for prompt engineering. In particular, we devote our effort to discussing factors which affect the quality of LLM experimental data.

2.1 A Brief History of LLMs
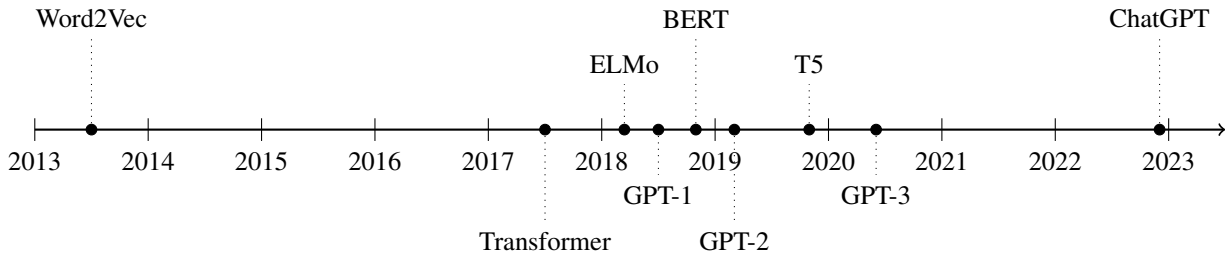
Figure 1: Timeline of Milestones in LLMs' History: 2013–2023

Figure 1 presents a timeline of pivotal events in the evolution of LLMs. The progression began with Word2Vec in the middle of 2013 (Mikolov et al., 2013), which established a foundation for learning semantic representations of words. The rapid evolution of LLMs can then be traced back to the introduction of the Transformer architecture (Vaswani et al., 2017), which employed self-

4

attention to overcome recurrent neural network (RNN) bottlenecks in processing long sequences. By parallelizing data handling instead of processing sequences step by step, Transformers significantly boosted efficiency and paved the way for subsequent breakthroughs.

In early 2018, ELMo (Peters et al., 2018) advanced the field by introducing deep contextualized word representations. Unlike static embeddings, ELMo could generate dynamic word vectors that capture context-dependent nuances, further enriching semantic understanding. Building on this foundation, OpenAI released GPT-1 in the middle of 2018, which showcased the potential of generative pretraining (Radford et al., 2018). Shortly afterward, Google released BERT (Devlin et al., 2019), which demonstrated the powerful generalization capability of pretrained Transformer-based model on various tasks. The momentum continued into early 2019 with the unveiling of GPT-2, a model that significantly expanded the generative capabilities. Later in 2019, Google's T5 introduced a unified text-to-text framework based on an encoder-decoder architecture, highlighting the versatility of these models in various NLP tasks.

In early 2020, GPT-3 was released (Brown et al., 2020), whose unprecedented scale (175 billion parameters) and few-shot learning prowess spurred a surge in both research and commercial applications. Most recently, the emergence of ChatGPT in late 2022 had not only popularized interactive dialogue systems but also signaled a broader convergence towards decoder-only architectures in the field of LLMs.

In summary, the evolution of large language models over the past decade reflects a series of groundbreaking innovations, from improved word embeddings and self-attention mechanisms to the development of versatile and interactive generative models.

2.2 How to Type LLMs

In this section, we type LLMs by different features and discuss their implications for research. Table A.1 presents the description of the representative models by types.

Open-sourced vs. proprietary model. LLMs can be broadly classified into two groups according to their accessibility. Open-sourced models, such as Meta's Llama, publicly disclose their architectures and weights. This openness enables researchers to inspect, modify, and fine-tune these models for tailored research use. In contrast, proprietary models—like OpenAI's GPT series, Anthropic's Claude, and Google's Gemini—offer access exclusively through black-box interfaces such as application programming interfaces (APIs), limiting insight into their internal mechanisms. The divergence is stark: open-sourced systems promote academic collaboration and experimental transparency, whereas proprietary offerings emphasize commercial robustness and performance.

Model size. Model size, measured in the total number of parameters, serves as a proxy for the complexity of the model and its ability to capture intricate language patterns. Larger models have

been shown to deliver enhanced performance on complex tasks and exhibit improved zero-shot and few-shot learning. However, these benefits come at significant practical costs, consistent with the scaling laws (Kaplan et al., 2020). Models with increased parameter counts typically require more computational resources for local deployment and correlate with higher API call charges. For instance, while a small variant of Meta's Llama (8B parameters) may run efficiently on a single GPU or even a standard workstation, a larger variant (405B parameters) might require a dedicated cluster of high-end GPUs—such as around 16 NVIDIA A100 with 80GB RAM, each costing roughly $15,000; For API calls, larger models such as GPT-4o incur higher per-token charges ($2.50 / 1M tokens input) compared to smaller counterparts GPT-4o mini ($0.150 / 1M tokens input), reflecting their increased computational demands during inference.[3]

Notice that model architectures, notably dense and mixture of experts (MoE) structures, are closely related to model size. In dense models, all parameters are activated for every input, leading to a predictable scaling of the computational cost with increased model size. In contrast, MoE models incorporate a gating mechanism that activates only a subset of parameters for a given input, enabling effective scaling to extremely large parameter counts while potentially reducing the computational burden per inference (Shazeer et al., 2017; Fedus et al., 2022). In particular, mainstream models such as GPT-3 employ dense architectures, while models like Switch Transformers, DeepSeek-V3, and Doubao-1.5-pro adopt MoE structures.

In summary, the selection between large and small models should be guided by the specific demands of the target task and the associated resource constraints. In particular, tasks characterized by straightforward, well-defined input–output mappings and limited semantic complexity, such as sentiment analysis, spam detection, or routine text classification, are generally well-served by smaller models. These applications benefit from lower latency, reduced computational cost, and simpler deployment requirements. In contrast, tasks that require nuanced language understanding, creative generation, or complex reasoning, such as multi-turn dialogue, detailed summarization of long documents, creative writing, and multi-hop question answering, tend to gain significantly from the richer representations and in-context learning capabilities of large models (Brown et al., 2020; Bommasani et al., 2021).

Context length. Context length is defined as the maximum number of tokens that an LLM can process in a single forward pass. Earlier models, such as GPT-3.5, typically support input windows of approximately 4,096 tokens. However, recent developments have significantly extended this capacity. Newer systems, for example Anthropic's Claude 3.5 Sonnet, can handle contexts up

---

[3]Regarding the economics of LLMs, Bergemann et al. (2025) propose a theoretical framework to study how providers of LLMs can optimally price access to their services. Their framework accounts for differences in user needs, such as the volume of tasks and sensitivity to accuracy, as well as the costs of processing inputs, outputs, and customizing models. The authors show that tiered pricing strategies, including usage-based fees and upfront charges for customization, align with profit-maximizing principles under user heterogeneity.

to 200,000 tokens. This extended capacity is advantageous in applications requiring sustained coherence, such as document summarization, multi-turn dialogue, and code generation. Furthermore, when inputs exceed a certain threshold, there is a risk that critical information provided early in the text may be diluted or overlooked, which could affect overall performance (Xiao et al., 2024; Liu et al., 2024; Wu et al., 2025).

Human alignment. Human alignment is a critical process that tailors the outputs of language models to be consistent with human values, ethical norms, and realistic decision-making patterns (Ouyang et al., 2024). In practice, alignment involves several complementary approaches. A widely adopted method is *reinforcement learning from human feedback* (RLHF), where human evaluators provide feedback on model output, and this feedback is used as the binary classification objective to train a reward model. The reward model is subsequently used to generate the preference score for model completion as reinforcement signal to adjust the model behavior. This process not only improves the safety and coherence of the generated text but also helps the model better approximate human reasoning under uncertainty, a feature of particular importance in LLM experiments.

Specifically, the importance lies in its dual role: it mitigates the risk of generating outputs that are ethically questionable or behaviorally unrealistic and improves the model's ability to simulate authentic human decision-making processes. However, alignment is not without cost. Overly aggressive alignment can lead to *alignment tax*, where models tailor their output too closely to the expected prompt structure rather than engaging in genuine reasoning. This is analogous to the experimenter effect in experimental economics. Furthermore, while alignment reduces harmful biases present in training data, it can inadvertently filter out legitimate variability in human behavior (Chaudhari et al., 2025). This issue is highlighted in recent discussions on pluralism in RLHF, where diverse human feedback is emphasized to avoid overly homogenized outputs that only reflect a subset of human perspectives (González Barman et al., 2025).

Base vs. chat versions. Some LLMs are released in two variants: base and chat (or instruct) models. Base models are the foundational pre-trained networks, developed on vast corpora without any specific tuning for interactive dialogue. They possess broad linguistic capabilities but require additional prompt engineering or fine-tuning to excel in conversational settings. In contrast, chat models are derived from base models through further instruction-based or supervised fine-tuning on dialogue-centric datasets. This additional training improves their ability to understand multi-turn conversations, adhere to stylistic guidelines, and provide contextually appropriate responses. Although both types share a common underlying architecture, the tuning in chat models makes them more accessible.

Unimodal vs. multimodal models. Some large language models are designed exclusively for

text-based interaction. For instance, Meta's Llama is a unimodal model that processes only textual input. In contrast, multimodal models integrate diverse input types, such as images along with text, to provide a richer and more context-aware interaction. OpenAI's GPT-4, for example, has a multimodal variant that can process both text and images, whereas its text-only counterpart powers ChatGPT. Similarly, models such as DeepMind's Flamingo (Alayrac et al., 2022) and Google's PaLM-E (Driess et al., 2023) are multimodal, trained on combined visual and textual data. Although unimodal models often excel at tasks where input and output are solely linguistic, multimodal systems offer enhanced user engagement by accommodating multiple data forms, which can be particularly advantageous in applications requiring visual context or richer interactivity. As current LLM experiments focus on text-based interactions, we discuss text data in this paper. However, it should be noted that future LLM experiments may incorporate multimodal capabilities.

Reasoning vs. non-reasoning models. Reasoning models incorporate additional training or prompting strategies to enable multi-step inference. Typically, they undergo reinforcement learning to develop an internal chain of thought that allows them to break down complex problems into sequential steps. In contrast, non-reasoning models are generally refined through supervised fine-tuning and produce outputs in a single pass without explicit multi-step reasoning. The advantages of reasoning models include improved performance on tasks that require multi-step inference and greater capability in handling ambiguous or complex queries (Wei et al., 2022). However, these benefits are often accompanied by higher computational costs and increased latency. Among current models, many of OpenAI's offerings illustrate this distinction. For instance, the OpenAI-o1 and DeepSeek-R1 have been refined with chain-of-thought (CoT) techniques and reinforcement learning, which enhances their ability to mimic human-like decision processes. In contrast, earlier models such as GPT-3.5 and the widely used Llama series from Meta remain primarily non-reasoning.

Looking ahead, the integration of CoT capabilities into a broader array of models appears promising. As research continues to explore methods for embedding structured reasoning within LLMs, reasoning models may become increasingly central to applications that seek to simulate human decision-making. This trend underscores the potential for future systems to more closely approximate the processes underlying human judgment and decision-making.

2.3   How to Prompt LLMs

In an interaction with LLMs using API calls or web-based chat interfaces, there are generally two key aspects that users need to adjust. The first is the specification of hyperparameters, which govern the model's output behavior. The second aspect involves the formulation of the prompt

itself. Proper tuning of both the hyperparameters and the prompt is essential to obtain meaningful and reliable outputs. This section will discuss these two critical components.

2.3.1 How to Set Up LLMs: Hyperparameter Specification

A critical step in employing these models for LLM experiments is the careful configuration of hyperparameters. Here we focus on the discussion of temperature, decoding methods, and the specification of roles engaged in the chat template. These hyperparameters are typically set through the API call or the user interface provided by the platform, where they can be adjusted before initiating the model's response generation. While parameters including temperature and decoding methods control the randomness of output, the specification of roles helps define the context and task instructions, and thus align model responses with the research objectives.

Randomness in outputs: temperature and sampling-based decoding. The temperature parameter is central to controlling the randomness of LLMs' output. Technically, temperature scales the model's logits—raw scores computed for each potential token—prior to applying the softmax function.[4] Lower temperature values sharpen the probability distribution, making high-probability tokens far more likely to be selected. This results in outputs that are more deterministic, coherent, and consistent. In contrast, higher temperatures flatten the distribution, increasing the chances of selecting less likely tokens and promoting more diverse, or unexpected outputs. In practice, temperature selection involves a fundamental trade-off between predictability and exploration. For tasks requiring strict adherence to factual or logical structure (e.g., legal document generation), low temperatures minimize variability and reduce error rates. Conversely, higher temperatures are often necessary for open-ended tasks where diverse outputs are desirable (e.g., generating multiple hypotheses in exploratory research).

Beyond temperature, the decoding strategy further governs randomness by dictating how tokens are selected from the probability distribution. Probabilistic sampling leverages temperature-adjusted probabilities to introduce controlled variability, while deterministic methods, e.g., greedy decoding, select the most probable token at each step. When replicability is critical, pairing a fixed random seed with temperature mitigates stochasticity in closed-system environments, due to inherent non-determinism in hardware operations or parallelized computations. For instance, even with greedy decoding, subtle numerical inconsistencies across GPU architectures may yield divergent outputs. Together, these two parameters allow researchers to fine-tune the balance between randomness and consistency in LLMs' responses.

Role specification. The role parameter—typically labeled as "system" and "user"—are now a

---

[4]Specifically, $p_i = \frac{\exp\left(\frac{z_i}{T}\right)}{\sum_j^n \exp\left(\frac{z_j}{T}\right)}$, where p_i denotes the probability of choosing the token, z_i represents the logit of token i and T is the temperature. When $T = 0$, $p_i = 1$ if $z_i = \max\{z_1, \cdots, z_n\}$ and $p_i = 0$ if $z_i \neq \max\{z_1, \cdots, z_n\}$. When $T \to \infty$, $p_i \to \frac{1}{n}$.

standard design pattern in most LLMs' APIs such as OpenAI's ChatGPT and Anthropic's Claude. These roles structure the conversation by assigning distinct responsibilities to each type of message. The "system" message, for example, provides high-level context and guidelines that prime the model to behave in a certain way, while the "user" messages carry the specific queries or commands. Essentially, most modern LLMs' APIs use this role-based approach because it improves clarity and context management, leading to output that is more consistent and aligned with the specific objects. During the training and fine-tuning process, models learn to interpret these roles differently. For example, the system prompt "teaches" the model the expected style and operational boundaries, and this helps guide subsequent responses.

Taking an LLM experiment as an example, researchers should choose the role content based on the intended function: the system message may include instructions on characteristics and tone, whereas the user message should clearly demonstrate the decision-making task. If the roles are selected incorrectly, the output might become less controlled, more prone to hallucinations, less consistent in style.

2.3.2   How to Improve Answers of LLMs: Prompt Engineering

Using natural language prompts to interact with LLMs enables researchers to engage with sophisticated AI models without requiring expertise in programming languages. It broadens the accessibility of advanced computational techniques, and allow the focus of individuals to shift from the intricacies of code implementation toward more important tasks.

At its core, prompt engineering is a technical process that involves designing the input text (commonly called prompt) to guide the probabilistic output of the model in a desired direction, and a recent survey of prompt engineering is provided by Sahoo et al. (2024). Specifically, a prompt is a sequence of tokens that establishes the context for the prediction of the next token of the model, using the statistical patterns encoded during the training and fine-tuning phases. Through a careful selection of vocabulary, structure, and context, researchers manipulate the conditional probabilities that govern the model response. This process is based on the internal representations and alignment mechanisms of the model, where subtle variations in the prompt formulation can possibly alter the generated output. While there is no single canonical approach, systematic strategies, such as those outlined in recent guidelines, offer empirical rules to enhance output quality.

Given that prompt engineering is inherently task-specific, there is no universal set of deterministic rules that guarantee optimal performance across all scenarios. However, the nature of training and alignment gives rise to several empirically derived guidelines that can significantly enhance the quality of output in a task-dependent manner. For example, OpenAI recommends a range of strategies and tactics, from writing clear and detailed instructions to systematically testing prompt

modifications, to better harness the probabilistic nature of these models.[5] Specially, these strategies include 1) writing clear instructions, 2) providing reference text, 3) splitting complex tasks into simpler subtasks, 4) giving the model time to "think", 5) use external tools, and 6) test changes systematically. It is important to note that these prompt engineering strategies require adaption in LLM experiments due to differences in tasks and purposes, which will be discussed in the next section.

3 Experimental Economics: From Humans to LLMs

In this section, we begin by outlining some principles drawn from the practices of experimental economics with human subjects, then review the procedures of existing research on LLM experiments.

3.1 Human Experiments: Key principles

To review key principles for human experiments, we resort to classical textbooks and survey papers that offer comprehensive discussions on experimental methodology. We choose two textbooks, Davis and Holt (1993) and Friedman and Sunder (1994), which have high citation counts.[6] Croson (2005) is also included for its detailed comparison in experimental methods between economics and psychology. Two students who have taken the Ph.D. level experimental economics class independently read these three references and summarize the key methods for experimental design and implementation. We take the union of these principles and then distill them based on two criteria: 1) they are mentioned in at least two of the three sources, and 2) they have potential applicability to LLM experiments. Table 1 provides a summary of key guidelines from these three sources.

First, in the experimental design stage, we focus on the following principles.

1. *Follow the induced value theory.* For economic experiments, a central consideration is the alignment of monetary incentives with induced value theory. Specifically, payment schemes must adhere to principles such as monotonicity, salience, and dominance (Smith, 1982). Induced value theory ensures a linkage between participants' payoff and their decisions, thereby eliciting truthful responses. In economic experiments, cash is typically used as the reward medium, instead of other methods, e.g., grades or extra credit points. Moreover,

---

[5]The detailed explanations and examples for these strategies and tactics can be found at OpenAI's website: https://platform.openai.com/docs/guides/prompt-engineering/strategy-write-clear-instructions, last accessed on May 25, 2025.

[6]As of April 9, 2025, the Google Scholar citation counts for Davis and Holt (1993) and Friedman and Sunder (1994) were 3,197 and 1,442, respectively.

Table 1: Key Principals for Human Experiments

	DH1993	FS1994	Croson2005
Panel A: Basic Rules in Experimental Design			
Follow the induced value theory	✓	✓	✓
Avoid deception	✓	✓	✓
Minimize experimenter demand effects	✓	✓	✓
Panel B: Execution Rules in Experimental Implementation			
Use context-neutral instructions	✓	✓	✓
Make instructions comprehensible		✓	✓
Provide illustrative examples		✓	✓
Incorporate comprehension assessments		✓	✓
Prevent inter-subject communication		✓	✓

Notes: DH1993 refers to Davis and Holt (1993), FS1994 refers to Friedman and Sunder (1994) and Croson2005 refers to Croson (2005). A checkmark indicates that the corresponding rule is discussed in the reference.

though negative payment is possible in theory, subjects' payments are always positive in practice.

2. *Avoid deception.* Economic experiments strictly prohibit deception on participants, such as the inconsistency between the actual and prespecified payment rules as well as the provision of false feedback. Deception undermines the trust of participants in experimenters, thus potentially changing their behavior in future experiments and contaminating the subject pool. For example, deception on the payoff rule will weaken the link between the payoff and the participants' decisions due to distrust, leading to false reports (Croson, 2005).

3. *Minimize experimenter demand effects.* Experimenter demand effects also pose a threat to the validity of experimental results. Specifically, The effects occur when participants discern the experimental objectives and correspondingly modify their behavior. For example, Brañas-Garza (2007) includes "Note that he relies on you" within the instruction and finds that it leads to more generous behavior in behavioral games. Using non-deceptive obfuscation to conceal experimental objectives in certain cases is emphasized as a means of reducing the effects (Zizzo, 2010).

In summary, all of these principles are designed to guarantee truthful reports from human subjects. However, their applicability to LLM experiments might be limited, as LLMs generate outputs by sampling from probability distributions learned during the training processes, rather than through subjective decision-making.

Second, during the experimental implementation stage, most rules emphasize the clarity of the

instruction to ensure participant comprehension.

1. *Use context-neutral instructions.* Non-neutral language may additionally introduce participants' preferences and attitudes as confounders, thereby compromising the reliability of experimental results. For example, as shown in the instructions from Plott and Smith (1978) in Appendix B, participants are simply informed that they are trading a generic commodity without reference to specific categories such as fruits or meat. This approach eliminates the possibility that participants' choices are influenced by personal preferences for particular goods.

2. *Make instructions comprehensible.* The instructions should be easily understood. Meanwhile, the payment rules, as well as the available choices and actions, should be clearly outlined in the instructions.

3. *Provide illustrative examples.* In the case of complex tasks, illustrative examples are often used to enhance comprehension.

4. *Incorporate comprehension assessments.* Understanding questions serves as an effective method to assess the participants' comprehension of experimental instructions. When the subject pool is sufficiently large, as in online experiments, individuals who fail these questions can be immediately excluded from experiments. In other cases, the accuracy rate on the understanding questions can be taken into account in the analysis to control for variations in comprehension.

5. *Prevent inter-subject communication.* Prohibiting communication among participants is a standard practice at all stages of experiments, except that communication itself is a focus of the experimental design. Communication can lead participants to infer the experimental objectives by learning from others. Moreover, it compromises the independence of the observations, thus reducing the statistical power of the analysis.

In the implementation stage, instructions are particularly relevant for LLM experiments, as the instructions provided for human participants closely parallel the instructions for LLMs. Additionally, to elaborate how the principles in this subsection are embedded in the experimental instructions, we explain the instructions from Plott and Smith (1978) in Appendix B, which are also reprinted in Friedman and Sunder (1994).

3.2 LLM Experiments: An Emerging Era

Next, we summarize 12 papers involving LLM experiments that have been published in leading journals or conferences, including *Proceedings of the National Academy of Sciences* (*PNAS*), *Marketing Science*, *Manufacturing & Service Operations Management*, and *Nature Human Behaviour*,

among others. Although there is a rapidly growing literature on this topic (e.g., Leng and Yuan, 2023; Kim et al., 2024; Xie et al., 2024; Deng et al., 2025), we decide to include only published papers because their experimental procedures are finalized and many of them provide the code which enables us to replicate their results. Table 2 summarizes these studies.[7]

Now we apply what we have learned from both artificial intelligence and human experiments to evaluate these studies. We find that these studies exhibit a large degree of heterogeneity in the choice of parameters in the prompts (Table 3), as summarized below.

1. *Temperature.* Temperature is a unique design parameter raised by LLM experiments. Take the models in GPT family as an example, three out of twelve studies set temperature at 0, five choose 1, one chooses 0.7, and two studies have tried multiple values such as 0 and 1 and do not find significant differences across temperatures (Chen et al., 2023; Kosinski, 2024).

2. *Persona.* Prior experimental studies have shown that individual choices are affected by their demographics (Croson and Gneezy, 2009; Dohmen et al., 2011). Although most LLM experiments evaluate the performance without assigning specific demographics, some examine the impact of demographics with different purposes. To compare with the choices of undergraduate students in the games, Brookins and DeBacker (2024) explicitly assign LLMs the persona of an undergraduate student. To explore whether LLMs can exhibit human-like behavioral heterogeneity across groups, Chen et al. (2023) and Mei et al. (2024) vary different demographics such as gender, ethnicity, and occupation.

3. *Incentive.* Four studies incorporate a hypothetical incentive scheme into the prompt, while six do not. The remaining two studies apply the incentive to some tasks and omit it in others.

4. *Illustrative example and understanding question.* Nearly all LLM experiments do not in-

---

[7]First, in terms of model choice, all these papers utilize GPT models which are considered as the most popular. Specifically, only Binz and Schulz (2023) use GPT-3, and all subsequent publications involve GPT-3.5 or more advanced versions of GPT. Macmillan-Scott and Musolesi (2024) and Wang et al. (2024) also evaluate the performance of LLMs in other model families, such as Llama 2. Next, given that psychologists and economists measure individuals' decision-making abilities using different tasks, the evaluation task for LLM also varies across studies with some overlapping tasks. Specifically, Binz and Schulz (2023); Hagendorff et al. (2023); Webb et al. (2023) and Kosinski (2024) explore the cognitive abilities of LLMs using a set of tasks such as the cognitive reflection test (CRT) and Raven's progressive matrices, both of which are often examined in cognitive psychology (e.g., Raven, 2000; Frederick, 2005) and recently in economics (e.g., Gill and Prowse, 2016; Alan and Ertac, 2018). Motivated by key decision-making biases observed in human experiments (Kahneman and Tversky, 1973; Kahneman et al., 1982; Arkes and Blumer, 1985), several studies also investigate whether LLMs exhibit behavioral biases such as base rate neglect, conjunction fallacy, and sunk cost effect (Binz and Schulz, 2023; Suri et al., 2024; Macmillan-Scott and Musolesi, 2024; Chen et al., 2025b; Wang et al., 2025). Since the fundamental assumption of microeconomic theory is that individuals are rational, that is, maximize their utility given budget constraints, some studies evaluate LLMs' economic rationality (Chen et al., 2023) and their underlying preferences (Chen et al., 2023; Goli and Singh, 2024; Mei et al., 2024). Beyond these tasks, which mainly focus on individual choices, studies also test LLMs' abilities in strategic interactions, measured by their choices in prisoners' dilemma and public goods games (Brookins and DeBacker, 2024; Mei et al., 2024).

Table 2: The Description of Representative LLM Experiments

Study	Journal/Conference	Model	Representative task
Binz and Schulz (2023)	PNAS	GPT-3 (ada, babbage, curie, davinci)	Linda problem and CRT
Chen et al. (2023)	PNAS	GPT-3.5-Turbo	Budgetary tasks
Hagendorff et al. (2023)	Nature Computational Science	GPT-1, GPT-2XL, GPT-3 (ada, babbage, curie, davinci-001, davinci-002, davinci-003), ChatGPT-3.5, ChatGPT-4	Semantic illusions and CRT
Webb et al. (2023)	Nature Human Behaviour	GPT-3 (text-davinci-003, code-davinci-002, text-davinci-002, davinci), GPT-4	Raven's Progressive Matrices
Brookins and DeBacker (2024)	Economics Bulletin	GPT-3.5-Turbo	Dictator game and prisoner's dilemma game
Suri et al. (2024)	Journal of Experimental Psychology: General	GPT-3.5, ChatGPT-4, Bard	Linda problem
Goli and Singh (2024)	Marketing Science	GPT-3.5-Turbo, GPT-4	Intertemporal choice task
Macmillan-Scott and Musolesi (2024)	Royal Society Open Science	GPT-3.5, GPT-4, Bard, Claude 2, Llama 2 (7B, 13B, 70B)	Linda problem and Wason task
Mei et al. (2024)	PNAS	GPT-3.5-Turbo, GPT-4, ChatGPT (subscription-based Web version (Plus), freely available Web version (Free))	Dictator game, ultimatum game, trust game, bomb risk game, public goods game, prisoner's dilemma game
Kosinski (2024)	PNAS	GPT-1, GPT-2XL, GPT-3 (ada, babbage, curie, davinci-001, davinci-002, davinci-003), BLOOM, ChatGPT-3.5-turbo, ChatGPT-4	Unexpected contents task and unexpected transfer task
Wang et al. (2024)	COLM	GPT-3.5-Turbo, GPT-4, PaLM 2, Llama 2 70B	Linda problem
Chen et al. (2025b)	Manufacturing & Service Operations Management	GPT-3.5-Turbo, GPT-4	Linda problem and CRT

Notes: PNAS: *Proceedings of the National Academy of Sciences*. COLM: *Conference on Language Modeling*. CRT: Cognitive reflection task.

Table 3: The Prompt Parameter Specification of Representative LLM Experiments

Study	Temperature	Persona	Incentive	Example	Understanding question	Dialogue type	Answer type	Invalid answers
Binz and Schulz (2023)	0		Both			Multi-turn	Both	
Chen et al. (2023)	0, 0.5, 1	Both	Y		Y	Single-turn	Both	
Hagendorff et al. (2023)	0					−	Open	Y
Webb et al. (2023)	0					Multi-turn	Both	
Brookins and DeBacker (2024)	1	Y	Y			−	Both	Y
Suri et al. (2024)	−					−	Both	
Goli and Singh (2024)	1		Y			−	Choice	
Macmillan-Scott and Musolesi (2024)	1					−	Both	Y
Mei et al. (2024)	1	Both	Y	Both		Multi-turn	Both	
Kosinski (2024)	0, 1					−	Open	
Wang et al. (2024)	0.7			Both		−	Both	
Chen et al. (2025b)	1		Both			−	Both	Y

Notes: In terms of temperature, we only summarize the temperature of models in GPT family for comparison, and "−" represents the parameter is not revealed in the study. "Y" represents the parameter or feature is specified in the study, while the black represents the opposite. "Both" represents that the study includes the parameter or feature in some experimental conditions or tasks and not in the others. For dialogue type, "−" represents the feature is not applicable, as each LLM agent only needs to answer one question. For answer type, "Open" represents the open-ended type and "Choice" represents the multiple-choice type. "Invalid answers" describes whether the study reports the proportion of invalid answers.

clude examples and understanding questions. We conjecture this is because the tasks used in these studies are simple and easy to understand. However, we notice the use of illustrative examples (Mei et al., 2024; Wang et al., 2024) and understanding questions (Chen et al., 2023) in some studies. Taking the latter as an example, the rationale lies in its goal to compare the performance of LLMs with human subjects, and understanding questions are employed in human experiments.

5. *Single-turn vs. multi-turn dialogue.* Among these studies, four of them involve situations that each LLM agent needs to answer multiple questions (Binz and Schulz, 2023; Chen et al., 2023; Webb et al., 2023; Mei et al., 2024), while the remaining studies ask each agent only to answer one question. In practice, both single-turn dialogue and multi-turn dialogue can be applied to instruct LLMs to collect answers for multiple questions. Figure A.1 provides an illustration of these methods. Within these three studies, one utilizes the single-turn dialogue, while the other two employ the multi-turn dialogue. Notably, Mei et al. (2024) pose multiple questions to each LLM agent and provide feedback to elicit its strategy in several games. Such feedback can only be delivered using the multi-turn dialogue, thus single-turn dialogue is not feasible due to the experimental design.

6. *Open-ended vs. multiple-choice answer type.* Open-ended answer type prompts responses without predefined options, whereas multiple-choice type explicitly lists a finite set of choices for LLMs to select from within the prompts. Of the studies reviewed, two use the open-ended answer type, one provides multiple choices, and the remaining nine incorporate both types.

7. *Invalid answer.* There are instances in which models fail to generate valid answers. For example, LLMs may produce statements such as "As an AI language model, I cannot participate in surveys nor accept tokens as payment." (Goli and Singh, 2024) or "As an AI language model, I am not capable of making decisions on my own." (Chen et al., 2023). Another type of invalid answers features that LLMs generate elaborative explanatory content without explicitly providing the requested outputs. In handling these invalid responses, four studies report the proportion of them in the analysis, whereas the others do not mention the proportion.

4   Study 1: Economic Rationality

In this section, we implement the budgetary tasks of Chen et al. (2023) to examine the extent to which the considerations discussed above influence the performance of LLMs. Economic rationality is a central assumption in economics and is widely evaluated through various domains of decision-making (e.g., Andreoni and Miller, 2002; Choi et al., 2007, 2014; Echenique et al., 2011; Kim et al., 2018).

4.1 Experimental Design

We choose four representative LLMs to carry out our tasks: (1) GPT-4o-2024-11-20, the advanced LLM developed by OpenAI, which is widely regarded as the pioneering leader in the field, (2) DeepSeek-V3-0324, the open-sourced model with the largest parameter size (Maslej et al., 2025), (3) Llama-3.1-8B-Instruct, a widely studied model in artificial intelligence and the most downloaded text-generation model released since 2024, ranked by Hugging Face, and (4) Qwen2.5-7B-Instruct, an emerging model that has gained increasing attention in the field of artificial intelligence since early 2025 (e.g., Gandhi et al., 2025; Muennighoff et al., 2025; Xie et al., 2025). All four models are prompted via the API using Python.

4.1.1 Budgetary Decision Tasks

We use standard budgetary decision tasks in risk preference (Choi et al., 2007, 2014) and social preference (Andreoni and Miller, 2002) to measure economic rationality. In the risk preference domain, the agent allocates 100 points to two accounts, where the exchange rates between points and payoffs differ. She receives the reward from a random account with probability 50%. For the social preference, the agent allocates 100 points between herself and another randomly matched agent. These points are converted into payoffs with distinct exchange rates for the two agents. We construct 25 rounds of tasks in each simulation, with budget lines that vary between rounds.

4.1.2 Baseline Condition

In the baseline condition, we closely align the experimental procedures with those commonly used in human experiments. The temperature is set to the default value of each LLM,[8] and there is no specific persona included in the system message. We outline the payment rule in which one round is randomly selected for payment to hypothetically incentivize LLMs. In the user message, we also provide a demonstration example and evaluate the degree of comprehension using understanding questions. In each simulation, we use a multi-turn approach, where each round of question is asked separately, with the preceding dialogue included. Thus, each LLM makes one decision per round. Additionally, we use the open-ended answer type and restrict the output to the JSON format to streamline data collection. When invalid answers occur, we exclude them and continue until 25 valid answers are collected in a simulation.[9] Meanwhile, the proportion of invalid

---

[8]The default temperatures are 1 for GPT (https://platform.openai.com/docs/api-reference/responses/create), 0.3 for DeepSeek (https://huggingface.co/deepseek-ai/DeepSeek-V3-0324), 0.6 for Llama (https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct/blob/main/generation_config.json) and 0.7 for Qwen (https://huggingface.co/Qwen/Qwen2.5-7B-Instruct/blob/main/generation_config.json), all last accessed on May 25, 2025.

[9]Invalid answers include instances where (1) LLMs refuse to provide answers, (2) the outputs do not adhere to the specified formats, i.e., JSON format in Study 1 or number highlighted using "[[]]" in Study 2, and (3) the answers fail to meet task requirements, e.g., the total points allocated to two accounts do not sum to 100 in Study 1 or the number

answers is also documented. We repeat the above procedure 100 times to generate 100 simulations, ensuring sufficient statistical power for our subsequent comparisons between conditions.[10]

4.1.3 Experimental Conditions

Based on the baseline condition, we implement several variations to investigate the impact of the considerations we have summarized. In particular, we focus on the following three factors which suggest that the corresponding variation may influence LLMs' performance. Each condition includes 100 simulations for each model. Appendix D provides detailed prompts for the baseline condition and these variations.

1. *Persona.* We separately assign LLMs a range of demographics, including gender, age, education level and ethnicity as in Chen et al. (2023), as well as occupations as in Mei et al. (2024).[11] Each specific persona corresponds to a unique condition.

2. *Dialogue type.* Instead of using multi-turn dialogue, we consider a single-turn dialogue format that incorporates all 25 tasks in one prompt, instructing LLMs to make 25 decisions simultaneously.

3. *Answer type.* We change the answer type from open-ended to multiple-choice (Kim et al., 2018; Chen et al., 2025a). Specifically, the continuous budget set is discretized into 21 allocation options with a step size of 5 points, and LLMs are required to select one option for each task.

4.2 Results for LLMs' Rationality

In this section, we first report on the economic rationality of LLMs in the baseline condition, followed by the results under experimental conditions. To evaluate the degree of economic rationality, we use the classic measure in revealed preference analysis, the critical cost efficiency index (CCEI) introduced by Afriat (1972). CCEI scores range from 0 to 1, with higher scores indicating greater economic rationality. A CCEI score of 1 indicates that the dataset satisfies the generalized axiom of revealed preference (GARP) and is consistent with utility maximization.

---

of points falls outside the specified range in both studies.

[10]By assuming equal variance and specifying the minimum detectable difference in CCEI scores of 0.01, we use the standard deviation obtained in Chen et al. (2023) and compute the minimum sample size for each condition, which is 24 in the risk preference and 18 in the social preference, respectively.

[11]We vary the persona by modifying the role specified in the system prompt, using statements such as "I want you to act as a male decision maker" and "I want you to act as a mathematician" (Chen et al., 2023). For variations in occupations, according to the method of Mei et al. (2024), we additionally provide the description of a list of core and supplementary tasks associated with each occupation.

4.2.1 Rationality in Baseline Condition

Before discussing the economic rationality of LLMs, we compute the proportion of invalid answers. Among the four models, GPT, DeepSeek and Qwen consistently produce valid answers across all 100 simulations in each domain. Only Llama generates invalid answers in one simulation in the risk preference. These findings suggest that LLMs are generally capable of following instructions and providing the required outputs in budgetary decision tasks.

Next, we focus on the economic rationality of LLMs. The average CCEI scores of GPT, DeepSeek, Llama, and Qwen are 1.000, 1.000, 0.953, and 0.980 in the risk preference domain, and 0.994, 0.999, 0.968, and 0.994 in the social preference domain.[12] In human experiments using comparable budgetary tasks, the average CCEI ranges from 0.881 to 0.980 in the risk preference (Choi et al., 2007, 2014; Halevy et al., 2018; Chen et al., 2023, 2025a), and from 0.860 to 0.967 in the social preference (Fisman et al., 2015, 2023; Li et al., 2017, 2022; Chen et al., 2023). Among the models evaluated, Llama and Qwen may fall within these ranges, while the other two models exhibit higher levels of rationality. We take human decisions in the baseline condition of Chen et al. (2023) as a reference and depict the cumulative distribution of CCEI scores for human subjects and the four LLMs in Figure 2a. The results indicate that GPT and DeepSeek outperform human subjects in both domains, Qwen surpasses humans in the social preference, while Llama performs worse than humans in the risk preference (all $p < 0.01$, two-sided two-sample t-tests).[13]

To investigate whether the responses of LLMs can be distinguished from those of humans, we conduct a form of the Turing test proposed in Mei et al. (2024). In each comparison, we randomly draw one CCEI score from the LLM's distribution and one from the human distribution, and evaluate whether these two scores are equally likely to be generated under the human distribution and, if not, which one is more likely to be generated under the human distribution. We repeat this process 10,000 times to compute the corresponding probabilities. The results show that all four LLMs pass the Turing test, as they are estimated more likely or equally likely to be human in over half of the comparisons (Figure 2b). Meanwhile, our randomly simulated agents fail the Turing test.

---

[12]To validate the sufficient power of detecting GARP violations, simulated CCEI is generated using random allocation across 25 decisions for 100 simulated agents (Bronars, 1987). The simulated CCEI is significantly lower than all four LLMs in both preference domains (all $p < 0.01$).

[13]We employ two-sided tests throughout all analyses. Unless otherwise specified, two-sample t-tests are used in this paper. Meanwhile, we conduct multiple hypotheses testing using the false discovery rate (FDR) method (Benjamini and Hochberg, 1995) in 8 tests (4 models × 2 preferences), and report corrected p values. This approach is also applied in the comparison in decisions between LLMs and humans in Study 2, where we regard the 20 tests (4 models × 5 scenarios) as a family for the correction.
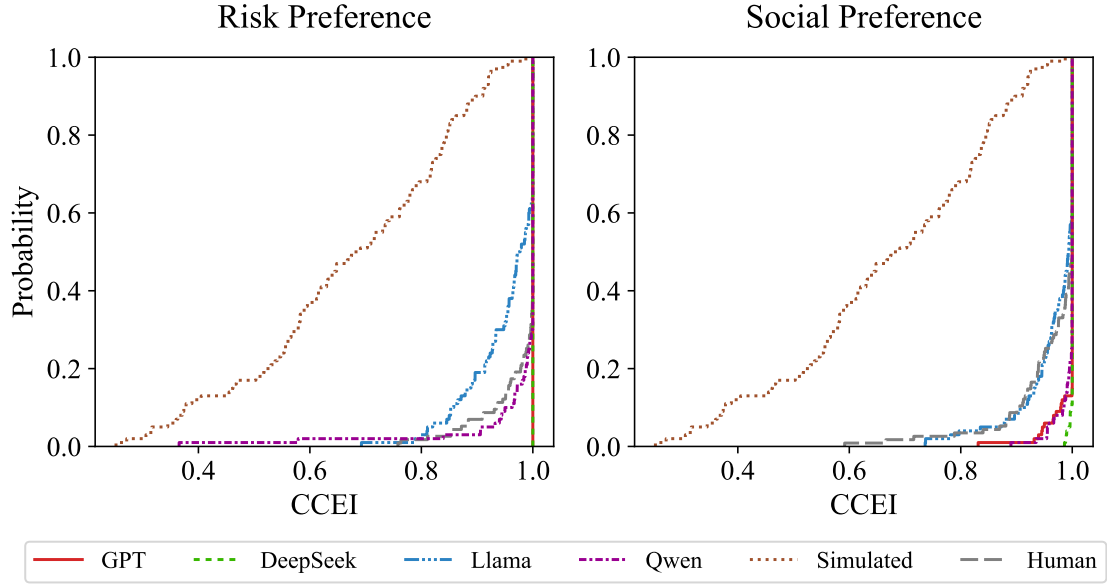
(a) Cumulative Distribution of CCEI Scores
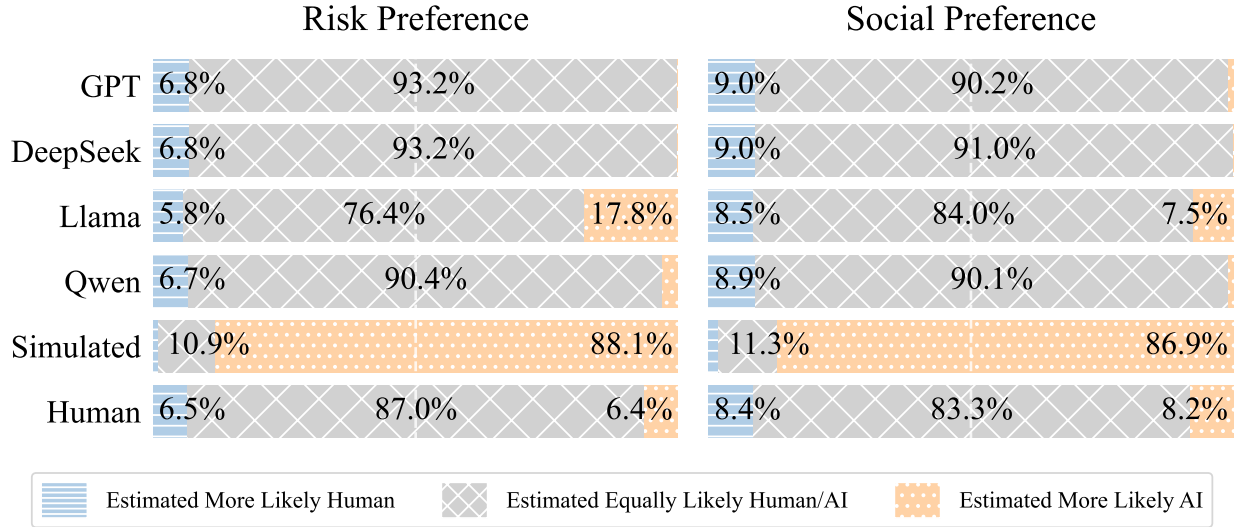
(b) Turing Test of CCEI Scores

Figure 2: Baseline Results of CCEI Scores.

Notes: Human data are from Chen et al. (2023). In Panel b, probabilities below 5% are not labeled with numbers. The white dashed lines represent the probability of 50%. In the simulated condition, each comparison involves one CCEI score randomly drawn from the simulated distribution and one from the human distribution. In the human condition, both CCEI scores in each comparison are randomly drawn from the human distribution.

4.2.2 Rationality in Experimental Conditions

To quantify to what extent LLMs respond significantly to the experimental variations, we first construct an overall sensitivity score as follows:

$$\lambda(\mathcal{S}) = \Sigma_{i \in \mathcal{I}, m \in \mathcal{M}, s \in \mathcal{S}} \frac{\mathbb{1}\{p_{ims} < 0.05\}}{|\mathcal{I}||\mathcal{M}||\mathcal{S}|}.$$

In the equation, $\mathcal{I}$ is the set of four LLMs, $\mathcal{M}$ denotes the set of preference domains in Study 1 or decision scenarios in Study 2. $\mathcal{S}$ represents the set of experimental conditions in each variation. p denotes the p value from two-sample t-tests, and the significance level is set at 0.05.[14] $\mathbb{1}(\cdot)$ is an indicator function. To further measure the sensitivity score under each preference domain (decision scenario) $m \in \mathcal{M}$, we also calculate $\lambda_m(\mathcal{S}) = \Sigma_{i \in \mathcal{I}, s \in \mathcal{S}} \frac{\mathbb{1}\{p_{ims} < 0.05\}}{|\mathcal{I}||\mathcal{S}|}$. By definition, both λ and λ_m belong to $[0, 1]$. The higher the values, the higher the sensitivity.

Persona. Using the persona specified in the system message, we calculate the average CCEI scores in different demographic conditions. Figure 3 presents the differences in CCEI between each experimental condition and the baseline condition. The results show that the rationality of all four models is not significantly influenced by the persona (all $p > 0.1$), leading to a sensitivity score $\lambda(\mathcal{S})$ of 0. In general, the findings suggest that the persona has no impact on the economic rationality of LLMs.

Single-turn vs. multi-turn dialogue. Compared to the baseline condition, the single-turn dialogue approach does not substantially affect the CCEI scores of GPT (risk: 0.999 vs. 1.000, $p > 0.1$, social: 1.000 vs. 0.994, $p < 0.01$) and DeepSeek (risk: 1.000 vs. 1.000, social: 1.000 vs. 0.999, $p < 0.01$) (Figure 3). However, when using the single-turn dialogue, the CCEI of Llama decreases from 0.953 to 0.841 in the risk preference and from 0.968 to 0.756 in the social preference; that of Qwen decreases from 0.980 to 0.739 in the risk preference and from 0.994 to 0.889 in the social preference (all $p < 0.01$). As a result, $\lambda(\mathcal{S})$ equals 75.0%. These declines may arise from the limitation of single-turn dialogue in supporting sequential reasoning processes.

Open-ended vs. multiple-choice answer type. As shown in Figure 3, compared to the baseline condition using the open-ended answer type, the multiple-choice answer type does not affect the CCEI scores for GPT (risk: 1.000 vs. 1.000, social: 0.996 vs. 0.994, $p > 0.1$) and DeepSeek (risk: 1.000 vs. 1.000, social: 0.994 vs. 0.999, $p = 0.060$). In contrast, under the multiple-choice answer type, the average CCEI scores in both domains decrease significantly for Llama (risk: 0.853 vs. 0.953, $p < 0.01$, social: 0.936 vs. 0.968, $p = 0.033$) and Qwen (risk: 0.902 vs. 0.980, social:

---

[14]Notably, we correct the p values by controlling FDR (Benjamini and Hochberg, 1995), where all tests within each variation are considered as a family for correction. In Study 1, we perform multiple hypothesis testing in 112 tests (4 models $\times$ 2 preferences $\times$ 14 conditions) for the persona; while for the dialogue type and answer type, similar analysis is conducted in 8 tests (4 models $\times$ 2 preferences). In Study 2, the number of tests within a family is 280 (4 models $\times$ 5 scenarios $\times$ 14 conditions) for the persona and 20 (4 models $\times$ 5 scenarios) for the answer type.
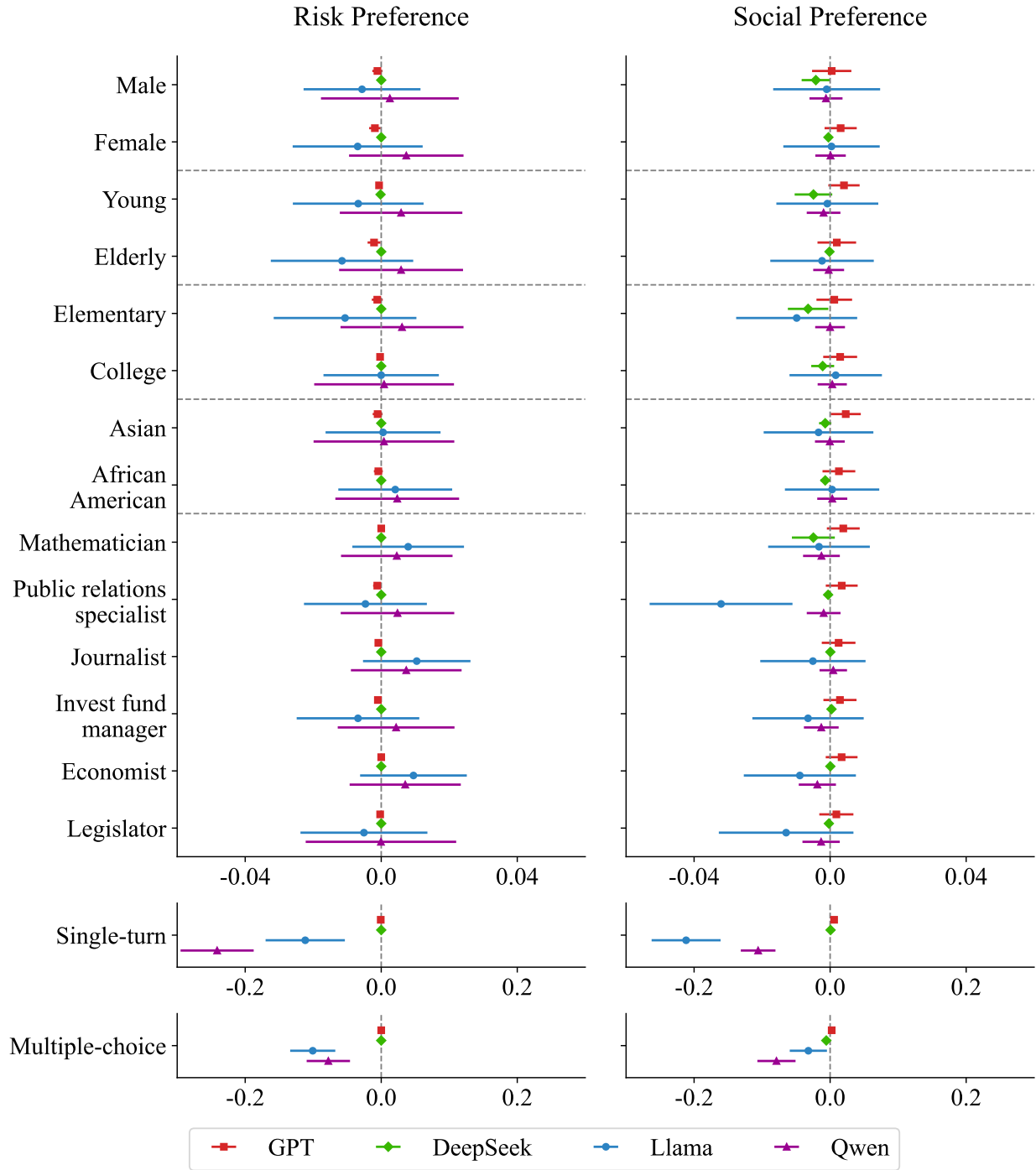
Figure 3: Comparison in Rationality of LLMs.

Notes: The markers represent differences in average CCEI scores of LLMs between experimental and baseline conditions. The error bars represent 95% confidence intervals.

0.916 vs. 0.994, both $p < 0.01$). This decrease mirrors the pattern observed in GPT-3.5-Turbo (Chen et al., 2023). Although the advancement from GPT-3.5-Turbo to GPT-4o over the past two years has eliminated the negative impact of discrete choices, this issue persists for smaller models such as Llama and Qwen.

Other variations. In addition to the three variations discussed above, we also examine the influences of other factors in Appendix C. In summary, temperature settings between 0 and 1 do not affect the rationality of any of the four LLMs. The inclusion of examples does not affect the rationality in the risk preference, but does affect in the social preference. The existence of hypothetical incentives has no substantial impact on the rationality of four LLMs, while the rationality of Qwen decreases surprisingly as stake sizes increase in both preferences.

5 Study 2: Economic Preferences

In this section, we repeat our exercise using another set of tasks assessing a range of preferences including risk, trust, fairness, altruism, and cooperation in Mei et al. (2024). This study complements the economic rationality measured in Study 1.

5.1 Experimental Design

We select the same four LLMs as in Study 1 and interact with them through the API using Python. The evaluated tasks include the dictator game, the ultimatum game, the public goods game, and the bomb risk game. In each game, the LLMs are required to make one decision, with the following exception—in the ultimatum game, the LLMs separately play the role of proposer, who suggests an allocation, and responder, who determines the minimum proposal they are willing to accept. As a result, we collect the decisions of LLMs across five distinct decision scenarios from the four games. A detailed description of these tasks can be found in Mei et al. (2024). To avoid learning effects and simplify the analysis, we focus solely on the first round of decisions in each scenario. LLMs' choices are required to be highlighted with "[[]]". Data collection continues until 100 valid responses are obtained per scenario.

In the baseline condition, we adhere to the methodology in Study 1. Building upon this baseline, we implement two systematic variations concerning the persona and answer type. We do not examine the variation of dialogue type, as single-turn dialogue type is not feasible in these games. The prompts are detailed in Appendix E.

1. *Persona.* We independently vary gender, age, education level, and ethnicity following Chen et al. (2023), as well as occupations based on Mei et al. (2024).

2. *Answer type.* Given that the feasible choice sets in all scenarios, except the public goods

game, range from 0 to 100, we discretize this range into 21 options with a gap of 5. In the public goods game, where the feasible choice set ranges from 0 to 20, we maintain the same number of 21 options and reduce the gap to 1.

5.2 Results for LLMs' Preferences

5.2.1 Preferences in Baseline Condition

Figure 4a describes the cumulative distribution of LLMs' decisions in five scenarios. Human data from Mei et al. (2024) are also included as a reference. In the following, we compare the decisions between LLMs and human subjects for each scenario.

1. *Dictator game.* GPT, DeepSeek, and Qwen allocate more endowment to the other player in the dictator game, compared to humans (all $p < 0.05$).

2. *Proposer in ultimatum game.* Compared to human subjects, Qwen allocates more as the proposer in the ultimatum game ($p < 0.01$).

3. *Responder in ultimatum game.* All LLMs, except DeepSeek, are willing to accept less money as responders in the ultimatum game compared to human subjects (all $p < 0.01$). Notably, GPT consistently accepts as low as \$1 across all 100 simulations.

4. *Public goods game.* DeepSeek contributes more than human subjects in the public goods game, while Qwen contributes less (both $p < 0.01$).

5. *Bomb risk game.* GPT and Llama are more risk averse than human players; conversely, DeepSeek and Qwen dominantly choose to open 50 boxes—half of the total boxes—exhibiting less risk aversion than human subjects (all $p < 0.05$).

Together, the four LLMs exhibit divergent preferences within each scenario, implying considerable heterogeneity in decisions between models. Moreover, Figure 4a shows that decisions from open-sourced models with smaller parameter sizes, such as Llama and Qwen, exhibit greater dispersion, which suggests higher heterogeneity in the inferred preferences of these models. In contrast, the behavior of GPT and DeepSeek is more consistent.[15]

Moreover, GPT passes the Turing test in all games, whereas DeepSeek and Qwen each fail in one of the games, and Llama fails in two of them (Figure 4b). These results suggest that more

---

[15]For example, GPT consistently selects \$50 in the dictator game and sets the acceptance threshold at \$1 when acting as the responder in the ultimatum game. DeepSeek consistently contributes half of its endowment in the public goods game and opens half of the boxes in the bomb risk game. In a quantitative analysis, we calculate the normalized standard deviation of each LLM using $\tilde{\sigma}_i = \Sigma_{m \in \mathcal{M}} \frac{\sigma_{im}}{L_m |\mathcal{M}|}$. Here, σ_{im} denotes the standard deviation of LLM i in decision scenario $m \in \mathcal{M}$ in the baseline condition, and L_m represents the length of the feasible set interval of scenario m, which is 20 for the public goods game and 100 for the other scenarios. The normalized standard deviations $\tilde{\sigma}_i$ for GPT, DeepSeek, Llama, and Qwen are 0.054, 0.030, 0.218, and 0.110, respectively. For comparison, the corresponding value for human decisions is 0.231.
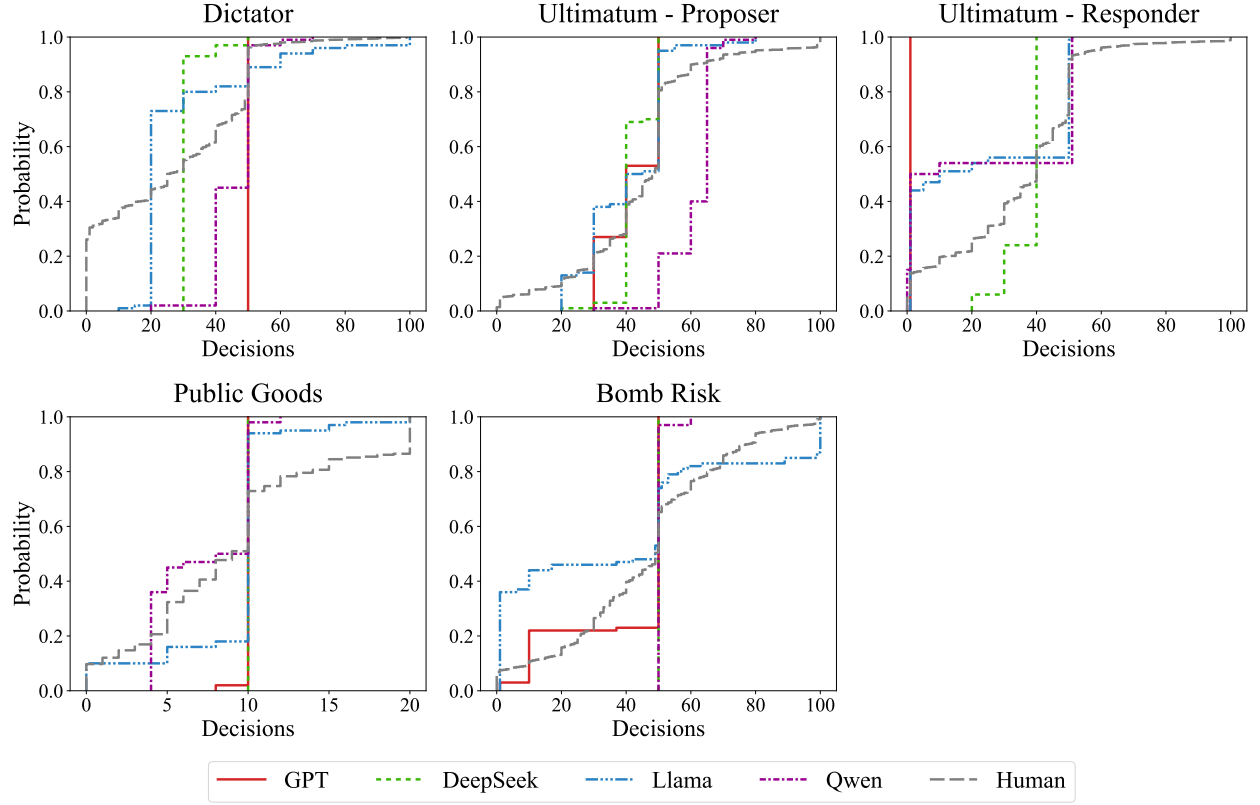
(a) Cumulative Distribution of Decisions
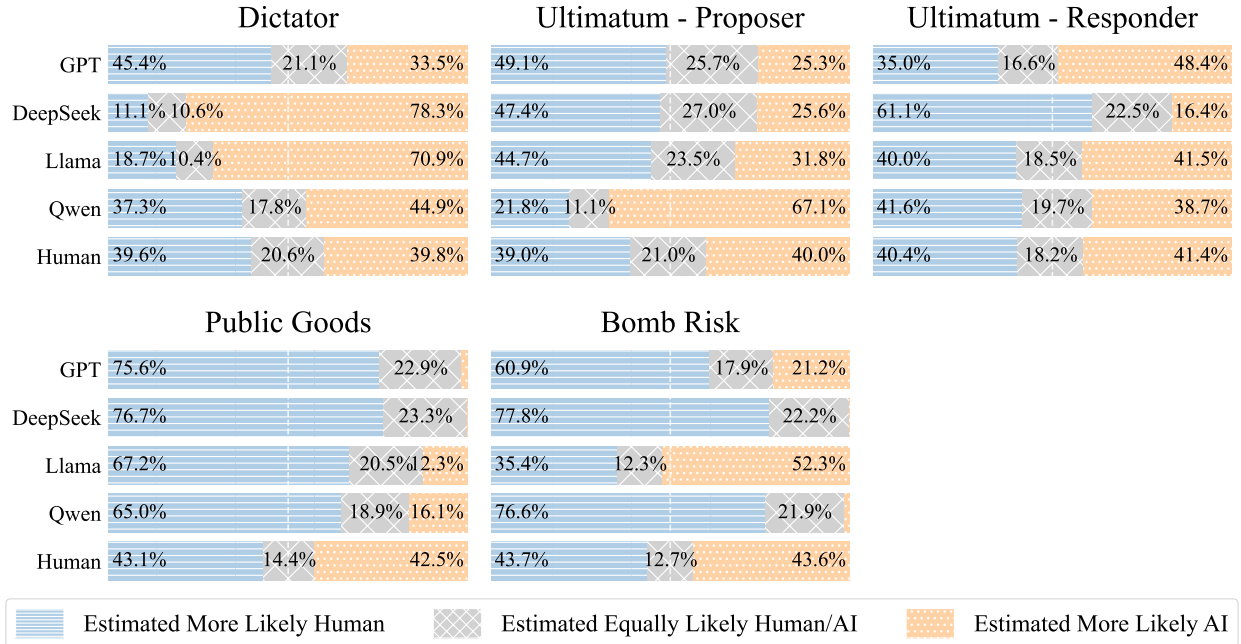
(b) Turing Test of Decisions

Figure 4: Baseline Results of Decisions.

Notes: Human data are from Mei et al. (2024). In Panel b, proportions below 5% are not labeled with numbers. The white dashed lines represent the probability of 50%. In the human condition, both CCEI scores in each comparison are randomly drawn from the human distribution.

advanced LLMs are capable of exhibiting human-like behavior, while less capable models fall short of this standard.

5.2.2 Preferences in Experimental Conditions

Persona. Two key observations emerge. First, the decisions of LLMs in behavioral games are more sensitive to personas than in budgetary tasks ($\lambda(\mathcal{S})$: 35.7% vs. 0%, $p < 0.01$, proportion test). Moreover, their choices in the public goods game are least likely to be influenced by personas among all games, with $\lambda_m(\mathcal{S})$ equal to 12.5%. Conversely, the sensitivity score increases significantly and reaches the highest when LLMs act as responders in the ultimatum game ($\lambda_m(\mathcal{S})$: 51.8% vs. 12.5%, $p < 0.01$, proportion test). These findings suggest that the impact of personas depends on the nature of the tasks evaluated.

Second, LLMs' outputs are more frequently influenced by occupational personas than by other demographic characteristics ($\lambda(\mathcal{S})$: 48.3% vs. 26.3%, $p < 0.01$, proportion test). This underscores that the effect of personas is dependent on the specific types of personas. Taken together, these observations suggest that we cannot draw definitive conclusions about how personas influence experimental results, as their impact varies across tasks and persona types.

Open-ended vs. multiple-choice answer type. Compared to the open-ended answer type in the baseline condition, responding within a restricted set of options significantly alters LLMs' outputs. Specifically, $\lambda(\mathcal{S})$ equals 55.0% under the multiple-choice answer type, indicating that the answer type significantly influences more than half of LLMs' decisions. Moreover, this value is marginally higher than that obtained under persona variations ($\lambda(\mathcal{S})$: 55.0% vs. 35.7%, $p = 0.084$, proportion test), demonstrating that LLMs' decisions are more sensitive to the answer type than personas.

6 Tactics for LLM Experiments

Drawing from key principles of human experiments, protocols in prior LLM research, and our own experimental findings, we present seven key tactics for conducting LLM experiments. These tactics serve two main purposes. First, they aim to align the procedures in LLM experiments with those in human experiments as closely as possible, thus facilitating the comparison in experimental results between LLM agents and human subjects. Second, with the burgeoning of LLM experiments, we hope these tips can be useful in improving the replicability of LLM experimental results. In the following, we will discuss each tactic by integrating insights from both experimental methods in economics and prompt engineering in artificial intelligence.
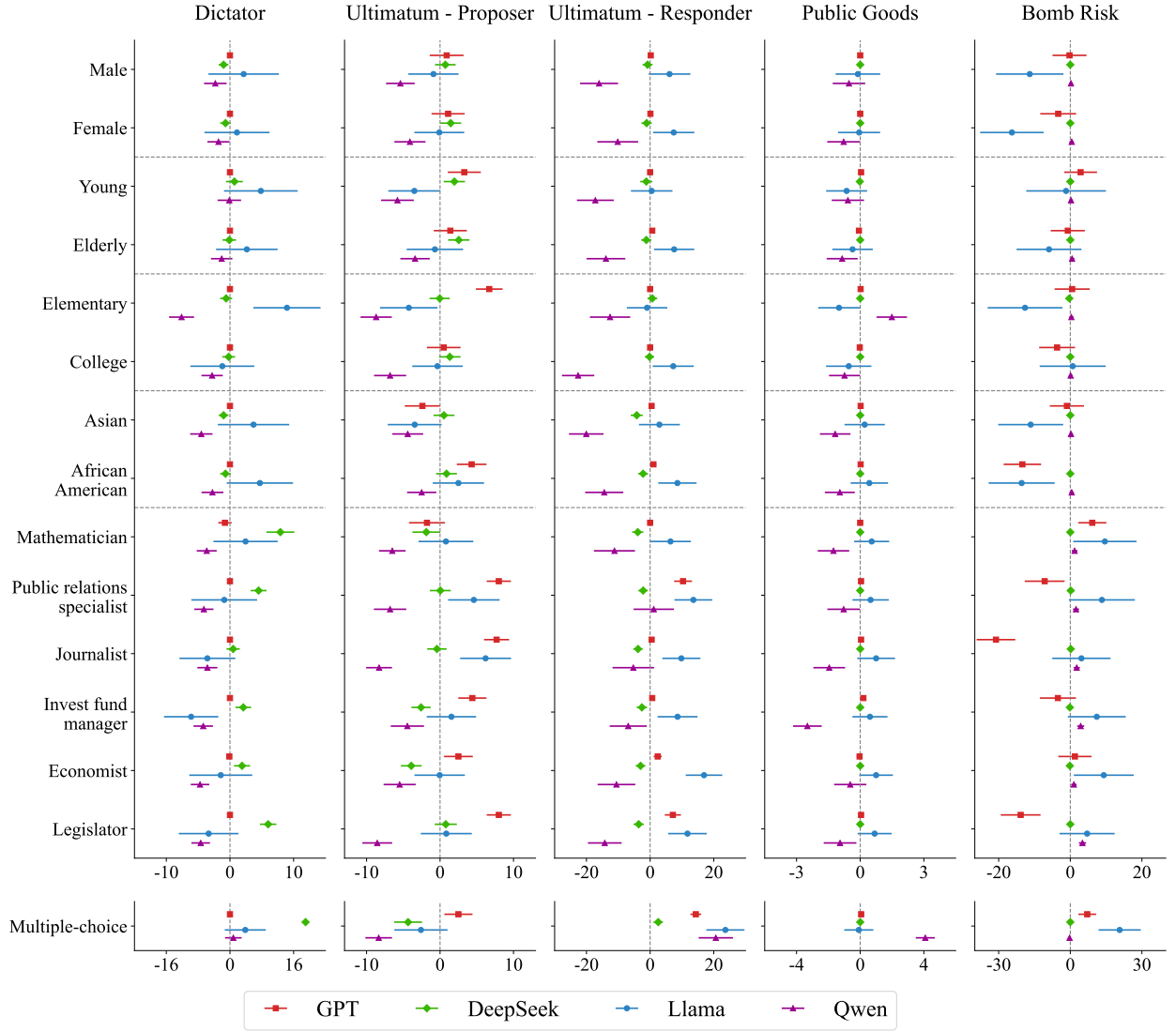
Figure 5: Comparison in Decisions of LLMs.

Notes: The markers represent differences in the mean value of LLMs' decisions between experimental and baseline conditions. The error bars represent 95% confidence intervals.

6.1 Tactics for Experimental Design

Temperature. The selection of temperature involves the trade-off between adopting the solution with higher probability and exploring a broader solution space (Renze and Guven, 2024). Human participants often exhibit a certain degree of heterogeneity. To mimic their behavior, we should allow the variation of responses between LLM agents. Meanwhile, both existing studies (Chen et al., 2023; Kosinski, 2024; Peeperkorn et al., 2024; Renze and Guven, 2024) and our results (Figure C.1) have shown that there are no significant behavioral differences in LLMs' outputs when different temperatures are used. Therefore, we suggest using the default temperatures recommended by model providers. When there are no default values, setting the temperature to 1 is a natural choice. This value does not scale the logits (see Section 2.3), thus best mirroring the application of softmax during the training process, which does not involve the temperature parameter (Radford et al., 2018). In addition, some model providers allow users to specify a random seed parameter. To enhance the replicability of results, researchers can set the seed when prompting LLMs via API. We summarize these discussions in the following.

Tactic 1: *Use the default temperature and fix the random seed.*

Persona. Following the conventions of experimental economics, where experimenters do not explicitly recruit participants with specific demographics unless this is part of the research design, we recommend not including specific personas in the prompts. Our two studies further support this tactic that the use of personas may introduce uncontrolled variables without consistently yielding significant or generalizable effects. In general, LLMs should be regarded as counterparts to representative individuals from the general human population. However, if the goal is to assess whether LLMs with diverse demographic characteristics respond differently to tasks (Chen et al., 2023; Mei et al., 2024; Zarifhonarvar, 2024; Wang et al., 2025), we can implement it by assigning different personas. The caveat here is that the comparison in results between such studies should be cautious due to the differences in selected tasks, evaluated models, and persona types.

Moreover, LLMs may fail to understand the meaning of assigned personas and under-represent the heterogeneity across different subgroups. For example, assigning male or female identities may cause LLMs' responses to shift in the same direction, compared to using no personas at all (Deng et al., 2025). LLMs with a persona also tend to resemble individuals without that identity describing those who possess it, rather than representing individuals with that identity. This probably comes from the nature of the training data: The demographic characteristics of the authors who generate these data are rarely disclosed. When such identities appear in the corpus, they are usually mentioned by others rather than self-identified by the individuals themselves (Wang et al., 2025).

Tactic 2*: Avoid assigning personas in prompts unless research requires.*

Incentive. Based on the induced value theory (Smith, 1976, 1982), the use of monetary incentives has become a standard practice in laboratory experiments, although the extent to which human behavior differs under real versus hypothetical incentives remains inconclusive (e.g., Holt and Laury, 2002; Laury et al., 2009; Brañas-Garza et al., 2023; Gneezy et al., 2024). Since real monetary rewards cannot be actually applied to LLMs, the feasible alternative is to instruct LLMs to maximize their hypothetical payoffs. Therefore, clear payment rules could be included in the prompts as in the instructions for human participants. However, unlike humans, LLMs may lack an intrinsic utility function; their data generation is driven by externally imposed optimization rules in training processes. This indicates that LLMs are less likely to report untruthfully, which is the issue that incentives are designed to address. Note that although LLMs may exhibit hallucination characterized by nonsensical or unfaithful responses, this phenomenon results from the training data, training processes, and decoding strategies (Huang et al., 2025) rather than strategic deception. Based on these facts and the experimental results (Figure C.1), we conjecture that LLMs' performance may change little when such hypothetical incentives are removed.

Tactic 3*: Use the typical payment scheme to be comparable to human experiments.*

6.2 Tactics for Experimental Implementation

Illustrative example and understanding question. In human experiments, particularly those using complex tasks, illustrative examples and understanding questions are routinely incorporated to facilitate comprehension. We apply the same principles for LLM experiments.

Tactic 4*: Incorporate examples and comprehension questions especially for complex tasks.*

Dialogue type. Since the structure of multi-turn dialogue is more similar to the sequential decision-making process of human subjects, we suggest using it for tasks involving multiple questions for a single agent. In contrast, although single-turn dialogue offers advantages in the efficiency of model inputs, it may constrain the economic rationality of certain models, particularly those with smaller parameter sizes (Figure 3).

Tactic 5*: Use multi-turn dialogue for tasks with multiple questions.*

Answer type. The construction of the answer format can influence model performance. For example, providing Likert scales or multiple choices may activate memory in the model training corpus, or enable models to infer experimental purposes and act in a demanding way (Kosinski, 2024). Furthermore, discretizing continuous feasible sets has been shown to significantly decrease the performance of models in both Chen et al. (2023) and our Study 1 (Figure 3). Consequently, the open-ended answer format is recommended rather than multiple-choice, unless the latter format is required by the nature of tasks. For example, the Linda problem requires multiple options with

inclusion relationships to measure the conjunction fallacy (Kahneman and Tversky, 1973).

Tactic 6: *Use an open-ended answer format whenever feasible.*

Invalid answer. In prior LLM experiments, the common practice is to exclude invalid answers from the dataset and continue data collection until the target sample size is reached (e.g., Chen et al., 2023; Brookins and DeBacker, 2024; Goli and Singh, 2024; Mei et al., 2024). However, the frequency of invalid answers also provides valuable information on model performance (Macmillan-Scott and Musolesi, 2024; Chen et al., 2025*b*). Therefore, although the content of invalid answers may remain unused, we believe that the proportion of such responses should be documented and reported as a performance metric.

Tactic 7: *Report the proportion of invalid responses.*

7   Conclusion

Integrating established principles from experimental economics with insights from artificial intelligence, this paper systematically investigates how key considerations from experimental design to implementation influence LLMs' outputs. We find that assigned personas have a considerable impact on economic preferences but do not affect economic rationality. Furthermore, the single-turn dialogue type reduces the economic rationality of two open-source models with small parameter sizes, Llama and Qwen, but not for the others, GPT and DeepSeek. Similarly, discretizing the continuous choice set into multiple choices leads to a reduction in rationality for Llama and Qwen, as well as substantial changes in decisions in behavioral games for all LLMs. These highlight the sensitivity of LLMs' performance to variations in experimental design and implementation. Based on these observations, we summarize seven tactics for designing and conducting LLM experiments. By providing a framework that can be applied consistently and be comparable across experiments, our study contributes to the replicability of experimental studies (Camerer et al., 2016, 2018).

We also acknowledge several limitations and outline directions for future research emerging from this work. First, we focus exclusively on the individual decision-making of LLMs. Extending this framework to accommodate multi-agent settings and human-AI hybrid experiments (Engel et al., 2023) will be important for broadening its applicability. Second, our study proposes a set of protocols and evaluates them in the context of text-based interactions. As LLMs' architectures continue to evolve, especially with the development of multimodal systems that integrate text, voice, image, and video data, further validation will be necessary to ensure the effectiveness of these protocols for next-generation models. Third, while we assess how design factors influence LLMs' performance, the underlying mechanisms are not understood sufficiently. Enhancing

the interpretability of LLMs' decision-making is crucial to validate whether observed behavior reflects genuine human-like reasoning or simply mimics patterns in training data (Olah et al., 2020; Bommasani et al., 2021; Black et al., 2022). Addressing these issues will improve experimental protocols and observations, develop interpretations and theories of LLMs' behavior, and support the effective application and integration of LLMs into everyday life.

References

Afriat, Sydney N. 1972. "Efficiency estimation of production functions." *International Economic Review*, 13(3): 568–598.

Alan, Sule, and Seda Ertac. 2018. "Fostering patience in the classroom: results from randomized educational intervention." *Journal of Political Economy*, 126(5): 1865–1911.

Alayrac, Jean-Baptiste, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millicah, Malcolm Reynolds, et al. 2022. "Flamingo: a visual language model for few-shot learning." In *Proceedings of the 36th International Conference on Neural Information Processing Systems*. 23716–23736.

Andreoni, James, and John Miller. 2002. "Giving according to GARP: an experimental test of the consistency of preferences for altruism." *Econometrica*, 70(2): 737–753.

Arkes, Hal R, and Catherine Blumer. 1985. "The psychology of sunk cost." *Organizational Behavior and Human Decision Processes*, 35(1): 124–140.

Benjamini, Yoav, and Yosef Hochberg. 1995. "Controlling the false discovery rate: a practical and powerful approach to multiple testing." *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1): 289–300.

Bergemann, Dirk, Alessandro Bonatti, and Alex Smolin. 2025. "The economics of large language models: token allocation, fine-tuning, and optimal pricing." In *Proceedings of the 26th ACM Conference on Economics and Computation*. 786.

Binz, Marcel, and Eric Schulz. 2023. "Using cognitive psychology to understand GPT-3." *Proceedings of the National Academy of Sciences*, 120(6): e2218523120.

Black, Sid, Lee Sharkey, Leo Grinsztajn, Eric Winsor, Dan Braun, Jacob Merizian, Kip Parker, Carlos Ramón Guevara, Beren Millidge, Gabriel Alfour, et al. 2022. "Interpreting neural networks through the polytope lens." *Working Paper*.

Bommasani, Rishi, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. "On the opportunities and risks of foundation models." *Working Paper*.

Brañas-Garza, Pablo. 2007. "Promoting helping behavior with framing in dictator games." *Journal of Economic Psychology*, 28(4): 477–486.

Brañas-Garza, Pablo, Diego Jorrat, Antonio M Espín, and Angel Sánchez. 2023. "Paid and hypothetical time preferences are the same: lab, field and online evidence." *Experimental Economics*, 26(2): 412–434.

Bronars, Stephen G. 1987. "The power of nonparametric tests of preference maximization." *Econometrica*, 55(3): 693–698.

Brookins, Philip, and Jason DeBacker. 2024. "Playing games with GPT: what can we learn about a large language model from canonical strategic games?" *Economics Bulletin*, 44(1): 25–37.

Brown, Tom, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. "Language models are few-shot learners." In *Proceedings of the 34th International Conference on Neural Information Processing Systems*. 1877–1901.

Brynjolfsson, Erik, Danielle Li, and Lindsey Raymond. 2025. "Generative AI at work." *Quarterly Journal of Economics*, 140(2): 889–942.

Camerer, Colin F, Anna Dreber, Eskil Forsell, Teck-Hua Ho, Jürgen Huber, Magnus Johannesson, Michael Kirchler, Johan Almenberg, Adam Altmejd, Taizan Chan, et al. 2016. "Evaluating replicability of laboratory experiments in economics." *Science*, 351(6280): 1433–1436.

Camerer, Colin F, Anna Dreber, Felix Holzmeister, Teck-Hua Ho, Jürgen Huber, Magnus Johannesson, Michael Kirchler, Gideon Nave, Brian A Nosek, Thomas Pfeiffer, et al. 2018. "Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015." *Nature Human Behaviour*, 2(9): 637–644.

Chang, Samuel, Andrew Kennedy, Aaron Leonard, and John A List. 2024. "12 best practices for leveraging generative AI in experimental research." *Working Paper*.

Charness, Gary, Brian Jabarian, and John A List. 2025. "The next generation of experimental research with LLMs." *Nature Human Behaviour*, 9(5): 833–835.

Chaudhari, Shreyas, Pranjal Aggarwal, Vishvak Murahari, Tanmay Rajpurohit, Ashwin Kalyan, Karthik Narasimhan, Ameet Deshpande, and Bruno Castro da Silva. 2025. "RLHF deciphered: a critical analysis of reinforcement learning from human feedback for LLMs." *ACM Computing Surveys*.

Chen, Mark, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde De Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. 2021. "Evaluating large language models trained on code." *Working Paper*.

Chen, Mingshi, Tracy Xiao Liu, You Shan, Shu Wang, Songfa Zhong, and Yanju Zhou. 2025*a*. "How general are measures of choice consistency? Evidence from experimental and scanner data." *Working Paper*.

Chen, Yang, Samuel N Kirshner, Anton Ovchinnikov, Meena Andiappan, and Tracy Jenkin. 2025*b*. "A manager and an AI walk into a bar: does chatGPT make biased decisions like we do?" *Manufacturing & Service Operations Management*, 27(2): 354–368.

Chen, Yiting, Tracy Xiao Liu, You Shan, and Songfa Zhong. 2023. "The emergence of economic rationality of GPT." *Proceedings of the National Academy of Sciences*, 120(51): e2316205120.

Chen, Zenan, and Jason Chan. 2024. "Large language model in creative work: The role of collaboration modality and user expertise." *Management Science*, 70(12): 9101–9117.

Choi, Syngjoo, Raymond Fisman, Douglas Gale, and Shachar Kariv. 2007. "Consistency and heterogeneity of individual behavior under uncertainty." *American Economic Review*, 97(5): 1921–1938.

Choi, Syngjoo, Shachar Kariv, Wieland Müller, and Dan Silverman. 2014. "Who is (more) rational?" *American Economic Review*, 104(6): 1518–1550.

Cobbe, Karl, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. 2021. "Training verifiers to solve math word problems." *Working Paper*.

Croson, Rachel. 2005. "The method of experimental economics." *International Negotiation*, 10(1): 131–148.

Croson, Rachel, and Uri Gneezy. 2009. "Gender differences in preferences." *Journal of Economic Literature*, 47(2): 448–474.

Davis, Douglas D, and Charles A Holt. 1993. *Experimental economics.* Princeton University Press.

Deng, Zebang, Rosemarie Nagel, Jubo Yan, and Tong Zhang. 2025. "Can LLM agents recognize demographic heterogeneity in economic games?" *Working Paper*.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. "BERT: pre-training of deep bidirectional transformers for language understanding." In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. 4171–4186.

Dohmen, Thomas, Armin Falk, David Huffman, Uwe Sunde, Jürgen Schupp, and Gert G Wagner. 2011. "Individual risk attitudes: measurement, determinants, and behavioral consequences." *Journal of the European Economic Association*, 9(3): 522–550.

Driess, Danny, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, et al. 2023. "PaLM-E: an embodied multimodal language model." In *Proceedings of the 40th International Conference on Machine Learning*. 8469–8488.

Echenique, Federico, Sangmok Lee, and Matthew Shum. 2011. "The money pump as a measure of revealed preference violations." *Journal of Political Economy*, 119(6): 1201–1223.

Ellis, Keaton, Shachar Kariv, and Erkut Ozbay. 2024. "Predicting and Understanding Individual-Level Choice Under Risk." *Working Paper*.

Engel, Christoph, Max RP Grossmann, and Axel Ockenfels. 2023. "Integrating machine behavior into human subject experiments: a user-friendly toolkit and illustrations." *Working Paper*.

Fedus, William, Barret Zoph, and Noam Shazeer. 2022. "Switch transformers: scaling to trillion parameter models with simple and efficient sparsity." *Journal of Machine Learning Research*, 23(120): 1–39.

Fedyk, Anastassia, Ali Kakhbod, Peiyao Li, and Ulrike Malmendier. 2025. "AI and perception biases in investments: an experimental study." *Working Paper*.

Filippas, Apostolos, John J Horton, and Benjamin S Manning. 2024. "Large language models as simulated economic agents: What can we learn from homo silicus?" In *Proceedings of the 25th ACM Conference on Economics and Computation*. 614–615.

Fisman, Raymond, Pamela Jakiela, Shachar Kariv, and Daniel Markovits. 2015. "The distributional preferences of an elite." *Science*, 349(6254): aab0096.

Fisman, Raymond, Pamela Jakiela, Shachar Kariv, and Silvia Vannutelli. 2023. "The distributional preferences of Americans, 2013–2016." *Experimental Economics*, 26(4): 727–748.

Fréchette, Guillaume R, Kim Sarnoff, and Leeat Yariv. 2022. "Experimental economics: Past and future." *Annual Review of Economics*, 14(1): 777–794.

Frederick, Shane. 2005. "Cognitive reflection and decision making." *Journal of Economic Perspectives*, 19(4): 25–42.

Friedman, Daniel, and Shyam Sunder. 1994. *Experimental methods: a primer for economists.* Cambridge University Press.

Gandhi, Kanishk, Ayush Chakravarthy, Anikait Singh, Nathan Lile, and Noah D Goodman. 2025. "Cognitive behaviors that enable self-improving reasoners, or, four habits of highly effective stars." In *Second Conference on Language Modeling*.

Gill, David, and Victoria Prowse. 2016. "Cognitive ability, character skills, and learning to play equilibrium: a level-k analysis." *Journal of Political Economy*, 124(6): 1619–1676.

Gneezy, Uri, Yoram Halevy, Brian Hall, Theo Offerman, and Jeroen van de Ven. 2024. "How real is hypothetical? a high-stakes test of the allais paradox." *Working Paper*.

Goli, Ali, and Amandeep Singh. 2024. "Frontiers: can large language models capture human preferences?" *Marketing Science*, 43(4): 709–722.

González Barman, Kristian, Simon Lohse, and Henk W de Regt. 2025. "Reinforcement learning from human feedback in LLMs: whose culture, whose values, whose perspectives?" *Philosophy & Technology*, 38(2): 1–26.

Hagendorff, Thilo, Sarah Fabi, and Michal Kosinski. 2023. "Human-like intuitive behavior and reasoning biases emerged in large language models but disappeared in chatGPT." *Nature Computational Science*, 3(10): 833–838.

Halevy, Yoram, Dotan Persitz, and Lanny Zrill. 2018. "Parametric recoverability of preferences." *Journal of Political Economy*, 126(4): 1558–1593.

Hansen, Anne Lundgaard, John J Horton, Sophia Kazinnik, Daniela Puzzello, and Ali Zarifhonarvar. 2024. "Simulating the survey of professional forecasters." *Working Paper*.

Holt, Charles A, and Susan K Laury. 2002. "Risk aversion and incentive effects." *American Economic Review*, 92(5): 1644–1655.

Holzmeister, Felix, Magnus Johannesson, Colin F Camerer, Yiling Chen, Teck-Hua Ho, Suzanne Hoogeveen, Juergen Huber, Noriko Imai, Taisuke Imai, Lawrence Jin, et al. 2025. "Examining the replicability of online experiments selected by a decision market." *Nature Human Behaviour*, 9(2): 316–330.

Huang, Lei, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. 2025. "A survey on hallucination in large language models: principles, taxonomy, challenges, and open questions." *ACM Transactions on Information Systems*, 43(2): 1–55.

Kahneman, Daniel, and Amos Tversky. 1973. "On the psychology of prediction." *Psychological Review*, 80(4): 237.

Kahneman, Daniel, Paul Slovic, and Amos Tversky. 1982. *Judgment under uncertainty: heuristics and biases.* Cambridge university press.

Kaplan, Jared, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. "Scaling laws for neural language models." *Working Paper*.

Kim, Alex G, and Valeri V Nikolaev. 2024. "Context-based interpretation of financial information." *Journal of Accounting Research*, forthcoming.

Kim, Hyuncheol Bryant, Syngjoo Choi, Booyuel Kim, and Cristian Pop-Eleches. 2018. "The role of education interventions in improving economic rationality." *Science*, 362(6410): 83–86.

Kim, Jeongbin, Matthew Kovach, Kyu-Min Lee, Euncheol Shin, and Hector Tzavellas. 2024. "Learning to be homo economicus: Can an LLM learn preferences from choice." *Working Paper*.

de Kok, Ties. 2025. "ChatGPT for textual analysis? How to use generative LLMs in accounting research." *Management Science*, forthcoming.

Kosinski, Michal. 2024. "Evaluating large language models in theory of mind tasks." *Proceedings of the National Academy of Sciences*, 121(45): e2405460121.

Laury, Susan K, Melayne Morgan McInnes, and J Todd Swarthout. 2009. "Insurance decisions for low-probability losses." *Journal of Risk and Uncertainty*, 39: 17–44.

Leng, Yan. 2024. "Can LLMs mimic human-like mental accounting and behavioral biases?" In *Proceedings of the 25th ACM Conference on Economics and Computation*. 581.

Leng, Yan, and Yuan Yuan. 2023. "Do LLM agents exhibit social behavior?" *Working Paper*.

Leng, Yan, Yunxin Sang, and Ashish Agarwal. 2024. "Reduce preference disparity between LLMs and humans: calibration to compose human-like ensembles." *Working Paper*.

Li, Jing, Lawrence P Casalino, Raymond Fisman, Shachar Kariv, and Daniel Markovits. 2022. "Experimental evidence of physician social preferences." *Proceedings of the National Academy of Sciences*, 119(28): e2112726119.

Li, Jing, William H Dow, and Shachar Kariv. 2017. "Social preferences of future physicians." *Proceedings of the National Academy of Sciences*, 114(48): E10291–E10300.

Liu, Nelson F, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. "Lost in the middle: how language models use long contexts." *Transactions of the Association for Computational Linguistics*, 12: 157–173.

Liu, Ryan, Jiayi Geng, Joshua C Peterson, Ilia Sucholutsky, and Thomas L Griffiths. 2025. "Large language models assume people are more rational than we really are." In *The Thirteenth International Conference on Learning Representations*.

Macmillan-Scott, Olivia, and Mirco Musolesi. 2024. "(Ir)rationality and cognitive biases in large language models." *Royal Society Open Science*, 11(6): 240255.

Maslej, Nestor, Loredana Fattorini, Raymond Perrault, Yolanda Gil, Vanessa Parli, Njenga Kariuki, Emily Capstick, Anka Reuel, Erik Brynjolfsson, John Etchemendy, et al. 2025. "The AI Index 2025 Annual Report." AI Index Steering Committee, Institute for Human-Centered AI, Stanford University, Stanford, CA.

Mei, Qiaozhu, Yutong Xie, Walter Yuan, and Matthew O. Jackson. 2024. "A turing test of whether ai chatbots are behaviorally similar to humans." *Proceedings of the National Academy of Sciences*, 121(9): e2313925121.

Meng, Juanjuan. 2024. "AI emerges as the frontier in behavioral science." *Proceedings of the National Academy of Sciences*, 121(10): e2401336121.

Mikolov, Tomas, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. "Efficient estimation of word representations in vector space." In *The First International Conference on Learning Representations*.

Muennighoff, Niklas, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. 2025. "s1: simple test-time scaling." *Working Paper*.

Niederle, Muriel. 2025. "Experiments: Why, How, and A Users Guide for Producers as well as Consumers." *Working Paper*.

Niu, Yimeng, Jing Wu, Shenyang Jiang, and Zhibin Jiang. 2025. "The bullwhip effect in servitized manufacturers." *Management Science*, 71(1): 1–20.

Noy, Shakked, and Whitney Zhang. 2023. "Experimental evidence on the productivity effects of generative artificial intelligence." *Science*, 381(6654): 187–192.

Olah, Chris, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. 2020. "Zoom In: An Introduction to Circuits." *Distill*. https://distill.pub/2020/circuits/zoom-in.

Ouyang, Shumiao, Hayong Yun, and Xingjian Zheng. 2024. "How ethical should AI be? How AI alignment shapes the risk preferences of LLMs." *Working Paper*.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. "BLEU: a method for automatic evaluation of machine translation." In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. 311–318.

Peeperkorn, Max, Tom Kouwenhoven, Dan Brown, and Anna Jordanous. 2024. "Is temperature the creativity parameter of large language models?" In *International Conference on Computational Creativity 2024*.

Peters, Matthew E, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. "Deep contextualized word representations." In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. 2227–2237.

Plott, Charles R, and Vernon L Smith. 1978. "An experimental examination of two exchange institutions." *Review of Economic Studies*, 45(1): 133–153.

Radford, Alec, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. "Improving language understanding by generative pre-training." OpenAI. Technical report.

Raven, John. 2000. "The Raven's progressive matrices: change and stability over culture and time." *Cognitive psychology*, 41(1): 1–48.

Razavi, Amirhossein, Mina Soltangheis, Negar Arabzadeh, Sara Salamat, Morteza Zihayat, and Ebrahim Bagheri. 2025. "Benchmarking prompt sensitivity in large language models." *Working Paper*.

Renze, Matthew, and Erhan Guven. 2024. "The effect of sampling temperature on problem solving in large language models." In *Findings of the Association for Computational Linguistics: EMNLP 2024*. 7346–7356.

Sahoo, Pranab, Ayush Kumar Singh, Sriparna Saha, Vinija Jain, Samrat Mondal, and Aman Chadha. 2024. "A systematic survey of prompt engineering in large language models: techniques and applications." *Working Paper*.

Shazeer, Noam, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. 2017. "Outrageously large neural networks: the sparsely-gated Mixture-of-Experts layer." In *The Fifth International Conference on Learning Representations*.

Siano, Federico. 2025. "The news in earnings announcement disclosures: Capturing word context using LLM methods." *Management Science*, forthcoming.

Smith, Vernon L. 1976. "Experimental economics: induced value theory." *American Economic Review*, 66(2): 274–279.

Smith, Vernon L. 1982. "Microeconomic systems as an experimental science." *American Economic Review*, 72(5): 923–955.

Sun, Fengfei, Ningke Li, Kailong Wang, and Lorenz Goette. 2025. "Large Language Models are overconfident and amplify human bias." *Working Paper*.

Suri, Gaurav, Lily R Slater, Ali Ziaee, and Morgan Nguyen. 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.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. "Attention is all you need." In *Proceedings of the 31st International Conference on Neural Information Processing Systems*. 6000–6010.

Wang, Angelina, Jamie Morgenstern, and John P Dickerson. 2025. "Large language models that replace human participants can harmfully misportray and flatten identity groups." *Nature Machine Intelligence*, 7: 400–411.

Wang, Pengda, Zilin Xiao, Hanjie Chen, and Frederick L Oswald. 2024. "Will the real Linda please stand up... to large language models? Examining the representativeness heuristic in LLMs." In *First Conference on Language Modeling*.

Webb, Taylor, Keith J Holyoak, and Hongjing Lu. 2023. "Emergent analogical reasoning in large language models." *Nature Human Behaviour*, 7(9): 1526–1541.

Wei, Jason, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. "Chain-of-thought prompting elicits reasoning in large language models." *Advances in Neural Information Processing Systems*, 35: 24824–24837.

Wu, Wenhao, Yizhong Wang, Guangxuan Xiao, Hao Peng, and Yao Fu. 2025. "Retrieval head mechanistically explains long-context factuality." In *The Thirteenth International Conference on Learning Representations*.

Xiao, Guangxuan, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. 2024. "Efficient streaming language models with attention sinks." In *The Twelfth International Conference on Learning Representations*.

Xie, Tian, Zitian Gao, Qingnan Ren, Haoming Luo, Yuqian Hong, Bryan Dai, Joey Zhou, Kai Qiu, Zhirong Wu, and Chong Luo. 2025. "Logic-RL: unleashing llm reasoning with rule-based reinforcement learning." *Working Paper*.

Xie, Yutong, Yiyao Liu, Zhuang Ma, Lin Shi, Xiyuan Wang, Walter Yuan, Matthew O Jackson, and Qiaozhu Mei. 2024. "How different AI chatbots behave? Benchmarking large language models in behavioral economics games." *Working Paper*.

Zarifhonarvar, Ali. 2024. "Evidence on Inflation Expectations Formation Using Large Language Models." *Working Paper*.

Zhou, Eric, and Dokyun Lee. 2024. "Generative artificial intelligence, human creativity, and art." *PNAS Nexus*, 3(3): pgae052.

Zhu, Jian-Qiao, Haijiang Yan, and Thomas L Griffiths. 2025. "Language models trained to do arithmetic predict human risky and intertemporal choice." In *The Thirteenth International Conference on Learning Representations*.

Zizzo, Daniel John. 2010. "Experimenter demand effects in economic experiments." *Experimental Economics*, 13: 75–98.

Online Appendices

A Additional Figures and Tables

Prompt 1

user:
Q1
Q2
Q3

(a) Single-turn dialogue

Prompt 1 ⇨ Prompt 2 ⇨ Prompt 3

user:
Q1

user:
Q1
Assistant:
A1
user:
Q2

user:
Q1
Assistant:
A1
user:
Q2
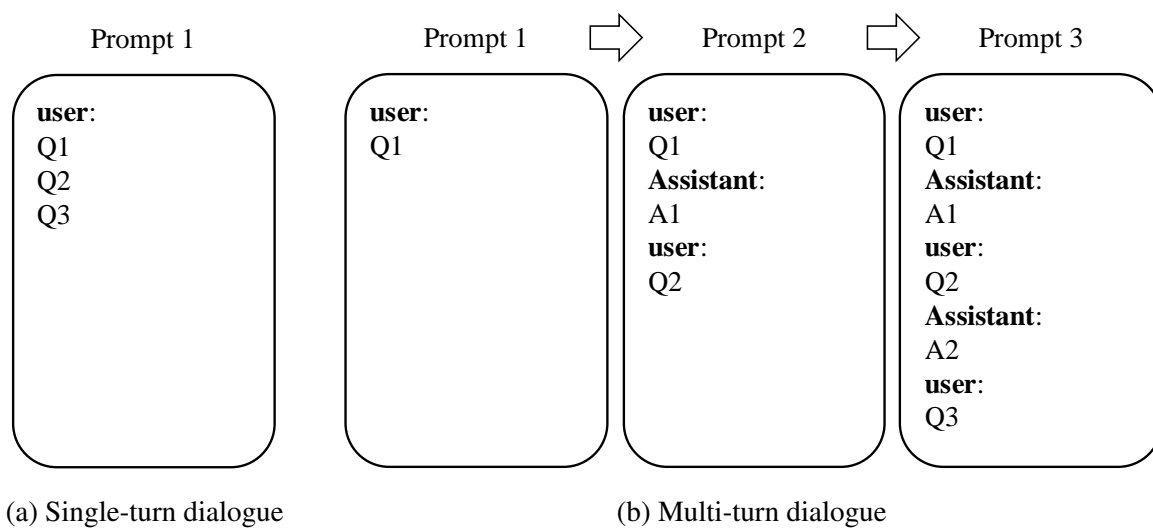**Assistant**:
A2
user:
Q3

(b) Multi-turn dialogue

Figure A.1: Illustration of Dialogue Type

Table A.1: Main Features of Representative LLMs

Type	Provider	Model	# Parameter	Context length	Chat vs. base	Human alignment	Unimodal vs. multimodal	Reasoning	Release date
Proprietary	OpenAI	GPT-3.5	Unknown	4K	Chat	Yes	Unimodal	No	Mar 2022
		GPT-4		8K			Multimodal	No	Mar 2023
		GPT-4o-mini		128K			Multimodal	No	Jul 2024
		OpenAI o3		200K			Multimodal	Yes	Apr 2025
	Google	PaLM 2		8K			Unimodal	No	May 2023
		Gemini 2.0 Flash		1M			Multimodal	Yes	Feb 2025
	Anthropic	Claude		9K			Unimodal	No	Mar 2023
		Claude Opus 4		200K			Multimodal	Yes	May 2025
Open-sourced	Meta	LLaMA-65B	65B	2K	Base	No	Unimodal	No	Feb 2023
		Llama-3.1-8B	8B	128K	Base&Chat	Yes (Chat version)	Unimodal	No	Jul 2024
	Alibaba Cloud	Qwen2.5-7B	7B	128K	Base&Chat	Yes (Chat version)	Unimodal	No	Sept 2024
		Qwen3-32B	32B	128K	Base&Chat	Yes (Chat version)	Unimodal	Yes	Apr 2025
	Deepseek	DeepSeek-V2	236B	128K	Base&Chat	Yes (Chat version)	Unimodal	No	May 2024
		DeepSeek-V3	671B	128K	Base&Chat	Yes (Chat version)	Unimodal	No	Dec 2024
	Zhipu AI	GLM-130B	130B	2K	Base	No	Unimodal	No	Aug 2022
		GLM-4-32B-0414	32B	32K	Base&Chat	Yes (Chat Version)	Unimodal	No	Apr 2025

Notes: All information is based on the initial release version. "Unknown" means the information is not disclosed by the provider.

B Sample Instructions

In this section, we will take the instructions in Plott and Smith (1978) as an illustrative example to explain the rules listed in Section 3.1. The instructions are also reprinted in Friedman and Sunder (1994) in the appendix. The whole original instructions include four parts: General instructions, specific instructions for sellers, specific instructions for buyers, and market organizations. For simplicity concern, we include the first two parts to illustrate. Additional explanations are added in square brackets with underlines.

"*General*

*This is an experiment in the economics of market decision making. Various research foundations have provided funds for the conduct of this research. The instructions are simple, and if you follow them carefully and make good decisions you might earn a considerable amount of money which will be paid to you in cash after the experiment. [**Make instructions comprehensible. Follow the induced value theory.**]*

*In this experiment we are going to simulate a market in which some of you will be buyers and some of you will be sellers in a sequence of market days or trading periods. Two kinds of sheets will now be distributed—information for buyers and information for sellers. The sheets are identified and numbered. The number is only for data-collecting purposes. If you have received sellers' information, you will function only as a seller in this market. Similarly, if you have received buyers' information, you will function only as a buyer in this market. The information you have received is for your own private use. Do not reveal it to anyone. [**Prevent inter-subject communication.**]*

*This is a one commodity market in which there is no product differentiation. [**Use context-neutral instructions.**] That is, each seller produces a product which is similar in all respects to the products offered by the other sellers. A seller is free to sell to any buyer or buyers. Likewise, a buyer may purchase from any seller or sellers.*

Specific Instructions for Sellers

During each market period, you are free to produce and sell any of the amounts listed on your information sheet. Assume that you produce only for immediate sale—there are no inventories. The dollar amounts listed in column 2 on your information sheet are your costs of producing that quantity.

Your payoffs are computed as follows: At the beginning of the experiment you will receive starting capital of $0.30. If you are able to make any sales, you will receive the difference between your sales revenue and your cost. For example, if you were to sell two units at $100 each, total revenue would be $200. Suppose your information sheet indicated that the cost of producing two units was $190. Your total profit would then be $200−$190 = $10 for the trading period. If you sold two

*units for less than $190 you would incur a loss. Column 3 will be useful to a seller in deciding at any time during a given trading period whether to sell an additional unit. Suppose a seller has already sold one unit at a profit, and wants to know if he should sell a second unit. If the additional cost of producing the second unit is $10, then he will lose money on that unit if he sells it at any price below $10. [**Provide illustrative examples.**] Obviously, these figures are illustrative only and should not be assumed to apply to the actual sellers in this experiment. [**Minimize experimenter demand effects.**]*

All of your profits will be added to your starting capital, and any losses you might incur will be subtracted. Your total payoffs will be accumulated over several trading periods and the total amount will be paid to you after the experiment."

C   Additional Analysis in Study 1

Temperature   We examine the impact of temperature from 0 to 1 with a step of 0.1, and plot the differences in average CCEI scores between each condition and the baseline with the temperature of 1 in Figure C.1. We do not observe any significant effect of temperature on CCEI scores in both risk and social preference domains (all $p > 0.1$).

Incentive   As induced value theory emphasizes the crucial role of monetary incentives in human experiments, we assess whether the existence of hypothetical incentives and stake sizes affect economic rationality in LLM experiments. We perform the following conditions independently. First, we remove the payment scheme from the prompts. Second, we enlarge the stakes by 10, 100 and 1,000 times separately to examine the effect of stake sizes. Section D.2.2 details the prompts and Figure C.1 presents the results.

Removing incentive slightly reduces the CCEI of GPT by 0.004 in risk preference ($p = 0.023$). Although statistically significant, the difference is relatively small in magnitude. Furthermore, when enlarging the stake size by 10, 100 and 1000 times, the CCEI scores of Qwen are 0.932, 0.899, 0.802 in the risk preference, and 0.991, 0.968, 0.935 in the social preference, respectively. Most of these scores are significantly lower than CCEI scores in the baseline condition (all $p < 0.05$ except 0.991). Additionally, larger stake sizes result in lower rationality of Qwen (risk: stake size$\times$10 vs. stake size$\times$100, $p = 0.049$, stake size$\times$100 vs. stake size$\times$1000, $p < 0.01$; social: stake size$\times$10 vs. stake size$\times$100, $p = 0.022$, stake size$\times$100 vs. stake size$\times$1000, $p = 0.061$). The above findings suggest that the existence of incentives does not substantially affect the economic rationality of LLMs; while the rationality of Qwen substantially decreases as stake sizes increase.

Example   We evaluate the impact of the example by removing it from the baseline condition. As illustrated in Figure C.1, no significant differences are observed in the risk preference domain. In terms of social preference, GPT generates slightly higher CCEI when removing the example, while DeepSeek and Qwen result in lower CCEI scores compared to the baseline (all $p < 0.01$).
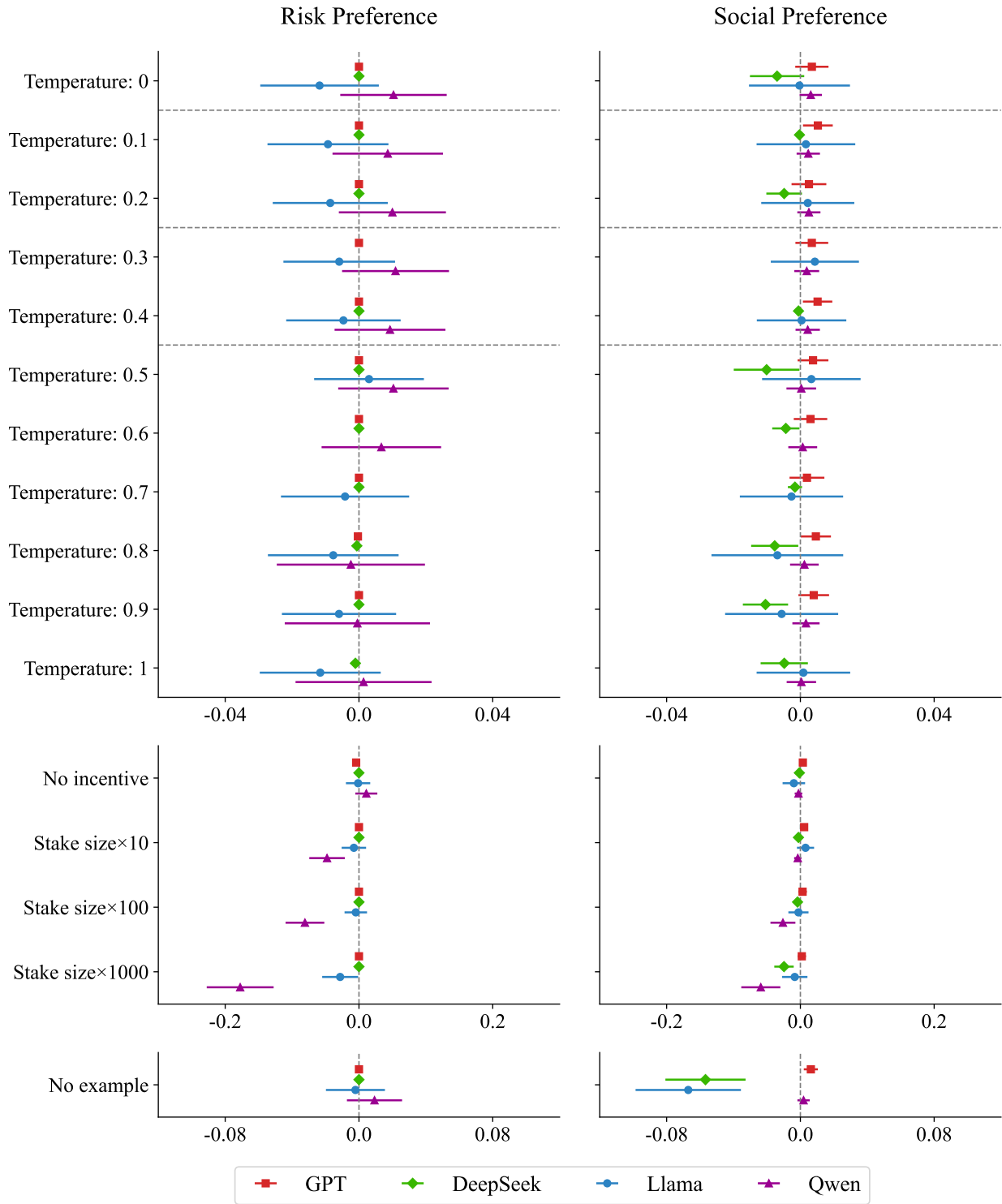
Figure C.1: Comparison in Rationality of LLMs.

Notes: The baseline condition is omitted in the figure, where the corresponding temperature is 1 for GPT, 0.3 for DeepSeek, 0.6 for Llama and 0.7 for Qwen. The markers represent differences in average CCEI scores between experimental and baseline conditions. The error bars represent 95% confidence intervals.

D   Prompts in Study 1

D.1   Risk Preference

D.1.1   Baseline Condition

In the first round of each simulation, the prompts used are listed below. {} represents the randomly generated number in each round.

System Message.   I want you to act as a human decision maker. You will be given 25 rounds of decision-making tasks and will be responsible for making decisions. I will randomly choose one of your 25 decisions to determine your payoff as described in the task. You should use your best judgment to come up with solutions that maximize your payoff. My first request is "You must provide your answers in every round."

User Message.   In every round, you have 100 points that need to be invested between Asset A and Asset B. You have a 50% chance to get the return from Asset A or the other 50% chance to get the return from Asset B. First please tell me the number of points for investing Asset A, then please tell me the number of points for investing Asset B.

Here is an example: [Suppose that you invest 90 points to Asset A, and 10 points to Asset B. In this round, investing every 1 point for Asset A returns 0.8 dollars, and investing every 1 point for Asset B returns 0.2 dollars. Your payoff will be 90*0.8=72 dollars from Asset A with a 50% chance and 10*0.2=2 dollars from Asset B with another 50% chance.]

Please provide a Markdown code snippet that adheres to the following format: It must begin with "```json" and end with "```", and should not include any notes or comments. Additionally, the JSON content must conform to the following specified schema.

```json
{
"Points for investing Asset A": float,
"Points for investing Asset B": float
}```

In this round, investing every 1 point for Asset A returns {} dollars, and investing every 1 point for Asset B returns {} dollars. What is your allocation?

Please output the points for investing Asset A and the points for investing Asset B using the specified schema as shown above.

In subsequent rounds after the first round of each simulation, the system prompts remain the same, and the input incorporates both the prior user message and LLMs' responses (Figure A.1b). The new question in the user message for the current round is listed below.

**User Message**  In this round, investing every 1 point for Asset A returns {} dollars, and investing every 1 point for Asset B returns {} dollars. What is your allocation?
Please output the points for investing Asset A and the points for investing Asset B using the specified schema as shown above.

### D.1.2  Experimental Conditions

#### Persona

For personas except occupations, we only vary the first sentence in the system message of the baseline condition, i.e., "I want you to act as a human decision maker" (Chen et al., 2023). The variations are listed below.

- Male: I want you to act as a male decision maker.
- Female: I want you to act as a female decision maker.
- Young: I want you to act as a young child decision maker.
- Elderly: I want you to act as an elderly decision maker.
- Elementary: I want you to act as a decision maker with an elementary school education.
- College: I want you to act as a decision maker with a college education.
- Asian: I want you to act as an Asian decision maker.
- African American: I want you to act as an African American decision maker.

For occupations, in addition to changing the role, the description of core tasks and supplemental tasks are also included in the system message (Mei et al., 2024). The example of the mathematician is as follows.

I want you to act as a mathematician.
You core tasks include:
* Address the relationships of quantities, ...
* Disseminate research by writing reports, ...
* Maintain knowledge in the field by ...
...

Your supplemental tasks include:

* Design, analyze, and decipher encryption systems ... "

**Dialogue Type**

In single-turn dialogue, the system message is the same as in the baseline condition (Section D.1.1), while LLMs respond to 25 rounds of questions in a single prompt (Figure A.1a) for each simulation. The user prompt is as follows.

**User Message**   In every round, you have 100 points that need to be invested between Asset A and Asset B. You have a 50% chance to get the return from Asset A or the other 50% chance to get the return from Asset B. First please tell me the number of points for investing Asset A, then please tell me the number of points for investing Asset B.

Here is an example: [Suppose that you invest 90 points to Asset A, and 10 points to Asset B. In this round, investing every 1 point for Asset A returns 0.8 dollars, and investing every 1 point for Asset B returns 0.2 dollars. Your payoff will be 90*0.8=72 dollars from Asset A with a 50% chance and 10*0.2=2 dollars from Asset B with another 50% chance.]

Please provide a Markdown code snippet that adheres to the following format: It must begin with "``` json" and end with "```", and should not include any notes or comments. Additionally, the JSON content must conform to the following specified schema.

```json
{
"Points for investing Asset A": float,
"Points for investing Asset B": float
}```

1. In this round, investing every 1 point for Asset A returns {} dollars, and investing every 1 point for Asset B returns {} dollars. What is your allocation?

2. In this round, investing every 1 point for Asset A returns {} dollars, and investing every 1 point for Asset B returns {} dollars. What is your allocation?

...

25. In this round, investing every 1 point for Asset A returns {} dollars, and investing every 1 point for Asset B returns {} dollars. What is your allocation?

Please output the points for investing Asset A and the points for investing Asset B in all 25 rounds using the specified schema as shown above. Namely, you should output 25 JSON, each conforming to the specified schema as shown above.

**Answer Type**

In multiple-choice conditions, the system message is the same as in the baseline condition. The user message is as follows.

> **User Message.** In every round, you will be presented with 21 options, each represented in the form (M, N), which means investing M points for Asset A and N points for Asset B. You have a 50% chance to get the return from Asset A or the other 50% chance to get the return from Asset B.
>
> Here is an example: [Suppose that you invest 90 points to Asset A, and 10 points to Asset B. In this round, investing every 1 point for Asset A returns 0.8 dollars, and investing every 1 point for Asset B returns 0.2 dollars. Your payoff will be 90*0.8=72 dollars from Asset A with a 50% chance and 10*0.2=2 dollars from Asset B with another 50% chance.]
>
> Please provide a Markdown code snippet that adheres to the following format: It must begin with "```json" and end with "``` ", and should not include any notes or comments. Additionally, the JSON content must conform to the following specified schema.
>
> ```json
> {
> "Points for investing Asset A": float,
> "Points for investing Asset B": float
> }```
>
> In this round, investing every 1 point for Asset A returns {} dollars, and investing every 1 point for Asset B returns {} dollars. There are 21 options, which are (0,100), (5,95), (10,90), (15,85), (20,80), (25,75), (30,70), (35,65), (40,60), (45,55), (50,50), (55,45), (60,40), (65,35), (70,30), (75,25), (80,20), (85,15), (90,10), (95,5), (100,0). Which is the best?
>
> Please output the points for investing Asset A and the points for investing Asset B using the specified schema as shown above.

**Other Variations**

**No Example.** To exclude the example, we remove the second paragraph from the user message in the baseline condition.

**No Incentive.** Compared to the baseline condition, we keep user message as the same while changing the system message as follows.

**System Message.** I want you to act as a human decision maker. You will be given 25 rounds of decision-making tasks and will be responsible for making decisions. You should use your best judgment to come up with solutions that you like most. My first request is "You must provide your answers in every round."

**Stake Size.** Based on the baseline condition, we scale the stake size by multiplying the number in {} within the user message by 10, 100 and 1,000. Other prompts remain the same.

## D.2    Social Preference

### D.2.1    Baseline Condition

In the first round of each simulation, the prompts used are listed below. {} represents the randomly generated number in each round.

**System Message.**   I want you to act as a human decision maker. You will be given 25 rounds of decision-making tasks and will be responsible for making decisions. I will randomly choose one of your 25 decisions to determine your payoff as described in the task. You should use your best judgment to come up with solutions that you like most. My first request is "You must provide your answers in every round."

**User Message.**   In every round, you are randomly matched with a new anonymous subject and there is no feedback across rounds. You have 100 points that need to be allocated between yourself and the other one. You will get return from the points allocated to yourself and the other one will get return from the points allocated to him/her. First please tell me the number of points allocated to yourself, then please tell me the number of points allocated to the other one.

Here is an example: [Suppose that you allocate 90 points to yourself, and 10 points to the other. In this round, allocating every 1 point to yourself returns 0.8 dollars, and allocating every 1 point to the other returns 0.2 dollars. Your payoff will be 90*0.8=72 dollars, and the other's payoff will be 10*0.2=2 dollars.]

Please provide a Markdown code snippet that adheres to the following format: It must begin with "``` json" and end with "```", and should not include any notes or comments. Additionally, the JSON content must conform to the following specified schema.
``` json
{
"Points allocated to yourself": float,
"Points allocated to the other one": float
```

}```

In this round, allocating every 1 point to yourself returns {} dollars, and allocating every 1 point to the other returns {} dollars. What is your allocation?

Please output the points allocated to yourself and points allocated to the other one using the specified schema as shown above.

In subsequent rounds after the first round of each simulation, the system prompts remain the same, and the input incorporates both the prior user message and LLM's responses (Figure A.1b). The new question in the user message for the current round is listed below.

User Message In this round, allocating every 1 point to yourself returns {} dollars, and allocating every 1 point to the other returns {} dollars. What is your allocation?

Please output the points allocated to yourself and points allocated to the other one using the specified schema as shown above.

D.2.2 Experimental Conditions

Persona

The variation of personas is the same as that introduced in Section D.1.2.

Dialogue Type

In single-turn dialogue, the system message is the same as in the baseline condition (Section D.2.1), while LLMs respond to 25 rounds of questions in a single prompt (Figure A.1a) for each simulation. The user prompt is as follows.

User Message In every round, you are randomly matched with a new anonymous subject and there is no feedback across rounds. You have 100 points that need to be allocated between yourself and the other one. You will get return from the points allocated to yourself and the other one will get return from the points allocated to him/her. First please tell me the number of points allocated to yourself, then please tell me the number of points allocated to the other one.

Here is an example: [Suppose that you allocate 90 points to yourself, and 10 points to the other. In this round, allocating every 1 point to yourself returns 0.8 dollars, and allocating every 1 point to the other returns 0.2 dollars. Your payoff will be 90*0.8=72 dollars, and the other's payoff will be 10*0.2=2 dollars.]

Please provide a Markdown code snippet that adheres to the following format: It must begin with "```json" and end with "```", and should not include any notes or

comments. Additionally, the JSON content must conform to the following specified schema.

```json
{
"Points allocated to yourself": float,
"Points allocated to the other one": float
}```

1. In this round, allocating every 1 point to yourself returns {} dollars, and allocating every 1 point to the other returns {} dollars. What is your allocation?

2. In this round, allocating every 1 point to yourself returns {} dollars, and allocating every 1 point to the other returns {} dollars. What is your allocation?

...

25. In this round, allocating every 1 point to yourself returns {} dollars, and allocating every 1 point to the other returns {} dollars. What is your allocation?

Please output the points allocated to yourself and points allocated to the other one in all 25 rounds using the specified schema as shown above. Namely, you should output 25 JSON, each conforming to the specified schema as shown above.

## Answer Type

In multiple-choice conditions, the system message is the same as in the baseline condition. The user message is as follows.

**User Message.** In every round, you are randomly matched with a new anonymous subject and there is no feedback across rounds. You will be presented with 21 options, each represented in the form (M, N), which means allocating M points to yourself and N points to the other one. First please tell me the number of points allocated to yourself, then please tell me the number of points allocated to the other one.

Here is an example: [Suppose that you allocate 90 points to yourself, and 10 points to the other. In this round, allocating every 1 point to yourself returns 0.8 dollars, and allocating every 1 point to the other returns 0.2 dollars. Your payoff will be 90*0.8=72 dollars, and the other's payoff will be 10*0.2=2 dollars.]

Please provide a Markdown code snippet that adheres to the following format: It must begin with "```json" and end with "``` ", and should not include any notes or comments. Additionally, the JSON content must conform to the following specified schema.

```json
```

{

"Points allocated to yourself": float,

"Points allocated to the other one": float

}```

In this round, allocating every 1 point to yourself returns {} dollars, and allocating every 1 point to the other returns dollars. There are 21 options, which are (0,100), (5,95), (10,90), (15,85), (20,80), (25,75), (30,70), (35,65), (40,60), (45,55), (50,50), (55,45), (60,40), (65,35), (70,30), (75,25), (80,20), (85,15), (90,10), (95,5), (100,0). Which is the best?

Please output the points for yourself and points allocated to the other one using the specified schema as shown above.

Other Variations

No Example. To exclude the example, we remove the second paragraph from the user message in the baseline condition.

No Incentive. Compared to baseline condition, we keep user message as the same while changing the system message as follows.

> **System Message.** I want you to act as a human decision maker. You will be given 25 rounds of decision-making tasks and will be responsible for making decisions. You should use your best judgment to come up with solutions that you like most. My first request is "You must provide your answers in every round."

Stake Size. Based on the baseline condition, we scale the stake size by multiplying the number in {} within the user message by 10, 100 and 1,000. Other prompts remain the same.

E Prompts in Study 2

E.1 Dictator Game

E.1.1 Baseline Condition

In the baseline condition, we follow the approach in Mei et al. (2024) to first input "Hi, let's play a game", followed by the formal question. Two user messages are prompted sequentially. Such procedure is applied throughout the Study 2.

> **System Message.** You are a helpful assistant.
>
> **User Message 1.** Hi, let's play a game.
>
> **User Message 2.** You are paired with another player. Your role is to decide how to divide $100 and the other player simply receives your choice.
>
> Here is an example: [Suppose that you decide to keep $60 and give the other player $40. Your payoff will be $60, and the other player's payoff will be $40.]
>
> How would you like to divide the money? Please only give one concrete choice and highlight the amount you give to the other player in [[]] (such as [[$x]]).

E.1.2 Experimental Conditions

Persona

For persona except occupations, we only vary the prompts in the system message of the baseline condition, i.e., "You are a helpful assistant" (Chen et al., 2023). The variations are listed below.

- Male: You are a helpful male assistant.
- Female: You are a helpful female assistant.
- Young: You are a helpful young child assistant.
- Elderly: You are a helpful elderly assistant.
- Elementary: You are a helpful assistant with an elementary school education.
- College: You are a helpful assistant with a college education.
- Asian: You are a helpful Asian assistant.
- African American: You are a helpful African American assistant.

For occupations, in addition to changing the role, the description of core tasks and supplemental tasks are also included in the system message (Mei et al., 2024). The example of the mathematician is as follows.

You are a mathematician.

You core tasks include:

* Address the relationships of quantities, ...

* Disseminate research by writing reports, ...

* Maintain knowledge in the field by ...

...

Your supplemental tasks include:

* Design, analyze, and decipher encryption systems ... "

Answer Type

The System Message and User Message 1 is the same as those in the baseline condition.

User Message 2. You are paired with another player. Your role is to decide how to divide $100 and the other player simply receives your choice.

Here is an example: [Suppose that you decide to keep $60 and give the other player $40. Your payoff will be $60, and the other player's payoff will be $40.]

You will be presented with 21 options, each represented in the form (M), which means giving $M to the other player. The 21 options are (0), (5), (10), (15), (20), (25), (30), (35), (40), (45), (50), (55), (60), (65), (70), (75), (80), (85), (90), (95), (100). How would you like to divide the money? Please only give one concrete choice and highlight the amount you give to the other player in [[]] (such as [[$x]]).

E.2 Proposer in Ultimatum Game

E.2.1 Baseline Condition

System Message. You are a helpful assistant.

User Message 1. Hi, let's play a game.

User Message 2. This is a two-player game. You are the Proposer, and the other player is the Responder. As the Proposer, you propose how to divide $100 and the Responder chooses either Accept or Reject. If accepted, the two of you will earn as described by the accepted proposal accordingly. If rejected, then both of you will earn $0.

Here is an example: [Suppose that you propose to keep $60 and give the Responder $40. The Responder accepts your proposal. Your payoff will be $60, and the Responder's payoff will be $40.]

How much would you like to propose to give to the Responder? Please only give one concrete choice, and highlight the amount with [[]] (such as [[$x]]).

E.2.2 Experimental Conditions

Persona

The variation of personas is the same as that introduced in Section E.1.2.

Answer Type

The System Message and User Message 1 is the same as those in the baseline condition.

User Message 2. This is a two-player game. You are the Proposer, and the other player is the Responder. As the Proposer, you propose how to divide $100 and the Responder chooses either Accept or Reject. If accepted, the two of you will earn as described by the accepted proposal accordingly. If rejected, then both of you will earn $0.

Here is an example: [Suppose that you propose to keep $60 and give the Responder $40. The Responder accepts your proposal. Your payoff will be $60, and the Responder's payoff will be $40.]

You will be presented with 21 options, each represented in the form (M), which means proposing $M to the Responder. The 21 options are (0), (5), (10), (15), (20), (25), (30), (35), (40), (45), (50), (55), (60), (65), (70), (75), (80), (85), (90), (95), (100). How much would you like to propose to give to the Responder? Please only give one concrete choice, and highlight the amount with [[]] (such as [[$x]]).

E.3 Responder in Ultimatum Game

E.3.1 Baseline Condition

System Message. You are a helpful assistant.

User Message 1. Hi, let's play a game.

User Message 2. This is a two-player game. You are the Responder, and the other player is the Proposer. The Proposer proposes how to divide $100 and you, as the Responder, choose either Accept or Reject. If accepted, the two of you will earn as described by the accepted proposal accordingly. If rejected, then both of you will earn $0.

Here is an example: [Suppose that the Proposer proposes to keep $60 and give you $40. You accept the proposal. The Proposer's payoff will be $60, and your payoff will be $40.]

As the Responder, what is the minimal amount in the proposal that you would accept? Please only give one concrete choice, and highlight the amount with [[]] (such as [[$x]]).

E.3.2 Experimental Conditions

Persona

The variation of personas is the same as that introduced in Section E.1.2.

Answer Type

The System Message and User Message 1 is the same as those in the baseline condition.

> **User Message 2.** This is a two-player game. You are the Responder, and the other player is the Proposer. The Proposer proposes how to divide $100 and you, as the Responder, choose either Accept or Reject. If accepted, the two of you will earn as described by the accepted proposal accordingly. If rejected, then both of you will earn $0.
>
> Here is an example: [Suppose that the Proposer proposes to keep $60 and give you $40. You accept the proposal. The Proposer's payoff will be $60, and your payoff will be $40.]
>
> You will be presented with 21 options, each represented in the form (M), where $M is the minimal amount in the proposal that you would accept. The 21 options are (0), (5), (10), (15), (20), (25), (30), (35), (40), (45), (50), (55), (60), (65), (70), (75), (80), (85), (90), (95), (100). As the Responder, what is the minimal amount in the proposal that you would accept? Please only give one concrete choice, and highlight the amount with [[]] (such as [[$x]]).

E.4 Public Goods Game

E.4.1 Baseline Condition

> **System Message.** You are a helpful assistant.
>
> **User Message 1.** Hi, let's play a game.
>
> **User Message 2.** In this public good game, you and 3 others will choose how much to contribute towards a water cleaning project. Each of you gets $20 per round to contribute between $0 and $20. The project has a 50% return rate. Your payoff relies on both your and others' contributions. Everyone benefits from the group's total contribution. Your payoff in each round equals the amount you didn't contribute (endowment - contribution) plus everyone's benefit (total contributions * 50% return rate).
>
> Here is an example: [Suppose that you contribute $12, and total group contributions are $20. Your payoff will be ($20-$12) + $20*50% = $18]

We will play a total of 3 rounds, in the first round, how much of the $20 would you like to contribute? Please only give one concrete number and highlight it with [[]] (e.g., [[x]]).

E.4.2 Experimental Conditions

Persona

The variation of personas is the same as that introduced in Section E.1.2.

Answer Type

The System Message and User Message 1 is the same as those in the baseline condition.

User Message 2.   In this public good game, you and 3 others will choose how much to contribute towards a water cleaning project. Each of you gets $20 per round to contribute between $0 and $20. The project has a 50% return rate. Your payoff relies on both your and others' contributions. Everyone benefits from the group's total contribution. Your payoff in each round equals the amount you didn't contribute (endowment - contribution) plus everyone's benefit (total contributions * 50% return rate).

Here is an example: [Suppose that you contribute $12, and total group contributions are $20. Your payoff will be ($20-$12) + $20*50% = $18]

You will be presented with 21 options, each represented in the form (M), which means contributing $M. The 21 options are (0), (1), (2), (3), (4), (5), (6), (7), (8), (9), (10), (11), (12), (13), (14), (15), (16), (17), (18), (19), (20). We will play a total of 3 rounds, in the first round, how much of the $20 would you like to contribute? Please only give one concrete number and highlight it with [[]] (e.g., [[x]]).

E.5 Bomb Risk Game

E.5.1 Baseline Condition

System Message.   You are a helpful assistant.

User Message 1.   Hi, let's play a game.

User Message 2.   There are 100 boxes, and one bomb has been randomly placed in 1 of 100 boxes. You can choose to open 0-100 boxes at the same time. If none of the boxes you open has the bomb, you earn dollars that are equal to the number of boxes you open. If one of the boxes you open has the bomb, you earn zero dollars.

Here is an example: [Suppose that you open 60 boxes, and one of the box has the bomb. Your payoff will be $0.]

How many boxes would you open? Please only give one concrete number and highlight it with [[]] (such as [[x]]).

E.5.2 Experimental Conditions

Persona

The variation of personas is the same as that introduced in Section E.1.2.

Answer Type

The System Message and User Message 1 is the same as those in the baseline condition.

User Message 2. There are 100 boxes, and one bomb has been randomly placed in 1 of 100 boxes. You can choose to open 0-100 boxes at the same time. If none of the boxes you open has the bomb, you earn dollars that are equal to the number of boxes you open. If one of the boxes you open has the bomb, you earn zero dollars.

Here is an example: [Suppose that you open 60 boxes, and one of the box has the bomb. Your payoff will be $0.]

You will be presented with 21 options, each represented in the form (M), which means opening M boxes. The 21 options are (0), (5), (10), (15), (20), (25), (30), (35), (40), (45), (50), (55), (60), (65), (70), (75), (80), (85), (90), (95), (100). How many boxes would you open? Please only give one concrete number and highlight it with [[]] (such as [[x]]).